Module 3

Psychological Scaling

"*Measurement* essentially is concerned with the methods used to provide quantitative descriptions of the extent to which individuals manifest or possess specified characteristics" (Ghiselli, Campbell, & Zedeck, 1981, p. 2). "*Measurement* is the assigning of numbers to individuals in a systematic way as a means of representing properties of the individuals" (Allen & Yen, 1979, p. 2). "'*Measurement*' consists of rules for assigning symbols to objects so as to (1) represent quantities of attributes numerically (scaling) or (2) define whether the objects fall in the same or different categories with respect to a given attribute (classification)" (Nunnally & Bernstein, 1994, p. 3).

No matter which popular definition of the term **measurement** you choose, several underlying themes emerge. First, we need to be able to quantify the attribute of interest. That is, we need to have numbers to designate how much (or little) of an attribute an individual possesses. Second, we must be able to quantify our attribute of interest in a consistent and systematic way (i.e., standardization). That is, we need to make sure that if someone else wants to replicate our measurement process, it is systematic enough that meaningful replication is possible. Finally, we must remember that we are measuring attributes of individuals (or objects), not the individuals per se. This last point is particularly important when performing high-stakes testing or when dealing with sensitive subject matter. For example, if we disqualify a job candidate because he or she scored below the established cutoff on a preemployment drug test, we want to make sure that the person is not "labeled" as a drug addict. Our tests are not perfect and whenever we set a cutoff on a test, we may be making an error by designating someone as

above or below the cutoff. In the previous example, we may be mistakenly classifying someone as a drug user when, in fact, he or she is not.

Levels of Measurement

As the definition of Nunnally and Bernstein (1994) suggests, by systematically measuring the attribute of interest we can either classify or scale individuals with regard to the attribute of interest. Whether we engage in classification or scaling depends in large part on the level of measurement used to assess our **construct.** For example, if our attribute is measured on a **nominal scale** of measurement, then we can only classify individuals as falling into one or another mutually exclusive category. This is because the different categories (e.g., men versus women) represent only **qualitative** differences. Say, for example, we are measuring the demographic variable of race. An individual can fall into one of several possible categories. Hence, we are simply classifying individuals based on self-identified race. Even if we tell the computer that Caucasians should be coded 0, African Americans 1, Hispanics 2, Asian Americans 3, and so on, that does not mean that these values have any quantitative meaning. They are simply labels for our racial categories.

For example, the first author once had an undergraduate student working on a research project with him. She was asked to enter some data and run a few Pearson correlation coefficients. The student came back very excited that she had found a significant relationship between race and our outcome variable of interest (something akin to job performance). Race had a coding scheme similar to that described previously. When the student was asked to interpret the correlation coefficient, she looked dumbfounded, as well she should, because the correlation coefficient was not interpretable in this situation, as the variable race was measured at the nominal level.

On the other hand, we may have a variable such as temperature that we can quantify in a variety of ways. Assume we had 10 objects and we wanted to determine the temperature of each one. If we did not have a thermometer, we could simply touch each one, assuming it was not too hot or too cold, and then rank order the objects based on how hot or cold they felt to the touch. This, of course, is assuming that the objects were all made of material with similar heat transference properties (e.g., metal transfers heat, or cold, much better than wood). This would represent an **ordinal scale** of measurement where objects are simply rank ordered. You would not know how much hotter one object is than another, but you would know that A is hotter than B, if A is ranked higher than B. Is the ordinal level of measurement

sufficient? In some cases, it is. For example, if you want to draw a bath for your child, do you need to know the exact temperature? Not really, you just need to be careful not to scald your child.

Alternatively, we may find a thermometer that measures temperature in degrees Celsius and use it to measure the temperature of the 10 items. This device uses an **interval scale** of measurement because we have equal intervals between degrees on the scale. However, the zero point on the scale is arbitrary; 0 degrees Celsius represents the point at which water freezes at sea level. That is, zero on the scale does not represent "true zero," which in this case would mean a complete absence of heat. However, if we were to use a thermometer that used the Kelvin scale, we would be using a **ratio scale** of measurement because zero on the Kelvin scale does represent "true zero" (i.e., no heat).

When we measure our construct of interest at the nominal (i.e., qualitative) level of measurement, we can only classify objects into categories. As a result, we are very limited in the types of data manipulations and statistical analyses we can perform on the data. Referring to the previous module on descriptive statistics, we could compute frequency counts or determine the modal response (i.e., category), but not much else. However, if we were at least able to rank order our objects based on the degree to which they possess our construct of interest (i.e., we have **quantitative** data), then we could actually scale our construct. In addition, higher levels of measurement allow for more in-depth statistical analyses. With ordinal data, for example, we can compute statistics such as the median, range, and interquartile range. When we have interval-level data, we can calculate statistics such as means, standard deviations, variances, and the various statistics of shape (e.g., skew and kurtosis). With interval-level data, it is important to know the shape of the distribution, as different-shaped distributions imply different interpretations for statistics such as the mean and standard deviation.

Unidimensional Scaling Models

In psychological measurement, we are typically most interested in **scaling** some characteristic, trait, or **ability** of a person. That is, we want to know how much of an attribute of interest a given person possesses. This will allow us to estimate the degree of inter-individual and intra-individual differences (as discussed in Module 1) among the subjects on the attribute of interest. This measurement process is usually referred to as **psychometrics** or *psychological measurement*. However, we can also scale the stimuli that we give to individuals, as well as the responses that individuals provide. Scaling

of stimuli and responses is typically referred to as *psychological scaling*. Scaling of stimuli is more prominent in the area of psychophysics or sensory/perception psychology that focuses on physical phenomena and whose roots date back to mid–19th century Germany. It was not until the 1920s that Thurstone began to apply the same scaling principles to scaling psychological attitudes. In addition, we can attempt to scale several factors at once. This can get very tricky, however. So more often than not, we hold one factor constant (e.g., responses), collapse across a second (e.g., stimuli), and then scale the third (e.g., individuals) factor.

For example, say we administered a 25-item measure of social anxiety to a group of schoolchildren. We would typically assume all children are interpreting the response scale (e.g., a scale of 1–7) for each question in the same way (i.e., responses are constant), although not necessarily responding with the same value. If they did all respond with exactly the same value, then we would have no variability and thus the scale would be of little interest to us because it would have no predictive value. Next, we would collapse across stimuli (i.e., get a total score for the 25 items). As a result, we would be left with scaling children on the construct of social anxiety.

Many issues (besides which factor we are scaling) arise when performing a scaling study. One important factor is who we select to participate in our study. When we scale people (*psychometrics*), we typically obtain a random sample of individuals from the population that we wish to generalize. In our preceding example, we would want a random sample of school-aged children so that our results generalize to all school-aged children. Conversely, when we scale stimuli (*psychological scaling*), we do not want a random sample of individuals. Rather, the sample of individuals we select should be purposefully and carefully selected based on their respective expertise on the construct being scaled. That is, they should all be **subject matter experts (SMEs).** In our preceding example, we would want experts on the measurement of social anxiety, particularly as it relates to children in school settings, to serve as our SMEs. Such SMEs would likely include individuals with degrees and expertise in clinical, school, developmental, counseling, or personality psychology.

Another difference between psychometrics and psychological scaling is that with psychometrics we ask our participants to provide their individual feelings, **attitudes,** and/or personal ratings toward a particular topic. In doing so, we will be able to determine how individuals differ on our construct of interest. With psychological scaling, however, we typically ask participants (i.e., SMEs) to provide their professional judgment of the particular stimuli, regardless of their personal feelings or attitudes toward the topic or stimulus. This may include ratings of how well different stimuli represent the construct and at

what level of intensity the construct is represented. Thus, with psychometrics, you would sum across items (i.e., stimuli) within an individual respondent in order to obtain his or her score on the construct. With psychological scaling, however, the researcher would sum across raters (SMEs) within a given stimulus (e.g., question) in order to obtain rating(s) of each stimulus. Once the researcher was confident that each stimulus did, in fact, tap into the construct and had some estimate of the level at which it did so, only then should the researcher feel confident in presenting the now scaled stimuli to a random sample of relevant participants for psychometric purposes.

The third category of responses, which we said we typically hold constant, also needs to be identified. That is, we have to decide in what fashion we will have subjects respond to our stimuli. Such response options may include requiring our participants to make comparative judgments (e.g., which is more important, A or B?), subjective evaluations (e.g., strongly agree to strongly disagree), or an absolute judgment (e.g., how hot is this object?). Different response formats may well influence how we write and edit our stimuli. In addition, they may also influence how we evaluate the quality or the "accuracy" of the response. For example, with absolute judgments, we may have a standard of comparison, especially if subjects are being asked to rate physical characteristics such as weight, height, or intensity of sound or light. With attitudes and psychological constructs, such "standards" are hard to come by.

There are a few options (e.g., Guttman's Scalogram and Coomb's unfolding technique) for simultaneously scaling people and stimuli, but more often than not we scale only one dimension at a time. However, we must scale our stimuli first (or seek a well-established measure) before we can have confidence in scaling individuals on the stimuli. Advanced texts such as Nunnally and Bernstein (1994), Ghiselli et al. (1981), and Crocker and Algina (1986) all provide detailed descriptions of different scaling methods for scaling stimuli and response data at a variety of different levels of measurement. We refer you to these advanced texts for more detailed explanations. In the following discussion, we will provide only a general overview of the major unidimensional scaling techniques.

We can scale stimuli at a variety of different measurement levels. At the nominal level of measurement, we have a variety of sorting techniques. In this case, SMEs are asked to sort the stimuli into different categories based on some dimension. For example, our SMEs with expertise in the social anxiety of school-aged children might be asked to sort a variety of questions according to whether the items are measuring school-related social anxiety or not. In doing so, we are able to determine which items to remove and which to keep for further analyses when our goal is to measure school-related social anxiety.

At the ordinal level of measurement, we have the Q-sort method, paired comparisons, Guttman's Scalogram, Coomb's unfolding technique, and a variety of rating scales. The major task of SMEs is to rank order items from highest to lowest or from weakest to strongest. Again, our SMEs with expertise in school-related social anxiety might be asked to sort a variety of questions. However, instead of a simple "yes" and "no" sorting, in terms of whether the questions measure social anxiety or not, the SMEs might be asked to sort the items in terms of the extent to which they measure social anxiety. So, for example, an item that states, "I tend to feel anxious when I am at school" would likely get a higher ranking than an item that states, "I tend to have few friends at school." While both items may be tapping into social anxiety, the first item is clearly more directly assessing school-related social anxiety.

At the interval level of measurement, we have direct estimation, the method of bisection, and Thurstone's methods of comparative and categorical judgments. With these methods, SMEs are asked not only to rank order items but also to actually help determine the magnitude of the differences among items. With Thurstone's method of comparative judgment, SMEs compare every possible pair of stimuli and select the item within the pair that is the better item for assessing the construct. Thurstone's method of categorical judgment, while less tedious for SMEs when there are many stimuli to assess in that they simply rate each stimulus (not each pair of stimuli), does require more cognitive energy for each rating provided. This is because the SME must now estimate the actual value of the stimulus.

Multidimensional Scaling Models

With unidimensional scaling, as described previously, subjects are asked to respond to stimuli with regard to a particular dimension. For example, a consumer psychologist might ask subjects how they would rate the "value" of a particular consumer product. With **multidimensional scaling (MDS)**, however, subjects are typically asked to give just their general impression or broad rating of similarities or differences among stimuli. For example, subjects might be asked to compare several different types of products and simply rate which are similar or which they prefer the best overall. Subsequent analyses, using Euclidean spatial models, would "map" the products in multidimensional space. The different multiple dimensions would then be "discovered" or "extracted" with multivariate statistical techniques, thus establishing which dimensions the consumer is using to distinguish the products. MDS can be particularly useful when subjects are unable to articulate "why" they like a stimulus, yet they are confident that they prefer one stimulus to another.

A Step-by-Step Scaling Example

Let us now work through our earlier example on school-related social anxiety in school-aged children from start to finish. What would be the first step in conducting a study where you wanted to develop a measure to assess school-related social anxiety in school-aged children? Well, our first step is to make sure we have a clear definition of what we mean by our construct of school-related social anxiety. Everyone who hears this term may have a slightly different impression of what we would like to assess. Therefore, we need to be able to present our SMEs with a single definition of what we are trying to assess when we talk about this construct. In this case, we will start with, "School-related social anxiety refers to the uneasiness school-aged children experience when they are in school-related social settings, but that may not be manifested in nonschool social settings such as at home or with friends outside of school. Such uneasiness may include feelings of isolation, physical stressors, and other such psychological and physical symptoms." Okay, it is not great, but it is a start. What next? Now we need to start developing items to assess our construct. Who should do that? Ah, yes, our infamous SMEs. Who should serve as SMEs in this instance and how many do we need? We stated earlier that, ideally, we would want to use school psychologists, clinical psychologists, counseling psychologists, developmental psychologists, and/or personality theorists. It may be difficult, however, to convince such individuals to participate in the item generation stage of the study. Therefore, it may be more practical and realistic for you, the researcher, and some colleagues and/or research assistants to generate potential items and then reserve the SMEs to provide actual ratings on the items you generate.

How many items do we need? Unfortunately, there is no easy answer to this question. The best response is, "The more the better." Ideally, you would want to generate at least twice as many items as you hope to have on your final scale. Therefore, if you want a 25-item scale of school-related social anxiety, you should generate at least 50 items. Now that we have our 50 or more items, it is time to bring in our SMEs. Again, how many SMEs do we need? Ideally, it would be nice to have "lots" of them; in reality, we may be lucky to get four or five. At a minimum, you need to have more than two in order to obtain variability estimates. Any number beyond two will be advantageous, within reason of course. This is also the step where we need to select one of the scaling models. Remember, these "models" are simply standardized procedures that will allow us to attach meaningful numbers to the responses our subjects will ultimately provide. Thus, we need not get too anxious (pardon the pun) over which method we choose to scale social

anxiety. One prominent scaling procedure, which we touched on briefly, is Likert scaling, so we will use that.

Before we jump into scaling our stimuli, however, we need to know what type of responses we want our subjects to provide. In fact, this would probably be good to know as we are writing our questions. Remember, we pointed out earlier that these might include evaluative judgments, degree of agreement, frequency of occurrence, and so on. Which one we choose is probably not as critical as the fact that all of our items are consistent with the response scale we choose. For example, we do not want to mix questions with statements. In this case, we will go with the degree of agreement format because this is common with Likert-type scales. With most **Likert scales,** we usually have a four- or five-option response scale ranging from strongly agree to strongly disagree (e.g., 1 = Strongly Disagree, 2 = Disagree, 3 = Undecided, 4 = Agree, and 5 = Strongly Agree). With an odd number of scale values, we have an undecided or neutral option in the middle. With an even number of scale values, we force the respondent to agree or disagree (sometimes called a forced format or choice scale). So should we use four or five options? It is mostly a matter of preference; be aware, however, which one you choose can affect the interpretation of your scores.

So far, we have defined our construct, generated items, and decided on a response scale. Now it is time to let our SMEs loose on the items. Remember, the SMEs are providing their professional judgment as to how well each item represents the construct or to what degree it represents the construct, regardless of their personal feelings. Once we have the SME ratings, how do we use these to decide which items to retain? Well, we could compute statistics such as item-total correlations. That is, we could determine how well a given item correlates with the total score on the scale. If it correlates poorly, then we would likely discard that item. What constitutes "poorly"? There is no hard and fast rule, but you would probably only want to retain items that had at least a .50 or higher correlation with the total scale. You may also want to look at the variability in ratings provided by the SMEs. If the ratings for a given item do not differ much, then the SMEs are being consistent in their ratings, which is a good thing, but from a psychometric standpoint, too little variability leaves us unable to compute certain statistics (i.e., correlation coefficients). Ultimately, which items to keep and which to remove is a professional judgment call. However, in practical terms, remember you wanted a 25-item scale. So why not choose the top 25 items in terms of their item-total correlation and discriminability? Some of these items may, of course, still require further editing before being implemented.

Finally, you are ready to administer your newly developed Likert scale to actual subjects. How many subjects do you need? For the psychometric portion of the study (estimating reliability and validity, as discussed later in the

book), the answer is again, "the more the better." Realistically, though, we need to have enough for our statistics to be meaningful. That usually means at least 100 subjects. For evaluating research questions and hypotheses, many factors come into play in determining appropriate sample size. In that instance, most researchers now conduct power and/or precision analyses to determine the most desirable sample size for their particular situation.

An individual's score on the scale will be the sum or mean of his or her responses to the 25 items. Remember that you may have some items that have reverse meaning (e.g., they were really assessing social calmness, not social anxiety). These items will need to be reverse scored. That is, what was a 1 is now a 5, a 2 becomes a 4, 3 stays 3, 4 becomes 2, and 5 becomes 1. This reverse scoring of reversed items should be done before the summated total score is obtained. Now that you have created and evaluated your school-related social anxiety scale, you are ready to carry out the psychometric studies that we discuss in Modules 5 through 8.

Concluding Comments

We began by looking at several definitions of measurement and examining the key elements of psychological measurement. Next, we discussed the different levels of measurement that our psychological scales can assess. Then we talked about key issues distinguishing psychometrics from psychological scaling. We next provided an overview to the different unidimensional scaling models and how they relate to the different levels of measurement. Finally, we worked through a realistic step-by-step example of what an applied scaling project might look like. We also briefly touched on multidimensional scaling. In the final analysis, the key is first to have confidence in your stimuli and responses and then move on to scale individuals. This is the crux of the psychometric process, which is the topic of the remainder of this book.

Practical Questions

1. What is the difference between scaling and classification?
2. What is the difference between psychometrics and psychological scaling?
3. Why do you think it is so difficult to scale more than one dimension (i.e., people, stimuli, and responses) at once?
4. Why is it important to know the level of measurement of our data before we begin the scaling process?
5. How would we scale multiple dimensions at one time?

Case Studies

CASE STUDY 3.1: SCALING STUDY IN CONSUMER PSYCHOLOGY

Benjamin, a senior who had a dual major in psychology and marketing, decided he wanted to do his undergraduate honors thesis in the area of consumer psychology. Specifically, he was interested in determining how well young children were able to recall a series of visual only (e.g., magazine advertisements), auditory only (e.g., radio commercials), and combined visual and auditory (e.g., television commercials) advertisements for Lego® building toys. He had learned in his undergraduate tests and measurements class that most of the time we were interested in looking at individual differences within our subjects. In this case, would it be who was able to remember one type of advertisement better than another? That didn't really seem to be the issue of major concern here. Why would advertisers be interested in the type of preadolescent who remembered one type of advertisement better than others his age? Maybe it would allow advertisers to target their product to specific children (e.g., those who watch PBS programming versus those who watch network or cable programming).

On second thought, Benjamin wondered whether the real issue was which method of advertising was most likely to be remembered by a "typical" child. If so, it seemed as if he should really be more interested in scaling different types of advertising modalities (i.e., stimuli) than in scaling subjects. By doing so, advertisers could determine which modalities would produce the best recall and thus how to most effectively spend their advertising dollars. As Benjamin thought some more, he began to wonder if it was the response that was really of most interest. That is, who cares if the child recalls the advertisement or not, isn't the bigger issue whether the child (or his or her parents) actually buys the toy (i.e., their response)? Maybe he needed to scale the responses children have to the different modes of advertisement, not the subjects or stimuli. Suddenly, it all seemed rather confusing. So, it was off to his advisor's office to get some advice and direction.

Questions to Ponder

1. What type of scaling should Benjamin be most concerned with? Subject, stimulus, or response? Why?
2. Who should Benjamin get to serve as subjects for his study?
3. Would he be better served with a random sample of children or with a relatively homogeneous group of subject matter experts (SMEs) for his scaling study?
4. What level of measurement data is Benjamin dealing with?
5. Will Benjamin actually have to do several scaling projects to get the information he needs?

CASE STUDY 3.2: A CONSULTING PROJECT ON PERFORMANCE ASSESSMENT

Jennifer, a graduate student who had just completed her first year in an industrial and organizational psychology PhD program, was excited because she had just gotten her first consulting job. She was to develop a performance appraisal form to assess workers in her uncle's small domestic cleaning service. There were a total of 15 "maids" and two office supervisors. Her uncle wanted to know which maids should receive a pay raise and how much he should give each of them. He wanted to make sure, however, that their raises were performance based. So, he contracted with Jennifer to create an easy-to-use performance appraisal form that he and his two office supervisors could use to assess each maid and ultimately use that information to determine the size of the raise for each maid.

Jennifer first conducted a literature search to see if she could find an existing performance assessment form that would fit the bill. While some existing forms looked like they might work, it seemed like no matter which one she chose she would have to do some significant modifications. She also noticed that different forms used different points of reference. For example, some performance appraisal forms used an absolute scale (e.g., below standard . . . at standard . . . above standard), while others used a relative scale (e.g., below average . . .

average . . . above average). Some used a paired comparison technique. That is, who is the better performer? Maid A or Maid B? Maid A or Maid C? Also, some scales had three categories or anchors, others five, some seven, and one was on a scale of 1–100. There was even one that had no numbers or words at all; it was simply a series of faces ranging from a deep frown to a very big grin. A bit overwhelmed and a little unsure of how to proceed, Jennifer decided to seek the advice of the professor who would be teaching her performance assessment class next semester.

Questions to Ponder

1. What difference (if any) does it make if Jennifer uses an absolute or a relative rating scale?
2. Should Jennifer just develop her own scale or try to use an existing measure?
3. What issues should Jennifer be concerned with if she modifies an existing scale?
4. Is Jennifer more interested in scaling responses, stimuli, or subjects? Explain.
5. Who should serve as the raters in this case? The supervisors? Her uncle? The respective clients?
6. Would the decision in terms of who will serve as raters affect which type of scale is used (e.g., relative versus absolute versus paired comparison)?

Exercises

EXERCISE 3.1: LEVELS OF MEASUREMENT

OBJECTIVE: To practice defining and identifying different levels of measurement.

Identify two psychological constructs of interest. These could include personality (e.g., anxiety, extraversion), intelligence (e.g., verbal, quantitative, space visualization), or more ephemeral (e.g., infatuation, anger)

constructs. Then, identify existing measures of these two constructs and, based on the description of the different levels of measurement discussed in the overview to this module (i.e., nominal, ordinal, interval, and ratio), or your main text, discuss the level of measurement of the constructs. Finally, identify how you would measure the construct at two different alternative levels of measurement.

For example, assume you wanted to look at extraversion and you chose the NEO personality inventory. Most would identify extraversion as being measured at the interval level of measurement using this instrument. Thus, how might you measure extraversion at a nominal, ordinal, or ratio level of measurement?

EXERCISE 3.2: CONDUCTING A SCALING STUDY

OBJECTIVE: To provide practice in conducting a scaling study.

Outline a scaling study similar to the example that is provided at the end of the overview section of the module. Select a construct (other than school-related social anxiety) and answer the following questions:

1. What is the definition of your construct?
2. Who is going to generate items to measure the construct? How many items do they need to generate? Why?
3. What scaling model would be most appropriate for your example?
4. Who is going to serve as SMEs to rate the items?
5. On what basis are you going to select the items to keep for the final version of your scale?
6. Who are going to serve as subjects for your study? How many subjects do you need?

EXERCISE 3.3: SCALING ITEMS

OBJECTIVE: To practice scaling items.

Using the data from the "Bus Driver.sav" data set, scale the 10 task items on the three dimensions of "Frequency," "Relative Time Spent," and "Importance." Use Table 3.1 to fill in the mean task ratings across the three dimensions. In order for an item to be "retained" for further

consideration, the task must, on average, be carried out at least "regularly" (i.e., 3.0 or higher) in terms of frequency, fall between "little" and "moderate" (i.e., 2.5 or higher) in terms of relative time spent, and be rated as "very important" (i.e., 4.0 or higher) in terms of importance. Given these criteria, which of the 10 tasks meet all three of these criteria and, thus, should be retained?

Table 3.1 Summary of Task Ratings

Task Number	*Average Frequency Rating (≥ 3.0)*	*Average Relative Time Spent Rating (≥ 2.5)*	*Average Importance Rating (≥ 4.0)*
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

Internet Web Site References

3.1. http://web.uccs.edu/lbecker/SPSS/scalemeas.htm

This Web page provides information on the nominal, ordinal, interval, and ratio scales of measurement, along with possible arithmetic operations and example statistics for each.

3.2. http://www.fs.fed.us/rm/pubs_rm/rm_rp293.pdf

This Web page displays a paper in PDF format by Thomas Brown and Terry Daniel that provides theoretical and descriptive background on psychological scaling

and a specific computer program for psychological scaling called RMRATE. Their paper also provides a detailed description of the application of these procedures to an applied setting.

3.3. http://trochim.human.cornell.edu/kb/measlevl.htm

The Web page also provides information on levels of measurement.

3.4. http://trochim.human.cornell.edu/kb/scaling.htm

This Web page presents an overview of scaling, with links to issues in scaling and explanations of the Thurstone, Likert, and Guttman scaling methods.

Further Readings

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory* (pp. 45–66). Belmont, CA: Wadsworth.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences* (pp. 391–420). New York: W. H. Freeman.

Guildford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed., pp. 31–82). New York: McGraw-Hill.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

van der Ven, A. H. G. S. (1980). *Introduction to scaling*. New York: Wiley.