**Scaling[[1]](#footnote-1)©**

Psychologists commonly employ surveys (questionnaires and interviews) to gather data on their human subjects, in both nonexperimental and experimental research. Scaling involves the construction of instruments for the purpose of measuring abstract concepts such as intelligence, hypomania, ethical ideology, misanthropy, political conservatism, and so on. We shall deal primarily with unidimensional scaling methods -- that is, measuring concepts for which we are comfortable assuming that only a single dimension is involved. I shall briefly touch on the use of exploratory factor analysis to investigate the dimensionality of a scale.

**Thurstone Scales**

**Defining the Concept.** Using the Method of Equal-Appearing Intervals, the first step is to define the concept that you wish to measure. Suppose that I want to measure misanthropy. My definition of misanthropy is a dislike of human beings as a species. People who score high on the scale I wish to construct will have an unfavorable attitude about humans, while people who score low will have a favorable attitude about humans.

**Generating Potential Scale Items.** After defining the concept, I shall, with the assistance of my associates, generate a large number (80 - 100) of statements that are worded in a similar fashion but which differ with respect to whether agreement with the statement would indicate having a large, medium, or low amount of the measured attribute. For example, I might generate the following statement as one for which I expect that agreement would indicate great misanthropy: “Most humans are despicable creatures.” On the other hand, agreement with this next statement would seem to indicate low misanthropy: “Humans are basically good.”

**Rating the Potential Scale Items.** Next I get a group of judges to rate each of the statements on an 11-point scale, where a rating of 1 indicates that they believe that agreeing with the statement would indicate that the respondent has a very low amount of the attribute being measured, and a rating of 11 indicates that they believe that agreeing with the statement would indicate that the respondent has a very high amount of the attribute being measured. I would encourage the judges to use the entire response scale -- that is, they should assign some statements to each of the 11 possible scale values.

**Computing the Scale Score Values for the Potential Items.** I enter the data into a statistical package and find, for each potential item, the median rating and some measure of the variability of the ratings (such as the interquartile range or the standard deviation). The medians will be the scale score values for the potential items. I prepare a table which includes each potential item, its median, and its variability. I sort the items by the median, from low to high, and then within each group of items with identical medians, I sort the items from low to high on variability.

**Selecting the Final Scale Items.** I want to select one item for each possible scale score value. Look at the table in the document [Thurstone Scaling](http://core.ecu.edu/psyc/wuenschk/docs2210/Thurstone.htm) for an example (from Trochim’s Internet document on this topic) of item selection.

**Administering the Final Scale.** The selected final items are presented to the respondents in a random order (not sorted in order of the items’ scale values). Respondents are asked to indicate, for each item, whether they agree or disagree with the statement. Each respondent’s score is computed by finding the mean scale score (the median from the judges’ ratings) of the items with which the respondent agreed.

**Guttman Scaling**

Statements on a Guttman scale are selected so that they create an ordered sequence. Agreeing with the first statement indicates that the respondent has at least a little of the measured attribute. Agreeing with the second statement indicates that the respondent has at least a little more of the measured attribute, and so on. The items are written so that any respondent who agrees with the nth statement should also agree with all of the preceding statements.

To construct a Guttman scale you would:

* Define the concept you wish to measure.
* Seek assistance in writing a large number of statements endorsement of which would indicate a respondent’s position on the measured dimension.
* Ask judges to indicate for each statement whether or not a person who has a high amount of the measured attribute would agree with the statement.
* Employ the tedious procedure (a scalogram analysis, best done with specialized computer software) to select items that meet or approximate the requirement that the statements form an ordered sequence. This analysis will help you select the items to be included on the scale and to assign a scale score to each item.
* To score the scale, sum, for each respondent, the scale scores for each statement to which he or she has agreed.

Here is an example of the sort of Guttman scale item that one might construct to measure respondent’s attitude about immigration (courtesy of [Trochim](http://www.socialresearchmethods.net/kb/scalgutt.htm)). The respond would be asked to indicate, for each statement, whether or not e agreed with it.

* I believe that this country should allow more immigrants in.
* I would be comfortable with new immigrants moving into my community.
* It would be fine with me if new immigrants moved onto my block.
* I would be comfortable if a new immigrant moved next door to me.
* I would be comfortable if my child dated a new immigrant.
* I would permit a child of mine to marry an immigrant.

One would expect that a person who would agree with the nth statement above would also agree with each of the preceding statements. In the most simple case the scale would be scored by adding up the number of agreements. A scalogram analysis can be employed to produce weights for each of the items (scale scores), in which case the total score would be the sum of the weights of the items that were endorsed.

See Hervé Abdi’s document [Guttman Scaling](http://www.utdallas.edu/~herve/abdi-GuttmanScaling2010-pretty.pdf), for another example of a Guttman scale.

**Likert Scales**

This is my favorite type of response scale for survey items. The statements used in a Likert scale can be very much like those used in a Thurstone scale, but instead of a dichotomous Disagree/Agree response scale, the respondent has a multi-point response scale of agreement. The response scale may have from 4 to 9 response options. Because I have used 5-point optical scanning response forms in my research, I have most often used this response scale:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A | B | C | D | E |
| strongly disagree | disagree | no opinion | agree | strongly agree |

**Generating Potential Items.** As with Thurstone scaling, you should start by defining the concept you wish to measure and then generate a large number of potential items. It is a good idea to recruit colleagues to help you generating the items. Some of the items should be worded such that agreement with them represents being high in the measured attribute and others should be worded such that agreement with them represents being low in the measured attribute.

**Evaluating the Potential Items.** It is a good idea to get judges to evaluate your pool of potential items. Ask each judge to evaluate each item using the following scale:

1 = agreeing indicates the respondent is very low in the measured attribute

2 = agreeing indicates the respondent is below average in the measured attribute

3 = agreeing does not tell anything about the respondent’s level of the attribute

4 = agreeing indicates the respondent is above average in the measured attribute

5 = agreeing indicates the respondent is very high in the measured attribute

Analyze the data from the judges and select items with very low or very high averages (to get items with good discriminating ability) and little variability (indicating agreement among the judges).

Alternatively, you could ask half of the judges to answer the items as they think a person low in the attribute to be measured would, and the other half to answer the items as would a person high in the attribute to be measured. You would then prefer items which best discriminated between these two groups of judges -- items for which the standardized difference between the group means is greatest.

Judges can also be asked whether any of the items were unclear or confusing or had other problems.

**Pilot Testing the Items.** After you have selected what the judges thought were the best items, you can administer the scale to respondents who are asked to answer the questions in a way that reflects their own attitudes. It is a good idea to do this first as a pilot study, but if you are impatient like me you can just go ahead and use the instrument in the research for which you developed it (and hope that no really serious flaws in the instrument appear). Even at this point you can continue your evaluation of the instrument -- at the very least, you should conduct an item analysis (discussed below), which might lead you to drop some of the items on the scale.

**Scoring the Items.** The most common method of creating a total score from a set of Likert items is simply to sum each person’s responses to each item, where the responses are numerically coded with 1 representing the response associated with the lowest amount of the measured attribute and N (where N = the number of response options) representing the response associated with the highest amount of the measured attribute. For example, for the response scale I showed above, A = 1, B = 2, C = 3, D = 4, and E = 5, assuming that the item is one for which agreement indicates having a high amount of the measured attribute.

You need to be very careful when using a computer to compute total scores. With some software, when you command the program to compute the sum of a certain set of variables (responses to individual items), it will treat missing data (items on which the respondent indicated no answer) as zeros, which can greatly corrupt your data. If you have any missing data, you should check to see if this is a problem with the computer software you are using. If so, you need to find a way to deal with that problem (there are several ways, consult a statistical programmer if necessary).

I generally use means rather than sums when scoring Likert scales. This allows me a simple way to handle missing data. I use the SAS (a very powerful statistical analysis program) function NMISS to determine, for each respondent, how many of the items are unanswered. Then I have the computer drop the data from any subject who has missing data on more than some specified number of items (for example, more than 1 out of 10 items). Then I define the total score as being the mean of the items which were answered. This is equivalent to replacing a missing data point with the mean of the subject’s responses on the other items in that scale -- if all of the items on the scale are measuring the same attribute, then this is a reasonable procedure. This can also be easily done with SPSS.

If you have some items for which agreement indicates a low amount of the measured attribute and disagreement indicates a high amount of the measured attribute (and you should have some such items), you must remember to **reflect** (reverse score) the item prior to including it in a total score sum or mean or an item analysis. For example, consider the following two items from a scale that I constructed to measure attitudes about animal rights:

* Animals should be granted the same rights as humans.
* Hunters play an important role in regulating the size of deer populations.

Agreement with the first statement indicates support for animal rights, but agreement with the second statement indicates nonsupport for animal rights. Using the 5-point response scale shown above, I would reflect scores on the second item by subtracting each respondent’s score from 6.

**Item Analysis.** You should have already read my document [Cronbach’s Alpha and Maximized Lambda4](http://core.ecu.edu/psyc/wuenschk/MV/alpha.pdf), which explains how to use SPSS to do an item analysis. If you will look at the output provided by SPSS, you will see that it gives you a corrected item-total correlation for each item in a scale. This is the correlation between score on the individual item and the sum of the scores on all the other items. If all of the items are measuring the same attribute, then all of the item-total correlations should be high. If you see that an item has a low item-total correlation, then you should consider dropping it from the scale. SPSS also tells you what Cronbach’s alpha would be if you dropped an item from the scale. If you see an item for which dropping it would increase alpha, you should consider dropping it. If you happen to see an item with a negative item-total correlation, that is probably an item that should have been reflected but which was not.

**Factor Analysis.** It may also be useful to conduct a factor analysis on the scale data to see if the scale really is unidimensional. Responses to the individual scale items are the variables in such a factor analysis. These variables are generally well correlated with one another. We wish to reduce the (large) number of variables to a smaller number of **factors** that capture most of the variance in the observed variables. Each factor is estimated as being a linear (weighted) combination of the observed variables. We could extract as many factors as there are variables, but generally most of those factors would contribute little, so we try to get just a few factors that capture most of the covariance. Our initial extraction generally includes the restriction that the factors be orthogonal, independent of one another.

Consider the analysis reported by Chia, Wuensch, Childers, Chuang, Cheng, Cesar-Romero, & Nava in the *Journal of Social Behavior and Personality*, 1994, *9*, 249-258. College students in Mexico, Taiwan, and the US completed a 45 item Cultural Values Survey. A principal components analysis (very similar to a factor analysis) produced seven components (each a linear combination of the 45 items) which explained in the aggregate 51% of the variance in the 45 items. We could have explained 100% of the variance with 45 components, but the purpose of conducting the PCA was to explain much of the variance with relatively few components. Imagine a **plot in seven dimensional space** with seven perpendicular (orthogonal) axes. Each axis represents one component. For each variable I plot a point that represents its loading (correlation) with each component. With luck I’ll have seven “clusters” of dots in this hyperspace (one for each component). I may be able to improve my solution by rotating the axes so that each one more nearly passes through one of the clusters. I may do this by an **orthogonal rotation** (keeping the axes perpendicular to one another) or by an **oblique rotation**. In the latter case I allow the axes to vary from perpendicular, and as a result, the components obtained are no longer independent of one another. This may be quite reasonable if I believe the underlying dimensions (that correspond to the extracted components) are correlated with one another.

With luck (or after having tried many different extractions/rotations), I’ll come up with a set of loadings that can be interpreted sensibly (that may mean finding what I expected to find). From consideration of which items loaded well on which components, I named the components Family Solidarity (respect for the family), Executive Male (men make decisions, women are homemakers), Conscience (important for family to conform to social and moral standards), Equality of the Sexes (minimizing sexual stereotyping), Temporal Farsightedness (interest in the future and the past), Independence (desire for material possessions and freedom), and Spousal Employment (each spouse should make decisions about his/her own job). Now, using weighting coefficients obtained with the analysis, I computed for each subject a score that estimated how much of each of the seven dimensions e had. These **component scores** were then used as dependent variables in 3 x 2 x 2, Culture x Sex x Age (under 20 vs. over 20) ANOVAs. US students (especially the women) stood out as being sexually egalitarian, wanting independence, and, among the younger students, placing little importance on family solidarity. The Taiwanese students were distinguished by scoring very high on the temporal farsightedness component but low on the conscience component. Among Taiwanese students the men were more sexually egalitarian than the women and the women more concerned with independence than were the men. The Mexican students were like the Taiwanese in being concerned with family solidarity but not with sexual egalitarianism and independence, but like the US students in attaching more importance to conscience and less to temporal farsightedness. Among the Mexican students the men attached more importance to independence than did the women.

Clearly the 45-item Cultural Values Scale employed in the research just discussed was not unidimensional. Consider another study in which I employed factor analysis. As part of her masters thesis, ECU graduate student Sunita Patel constructed a test which she called the SBS (Self‑Report of Behavior Scale). It was designed to measure respondents’ past aggressive behavior towards homosexuals. I factor analyzed subjects’ scores on the 21 items in Patel’s SBS. Although the instrument was designed to measure a single dimension, my analysis indicated that three dimensions were being measured. The first factor, on which 13 of the items loaded well, seemed to reflect avoidance behaviors (such as moving away from a gay, staring to communicate disapproval of proximity, and warning gays to keep away). The second factor (six items) reflected aggression from a distance (writing anti-gay graffiti, damaging a gay’s property, making harassing phone calls). The third factor (two items) reflected up-close aggression (physical fighting). Despite this evidence of three factors, item analysis indicated that the instrument performed well as a measure of a single dimension.  **Item-total correlations** were good for all but two items. **Cronbach’s alpha** was .91, a value which could not be increased by deleting from the scale any of the items.

[Return to Wuensch’s Research Designs Lesson Page](http://core.ecu.edu/psyc/wuenschk/DesignLessons.htm)

Copyright 2011, Karl L. Wuensch - All rights reserved.

1. © Copyright 2011, Karl L. Wuensch - All rights reserved. [↑](#footnote-ref-1)