Measuring Social Attitudes

A Handbook for Researchers and Practitioners

DANIEL J. MUELLER
Indiana University

TEACHERS COLLEGE PRESS
Teachers College, Columbia University
New York and London
1950
Thurstone Scale Construction

Louis Thurstone is considered by many social scientists to be the “father” of attitude scaling. Beginning in the late 1920s Thurstone and several colleagues published a series of articles and monographs that presented the logic of attitude measurement and described and illustrated Thurstone’s attitude-scaling methodology. These publications led directly to scholarly revolutions, first in the area of attitude measurement and scaling, and subsequently in the area of empirical research and theory regarding the formation, change, and effects of social attitudes. Indeed, in the decades immediately following the publication of attitude-scaling methodologies by Thurstone, and shortly thereafter by Likert, attitude grew to be the single most important construct in social-psychological research and theory.

Thurstone actually developed three separate, albeit related, attitude-scaling techniques: (1) the method of paired comparisons, (2) the method of equal-appearing intervals, and (3) the method of successive intervals (or graded dichotomies). All three methods use the judgments of a panel of judges (who may be any responsible persons) regarding the relative favorableness (positiveness) of attitude statements toward the attitudinal object. Favorableness values for each statement are computed from these judgments, and scale items are selected, based in large part upon these values.

The method of paired comparisons (Thurstone, 1927) requires that attitude statements be paired in every possible combination. Each judge then decides which statement in each pair is more favorable toward the attitudinal object. Since 20 statements require the judging of 190 pairs of items, and 40 statements result in 780 pairs, the cumbersoness of this method is readily apparent. The method of equal-appearing intervals, by

---

1 It should be acknowledged that in 1925 F. Allport and D. Hartman outlined an attitude-scaling technique that is clearly the forerunner of the Thurstone method. In fact, in his 1928 article, Thurstone indicated that his attitude-scaling technique was a refinement of the Allport and Hartman procedure.

2 See the description of the EDSTAT TSCALE computer program in Appendix B for computation of item favorableness values for Thurstone scales constructed by the successive intervals method.

THE METHOD OF EQUAL-APPEARING INTERVALS

As in Likert scale construction, the first consideration in Thurstone scaling is the identification and careful delimitation of the attitudinal object. The importance of careful delimitation of the attitudinal object was stressed in Chapter 2.

The next step is the accumulation of a pool of opinion statements about the attitudinal object. The criteria for writing, selecting, and editing these statements are almost identical to those outlined for Likert scaling. They will not be repeated here. The only significant difference in criteria is that whereas neutral items are disallowed in Likert scale construction, in Thurstone scaling some neutral items are required. Thurstone scales, ideally, incorporate a sampling of statements covering the entire evaluative spectrum—from statements extremely favorable to the attitudinal object, through statements that are neutral, to statements extremely unfavorable to the attitudinal object.

It is not possible to prescribe exactly the number of statements needed in
the initial pool. Thurstone used 130 statements in the pool of one of his earliest published scales. Since that time, however, it has been shown that highly reliable scales can be constructed from substantially smaller item pools, as long as the criteria outlined in Chapter 2 for item selection and editing are followed. It is also advisable to make a special effort to include some neutral and some extremely negative statements in the pool, because these types of statements typically appear most sparsely in attitude item pools. In a first attempt at Thurstone scaling, I would recommend beginning with a pool of 40 or 50 opinion statements.

The next step is to obtain favorability values for all statements in the pool. This is done by having a group of judges sort the statements into categories (Thurstone used 11 categories) on a continuum ranging from "extremely unfavorable" to "extremely favorable." Only the most and least favorable and the neutral categories are typically labeled. The categories are successively scored with "extremely unfavorable" having a score of 1 and "extremely favorable" an 11. The scale value of an item is simply the average categorization of that item by all judges. (Medians rather than means are recommended, for a reason we shall see shortly.)

Thurstone used 300 judges, but scales have since been successfully constructed with as few as 10 or 15 judges. Much more important than the number of judges is their degree of concentration on the task. I have found that the judges must be reminded over and over again that they are to respond solely on the basis of the favorableness or unfavorableness of the statements, not on the basis of their agreement or disagreement with the statements. Some judges seem to have enormous difficulty maintaining this distinction. In answer to this "response set" problem, Thurstone recommended eliminating the scores of judges who weren't paying attention or who couldn't maintain the appropriate set in making judgments. His rule of thumb was to reject all judgments of any judge who placed 30 or more statements (out of 130: 23 percent) into a single category. While this criterion for carelessness is somewhat arbitrary, it is important to adopt some procedure for the screening of nondiscriminating judges. In addition to extreme skewing of responses by judges, frequent disagreement with other judges signals a careless or inept judge—especially on items where there is strong consensus among the other judges. A further safeguard from irresponsible judges is to compute item medians rather than means when calculating scale values. If, for a particular item, the response of a careless judge is substantially distant from the bulk of the distribution of judgments, the median will be much less affected by this extreme score than will the mean.

There has been substantial debate in the scaling literature regarding criteria for selecting judges. Some evidence exists that judges with extreme attitudinal positions (either very positive or very negative) make fine discriminations among items close to their own position but tend not to discriminate much among the other items, placing them all into a few categories at the opposite end of the continuum. The lesson of these research findings seems to be: Try to use either judges with moderate attitudinal positions or judges from the entire attitudinal range.

Thurstone reproduced the pool items on cards and had judges sort them physically into piles. While this procedure has the advantage of allowing judges to easily check the internal consistency of each pile, it is an extremely cumbersome process. Subsequent researchers have tested a variety of alternative procedures. Three that are logistically more efficient and seem to work just as well as Thurstone's original method, from a psychometric point of view, are illustrated in Figure 4.1. None requires the physical sorting of items.

Another modification of Thurstone's method that appears to be perfectly acceptable is to reduce the number of favorableness categories from 11 to 9 (or even to 7 or 5). While such a reduction of categories has been shown to result in substantially reliable scales, very little is gained in efficiency. Even numbers of categories (e.g., 10, 8, or 6) are seldom used, since this precludes a neutral category.

One more item statistic is necessary before the final scale items can be selected from the pool: a measure of item response variability (i.e., average deviation, standard deviation, or interquartile range). Thurstone reasoned that if an item is ambiguous, there will be poor consensus among judges regarding its degree of favorableness. If there is poor consensus, the item will have large response variability. Hence, in an effort to avoid ambiguous items, those items with large response variability are excluded from the scale. Thurstone used the interquartile range (Q = 75th percentile - 25th percentile) as his measure of item variability, but standard deviation can quite readily be substituted, especially if you have access to a computer program that computes standard deviation but not one that computes Q.

Average deviation will also work for this purpose.

The final scale is constructed by selecting 20 to 25 statements whose median values are approximately equidistant. If a 9-category favorableness scale was used by judges and if 22 items are to be selected for the final scale, the items will need to be picked at scale intervals of approximately .36 (There are 8 units between 1.00 and 9.00; 8 x .22 = .36.) In fact, since no items will have median values as low as 1.00 or as high as 9.00, a slightly smaller interval size, perhaps around .33, should be used to select 22 equidistant items. If two items have the same or nearly the same medians, (Continued on page 40)
FIGURE 8.1
Three methods of administering statements to be judged in the Thurstone scale construction process that do not require the physical sorting of items.

a. For each statement, circle the number that indicates the degree of favorableness of the statement toward the attitudinal object.

<table>
<thead>
<tr>
<th>Extremely unfavorable</th>
<th>Neutral</th>
<th>Extremely favorable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9</td>
<td>1. Divorce should be encouraged for many unhappily married people.</td>
<td></td>
</tr>
<tr>
<td>1 2 3 4 5 6 7 8 9</td>
<td>2. Divorce brings happiness to some people, unhappiness to others.</td>
<td></td>
</tr>
<tr>
<td>1 2 3 4 5 6 7 8 9</td>
<td>3. Divorce weakens the moral fiber of our society.</td>
<td></td>
</tr>
</tbody>
</table>

b. For each statement, place a check mark on the adjacent line indicating the degree of favorableness of the statement toward the attitudinal object.*

<table>
<thead>
<tr>
<th>Extremely unfavorable</th>
<th>Neutral</th>
<th>Extremely favorable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9</td>
<td>1. Divorce should be encouraged for many unhappily married people.</td>
<td></td>
</tr>
<tr>
<td>1 2 3 4 5 6 7 8 9</td>
<td>2. Divorce brings happiness to some people, unhappiness to others.</td>
<td></td>
</tr>
<tr>
<td>1 2 3 4 5 6 7 8 9</td>
<td>3. Divorce weakens the moral fiber of our society.</td>
<td></td>
</tr>
</tbody>
</table>

c. For each statement, circle the letter that indicates the degree of favorableness of the statement toward the attitudinal object.

<table>
<thead>
<tr>
<th>Extremely unfavorable</th>
<th>Neutral</th>
<th>Extremely favorable</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCDEFGHIJK</td>
<td>1. Divorce should be encouraged for many unhappily married people.</td>
<td></td>
</tr>
<tr>
<td>ABCDEFGHIJK</td>
<td>2. Divorce brings happiness to some people, unhappiness to others.</td>
<td></td>
</tr>
<tr>
<td>ABCDEFGHIJK</td>
<td>3. Divorce weakens the moral fiber of our society.</td>
<td></td>
</tr>
</tbody>
</table>

* An 11 (or 9, or 7, and so forth) interval stencil or ruler is superimposed over each line to quantify each judge's rating.
generally the one with the smaller interquartile range (or standard deviation or average deviation) should be chosen—especially if the items have substantially different interquartile ranges.

The chosen items are printed on a form, with instructions to respondents to indicate either agreement or disagreement with each item. The items are usually randomized, but research has shown that printing them in the order of their favorableness values does not adversely affect responses. An example of a completed Thurstone scale appears in Figure 4.2. Each respondent's scale score is the median (or mean) of the values of all items to which he assents. If, for example, a respondent agrees with items 3, 6, 7, and 8 in Figure 4.2, his scale score is 7.25—the median (and in this case also the mean) of the scale values of the four items checked by the respondent.

In Figure 4.3 the items from Figure 4.2 are plotted by scale values. This figure illustrates a severe deficiency in the scale: there are almost no neutral or nearly neutral items. While this attitude scale may suffice for research studies of group comparisons, and to differentiate those respondents with a positive attitude toward open classrooms from those with a negative attitude, it will not discriminate well toward the center of the attitudinal continuum. A specific effort to incorporate more neutral and nearly neutral items in the original item pool would have resulted in a better Thurstone scale.

Figure 4.4 is a scale measuring attitude toward the Chinese. This scale was constructed by Thurstone and published in 1931. A rather elaborate set of instructions acquaints respondents with their task. Not only are respondents required to check (√) statements with which they agree; they are also to mark statements with which they disagree (with an "X") and for which they are undecided (with a "?"). Since only the checked statements are scored, it is not clear why the other responding procedures were used. If you were to plot the scale values of these 26 items, you would find that they describe a quite evenly spaced scale along the entire attitudinal continuum. Since the highest scale value is 11.5 and the lowest is 0.5, it appears that Thurstone's judges used a 13-point continuum, scored from 0 to 12, in sorting the items. In this scale, low scores indicate positive attitude, high scores negative attitude.

ADVANTAGES AND DISADVANTAGES OF THURSTONE SCALES

Constructing two equivalent forms of an attitude scale is relatively easier in Thurstone scaling than in most other scaling methods. The size of the initial item pool must, of course, be increased, but once the item statistics have been computed, items can be matched (or approximately matched) in

(Continued on page 45)
FIGURE 4.3
Items from Figure 4.2 plotted by approximate scale values.

9.0
13. It should be required by law that all schools adopt the open classroom model of instruction. (8.7)

8.5
17. I would sign a petition supporting the use of open classrooms in local elementary schools. (8.2)

8.0
3. Children taught in open classrooms are more creative. (7.9)
6. Open classrooms facilitate effective development in children. (7.6)

7.5
10. Open classrooms teach children a sense of responsibility. (7.4)
12. I would vote for a candidate who supports the open classroom concept. (7.1)

7.0
7. Open classrooms positively affect teacher attitude. (6.9)

6.5
14. Children learn more when they have more freedom. (6.6)

6.0

5.5
8. Education in open classrooms is no better and no worse than in ordinary classrooms. (5.0)

5.0

4.5
4. Open classrooms are too undisciplined for maximum learning. (3.6)

4.0
9. The noise level in open classrooms is too high to facilitate learning. (3.1)

3.5
15. Open classrooms result in higher anxiety for some pupils. (2.8)
16. Open classrooms are just another educational fad. (2.6)

3.0

2.5
11. Principals who promote open classrooms should be replaced. (2.0)

2.0
2. I would not want a child of mine to attend a school with open classrooms. (1.6)

1.5
1. Open classrooms lead to juvenile delinquency. (1.4)
5. Open classrooms are a communist plot. (1.1)

1.0

Note: Exact scale values appear in parentheses following each item.
Attitude Toward the Chinese

Try to indicate either agreement or disagreement for each statement. If you simply cannot decide about a statement, you may mark it with a question mark.

This is not an examination. There are no right or wrong answers to these statements. This is simply a study of people's attitudes toward the Chinese. Please indicate your own convictions by a check mark when you agree and by a cross when you disagree.

Put a check mark if you agree with the statement. 
Put a cross if you disagree with the statement.

1. I have no particular love or hate for the Chinese. (6.5)
2. I dislike the Chinese more every time I see one. (10.1)
3. The Chinese are pretty decent. (4.7)
4. Some Chinese traits are admirable, but on the whole I don't like them. (7.2)
5. The Chinese are superior to all other races. (0.5)
6. The Chinese, as part of the yellow race, are inferior to the white race. (8.7)
7. I like the Chinese. (3.5)
8. The more I know about the Chinese the better I like them. (2.8)
9. The Chinese are aptly described by the term "yellow devils." (11.0)
10. The high-class Chinese are superior to us. (1.8)
11. The Chinese are different but not inferior. (5.2)
12. I hate the Chinese. (11.5)
13. Chinese parents are unusually devoted to their children. (4.1)
14. Although I respect some of their qualities, I could never consider a Chinese as a friend. (7.7)

Note: Scale values appear in parentheses following each item.

scale values on the two forms. Of course, constructing a second form simply because it's easily done hardly seems an adequate reason for doing it. Only if an alternative form can facilitate one's research is this theoretical advantage a real advantage.

A second advantage that Thurstone scales enjoy over other types of attitude scales is the existence of a "zero" or neutral point. This allows for "absolute" interpretation of scale scores rather than only "relative" interpretation (in which scores have meaning only relative to the scores of other respondents). An attitude score at the neutral point on a Thurstone scale can truly be interpreted as a neutral attitude. In Likert scaling there is no clearly neutral point.

In overall comparison with Likert scaling the Thurstone procedure comes
off second best, though. Both types of scales, if carefully constructed, can achieve reliability coefficients in the .80s, but Likert scales can often match the reliability of Thurstone scales with slightly fewer items. The biggest drawback of Thurstone scaling, however, is the amount of effort required. The necessity for administration to a group of judges, totally separate from the administration to scale respondents, is enough to tip the balance, for most researchers, in favor of the Likert scaling method.

SUMMARY OF THURSTONE SCALING PROCEDURES

1. Identify the attitudinal object; delimit it quite specifically.
2. Develop a pool of about 50 belief items about the attitudinal object. Try to ensure that these range from extremely unfavorable to extremely favorable toward the attitudinal object. Neutral items must be included.
3. Obtain scale values for all items. These result from averaging (median is preferable to mean) the “favorableness” judgments of 10 or more judges, using a rating form such as is found in Figure 4.1. Make sure judges rate the relative favorableness (positiveness) of each item rather than indicating their own agreement or disagreement.
4. Compute a measure of dispersion of the judges’ ratings for each item (e.g., interquartile range, standard deviation, or average deviation). Eliminate items on which the judges don’t agree.
5. Select about 20 approximately equidistant items (based on scale values).
6. Randomize their order and print them in an administration form, as found in Figure 4.2.
7. Instruct respondents (those whose attitudes are to be measured) to mark the items with which they agree.
8. Average the scale values of all items marked to obtain an attitude score for each respondent.

In 1944 Louis Guttman introduced a scaling procedure designed to produce scales that are strictly unidimensional. When applied to attitude measurement, Guttman scales consist of opinion items similar to those in Likert and Thurstone scales. Guttman scale items are arranged by degree of positiveness or favorableness, just as are Thurstone scale items. What makes Guttman scaling unique is the extreme stress on unidimensionality. A respondent whose score on a Guttman scale places him at a particular point on the attitude continuum must agree with all items below (less positive than) his own scale position and must disagree with all items above his scale position. This characteristic of Guttman scaling is called reproducibility. Based upon a respondent’s scale score, his responses to all scale items are reproducible.

An example will best serve to illustrate this unique characteristic of Guttman scales. Below are five opinion items about abortion, arranged in order of positiveness:

1. Abortion is acceptable under any circumstances
2. Abortion is an acceptable mechanism for family planning
3. Abortion is acceptable in cases of rape
4. Abortion is acceptable if the fetus is found to be seriously malformed
5. Abortion is acceptable if the mother’s life is in danger

According to the Guttman scaling model, a respondent agreeing with item number 1 should agree with all items below number 1. A respondent who disagrees with item number 1 but agrees with item number 2 should also agree with items 3, 4, 5, and so forth. Whereas in Thurstone scaling items are graduated in favorableness based upon average judgments, in Guttman scaling there must be total agreement (among respondents) regarding the ordering of statements. Absolute enforcement of this criterion is virtually impossible in the attitude domain, especially over large numbers of respondents. A statistical coefficient of reproducibility is used to determine the extent of adherence to the reproducibility standard in each Guttman scale.
CONSTRUCTING GUTTMAN SCALES

Opinion items thought to be graduated in favorableness toward the attitudinal object are formulated and administered to respondents, generally using a simple "agree"-"disagree" response format. A response matrix of items by respondents is thus generated (see Table 5.1). For each respondent a total scale score is calculated, based upon the number of "agree" responses. For each item, the proportion of respondents making the point-getting ("agree") response is calculated. These figures are listed as marginal statistics in the item response matrix (Table 5.1).

Next, both items and respondents are rearranged in order of magnitude (Table 5.2). The transformed matrix exhibits a triangular pattern of positive item responses characteristic in Guttman scale analysis. This matrix facilitates the identification of items that are identical or nearly identical in "difficulty" and items with inconsistent response patterns, as judged by Guttman scaling criteria. This matrix is also used in the assessment of scale reproducibility.

TESTING FOR REPRODUCIBILITY

Guttman proposed a simple index of scale reproducibility (Rep): the proportion of responses following the prescribed response pattern, or

\[ \text{Rep} = 1 - \frac{\text{total number of errors}}{\text{total number of responses}} \]

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Proportion .50 .25 .63 .38 .88

The resultant coefficient indicates the proportion of all item responses that are "reproducible" from knowledge of total scale scores, across respondents. Since it is unrealistic to expect attitude (or other affective) scales to be perfectly reproducible, Guttman suggested .90 as the minimum acceptable level of reproducibility.

By Guttman's error-counting method, the data in Table 5.2 would yield a coefficient of reproducibility of .94, with 33 of 35 responses following the prescribed pattern. Guttman would assign one error for the response pattern of respondent C (11011), maintaining that it "should" be 11111, and one error for respondent F (00101), which should be 00111. Guttman's error-counting procedure, however, has been challenged, and alternate error-counting procedures have been suggested. One of the most popular alternate procedures is that of Goodenough and Edwards (Edwards, 1957). According to this procedure the number of response errors for respondent C would be calculated as two, with the "correct" response pattern being 01111. Likewise, the pattern for respondent F should be 00011, also resulting in two errors. The resultant coefficient of reproducibility is .89 (31 + 35). As you can readily see, the method of error counting selected has a dramatic impact upon the resultant coefficient of reproducibility and upon the commensurate judgment of scale quality.

To further confound the interpretation of reproducibility, it has been noted that scale reproducibility is, in part, dependent upon item response distributions, with items of moderate difficulty having the greatest potential for error. Thus scales with an excess of extreme items (items with
ASSessment of Guttman ScaLeS in Attitude Measurement

Hi as a Relative Index of this Difference:

In order to resolve this difficulty, Edwards (1957) proposed confirming which most respondents agree on which most respondents disagree.

The problem of attitude measurement is to construct scales covering all the attitudinal domains. Constructing such scales is not easy and requires careful consideration of the domain of the attitudes and the structure of the items.

In order to measure attitudes, the Guttman scaling technique (also known as GSA) was developed by Guttman himself. The technique relies on the assumption that there is a natural order to the items, and that respondents will agree more with items that are consistent with their overall attitude.

The Guttman scale is constructed by ordering items from least to most important, and then assigning a score to each respondent based on the number of items they agree with. The score is then used to place the respondent on a spectrum ranging from the most extreme to the least extreme positions on the attitude domain.

Guttman scales are useful for measuring attitudes because they provide a clear and unambiguous measure of an individual's position on the attitude scale. They are also easy to administer and interpret, and can be used to compare attitudes across different populations.

However, Guttman scales have limitations. They are sensitive to the order of the items, and can be influenced by the wording of the items. Additionally, the technique assumes that there is a natural order to the items, which may not always be the case.

Despite these limitations, Guttman scales remain a popular method for measuring attitudes, and continue to be used in a variety of settings.

For more detailed description of procedures for constructing and evalu-