INTER-JUDGE AGREEMENT AND INTRA-JUDGE CONSISTENCY IN JUDGING THURSTONE ATTITUDE-SCALE ITEMS

by

Charles Bates Jr.

B. S., Kansas State College
of Agriculture and Applied Science, 1954

A THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Psychology

KANSAS STATE COLLEGE
OF AGRICULTURE AND APPLIED SCIENCE

1956

TABLE OF CONTENTS

INTRODUCTION

Psychology is unique among the sciences in that from its outset as an experimental science a separate area of inquiry has existed dealing with measurement. Unlike other sciences, where concern with measurement has followed development of theory, psychology has maintained a persistent and healthy interest in measurement as such. The early work in measurement was concerned primarily with sensory functions in the area generally referred to as "psychophysics". Naturally, when investigators in other areas of psychology wished to measure or scale some phenomenon, they turned for their models to these previously successful psychophysical methods.

One of the first to adapt psychophysical methods to attitude scaling was Thurstone (9). He wished to develop a method for scaling attitudes to aid not only applied psychologists, but to advance theoretical considerations as well.

The first step in utilizing the Thurstone method of attitude scaling is to select a large number of items pertinent to the attitude area in question. These statement are selected to represent all points on the continuum, from extreme favorableness to extreme unfavorableness. After the items are selected, a large number of judges is asked to sort the items into a fixed number of categories (usually 9, 11, or 22) arranged in order of favorableness. From these categories, items are selected for the final scale. This is done by assigning numerical values to the categories and then determining the median scale value for each item. To weed out ambiguous items, the inter-quartile range of each item is computed, the assumption being that if a statement is unambiguous in meaning, the judges will tend to place it in one or a few adjacent categories and its Q value will be small. If, on the other hand, the statement is ambiguous, it will be placed in a large number of categories.

After such a scale is constructed, it is used to describe the positions of individual respondents on the attitude continuum. The individual is asked to check the items with which he agrees, and the median value of the items he checks is considered to represent his attitude position. The value of Thurstone's method must stand unchallenged for the contribution it has made to the field of attitude measurement.

Nevertheless, the utility of the method has sometimes been criticized. Likert (7) long ago reported that responses to his simpler scaling method correlated highly with responses to a Thurstone scale, and that the development of the scale involved much less work. More recently, Edwards (1) has subjected scales constructed by the Thurstone method of analysis by the Likert method. He finds that the middle, or "neutral", items in an 11-item Thurstone scale were non-differentiating. That is, a subject checking such an item could fall at either end of the continuum defined by the Likert method. Edwards called these middle items "catchall" categories. In a later investigation, (2), Edwards found that the use of the inter-quartile range in testing for ambiguity did not aid materially in selecting discriminating items. He also found that the inclusion of neutral items tends to lower the reliability of the scale. Other investigators have been concerned with the effect of the original judge's attitude on his judgment of the statements. Hovland and Sherif (5) found that judges who had strong feelings about the subject matter rated tended to see issues as polarized, and, for that reason, piled up statements in the extreme categories. They also found that when judges were not instructed to sort the items into a set number of categories, those judges with strong opinions used a fewer number of categories than did judges with moderate feelings.

More penetrating questions than these may be raised about the efficacy of Thurstone scales or, for that matter, about most contemporary phychological scales. One may ask, for example, whether different judges would rank the items in the same order, and whether a particular individual's judgments will be consistent over a series of replicated, independent comparisons of the same pair of items. Inter-judge disagreement and intra-judge inconsistency and intransitiveness would point to multidimensionality in the scale, and would raise serious questions about the meaning of an individual's score on such a scale.

METHOD

Eleven items were selected from the Hinckley _Scale of Attitudes toward Negroes_ (4), representing all points on the continuum, from extremely anti-Negro to extremely pro-Negro. The subjects were first asked to rank these items in terms of their favorableness to the Negro. It was emphasized in the verbal instructions that the subject was to protect himself into the role of the Negro in making these judgments. The written directions given the subjects were as follows:

> We are interested in finding out how people judge and compare statements made by other people. In order to do this we are asking classes of students to judge a number of different statements. The statements that we will be using today represent varying degrees of opinion about Negroes.
>
> The following statements are presented in a completely random, chance order. Please rank order the statements in terms of how favorable they are to the social position of the Negro. Place the figure 1 before the statement that seems to you to be the most favorable to the social position of the Negro. Place the figure 2 before the next most favorable statement, and so on, 3, 4, 5, 6, 7, 8, 9, 10, and 11, until you have ranked all eleven statements.

After this ranking was completed, the items were then presented randomly to the subject in groups of three (triads) and he was asked to rank each item of the triad in terms of its favorableness to the Negro. To utilize all possible triads from a total of eleven items, 165 separate trials would be required. This was felt to be too much work to require of a subject in one session. Therefore, triads were selected in such a way that each item was compared five times with every item that fell with six scale points of it on the Hinckley scale. This resulted in a total of 77 triads. The eleven Hinckley items and a sample triad are reproduced in Appendix A.

Two groups of subjects were used. For Group A (N=33) the material was presented in group form. The subjects first ranked the eleven items presented in random order on a mimeographed sheet. Following this, the triads were flashed on a screen by means of an opaque projector, and the subjects were asked to make their responses on prepared answer sheets. The first ten triads were presented for one minute and the exposure times was gradually decreased as the subjects gained experience with the task, so that for the last ten triads the exposure time was only 20 seconds. The entire procedure required approximately one hour. This method had two disadvantages; first, it made it impossible for the subjects to express any feelings they had while ranking the items. Secondly, it forced some subjects to speed up their responses to some triads and others to wait after completing certain triads. For these reasons it was decided to present the stimuli individually to a second group.

For Group B (N=47) individual presentation of triads the general procedure was the same as for Group A except that the subjects were allowed to work at their own pace, and were encouraged to express their feelings as the judgments progressed. In addition, the time required to rank the items in each triad was recorded with a standard stop watch.

The subjects used in the group presentation (Group A) were taken from a recitation section of "Biology in Relation tc Man" during the 1954 summer session. This class was largely made up of female elementary school teachers whose ages ranged from 18 to 40 years. Group B, the individual presentation group, was drawn from General Psychology classes at Kansas State College and was composed largely of freshmen. The age range here was from 18 to the middle 20's.

RESULTS

The first step in analyzing the data was to determine the mean value assigned to each of the eleven items by our subjects. This was done by summing the values assigned to each item and dividing by the number of subjects rating that item. Computation of these mean values provided a rank order of the items identical with that obtained by Hinckley's scaling method.

However, while this ordering obtained from the mean rankings was identical with the Hinckley ranking, only one subject in the entire group reproduced Hinckley's rank order. Table 1 below gives the mean rank and the range of ranks to which each item was assigned.

Table 1. Mean rank and range assigned to the eleven Hinckley items.

Items:	1	2	3	4	5	6	7	8	9	10	11
Rank	1.8	2.1	4.1	4.5	5.4	6.6	6.7	7.2	7.8	9.8	10.3
Range	1-4	1-6	1-9	2-9	3-11	2-9	2-11	5-11	4-11	6-11	6-11

The wide range of positions assigned by different subjects to the same items is reflected in the range of correlations between the subject's ranking of the items and the Hinckley ranking. Kendall's tau (6) was used to express the relationship of the subjects' ranking with the Hinckley ordering. This method of rank correlation requires only the assumption that the data fall into an ordinal scale. *Tau* is a function of the minimum number of the minimum number of interchanges between items required to transform one ranking into the other. It can be thought of as a "coefficient of disarray". Perfect agreement between ranks would yield a Tau of +1.00, perfect disagreement between ranks would yield a Tau of -1.00 and values increasing from -1.00 to +1.00 would indicate more and more agreement between ranks.

When we apply this method to these data, we find that variation among individuals is great. Taus for the eleven items ranged from +.16 to +1.00 with a median of +.71.

If Edwards (1) is correct in saying that the middle items on the Thurstone scale were "non-discriminating", then a measure of the relationship between a subject's ranking and the Hinckley ratings on the middle seven items should be lower than that for the entire set of eleven items. This was indeed the case. Rank correlation coefficients between the subjects' rankings and the Hinckley order of the neutral items ranged from -.33 to +1.00 with a median of +.43.

The above results indicate that agreement between individual subjects and the Hinckley ranking was not particularly high, especially on the middle seven items. The next question to ask then is: "How well do our judges agree among themselves in ranking the seven items?" In other words, a measure of inter-judge consistency was needed. Again it was felt necessary to use a method where the assumptions made about the data would be at a minimum. Such a method was found in the Coefficient of Concordance, or W, as developed by Kendall (6) p 81. This method requires only that the data fall into an ordinal scale. W ranges from 0 to 1.00, rather than from -1.00 to ≠1.00 as in most correlational techniques. As Kendall says:

> When more than two observors are involved agreement and disagreement are not symmetrical opposites. M observors may all agree but they cannot all disagree completely in the sense here considered. If, of three observors P, Q, and R, P disagrees with Q on a comparison and also disagrees with R, then Q and R must agree. (6) p. 81.

A separate W was computed for both Group A and Group B, the values of which were .68 and .71 respectively when computed on an eleven item basis. Since these values did not differ significantly, they were combined to give a total W of .66. To determine just how much of this inter-judge agreement was contributed by items at the extremes of the scale, a W was computed on the middle seven items. This, again, was done separately for the two groups. The W for Group A was .29 and for Group B .37. These two W's were again combined to give a total W for the middle seven items of .32.

A more basic question concerns intra-subject consistency and transitivity of judgments. That is, to what degree is a respondent reliable in reproducing his ordering of the stimuli. In order to test these assumptions the method of triads was used. Sihce each triad was judged 5 times, a perfectly consistent triad would be one where the subject reproduced his ordering every time. A triad that was split 4-1, or, in other words, had one judgment in disagreement with the others was defined as error, while a 3-2, split was thought to be a clearly inconsistent triad.

Two subjects out of a total number of 80 had no inconsistencies in their triad rankings but did have several pairs of items scored as error. The range of inconsistency was 0 to 16, with a median of 4. When error and inconsistency were grouped together, the range was found to be 5 to 33 with a median of 13. Of the 45 pairs of items, every pair was inconsistently judged by at least one subject with adjacent items on the Hinckley scale and neutral items more often inconsistently judged.

For subjects with great inconsistency, intransitivity is impossible to assess. However, for three subjects clear instances of intransitivity were found, e.g. subject 108, who judged item 7 more favorable than item 8 four times in five, judged 8 as more favorable than 11 four times in five, and judged item 11 as more favorable than 7 four times in five.

That this inconsistency did not result from simple carelessness is evidenced by analysis of the time required to make triad judgments. The judgment times for all triads by a subject were divided into quartiles and the time required to judge triads containing only consistently judged pairs was compared with the time required to judge triads containing pairs of items where error might have occurred and those containing inconsistently judged pairs. As is shown in Table 2, triads containing inconsistent pairs required a much longer judgment time than those containing pairs with possible error, and the latter required a significantly longer judgment time than did triads containing only consistently judged pairs.

Table 2. Decision time quartile required by triads expressed in seconds.

	Time required for decision fell in			
	short Q1	Q2	Q3	long Q4
Inconsistent Pairs	125	182	206	212
Possible Error	172	161	182	175
Consistent Pairs	321	240	196	148

DISCUSSION

What must one assume when he assesses a person's attitude by asking him to endorse those items on a Thurstone scale that best represent his position on the continuum in question? First, it must be assumed that the respondent will be consistent in his judgment of the items, i.e. that if he selects item A over item B in one instance, he will order the stimuli in a like manner the second time. Not only must one assume that the subject would be consistent in his judgments, but one must also assume transitivity in his ordering of the stimuli. For if item A is placed over item B, and item B is preferred to item C, then A must also be ranked over C, otherwise no single continuum exists.

These assumptions are not unique to Thurstone-type scales, but are implied by all psychological scales, and must be met if any conclusions are to be drawn about an individual's position on a continuum.

In this case it was found that the data do not meet the requirements of consistency and transitivity, nor do different individuals agree in their ranking of the items.

When such a condition occurs, the social scientist has one of two alternatives open to him. He may use a system of collecting data that does not impose a rank order on the data but which permits the discovery of underlying dimensions of the attribute, or he may call the deviations "error" and proceed as though the original attribute were really unidimensional. Coombs (3) has termed this the "dilemma of the social scientist". He says:

> The social scientist is faced by his dilemma when he chooses between mapping his data into a simple order and asking his data whether they satisfy a simple order. By selecting a strong enough system, the social scientist can always succeed in constructing a unidimensional scale of measurement, commonly an interval scale, thus requiring a portion of the data to be classified as error. By not requiring a strong system, the social scientist permits the data to determine whether a simple, unidimensional solution is adequate. (p. 486).

The Thurstone method would seem to choose the second horn of the dilemma, that of assuming error. An ordinal scale is forced upon the data by the method of scale construction. Consequently, the assumption must be made that any deviation from consistency and transitivity on the part of the respondent, and any differences between respondents, must be called error. As this study indicates, the magnitude of the "error" makes the technique next to useless in assessing the attitudes of any particular individual.

This does not, however, preclude the use of such a scale for group comparisons, although it could be argued that even here it is at best a rough, low-level measure. The justification for this contention lies in the reasons why subjects are intrasitive and inconsistent in their judgment of the items.

Among these reasons is the fact that one item may not have the same meaning for all judges. A particularly good example is i item 6 of the scale: "So great is the social range between the highly educated Negro and the 'nigger' that the race as a whole cannot be assigned to any one notch in the social scale." This may be interpreted as an item emphasizing the importance of attending to individual, rather than racial differences, and, as such, judged at one end of the continuum; or the word "nigger" in the statement may cause it to be placed at the unfavorable end of the scale. This specific criticism applies to many of the middle or neutral items in the scale; however, cases of disagreement about meaning were found involving items at the extremes of the scale as well.

Furthermore even if the items had the same meaning for all judges, judges might still disagree about the degree of favorableness or unfavorableness of an item. An example of this would be item 9: "Although the Negro is rather inferior mentally he has a deeper and fuller religious life than the white man, and thus has an emphatic claim upon our social approval." All of the respondents might agree that the Negro has a deep religious experience but this might raise or lower the judges' opinion of the Negro, depending on the religious orientation of the judge.

Similar problems of unidimensionality, of ambiguity, and of individual differences in estimation of the degree of favorableness or unfavorableness arise with most psychological tests. Such scales may be pragmatically useful in distinguishing groups whose members' responses cluster at opposite ends of the scale. However, it may be argued that psychologists ultimately must concern themselves with the processes that determine an individual's responses to each of the items, and with analyzing the underlying dimensions which combine to form the complex attributes currently measured by psychologists.

SUMMARY

The Hinckley Scale of Attitudes toward Negroes was given to a group of students and the subjects were asked to rank the eleven items according to their favorableness to the Negro. They were then given the items in groups of three, and were asked to rank the items in each triad according to the same criterion. The triads were so constructed that each item was compared five times with every item within six scale points of it.

When the mean rank assigned to each item was computed, it was found that the ranking of the items agreed exactly with the original Hinckley scale. However, only one of the 80 subjects duplicated the Hinckley order.in his ranking. Coefficients of disarray, or Tau's, were computed between each individual's ranking of the eleven items and the Hinckley ranking. Correlations ranged from +.16 to +1.00 with a median of +.71. When Tau's were computed on the middle or neutral items only, they ranged from -.33 to +1.00 with a median of +.43. Inter-judge agreement was examined with Kendall's coefficient of concordance. The coefficient, when computed for all eleven items, was +.66. It was +.30 when computed for the middle seven items only.

Only two subjects were consistent on all 45 triads, and great individual differences in consistency were found. Inconsistency occurred most frequently in the middle seven items. There was also a pronounced tendency for inconsistent triads to take longer judgment times than the consistent triads.

The evidence reported here supports the position that the assumptions underlying the Thurstone scaling technique imposes conditions on the data by assumption that are not inherent in the data. This raises serious questions about its ultimate utility as a technique of psychological measurement. To the extent that other measurement techniques require similar assumptions, they are also subject to question.

LITERATURE CITED

1. Edwards, Allen L. "A Critique of Neutral Items in Attitude Scales Constructed by the Method of Equal Appearing Intervals". Psychological Review, May, 1946, 53:159-169.

2. Edwards, Allen L. and Franklin P. Kilpatrick. "A Technique for the Construction of Attitude Scales." Journal of Applied Psychology, 1948, 32:374-384.

3. Festinger, Leon and Daniel Katz. Research Methods in the Behavioral Sciences. New York: The Dryden Press, Inc. 1953.

4. Hinckley, E. D. "The Influence of Individual Opinion on Construction of an Attitude Scale." Journal of Social Psychology, 1932, 3:283-296.

5. Hovland, C. I. and M. Sherif. "Judgmental Phenomena and Scales of Attitude Measurement I." Journal of Abnormal and Social Psychology, 1952, 47:822-832.

6. Kendall, M. G. Rank Correlation Methods. London: Charles Griffin and Company, Limited, 1948.

7. Likert, R. "A Technique for the Measurement of Attitudes." Archives of Psychology, 1932, No. 140.

8. Sherif, M. and C. I. Hovland. "Judgmental Phenomena and Scales of Attitude Measurement II." Journal of Abnormal and Social Psychology, 1953, 38:135-141.

9. Thurstone, L. L. and E. J. Chove. The Measurement of Attitude. Chicago: University of Chicago Press, 1929.

APPENDIX

The Hinckley Scale of Attitudes toward Negroes

1. I place the Negro on the same social basis as I would a mule.
2. No Negro has the slightest right to resent, or even question, the illegal killing of one of his race.
3. Negroes should not be allowed to mingle with whites in any way.
4. The white man has clearly shown the dominance of his race and should continue to exercise his power of leadership over the Negro.
5. Negroes must undergo many year of civilization before they may be said to reach the social level of the whites.
6. So great is the social range between the highly educated Negro and the "nigger" that the race as a whole cannot be assigned to any one notch in the social scale.
7. The Negro should not be condemned forever to a lower place than the whites, but to a different place.
8. The Negro should have the advantages of all social benefits of the white man but should be limited to his own race in the practice thereof.
9. Although the Negro is rather inferior mentally, he has a fuller and deeper religious life than the white man, and thus has an emphatic claim upon our social approval.
10. The Negro should be given the same educational opportunities as the white man.
11. I believe that the Negro is entitled to the same social privileges as the white man.

Triad Number 36

A. The Negro should not be condemned forever to a lower place than the whites, but to a different place.

B. Although the Negro is rather inferior mentally, he has a fuller and deeper religious life than the white man, and thus has an emphatic claim upon our social approval.

C. So great is the social range between the highly educated Negro and the "nigger" that the race as a whole cannot be assigned to any one notch on the social scale.

ACKNOWLEDGMENT

I wish to take this opportunity to thank Dr. Walter H. Crockett for his generous advice and counsel in the preparation of this thesis.

INTER-JUDGE AGREEMENT AND INTRA-JUDGE CONSISTENCY IN JUDGING THURSTONE ATTITUDE-SCALE ITEMS

by

Charles Bates Jr.

B. S., Kansas State College
of Agriculture and Applied Science, 1954

AN ABSTRACT OF A THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Psychology

KANSAS STATE COLLEGE
OF AGRICULTURE AND APPLIED SCIENCE

1956

The Hinckley Scale of Attitudes toward Negroes was given to a group of students and the subjects were asked to rank the eleven items according to their favorableness to the Negro. They were then given the items in groups of three and were asked to rank the items in each triad according to the same criterion. The triads were so constructed that each item was compared five times with every item within six scale points of it.

When the mean rank assigned to each item was computed, it was found that the ranking of the items agreed exactly with the original Hinckley scale. However, only one of the 80 subjects duplicated the Hinckley order in his ranking. Coefficients of disarray, or Tau's were computed between each individual's ranking of the eleven items and the Hinckley ranking. Correlations ranged from +.16 to +1.00 with a median of +.71. When Tau's were computed on the middle or neutral items only, they ranged from -.33 to +1.00 with a median of +.43. Inter-judge agreement was examined with Kendall's coefficient of concordance. The coefficient, when computed for all eleven items, was +.66. It was +.30 when computed for the middle seven items only.

Only two subjects were consistent on all 45 triads, and great individual differences in consistency were found. Inconsistency occurred most frequently in the middle seven items. There was also a pronounced tendency for inconsistent triads to take longer judgment times than the consistent triads.

The evidence reported here supports the position that the assumptions underlying the Thurstone scaling technique are not met by the data. That is, the Thurstone technique imposes conditions on the data by assumption that are not inherent in the data. This raises serious questions about its ultimate utility as a technique of psychological measurement. To the extent that other measurement techniques require similar assumptions, they are also subject to question.