PSYCHOMETRIKA—VOL. 21, NO. 1
MARCH, 1956

A METHOD OF SCALOGRAM ANALYSIS USING SUMMARY STATISTICS*

BERT F. GREEN
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

A method of Guttman scalogram analysis is presented that does not involve sorting and rearranging the entries in the item response matrix. The method requires dichotomous items. Formulas are presented for estimating the reproducibility of the scale and estimating the expected value of the chance reproducibility. An index of consistency is suggested for evaluating the reproducibility. An illustrative example is presented in detail. The logical basis of the method is discussed. Finally, several methods are suggested for dealing with non-dichotomous items.

Guttman's scaling method, known as scalogram analysis (4), has become popular among social scientists. However, current techniques for scalogram analysis are cumbersome. They all deal directly with the raw data in the form of an item response matrix that has a row for each respondent and a column for each item response category. An entry in the matrix indicates whether a particular respondent gave a particular item response. Various procedures have been described for rearranging the rows and columns of the item response matrix, as well as for combining response categories, so that a response "parallelogram" is achieved with few deviations. Suchman (12) described the scalogram board procedure in which the response matrix is represented by buckshot placed in small indentations in a set of removable slats. The sorting is accomplished by interchanging these slats in their frame. Methods for tabulating the response matrix on IBM equipment have been described by Noland (10), Ford (2), and Kahn and Bodine (6). Paper and pencil methods have been described by Guttman (3) and Marder (8).

These techniques are not automatic, but require keen judgment concerning the kind of sorting likely to pay off. Furthermore, the techniques are cumbersome since each attempts to evaluate the complete raw data matrix without the aid of summary statistics. For large numbers of respondents, the task is overwhelming. Moreover, it is difficult to deal with more than 10-20 items with these procedures. [A method, called H-technique, for combining items before making the scalogram analysis has been reported by Stouffer, Borgatta, Hays, and Henry (11)].

*Lois K. Anderson assisted the author materially in the many computations required for this paper. The research reported in this paper was supported in part by the Department of Economics and Social Sciences at M.I.T. and in part, jointly, by the Army, Navy and Air Force under contract with the Massachusetts Institute of Technology.

The purpose of this paper is to present a relatively simple method of scalogram analysis in which summary statistics are used to compute a close approximation to the scale *rep*. (In this paper "rep" is used for "reproducibility.") In this method, which requires dichotomous items, there is no limitation on the number of respondents; its application to large numbers of items is relatively easy. The method is particularly well-suited to punched-card techniques of processing data, since one must merely count the number of respondents who gave the positive response to each item, and the number of respondents giving certain specified combinations of responses. Obtaining these summary statistics is a simple, routine, completely objective matter.

In a sense, the method proposed here removes scalogram analysis from the list of subjective, slightly mystical techniques available only to experienced practitioners and places it on the list of objective methods available to any statistical clerk. The method also substantially reduces the time required for analysis. It gains these advantages at the expense of providing only an approximation to the "true" rep. Certain high-order scale errors are ignored. However the approximation appears to be a very close one.

The Method

All items must be dichotomous. In the mathematical notation we will let k be the number of items, N be the number of respondents, i be a subscript referring to item i (where the items are in any arbitrary order), and g be a subscript referring to item g in rank order.

Step 1. Designate the positive response to each item by referring to the item content. The positive response designations should be consistent with the investigator's hypothesis concerning the dimension being scaled.

Step 2. For each item tabulate n_i , the number of respondents who gave the positive response to the item.

Step 3. Arrange the items in rank order according to their popularities, (n_i/N), with the *least* popular item getting rank k, and the *most* popular item getting rank *1*. If there are any ties, adopt an arbitrary order.

Step 4. Tabulate $n_{g+1,\bar{g}}$ for $g = 1, 2, \cdots, k - 1$. This is the number of respondents who gave the positive response to item $g + 1$ and the negative response to item g. If it is easier to tabulate $n_{g+1,g}$ or $n_{\overline{g+1},g}$, then the following identities can be used:

$$n_{i\bar{j}} = n_i - n_{ij} \ ;$$

$$n_{\bar{i}j} = n_{i\bar{j}} + n_j - n_i \ .$$

Step 5. Use either of the following two methods for estimating the rep.

A. Tabulate $n_{g+2,g+1,\bar{g},\bar{g}-1}$ for $g = 2, 3, \cdots, k - 2$. This is the number of respondents who gave the positive response to item $g + 2$, *and* the positive response to item $g + 1$, *and* the negative response to item g, and the negative

response to item $g - 1$. Estimate the rep from the formula

$$\text{Rep}_A = 1 - \frac{1}{Nk}\sum_{g=1}^{k-1} n_{g+1,\bar{g}} - \frac{1}{Nk}\sum_{g=2}^{k-2} n_{g+2,g+1,\bar{g},\overline{g-1}} .$$

B. Tabulate $n_{g+2,\bar{g}}$ for $g = 1, 2, \cdots, k - 2$. This is the number of respondents who gave the positive response to item $g + 2$ *and* the negative response to item g. Estimate the rep from the formula

$$\text{Rep}_B = 1 - \frac{1}{Nk}\sum_{g=1}^{k-1} n_{g+1,\bar{g}} - \frac{1}{N^2k}\sum_{g=2}^{k-2} n_{g+2,\bar{g}} n_{g+1,\overline{g-1}} .$$

Rep_A and Rep_B should yield very similar estimates. The choice depends primarily on the relative ease of the alternative tabulations. Rep_A has the advantage that it is known to be an overestimate of the true (sample) rep.

The standard error of either Rep_A or Rep_B is approximately given by

$$\hat{\sigma}_{\text{Rep}} \approx \sqrt{\frac{(1 - \text{Rep})(\text{Rep})}{Nk}} .$$

Step 6. (*Optional*). Estimate the rep that would be expected by chance if the items had their observed popularities but were mutually independent. The rep of independent items is estimated by the formula

$$\text{Rep}_I = 1 - \frac{1}{N^2k}\sum_{g=1}^{k-1} n_{g+1} n_{\bar{g}} - \frac{1}{N^4k}\sum_{g=2}^{k-2} n_{g+2} n_{g+1} n_{\bar{g}} n_{\overline{g-1}} .$$

(Note that $n_{\bar{g}} = N - n_g$.)

Compute the *Index of Consistency*,

$$I = \frac{\text{Rep} - \text{Rep}_I}{1 - \text{Rep}_I} ,$$

where rep is either Rep_A or Rep_B . The index I will be unity if the items are perfectly scalable and has an expected value of zero when the items are independent. If the items show some negative correlation in the sample, I will be negative. If desired, label the set of items "scalable" if I is greater than .50.

Step 7. Give each respondent a scale score that is the number of items to which he gave the positive response.

Illustrative Example

A set of hypothetical data with $N = 20$, and $k = 6$ will be used as an example of the application of the method. The hypothetical data are shown in Table 1. The tabulations for Steps 2, 4, 5*A*, and 5*B* are also shown in Table 1. We have put the items and the respondents in rank order in Table 1 only to provide an easy comparison with the usual sorting techniques. In carrying out the tabulations of Steps 2, 4, and 5, it is not at all necessary that the

TABLE 1

Data and Tabulation for Illustrative Example
(+ = positive response; – = negative response)

Respondents	Items (i): 2	1	3	4	5	6	Scale Scores
	Rank Order (g): 6	5	4	3	2	1	
1	+	+	+	+	+	+	6
2	+	+	+	+	–	+	5
3	+	–	+	+	+	+	5
4	+	+	–	+	+	+	5
5	–	+	+	+	+	+	5
6	+	+	–	–	+	+	4
7	–	+	+	–	+	+	4
8	+	–	+	+	–	+	4
9	–	–	+	+	+	+	4
10	–	+	–	+	+	+	4
11	+	+	+	–	–	–	3
12	–	+	+	–	–	+	3
13	–	–	+	–	+	+	3
14	–	–	–	+	+	+	3
15	–	–	–	+	+	–	2
16	–	–	–	–	+	+	2
17	–	–	–	–	+	+	2
18	–	–	–	–	+	–	1
19	–	–	–	–	–	+	1
20	–	–	–	–	–	–	0
(Step 2) n_g	7	9	10	10	14	16	
(Step 6) $n_{\bar{g}}$	13	11	10	10	6	4	
(Step 4) $n_{g+1,\ \bar{g}}$	–	2	3	4	2	2	
(Step 5A) $n_{g+2,\ g+1,\ \bar{g},\ \overline{g-1}}$	–	–	1	2	0	–	
(Step 5B) $n_{g+2,\ \bar{g}}$	–	–	2	4	4	1	

raw data be arranged with either items or respondents in any particular order. It is quite possible to work directly with the individual response sheets, or with their punched-card equivalents.

From the formulas in Step 5, we compute

$$\text{Rep}_A = 1 - \frac{1}{120}(2 + 2 + 4 + 3 + 2) - \frac{1}{120}(0 + 2 + 1) = .867$$

$$\text{Rep}_B = 1 - \frac{1}{120}(2 + 2 + 4 + 3 + 2) - \frac{1}{2400}(1 \cdot 4 + 4 \cdot 4 + 4 \cdot 2) = .880.$$

(The actual reproducibility is .858. The large discrepancy between this figure and the estimates is due to the small N in this example.) For Step 6

we compute

$$\text{Rep}_I = 1 - \frac{1}{2400}(14 \cdot 4 + 10 \cdot 6 + 10 \cdot 10 + 9 \cdot 10 + 7 \cdot 11)$$

$$- \frac{1}{960{,}000}(10 \cdot 10 \cdot 6 \cdot 4 + 9 \cdot 10 \cdot 10 \cdot 6 + 7 \cdot 9 \cdot 10 \cdot 10) = .826$$

$$I_A = \frac{.867 - .826}{1 - .826} = .236.$$

Since I is less than .50, the set of items is not "scalable." The scale scores are shown in Table 1.

Justification of the Method

Ordering the items. The simplicity of the method of scalogram analysis presented in this paper is due largely to the use of popularity to rank the items. Guttman and his followers have used the order of items that yielded the highest rep. In a large majority of the cases, this order turns out to be the popularity order. In the other cases, the difference in the rep for the "best" (highest rep) order and for the popularity order is very small. Thus almost nothing will be lost and great simplification will be gained by using the popularity order. An arbitrary order may be adopted for tied items.

It is not surprising that the popularity order is usually the "best" order. In a perfect Guttman scale, the rank order of the items *must* correspond with the popularity order. For imperfect data, one would still expect the popularity order to be "best" if the scale errors are independent. Slight inversions might be expected if items had very similar popularities but the effect of these inversions would be small. Very peculiar error patterns would be required to make the popularity order markedly inferior to the "best" order.

Estimating Rep. The formulas for estimating the rep are based on an analysis of the scale errors in a pattern of item responses. In a perfect Guttman scale, the items can be arranged in a rank order so that a person who responds positively to (or endorses, or agrees with) any item also responds positively to all items of lower rank order. For a five-item scale, six ideal response patterns would be possible: [+++++], [−++++], [−−+++], [−−−++], [−−−−+], and [−−−−−]. In each ideal response pattern there is a dividing point, or cut, such that all item responses to the left of the cut are −, and all items responses to the right of the cut are +. The number of scale errors in any other response pattern is determined by placing a cut so that the number of +'s to the left of the cut and the number of −'s to the right of the cut are minimized. All such "misplaced" responses are errors. For example, the pattern [−−++−] would have its cut between items 4 and 3, (items are numbered in *decreasing* order from left to right, i.e.,

5, 4, 3, 2, 1) and one error would be counted. The pattern [− + + − −] could have its cut between items 5 and 4 or at the right of item 1. In either case there would be two errors.

In order to find a rule for counting the errors in any particular response pattern we must consider subpatterns of responses. First, consider a pair of *adjacent* items with the response subpattern (+ −); i.e., the response to the higher ranking item is + and the response to the lower ranking item is −. We would place the cut either to the right or to the left of this pair, and would have one error from the pair in either case. We would not place the cut between the two items, since this would yield two errors. Next, consider the reduced response pattern formed by deleting such a pair from a complete response pattern. Clearly the number of errors in the complete pattern is exactly one more than the number in the reduced pattern, for when we have determined the location of the cut in the reduced pattern, the two deleted items can and must be replaced together on either side of this cut. We may successively reduce a response pattern by eliminating pairs of adjacent items with (+ −) subpatterns, until there are no remaining errors in the reduced pattern. The number of pairs eliminated is then the number of scale errors in the complete response pattern. For example, in the response pattern [− + + − −] we first eliminate the pair (3, 2), (items are numbered 5, 4, 3, 2, 1) leaving (− + −); then we eliminate (4, 1), leaving (−). Hence there are two errors in the original pattern.

In practice the first step is simultaneously to eliminate from the complete response pattern *all* pairs of adjacent items with (+ −) subpatterns. These are first-order errors. Next, we simultaneously eliminate all (+ −) subpatterns from the reduced pattern; these are the second-order errors. We continue with third-, fourth-, and higher-order errors. Note that second-order errors are represented in the complete response pattern by a sub-pattern (+ + − −). The middle two items in this subpattern represent a first-order error, while the other two items represent the second-order error that appears when the first-order error is deleted. That is, second-order errors always straddle first-order errors. In the same way, third-order errors straddle second- and first-order errors, as in (+ + + − − −) or (+ + − + − −). In general, higher-order errors always straddle lower-order errors.

The formula for rep is

$$\text{Rep} = 1 - \frac{E}{Nk}, \tag{1}$$

where E is the total number of errors, N is the number of respondents, and k is the number of items. Then, we have

$$\text{Rep} = 1 - \frac{1}{Nk}\sum(+-) - \frac{1}{Nk}\sum(++--) - \text{terms of higher order.} \tag{2}$$

Now the sum of all adjacent item errors $(+-)$ is

$$n_{2\bar{1}} + n_{3\bar{2}} + \cdots + n_{k,\overline{k-1}} = \sum_{g=1}^{k-1} n_{g+1,\bar{g}} \,. \tag{3}$$

Similarly,

$$\sum (++--) = \sum_{g=2}^{k-2} n_{g+2,g+1,\bar{g},\overline{g-1}} \,. \tag{4}$$

Since the higher-order error patterns occur very infrequently, we will disregard all terms higher than second order. Substituting (3) and (4) in (2) we obtain the formula for Rep_A that we presented in Step 5*A*.

Instead of tabulating $\sum(++--)$, we may estimate this quantity. We shall assume that the two errors represented by the pattern $(++--)$ are independent. That is, the probability of a response pattern $(+\ 0\ -\ 0)$, where 0 symbolizes either response, is independent of the probability of a response pattern $(0\ +\ 0\ -)$. Then the product of these two probabilities is the probability of a response pattern $(++--)$. Under this assumption,

$$\frac{n_{g+2,g+1,\bar{g},\overline{g-1}}}{N} = \frac{n_{g+2,\bar{g}}}{N} \frac{n_{g+1,\overline{g-1}}}{N} \,. \tag{5}$$

Hence,

$$\sum_{g=2}^{k-2} n_{g+2,g+1,\bar{g},\overline{g-1}} = \frac{1}{N} \sum_{g=2}^{k-2} n_{g+2,\bar{g}} n_{g+1,\overline{g-1}} \,. \tag{6}$$

Substituting (3) in (2) and substituting (6) in (4) and the result in (2) we have the formula for Rep_B given in Step 5*B*.

A cruder approximation to rep can be obtained by ignoring the second-order error patterns, i.e., by omitting the last term from Rep_A or Rep_B . This approximation may be satisfactory in some cases, but it can be improved considerably by including the last term.

The formulas for Rep_A and Rep_B have been checked empirically using data published by Suchman (**12**, **13**). Table 2 gives the correct rep and our estimated Rep_A and Rep_B for ten scales of dichotomous items. The average discrepany is .002 for Rep_A and .003 for Rep_B . It should be pointed out that the actual reps are not necessarily those reported by Suchman since we have dichotomized all items and have not rejected any items from the scale. For larger numbers of items, the discrepancies may be slightly larger than those obtained here.

The errors in the proposed estimates of rep occur because terms of sixth and higher orders are neglected. It is possible to estimate the extent of the error by assuming that scale errors are independent. Both empirical and analytic results show that the error is about $(1 - \text{Rep})^3/\text{Rep}$.

The variance of Rep_A or Rep_B can be calculated. The calculations are straightforward but long, and lead to a complicated formula that will not

TABLE 2

Empirical Comparison of $\underline{\text{Rep}_A}$ and $\underline{\text{Rep}_B}$ with Correct Rep					
Scalogram No. and Page Reference to Suchman (12, 13)	k	N	Correct Rep	Rep_A	Rep_B
0. p. 118	12	100	0.895	0.897	0.898
1. p. 124	9	100	0.923	0.924	0.925
2. p. 126	14	100	0.870	0.871	0.876
3. p. 130	9	100	0.890	0.888	0.895
4. p. 134	5	100	0.962	0.962	0.964
5. p. 136	7	100	0.970	0.970	0.970
6. p. 138	7	100	0.929	0.929	0.929
7. p. 140	10	100	0.913	0.917	0.918
8. p. 146	8	100	0.825	0.831	0.829
9. p. 148	6	100	0.943	0.945	0.946

be presented here. A satisfactory approximation, suggested by Guttman (4), is given by the simple formula presented in Step 5 above. The factor of k in the denominator leads to very small sampling variances which tend to give the investigator a false sense of security. It should be noted that errors of unreliability are usually much larger than sampling errors in this situation. Indeed, the errors in our approximations for rep are of the same order of magnitude as the sampling standard deviation.

Chance Rep. One of the major advantages of the method of scalogram analysis described here is the ease with which the chance rep, Rep_I, can be obtained. Rep_I is the rep that would be expected by chance, if the items were actually independent, and is a function of the item popularities, or "marginals." To obtain Rep_I, we note that if the items are independent, then

$$n_{g+1,\bar{g}} = n_{g+1}n_{\bar{g}}/N; \qquad n_{g+2,g+1,\bar{g},\overline{g-1}} = n_{g+2}n_{g+1}n_{\bar{g}}n_{\overline{g-1}}/N^3.$$

Substituting these values in the formula for Rep_A, we obtain the formula for Rep_I shown in Step 6 above.

The hypothesis that Rep_A (or Rep_B) is not significantly larger than Rep_I can be tested by using the variance estimate presented in Step 5. Caution is necessary when such a test yields borderline significance because of the many approximations involved. It should be noted that a significant rep does not necessarily indicate a homogeneous scale. For homogeneity the intercorrelations of the items should be fairly high, as well as significantly nonzero. It is possible to construct an index of homogeneity that will be zero when Rep_A (or Rep_B) = Rep_I, and will be unity when $\text{Rep}_A = 1$. The index

I, presented in Step 6, has these properties. It will be affected very little by changes in the number of items, or in the item popularities. This index is very similar to Loevinger's (7) index of homogeneity and to Menzel's (9) coefficient of scalability, and is proposed only because it is easier to compute in the present circumstance.

According to Guttman (4), a set of items should meet several criteria in order to be considered "scalable," or homogeneous. These criteria were apparently generated from anxiety about the chance rep [Festinger (1)]. With the exception of the requirement about a random pattern of errors, each criterion is related to the problem of avoiding spurious results and achieving homogeneity. It seems natural to replace them by a single criterion concerning I: I should be .50 or more for scalability. This criterion appears to give roughly comparable results to the many criteria used heretofore, and will be helpful to those who desire to create a dichotomy of scales *vs.* nonscales. To this author, it seems preferable to evaluate a scale in terms of an index of consistency that varies along a continuum rather than in terms of an arbitrary dichotomy.

Scoring. Any scoring method may be used with the present method since none of the previous steps depends on the scale scores. The concep of rep refers to the reproduction of an individual's item responses from a knowledge of his scale score. The measure of rep is a measure of the success of this reproduction when the scores are so assigned that the number of errors is minimized. It follows that the logically consistent method of scoring respondents is to compare their responses with the scale types. Each scale type or perfect response pattern is assigned a number, usually the number of positive responses in the response pattern. Then each individual's response pattern is compared with the scale types. An individual's score is the number assigned to the scale type that his responses match with the fewest deviations.

When the scale is perfectly reproducible, the respondent's score as determined by this scale-type-comparison process is equivalent to the number of positive responses that he made. Happily, it has been found [Suchman (12)] that the number of positive responses is very highly correlated with the scale-type-comparison score, when the scale is not perfectly reproducible. It appears that very little precision will be lost in practice by using the simple scoring method of counting positive responses. Although the issue is somewhat academic for small numbers of items, the savings of time and trouble are sizable for large numbers of items.

Non-dichotomous Items

In order to use the present method of analysis with items that have more than two response alternatives, the investigator must dichotomize the items. Several possibilities exist. He may use some logical a priori considerations, perhaps depending on the content of the alternative item responses.

The splits might be determined in part by the popularities of the alternatives. It would be possible to use some method such as that proposed by Jackson (5) for attempting to choose the dichotomies that maximize rep. For example, each respondent could be given a provisional scale score, and each possible dichotomy of each item could be correlated with this score; the dichotomies with highest correlations or covariances would be chosen.

An alternative procedure is to use *all possible dichotomies*. That is, an n-alternative item is replaced by $(n - 1)$ dichotomous items, or pseudo-items. For this procedure, it is necessary to adjust the value of k in the formulas. Wherever k appears as a factor, in conjunction with N, it remains the number of original non-dichotomous items, but in the limits of summation, k becomes the total number of pseudo-items. The pseudo-item trick will provide a crude approximation to the rep of a set of non-dichotomous items. This approximation is always slightly too low, but the value of I should be more accurate since the same bias exists in both Rep_A (or Rep_B) and Rep_I .

REFERENCES

1. Festinger, L. The treatment of qualitative data by "scale analysis." *Psychol. Bull.*, 1947, **44**, 146–161.
2. Ford, R. N. A rapid scoring procedure for scaling attitude questions. *Publ. Opin. Quart.*, 1950, **14**, 507–532.
3. Guttman, L. The Cornell technique for scale and intensity analysis. *Educ. psychol. Measmt.*, 1947, **7**, 247–279.
4. Guttman, L. The basis for scalogram analysis. In S. A. Stouffer *et al.*: Measurement and prediction. Princeton, N. J.: Princeton Univ. Press, 1950. Pp. 60–90.
5. Jackson, J. M. A simple and more rigorous technique for scale analysis. In: A manual of scale analysis, Part II. Montreal: McGill Univ., Mimeo, 1949.
6. Kahn, L. H. and Bodine, A. J. Guttman scale analysis by means of IBM equipment. *Educ. psychol. Measmt.*, 1951, 11, 298–314.
7. Loevinger, Jane. A systematic approach to the construction and evaluation of tests of ability. *Psychol. Monogr.*, 1947, **61**, No. 4.
8. Marder, E. Linear segments: a technique for scalogram analysis. *Publ. Opin. Quart.*, 1952, **16**, 417–431.
9. Menzel, H. A new coefficient for scalogram analysis. *Publ. Opin. Quart.*, 1953, **17**, 268–280.
10. Noland, E. W. Worker attitude and industrial absenteeism: a statistical appraisal. *Amer. sociol. Rev.*, 1945, **10**, 503–510.
11. Stouffer, S. A., Borgatta, E. F., Hays, D. G., and Henry, A. F. A technique for improving cumulative scales. *Publ. Opin. Quart.*, 1952, **16**, 273–291.
12. Suchman, E. A. The scalogram board technique for scale analysis. In S. A. Stouffer *et al.*: Measurement and prediction. Princeton, N. J.: Princeton Univ. Press, 1950. Pp. 91–121.
13. Suchman, E. A. The utility of scalogram analysis. In S. A. Stouffer *et al.*: Measurement and prediction. Princeton, N. J.: Princeton Univ. Press, 1950. Pp. 122–171.

Manuscript received 11/8/54
Revised manuscript received 3/4/55