GUTTMAN SCALE ANALYSIS

David C. Blouin
Department of Experimental Statistics
Louisiana State University

Introduction

Guttman scale analysis may be considered as two inseparable endeavors: (1) to ascertain if a set of stimuli are unidimensional in nature and if so, (2) to rank order respondents along the same continuum based on their reaction to the stimuli. For example, suppose a set of four attitude statements (S1-S4) are constructed such that similarity of content exists (e.g. all concern attitude toward importance of college education) and that the stimuli have an intrinsic rank order (i.e. affirmative response to a statement indicates a more or less favorable attitude than an affirmative response to a second statement). Given such conditions, a Guttman type scale is tenable provided (1) a negative response to the lowest ordered stimulus, let’s say S1, precludes a positive response to S2-S4 and in turn, a negative response to S2 precludes affirmation of S3-S4 while a negative response to S3 precludes affirmation of S4 and (2) knowledge of an individual’s scale score equal to the number of affirmative responses predicts without error his responses to each item in the ordered set. Tables 1 and 2 below present a positive and negative illustration of the above conditions, respectively. Referring to Table 1, not only do the stimuli possess rank order along some continuum but also a scale score specifies an individual’s rank order along the same continuum, that is, an individual with a higher score ranks as high or higher on every statement in the ordered series than an individual with a lower scale score.

Table 1: True Guttman Scale Patterns*

<table>
<thead>
<tr>
<th>Respondent</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

* 0 refers to negative response, 1 refers to positive response, and S1-S4 are ordered such that S4 represents the strongest statement in the set.

Table 2: Guttman Scale Patterns with Non-scale Patterns

<table>
<thead>
<tr>
<th>Respondent</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

* Respondent 4 presents non-scale pattern since based on scale score equal to 3, 2 prediction errors would result, i.e. predict 0 for S4 when in fact 1 and predict 1 for S3 when in fact 0.

The remainder of the discussion centers primarily on the analytical definition of the output of the SAS Guttman procedure. The example output from the SAS USER’S GUIDE, 1979 Edition will be referenced throughout by bracketing the encircled number used to designate output. For example, [7] will refer to the marginal frequencies. The output is presented in three stages. The scalogram and statistics associated with the modified Goodenough technique (Menzel, 1953; Edwards, 1957; Torgerson, 1962; Magnusson, 1966) is presented first. The second and third stages present Proctor’s probabilistic formulation (1970; 1971) for the assignment of class scores (stage 2) and statistics associated with the adequacy of the model (stage 3).

Scalogram and Modified Goodenough Technique

The basic question conforming to the first endeavor as initially stated in the introduction is "Do the observed patterns of responses to the ordered set of stimuli conform to a true Guttman scale?" Crucial to an answer are the ordering of the stimuli and subsequent definitions of true type patterns, error patterns, scale scores, and prediction errors. The resulting statistics are based simply on the knowledge of marginal frequencies, that is, the number of respondents who gave an affirmative response to each item, namely 24, 34, 31, and 69, respectively. The true type patterns, in line with the illustration of Table 1, are thereby set to be T0 = (0 0 0 0), T1 = (0 0 0 1), T2 = (0 0 1 1), T3 = (0 1 1 1), and T4 = (1 1 1 1) with respective scale scores of 0, 1, 2, 3, and 4. All remaining patterns [10] are non-scale or error patterns, all of which possess a scale score equal to the number of affirmative responses. Prediction errors are based only on non-scale patterns and are determined as follows. All error patterns with the same scale score are compared to that true type pattern having that scale score. For example, the error pattern (0 1 0 1) was given by 7 respondents [10]. Based on the scale score of 2 and the true type pattern (0 0 1 1) corresponding to that scale score, two errors in prediction would be made per respondent yielding a total of 14 errors. Therefore, given the presence of non-scale patterns, item (0 1) responses are said to be non-reproducible based simply on the knowledge of scale scores.

These definitions lead to the scalogram output [1] and associated statistics, [2] - [6]. The scalogram is composed of item cells and marginal cells. Item cells give for each scale score the frequency of item (0 1) responses for all patterns with that scale score and the frequency of item (0 1) responses in error. For
example, 19 response patterns or respondents received a scale score of 3 (indicative of the true type pattern (0 1 1 1)) where for Q3, all responses were negative and 8 responses were positive, the latter being in error, denoted by the asterisk.

The marginal cells provide a variety of information. First, with respect to response pattern or respondent totals, the total number of respondents is given, namely (N = 111), along with the number of patterns or frequency of respondents for each scale score (e.g. 30 respondents received a scale score of 2). Notice that for this scale score corresponding to the true type pattern (0 0 1 1), a total of 18 prediction errors would be made, 7 for QC3 and 11 for Q17.

Next, with respect to item (0 1) responses, the information provided is (1) the total frequency of item (0 1) responses, (2) the modal response percentage, (3) the total number of item (0 1) responses in error and (4) the total number of errors. For example, referring to Q17, the total frequency of item (0 1) responses was 77 and 34, respectively. The modal response category being 0 has corresponding percentage of 69%. The number of item (0 1) responses in error was 7 and 14, respectively, yielding a total of 21 errors for Q17 extended to 70 errors across all items.

The remainder of the output, [2] to [6], is related to the scalogram output. The coefficient of reproducibility (CR) measures the predictability of the response patterns based simply on knowledge of the scale scores. The CR is equal to the proportion of correct predictions.

Let K = total number of items (i.e. K = 4)
N = total number of patterns or respondents (i.e. N = 111)
F = total number of prediction errors (i.e. F = 70)

CR = 1 - \( \frac{F}{K \times N} \) (i.e. CR = 0.8423)

Therefore, based on a CR = 0.8423, 374 correct predictions (84.23%) would be made while 70 error predictions (15.77%) would be made. The minimum marginal reproducibility (MMR) however is the CR that would be produced given only the modal response pattern. The modal responses are those given by the majority of the respondents. The modal response proportions as given in the scalogram for QC3, Q17, Q7 and Q23 are P1 = 0.78 (QC3 = 0), P2 = 0.69 (Q17 = 0), P3 = 0.52 (Q7 = 0) and P4 = 0.61 (Q23 = 1). Thus, the modal response pattern becomes (0 0 0 1).

Let Pi = modal response proportion for the ith item where i = 1, 2, ..., K = 4

\[
\text{MMR} = \frac{1}{K} \sum_{i=1}^{K} P_i \quad (\text{i.e. MMR} = 0.6554)
\]

Therefore, based on an MMR = 0.6554, 291 correct predictions (64.34%) would be made while 153 error predictions (35.66%) would be made. In and of itself, the CR is often misleading. This is because the lower bound of the CR is the MMR. Thus, given a CR equal to let's say 0.95, the scalability of the items does not necessarily follow as in the case of a correspondingly high MMR (e.g. 0.90). Thus, two additional measures are (1) the percent improvement (PI), that is, the difference between the CR and the MMR (i.e. PI = 0.1869) and (2) the coefficient of scalability (CS), namely the ratio of the PI to the maximum number of prediction errors possible, that is, 1 - CR. Thus, the CS as discussed by Menzel (1953) is 0.1869 divided by 0.8446 which gives a CS equal to 0.5425. The last of the output useful for item analysis considerations (Magnusson, 1966), consists of the correlation coefficients (i.e., the Pearson product-moment inter-correlations for all pairs of items and each item with the scale scores excluding that item).

Proctor's Probabilistic Formulation

A major justification for Proctor's formulation (1970; 1971) is to supply statistical formality for the assignment of respondent class scores [10] and evaluation of a Guttman scale model, such as reliability indices and tests of goodness-of-fit. Experiential evaluation such as a CR of 0.90 or better being acceptable (Guttman, 1950) and a CS between 0.60 and 0.65 being acceptable (Menzel, 1953) point to the attractiveness of a probability model for evaluation purposes.

The fundamental notions of Proctor's formulation are that individuals do in fact belong to one true Guttman type, yet non-scale patterns result from a "response error process" which intervene to produce an error response. The problem of class score assignment becomes one of determining which true type a respondent is most likely to belong to. The solution is one of conditional probability, that is, given an individual belongs to a particular true type, what is the probability that the individual elicits the observed response pattern.

The conditional probability model depends on two parameters, the misclassification parameter (alpha, "a") and the true proportion of the given true type (theta, "\theta"). The misclassification parameter refers to the intervening response error process and reflects the likelihood of specifying an item (0 1) response to be in error when in fact not. An assumption is made that the misclassification parameter is constant regardless of item and true type.

The notational scheme for model specification is given below.

Let K = number of items (i.e. K = 4)

\[
a = \text{misclassification parameter, the probability of incorrectly predicting an item (0 1) response}
\]

\[
1 - a = \text{the probability of correctly predicting an item (0 1) response}
\]
The ith response pattern, e.g. from [101], X1 = (0000), X2 = (0001), X3 = (0010), etc.

1 = i, 2, ..., I

I = 2^K = total number of possible response patterns (i.e. I = 16)

T_k = the kth true type pattern

K + 1 = total number of true type patterns, i.e. K + 1 = 5 where T_1 = (0000), T_2 = (0001), T_3 = (0011), T_4 = (0111), and T_5 = (1111)

θ_k = true proportion of the kth true type in the population

D_k = the number of item (0 1) responses that must be changed to modify the kth true type to match the ith response pattern with the probability of making an incorrect change equal to "a", i.e. 1 response must be changed to match T_4 = (0111) and X_6 = (0101)

K - D_k = the number of item (0 1) responses that do not need changing to modify the kth true type to match the ith response pattern with the probability of correct classification equal to "1 - a", i.e. 3 responses do not need changing to match T_4 = (0111) and X_6 = (0101)

As an example, the probability of X_6 = (0101) is given by

\[
Pr(X_6) = Pr(X_6 | T_1) + Pr(X_6 | T_2) + Pr(X_6 | T_3) + Pr(X_6 | T_4) + Pr(X_6 | T_5)
\]

\[
= a^2 (1 - a)^2 \theta_1 + a (1 - a)^2 \theta_2 + a^2 (1 - a)^2 \theta_3 + a (1 - a)^2 \theta_4 + \theta_5
\]

The conditional probability of X_6 = (0101) given the true type pattern T_4 = (0111) is given by the following:

\[
Pr(X_6 | T_4) = (1 - a)^3 \theta_4 + a^3 \theta_5
\]

The class score \( t_k \) assigned is the scale score associated with that true type which maximizes the "posterior probability" of the response pattern (Proctor 1970).

\[
t_k = \text{scale score of } k\text{th response pattern}
\]

Let \( t_k = \max \Pr(X_{i}|T_k) \)

Maximum likelihood estimates of the parameters and associated standard errors are obtained using "Fisher's method of scoring" (Proctor, 1971; Rao, 1965). Produced are maximum likelihood estimates of the misclassification parameter, called the "required probability of misclassification" (i.e. \( \alpha^* = 0.16 \)) and the true proportion of the true types (i.e. \( \theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*, \theta_5^* \)). Having the maximum likelihood estimates, the conditional probability, equation (1), and the probability model, equation (2), become

\[
(1) \quad Pr(X_{i}|T_k) = \alpha^* \frac{D_k}{K} (1 - \alpha^*)^K - D_k \theta_k^*
\]

\[
(2) \quad t_k = \sum_{\ell=1}^{K+1} \alpha^* \frac{D_\ell}{K} (1 - \alpha^*)^K - D_\ell \theta_\ell^*
\]

The standard errors of all parameter estimates are based on "Fisher's method of scoring." For example, the standard errors of \( \theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*, \theta_5^* \) are 0.06282, 0.07319, 0.06761, 0.09006, 0.04435, and 0.0254, respectively. Fisher's method of scoring produces \( N \) times the estimated variance-covariance matrix of parameter estimates, called the "information (single observation) matrix" [9]. The method redefines the model where

\[
(1) \quad \alpha^* = \frac{1}{K+1} \text{(i.e. } \alpha^* = 1/6.25088 = 0.16)\]

\[
(2) \quad \theta_{k+1} = 1 - \sum_{\ell=1}^{K} \alpha^* \frac{D_\ell}{K} (1 - \alpha^*)^K - D_\ell \theta_\ell^* \text{(i.e. } \theta_{5+1} = 0.09823)\]

The estimated variance-covariance matrix for parameter estimates is determined by dividing the "information (single observation) matrix" by \( N \). The estimated variance of \( \theta_k^* \) which is not produced can be determined by finding the variance of the linear combination given in equation (2) above and the estimated variance of \( \alpha^* \) can be determined as \( \text{Var}(\theta_k^*)/\text{Var}(\alpha^*) \). The standard errors of all estimates are then found simply by taking the square root of the variances.

The formulation of the probability model allows for the development of statistical criteria of the characteristics of the scale, in
particular, (1) the chi-square test of goodness-of-fit, (2) scale reliability, (3) flat reliability, and (4) the average score. Each of these measures is defined below.

The chi-square test is used to test the null hypothesis that the observed frequency of response patterns equals the expected frequency of the response patterns as specified by the probability model.

Let \( N = \) total frequency of response patterns (i.e. \( N = 111 \))

\( X^2 = \sum_{i=1}^{K} \left( \frac{N_i}{N} - \pi_i \right)^2 \)

\( \text{df} = I - K - 2 \)

Scale reliability (SR, [12]) is the square of the estimated correlation between the assigned scores, \( x_i \), and the true type scores, \( t_\ell \) (Magnusson, 1969).

By definition

\[ \text{Cov}(xt) = E(xt) - E(x) \cdot E(t) \]

\[ \text{Var}(x) = E(x^2) - E^2(x) \]

\[ \text{Var}(t) = E(t^2) - E^2(t) \]

By estimates

\[ E(xt) = \sum_{i=1}^{I} \sum_{\ell=1}^{K+1} x_i t_\ell \Pr(X_i \mid T_\ell) \]

\[ E(x) = \sum_{i=1}^{I} x_i \pi_i \]

\[ E(t) = \sum_{\ell=1}^{K+1} t_\ell \theta_\ell \]

\[ E(x^2) = \sum_{i=1}^{I} x_i^2 \pi_i \]

\[ E(t^2) = \sum_{\ell=1}^{K+1} t_\ell^2 \theta_\ell \]

Therefore

\[ \text{SR} = \frac{\text{Cov}^2(xt)}{\text{Var}(x) \cdot \text{Var}(t)} \]

Flat reliability (FR, [13]) is simply the scale reliability where the distribution of true type patterns in the population is assumed to be uniform, that is \[ 1 \]

\[ \Theta_\ell = \frac{1}{K+1} \]

\[ \text{FR} = \frac{1 - 4\alpha^2 (1 - \alpha^2) K - 1}{K + 2} \]

Finally, the average score (14) is the estimated mean of the true scores as given in equation (3) above.

Conclusion

The scalogram and associated output provide measures of the scalability of the items as a Guttman scale, [2] - [5], and measures for item analysis, [6]. Proctor's formulation however provides the probability basis for model evaluation thereby yielding statistical measures for evaluation purposes. Of interest for comparative purposes is the coefficient of reproducibility that would result from the assignment of class scores from the Proctor option. Based on these class scores [10], the number of prediction errors would be 40 (9%) and the coefficient of reproducibility would be equal to 91%.

References


