Homogeneity Analysis of Test Score Data: A Confrontation with the Latent Trait Approach

Dato N. M. de Gruijter
University of Leyden

In homogeneity analysis, or dual scaling, weights for item categories are obtained that maximize Cronbach's alpha. In this paper these weights are compared with the optimal scoring weights in the latent trait approach. This is done on the basis of data generated according to the two-parameter logistic model. As expected from a theoretical analysis, the homogeneity weights show less variation than the optimal weights of latent trait theory. It is argued that the homogeneity weights should not be used for item selection.

Homogeneity analysis, a data-analysis approach with a long history (see de Leeuw, 1973; Nishisato, 1980), became well-known with Guttman's principal components analysis of the Guttman scale (Guttman, 1950). In homogeneity analysis a common metric for item categories and units of measurement, mostly persons, is obtained. The quantifications of the item categories can be called weights and the quantifications for persons, scores. The weights and scores have symmetrical roles. When a person's score is defined as the average of the category weights endorsed, category weights are proportional to the average score of the persons who choose these categories. This reciprocal relationship is the basis for one of the algorithms for obtaining weights and scores. It is, however, possible to obtain weights without simultaneously deriving the scores.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 4, Fall 1984, pp. 385–390

0146-6216/84/040385-06$1.55

Once a solution is obtained, residual data can be computed and then a second solution can be obtained. This process can continue as long as a significant amount of variation can be explained. When this is done, the multivariate structure of the data is disclosed. However, dimensions other than the first can be spurious. This is the case, for example, in the homogeneity analysis of the essentially unidimensional Guttman scale: A multiplicity of weight vectors is obtained with weights that have a predictable relationship with the weights for the first component (Guttman, 1950). Also, other nonlinear latent trait models can be expected to produce additional components (McDonald & Ahlawat, 1974). These components are not discussed here, but attention is concentrated on the first component.

Lord (1958) demonstrated that the category weights of homogeneity analysis maximize coefficient alpha. This is one of the properties that makes such an analysis attractive, but the question arises how these "maxalpha" weights can be interpreted (McDonald, 1983) and whether they correspond to weights based on latent trait theory. Some preliminary results on this topic for data that conform to the Rasch model are given in Gifi (1980).

This paper reports an analysis for data generated according to the two-parameter logistic model. Only two item categories, correct and incorrect, are involved when there are no missing data, and in order not to complicate matters, it is assumed that this is the case. With binary data, the computational

burden can be diminished. This is done by introducing the item weight, the difference between the weight for correct and the weight for incorrect. Lord (1958) pointed to the fact that item weights can be obtained from a principal components analysis of the inter-item correlation matrix: The max-alpha weights equal the loadings on the first principal component of the correlation matrix divided by the corresponding item standard deviations. The category weights can be constructed from the item weights by stipulating that the average score on each item equals zero. For constant item weight, the weight for the correct category becomes more extreme as the item becomes more difficult.

Optimal Item Weights

Several definitions of best weights are possible. A general treatment of the weighting problem is given by McDonald (1968). In this section two weighting problems, the maximization of alpha and the maximization of reliability, are discussed for the situation with n congeneric measurements $X_i (i = 1, \ldots, n)$. In this situation the elements of the population variance-covariance matrix can be written as:

$$\sigma_{ii} = \beta_i^2 + \phi_i \quad (i = 1, \ldots, n) \quad , \tag{1}$$

$$\sigma_{ij} = \beta_i \beta_j \quad (i \neq j; i, j = 1, \ldots, n) \quad , \tag{2}$$

where β_i is the slope of the regression of measurement i on true score, and ϕ_i is the error variance.

The reliability of the weighted composite $\Sigma w_i X_i$ can be written as:

$$\rho = \frac{\mathbf{w}'(\boldsymbol{\Sigma} - \boldsymbol{\Phi})\mathbf{w}}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}} \quad , \tag{3}$$

where $\mathbf{w}$ is the vector with weights,
$\boldsymbol{\Sigma}$ is the matrix with observed score variances and covariances, and
$\boldsymbol{\Phi}$ is the diagonal matrix with error variances.

The "maxrho" weight vector, the vector with weights maximizing ρ, is given by Overall (1965) and Jöreskog (1971).

Maxrho weights should be distinguished from the maxalpha weights. The latter weights maximize

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\mathbf{w}'\boldsymbol{\Delta}\mathbf{w}}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}}\right) \quad , \tag{4}$$

where $\boldsymbol{\Delta}$ is the diagonal matrix with elements σ_{ii}. Both problems, however, require the maximization of a function:

$$\lambda = \frac{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}}{\mathbf{w}'\mathbb{V}\mathbf{w}} \quad . \tag{5}$$

In the maxalpha problem $\mathbb{V}$ equals $\boldsymbol{\Delta}$ and in the maxrho problem $\mathbb{V}$ equals $\boldsymbol{\Phi}$.

Differentiation of λ with respect to $\mathbf{w}$ and setting the derivative equal to zero results in the matrix equation:

$$(\boldsymbol{\Sigma} - \lambda\mathbb{V})\mathbf{w} = \mathbf{0} \quad . \tag{6}$$

Premultiplication of Equation 6 by $\mathbb{V}^{-1/2}$ and substituting $\mathbf{w} = \mathbb{V}^{-1/2}\mathbf{u}$ results in:

$$(\mathbb{V}^{-1/2}\boldsymbol{\Sigma}\mathbb{V}^{-1/2} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0} \quad . \tag{7}$$

The weight vector that maximizes ρ or α equals $\mathbb{V}^{-1/2}$ times the first principal component $\mathbf{u}$.

Inspection of the cells of $\mathbb{V}^{-1/2}\boldsymbol{\Sigma}\mathbb{V}^{-1/2}$ reveals that the optimal weights associated with the maximization of ρ are

$$w_i = \beta_i/\phi_i \quad , \tag{8}$$

that is, the weight of measurement i equals the slope divided by the error variance for this measurement.

In the maxalpha problem $\mathbb{V}^{-1/2}\boldsymbol{\Sigma}\mathbb{V}^{-1/2}$ equals the correlation matrix, $\mathbb{R}$, and the optimal weights are equal to the first principal component of this matrix divided by the standard deviations of the corresponding measurements. An impression of what the weights look like can be obtained from a related problem—that of finding the first principal component of $\mathbb{R}^*$, the correlation matrix with communalities in the diagonal. The loadings for this problem are

$$u_i^* = \beta_i/(\beta_i^2 + \phi_i)^{1/2} \quad , \tag{9}$$

with corresponding "weights,"

$$w_i^* = \beta_i/(\beta_i^2 + \phi_i) \quad . \tag{10}$$

Gifi (1980), quoting a result due to Bunch, Nielsen, and Sorenson (1978), noticed that there is a simple relationship between the loadings u_i^* and the principal components of $\mathbb{R}$. The first principal component of $\mathbb{R}$ is proportional to

$$u_i = \frac{u_i^*}{\lambda_1 - 1 + u_i^{*2}} \quad , \tag{11}$$

where λ_1 is the first eigenvalue of $\mathbb{R}$. From Equation 11 the final maxalpha weights are obtained. When λ_1 is relatively large, the maxalpha weights are approximately proportional to the w_i^*. Comparing Equations 8 and 9, it is clear that the maxalpha weights can be quite different from the maxrho weights. Further, from Equation 10 it might be concluded that the maxalpha weights are population dependent.

The Two-Parameter Logistic Model

The two-parameter logistic model can be written as:

$$P_i(\theta) = \Psi[a_i(\theta - b_i)] = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]} \quad , \tag{12}$$

where $P_i(\theta)$ is the probability of a correct response on item i given latent ability θ,
b_i is the difficulty parameter, and
a_i is the discrimination parameter of the item.

When the θs are estimated by the maximum likelihood method, the weighted sum $\Sigma w_i(\theta)x_i$ with item scores $x_i = 1$ for a correct response and $x_i = 0$ for an incorrect response, and with best weights

$$w_i(\theta) = P_i'(\theta)/[P_i(\theta)Q_i(\theta)] \quad , \tag{13}$$

with $Q_i(\theta) = 1 - P_i(\theta)$ and $P_i'(\theta) = \partial P_i(x)/\partial x$, evaluated at $x = \theta$, plays a role. Notice the resemblance between Equations 13 and 8: The best weight is equal to the local slope, divided by the local error variance (see also McDonald, 1982). From Equations 12 and 13, it is found that the best weights are independent of latent ability θ in the two-parameter logistic—the weights are equal to the discrimination parameters a_i.

An interesting question is whether the maxalpha weights correspond to the weights a_i from item response theory (IRT). An answer is provided by a simulation study. First, however, a theoretical analysis is presented. This analysis is based on the fact that the nonlinear tracelines of the two-parameter logistic model can be linearly approximated. Such an approximation seems especially reasonable when the spread of the ability parameters is relatively small.

Since the θs and bs are defined on an interval scale—$a^* = c^{-1}a$, $b^* = cb$ and $\theta^* = c\theta$ satisfy Equation 12 as well as the original parameters—it is convenient to choose the scale with $\mu_\theta = 0$ and $\sigma^2_\theta = 1$. Further, it is assumed that the θs are normally distributed; this distribution seems a reasonable approximation to the more or less symmetric distributions with not too heavy tails that frequently are encountered in practice. Finally, for the present the two-parameter logistic is approximated by the two-parameter normal ogive:

$$P_i(\theta) = \Phi[D^{-1}a_i(\theta - b_i)] \quad . \tag{14}$$

When the scaling constant D is set equal to 1.6, the difference between the logistic and normal densities is small, and in addition the cumulative curves are in close agreement for the middle region (Molenaar, 1974).

Given a standard normal distribution for the θs, Equation 14 can be linearly approximated as:

$$P_i(\theta) = \pi_i + \rho_i\phi(x_i)\theta \quad , \tag{15}$$

where $\pi_i = \Phi(-\rho_i b_i)$ is the item proportion correct, ϕ is the normal density, x_i equals $\rho_i b_i$, and

$$\rho_i = D^{-1}a_i/(D^{-2}a_i^2 + 1)^{1/2} \tag{16}$$

(McDonald, 1982, pp. 221–222) is the point-biserial between item i and the latent trait.

Clearly, the item weight can be approximated as:

$$w_i = \rho_i\phi(x_i)/\{\pi_i(1 - \pi_i)\} \quad . \tag{17}$$

The corresponding weight for the correct response is

$$w_{i+} = \rho_i\pi_i^{-1}\phi(x_i) \quad , \tag{18}$$

that is, the weight equals μ_{i+}, the mean θ given a correct response to item i (cf. de Gruijter & van der Kamp, 1984, p. 174). A similar result holds for the weight for incorrect.

Approximating $\phi(x)$ by the logistic density, the following is obtained:

$$\begin{aligned}\phi(x_i) &\approx D\psi(Dx_i) \\ &\approx D\Psi(Dx_i)\{1 - \Psi(Dx_i)\} \\ &\approx D\pi_i(1 - \pi_i) \quad .\end{aligned} \tag{19}$$

Substitution in Equation 17 results in:

$$w_i \approx D\rho_i \quad , \tag{20}$$

that is, the maxalpha weights are expected to be related to the point-biserials. From this, it is obvious that maxalpha weights will differ from the a_i, the best weights in the context of IRT. The size of the effect will depend on the general level of the as. With high as, the effect will be stronger. Or, for a constant level of a values, the effect will be stronger for more heterogeneous distributions.

A Simulation

In order to verify the extent to which the approximations in the previous section hold, a small simulation study was done with 15 items and 2,500 person parameters, randomly generated from the standard normal distribution. The relatively large sample size was chosen in order to avoid the possibility that effects become inconspicuous in the presence of sampling deviations. The first five items had $b = -1.75$, -1.0, 0.0, 1.0, and 1.75, respectively, and an $a = .5$. The next five items had the same b values, but an $a = 1.0$, and the last five items had $a = 1.5$. The particular choice of bs—symmetrical around zero—was expected to facilitate the appearance of the pattern of weights suggested in the previous section.

The total score, the sum of the item scores with $x = 1$ for a correct response and $x = 0$ for an incorrect response, is a reasonable and reliable measure of persons' abilities in many test theoretical applications. Therefore, satisfactory starting values for the category weights and consequently the item weights are easily obtained. Recalling the reciprocity property of weights and scores, for each category the average score of persons endorsing the category can be computed and the overall mean can be subtracted. The resulting weights might be biased, however, due to the fact that a score equal to 1.0 has been added to all scores of persons endorsing the correct category, whether the item discriminates or not. This spuriousness has been noticed in connection with the item-test correlation. Henrysson (1963) proposed to correct this correlation. In the numerator of the formula for the corrected correlation, one point is subtracted from the average score of the correct category and p, the item proportion correct, is subtracted from the overall mean. This results in alternative starting values for categories with corresponding starting values for the item weights equal to $\bar{x}_+ - 1 - \bar{x}_-$, where $\bar{x}_+$ is the average score of the persons with a correct response to the item, and $\bar{x}_-$ is the average of persons with an incorrect response. The choice of the starting values for the maxalpha weights further seems to be justified on the basis of results presented by Nishisato and Sheu (1980).

The results of the homogeneity analysis are given in Table 1, together with the corrected starting values. The item weights in the table were obtained by multiplying the outcomes from the analysis with a constant such that the average of the weights for items with $a = 1$, was equal to 1. From Table 1, it is clear that the homogeneity analysis underestimates high a values relative to low values as expected from the theoretical analysis.

Further, the starting values are very close to the final item weights, and alpha for the unweighted scores ($\alpha = .67$) is nearly as high as the maximum alpha ($\alpha = .69$). These results are to be expected with test items like the simulated items, with responses scored correct and incorrect.

Discussion

From this study, it is clear that item weights based on homogeneity analyses are not optimal in terms of IRT. In a simulation study based on the two-parameter logistic model, high a values were underestimated in a homogeneity analysis and low a values were overestimated. Neither can the weights be used as a basis for item selection as suggested by Nishisato (1980, p. 102). For item selection the information function is relevant, not the item discrimination parameter or the item-trait point-biserial.

Nevertheless, homogeneity analyses of test data can be useful. The strong point of homogeneity analyses lies where strong models do not fit, are computationally difficult, or are even nonexistent. For example, homogeneity analysis might be useful for multicategorical data, as an analysis in its own

Table 1

Item Weights from Homogeneity Analysis

Item	a	w	
		Starting Value	Final Value
1	0.5	.69	.68
2	0.5	.56	.57
3	0.5	.62	.62
4	0.5	.60	.62
5	0.5	.52	.52
6	1.0	.99	.99
7	1.0	.99	.97
8	1.0	.95	.95
9	1.0	1.00	1.01
10	1.0	1.07	1.08
11	1.5	1.53	1.45
12	1.5	1.30	1.26
13	1.5	1.24	1.21
14	1.5	1.26	1.22
15	1.5	1.37	1.31

right or as a provisional analysis before an analysis with a model based on strong assumptions, like the multicategory rating Rasch model.

References

Bunch, J. R., Nielsen, C. P., & Sorenson, D. C. (1978). Rank-one modification of the symmetric eigenproblem. *Numerische Mathematik, 31,* 31–48.

de Gruijter, D. N. M., & van der Kamp, L. J. Th. (1984). *Statistical models in psychological and educational testing*. Lisse, The Netherlands: Swets & Zeitlinger.

de Leeuw, J. (1973). *Canonical analysis of categorical data*. Unpublished doctoral dissertation, University of Leyden, The Netherlands.

Gifi, A. (1980). *Niet-lineaire multivariate analyse*. Leyden, The Netherlands: University of Leyden, Department of Data Theory.

Guttman, L. (1950). The principal components of scale analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, *Measurement and prediction* (pp. 312–361). Princeton NJ: Princeton University Press.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika, 28,* 211–218.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Lord, F. M. (1958). Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika, 23,* 291–296.

McDonald, R. P. (1968). A unified treatment of the weighting problem. *Psychometrika, 33,* 351–381.

McDonald, R. P. (1982). Some alternative approaches to the improvement of measurement in education and psychology: Fitting latent trait models. In D. Spearritt (Ed.), *The improvement of measurement in education and psychology* (pp. 213–232). Hawthorn, Victoria: The Australian Council for Educational Research.

McDonald, R. P. (1983). Alternative weights and invariant parameters in optimal scaling. *Psychometrika, 48,* 377–391.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27,* 82–99.

Molenaar, W. (1974). De logistische en de normale kromme. *Nederlands Tijdschrift voor de Psychologie, 29,* 415–420.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Nishisato, S., & Sheu, W-J. (1980). Piecewise method of reciprocal averages for dual scaling of multiple-choice data. *Psychometrika, 45,* 467–478.

Overall, J. E. (1965). Reliability of composite ratings. *Educational and Psychological Measurement, 25,* 1011–1022.

Author's Address

Send requests for reprints or further information to Dato N. M. de Gruijter, Educational Research Center, University of Leyden, Boerhaavelaan 2, 2334 EN Leyden, The Netherlands.