Behaviormetrika
Vol.32, No.2, 2005, 127–139

REPARAMETRIZATION OF HOMOGENEITY ANALYSIS TO ACCOMMODATE ITEM RESPONSE FUNCTIONS

Matthijs J. Warrens*, Willem J. Heiser*, and Dato N.M. de Gruijter**

Two test theoretical approaches to item analysis are compared, an approach based on homogeneity analysis and one based on item response theory. The literature on the relationship between the two approaches is briefly reviewed. The paper contains a contribution to the relationship between the two approaches for the case that the scores are dichotomous and a single latent variable is assumed to underlie the data. A loss function is proposed for modelling item response functions with two parameters, one for discrimination and one for difficulty. It turns out that the loss of the proposed loss function is related to loss of homogeneity. Demonstrations with simulated data are used to evaluate the proposed method.

1. Introduction

In this manuscript two test theoretical approaches to item analysis are distinguished, i.e. homogeneity analysis (abbreviated HA), and item response theory (abbreviated IRT). The optimal scaling technique, HA, is used for obtaining scores or quantifications of the multivariate categorical data, which are often visualized in low dimensional Euclidean space. There are several approaches to optimal scaling, e.g. the third type of quantification (Hayashi, 1952), dual scaling (Nishisato, 1980), multiple correspondence analysis (Greenacre, 1984), or HA (Gifi, 1990), leading to essentially equivalent solutions. The technique has been successfully applied to numerous kinds of categorical data from various fields. However, multicategorical data, and especially dichotomous scores, with an assumed latent dominance structure, are usually not analyzed with the optimal scaling technique, but with IRT models, either by translation families (Rasch, 1960), or by models with more parameters (Lord, 1952; Birnbaum, 1968).

IRT stands for a family of models which use item response functions (abbreviated IRFs) for explaining persons' probabilities of answering an item correct as a function of a latent variable (cf. van der Linden & Hambleton, 1997). The list of literature on the relationship between HA and IRT is relatively short. For the polytomous case, a contribution was made by Cheung & Mooi (1994), who made a comparison between the rating scale model (Andrich, 1978) and dual scaling on data comforming to Likert scales. For the dichotomous case, de Gruijter (1984) compared the discrimination parameter of the logistic two-parameter model (abbreviated 2-PM) to the item weight obtained by the difference

Key Words and Phrases: homogeneity analysis, two-parameter model, item response functions

∗ Department of Psychology, Leiden University

∗∗ ICLON, Leiden University

Mail Address: Methods and Statistics, Department of Psychology, Leiden University, Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands. E-mail: warrens@fsw.leidenuniv.nl

The manuscript is a completely revised and extended version of an unpublished paper by Heiser (1994). This paper was completed while the second author was research fellow at the Netherlands Institute in the Advanced Study in the Humanities and Social Sciences (NIAS) in Wassenaar, The Netherlands.

between the quantifications of the correct and incorrect categories.

In this paper a more formal formulation of the relationship between HA and IRT is attempted. A loss function for modelling the IRFs of the 2-PM in the unidimensional, dichotomous case is proposed. First, the next section is used to describe HA, using a formulation related to loss of homogeneity as defined by Gifi (1990). The IRT 2-PM is further described in section three. Then, in section four the loss function is proposed. The necessary conditions for its minimum are derived, and interpreted as properties of the fitted set of IRFs. These properties give a new insight in an old method, due to the main result of this paper, which is that, with an appropriate reparametrization, the loss of the proposed loss function is related to loss of homogeneity. In section five it is shown that the proposed loss function is related to, yet different from the linear item response approach of McDonald (1982). In section six some demonstrations are given. The discussion is in section seven.

2. Homogeneity analysis

HA is the name used here for Guttman's (1941) method for principal components analysis of categorical data. Guttman used the correlation ratio to define the objective of his method, which has evolved in the dual scaling methodology (Nishisato, 1980), but for the present purposes it is more convenient to use a formulation related to the Gifi (1990) reformulation, based on loss of homogeneity. With the general method an item can be formed by any number of mutually exclusive categories. Here, it will be sufficient to present HA for the unidimensional, dichotomous case.

Suppose the data of the problem, with n persons and m items, are collected in m binary vectors $\mathbf{z}_j$ $(j = 1, \ldots, m)$, of length n, containing 1 where a correct response occurred and -1 for an incorrect response. Let $\mathbf{G}_j$ be an indicator matrix, defined as the order $n \times 2$ matrix $\mathbf{G}_j = \{g_{rij}\}$ $(i = 1, \ldots n)$, where g_{rij} is a (0,1) variable, in which $r = +$ indexes the correct category, and $r = -$ the incorrect category. Thus the columns of $\mathbf{G}_j$ refer to the two possible responses, so that $g_{+ij} = 1$ (and $g_{-ij} = 0$) if subject i responded correctly on item j, while we code $g_{-ij} = 1$ (and $g_{+ij} = 0$) otherwise. Let the vector $\mathbf{y}_j = (y_{+j}, y_{-j})'$ contain the quantifications of the categories. Note that by choosing $\mathbf{y}_j^0 = (1, -1)'$, we obtain $\mathbf{z}_j = \mathbf{G}_j\mathbf{y}_j^0$, the initial coding of the data. Loss of homogeneity is defined as

$$\sigma(\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_m) = m^{-1} \sum_j \left\| \mathbf{G}_j\mathbf{y}_j - \mathbf{x} \right\|^2, \tag{1}$$

where $\mathbf{x}$ is the unknown n-vector of subject scores. The aim of the analysis is to find subject scores and category quantifications that minimize loss of homogeneity.

Stationary equations. The minimum of (1) over $\mathbf{y}_j$ is attained for

$$\mathbf{y}_j = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{x}, \tag{2}$$

where $\mathbf{D}_j = \text{diag}(\mathbf{G}_j'\mathbf{G}_j)$. The minimum of (1) over $\mathbf{x}$ is attained for

$$\mathbf{x} = m^{-1} \sum_j \mathbf{G}_j \mathbf{y}_j. \tag{3}$$

Stationary equations (2) and (3) are implemented in the algorithm HOMALS (Gifi, 1990), with identification constraints on the subject scores $\mathbf{x}$, which are put in standard scores, with zero mean and normalized as $\mathbf{x}'\mathbf{x} = n$. If $\mathbf{x}$ is centered, then also $\mathbf{D}_j\mathbf{y}_j = \mathbf{G}_j'\mathbf{x}$ from (2), for which it holds that $\mathbf{1}'\mathbf{D}_j\mathbf{y}_j = \mathbf{1}'\mathbf{G}_j'\mathbf{x} = 0$, that is, the different variables are centered. Thus, the technique does not reflect the first order moments of the variables, which are the common indications of item difficulties within the test theory framework.

3. The two-parameter IRT model

Over the last couple of years a vast amount of models has been developed in the field of IRT (cf. van der Linden & Hambleton, 1997). The IRFs of these models frequently form a family with a common shape, varying in one or a few parameters. One of the first models was the unidimensional 2-PM for dichotomous scores. With the 2-PM each IRF has two item parameters, one for location and one for discrimination. The normal ogive formulation of the 2-PM comes from Lord (1952), and Birnbaum (1968) later on proposed the logistic form of the 2-PM. In the latter form, the conditional probability of a correct response $Z_{ij} = 1$ of subject i on dichotomous item j is modeled as a logistic function of the latent variable θ, with

$$P_j(\theta) = \text{Prob}(Z_{ij} = 1|\theta_i, a_j, b_j) = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}, \tag{4}$$

where θ_i is the ability score of subject i, a_j the discrimination parameter, and b_j the difficulty parameter for item j. Equation (4) gives the IRF of item j as a function of θ. The incorrect response is modelled by $1 - P_j(\theta)$. For the Rasch model the discrimination parameters are equal, i.e., $a_j = a$ for all j. The class of IRFs that can be described by a discrimination and a difficulty parameter can be formulated as

$$P_j(\theta) = \Phi\left[d_j(\theta - \mu_j)\right], \tag{5}$$

where Φ is the common shape of the IRFs, for example, the logistic, the normal ogive, or any other monotonic function bounded by zero and one, and with d_j acting as the discrimination parameter and μ_j as the item difficulty or location paramater.

It is clear that different models correspond to different choices of Φ in (5). For each of them, a variety of fitting procedures has been developed, which all seem to have some advantages in some circumstances. However, it is also of interest to have an omnibus loss function that should be reasonable for various models in the class of (5), and could be applied in a broad range of circumstances. The formulation of such a badness-of-fit function is attempted in the next section.

4. A reparametrization

The optimal scaling technique has its counterpart for the IRT discrimination parameter. The fact that the optimal category weights maximize coefficient alpha (Lord, 1958),

has been exploited for applications in the test theory framework (see, e.g., Serlin & Kaiser, 1978). With dichotomous scores the maxalpha item weight is given by the difference between the quantifications of the correct and incorrect categories. However, as traditionally conceived, the optimal scaling approach lacks a counterpart for the difficulty parameter. At the end of section two it was shown that the technique does not reflect the first order moments of the variables, which are the common indications of item difficulties within the test theory framework. This makes a direct comparison of HA with the IRT approach to test theory a little difficult. A solution is proposed below.

Assume the validity of an IRT model. What is the information to be obtained by HA? We take the data $\mathbf{z}_j$ as indications of the positions of the subjects on the θ-scale. With dichotomous data the quantified variable $\mathbf{q}_j = \mathbf{G}_j\mathbf{y}_j$, can be arbitrarily transformed by $\mathbf{q}_j = d_j\mathbf{z}_j + \mu_j\mathbf{e}$, since there are two free parameters in the transformation, where $\mathbf{e}$ is a n-vector of ones. In this transformation d_j can be regarded as a discrimination parameter. Note that with this transformation a difficulty parameter μ_j is incorporated for each variable j. In order to make a comparison between HA and the IRT class of models in (5), we take $\mathbf{x} \approx \boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is used to denote a n-vector with elements θ_i. Hence

$$\mathbf{G}_j\mathbf{y}_j - \mathbf{x} = d_j\mathbf{z}_j + \mu_j\mathbf{e} - \boldsymbol{\theta} = d_j\mathbf{z}_j - (\boldsymbol{\theta} - \mu_j\mathbf{e}). \tag{6}$$

Taking the average sum of squared residuals of the right part of (6) gives

$$\tau\left(\boldsymbol{\theta}, d_1, \ldots, d_m, \mu_1, \ldots, \mu_m\right) = m^{-1}\sum_j \left\|d_j\mathbf{z}_j - (\boldsymbol{\theta} - \mu_j\mathbf{e})\right\|^2. \tag{7}$$

(For another formulation of (7) in a different context, see, Takane & Oshima-Takane, 2003). For each item in (7), the data vector, which has not necessarily zero mean, is rescaled by a factor d_j, and the rescaled item scores are compared with the unknown ability scores, after translation by an amount μ_j.

Stationary equations. Let $C_j = \{i|z_{ij} = 1\}$ and $I_j = \{i|z_{ij} = -1\}$; the minimum of (7) over d_j is attained for

$$\hat{d}_j = n^{-1}(\boldsymbol{\theta} - \mu_j\mathbf{e})'\mathbf{z}_j = n^{-1}\left[\sum_{i\in C_j}(\theta_i - \mu_j) - \sum_{i\in I_j}(\theta_i - \mu_j)\right], \tag{8}$$

which is a between-groups deviation. It can be expressed as a weighted deviation between the mean ability scores of the correct group ($i \in C_j$) and the incorrect group ($i \in I_j$) after translation:

$$\hat{d}_j = p_{+j}(\bar{\theta}_{+j} - \mu_j) - p_{-j}(\bar{\theta}_{-j} - \mu_j), \tag{9}$$

where p_{+j} is the proportion of subjects with correct responses, $p_{-j} = 1 - p_{+j}$, and where

$$\bar{\theta}_{+j} = (np_{+j})^{-1}\sum_{i\in C_j}(\theta_i - \mu_j),$$

with $\bar{\theta}_{-j}$ defined analogously. It follows from (9), and from the fact that two means are furthest apart when taken of two non-overlapping groups of values, that the items are

weighted in (7) according to their ability to discriminate the correct group from the incorrect group on the θ-scale. Thus d_j could be used as a discrimination diagnostic. When it becomes negative, the original item scores reverse the subject order compared to the major trend in the items.

The minimum of (7) over the item difficulty parameter μ_j is attained for

$$\hat{\mu}_j = n^{-1}(\boldsymbol{\theta} - d_j\mathbf{z}_j)'\mathbf{e} = \bar{\theta} - d_j(p_{+j} - p_{-j}), \tag{10}$$

where $\bar{\theta}$ is the mean ability score. Suppose that d_j is positive; if the correct responses are in the majority ($p_{+j} - p_{-j} > 0$), the difficulty parameter estimate $\hat{\mu}_j$ moves to the left of $\bar{\theta}$, and if the incorrect responses are in the majority, it moves to the right. These moves are proportional to the discrimination diagnostic d_j.

Unconstrained least squares estimates of the ability scores $\boldsymbol{\theta}$ can be obtained by

$$\tilde{\boldsymbol{\theta}} = m^{-1}\sum_j (d_j\mathbf{z}_j + \mu_j\mathbf{e}), \tag{11}$$

an average of linear transformations of the data vectors. Since the class in (5) is invariant under changes of origin, and since (7) is not invariant under arbitrary rescalings of $\boldsymbol{\theta}$, some identification constraints are needed. Several posibilities exist, but we require $\boldsymbol{\theta}$ to be in standard scores, i.e., $\bar{\theta} = 0$ and $\boldsymbol{\theta}'\boldsymbol{\theta} = n$. If the origin is chosen as the mean, $\bar{\theta} = 0$, then we may re-express (11), using (10), as

$$\tilde{\boldsymbol{\theta}} = m^{-1}\sum_j d_j\left[\mathbf{z}_j - (p_{+j} - p_{-j})\mathbf{e}\right],$$

a weighted average of the variables after they have been put in deviation from their mean. The standardized estimate $\hat{\boldsymbol{\theta}}$ is obtained by setting $\hat{\boldsymbol{\theta}} = n^{1/2}\tilde{\boldsymbol{\theta}}/\lambda$, with $\lambda = \left\|\tilde{\boldsymbol{\theta}}\right\|$. Stationary equations (8), (10), and (11) could be used to define an alternating least squares algorithm for minimizing (7), but this is not necessary due to the following result.

Proposition. Loss function (7) is equivalent to loss of homogeneity in (1), with the reparametrization $\boldsymbol{\theta} = \mathbf{x}$, $\mu_j = (y_{+j} + y_{-j})/2$, and $d_j = (y_{+j} - y_{-j})/2$.

Proof. Define $n \times 2$ matrices $\mathbf{F}_j$=($\mathbf{e}$ $\mathbf{z}_j$) and the 2-vectors $\mathbf{u}_j' = (\mu_j\ d_j)$ for $j = 1, \ldots, m$. The residuals in (7) can be written as $(\mathbf{F}_j\mathbf{u}_j - \boldsymbol{\theta})$. With $\boldsymbol{\theta} = \mathbf{x}$, the two functions are equivalent if $\mathbf{F}_j\mathbf{u}_j = \mathbf{G}_j\mathbf{y}_j$.

Define the 2×2 matrices $\mathbf{S} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ and $\mathbf{T} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix}$. They satisfy $\mathbf{TS} = \mathbf{I}$, and we have $\mathbf{G}_j = \mathbf{F}_j\mathbf{T}$ and $\mathbf{F}_j = \mathbf{G}_j\mathbf{S}$. If $\mathbf{y}_j = \mathbf{S}\mathbf{u}_j$ and $\mathbf{u}_j = \mathbf{T}\mathbf{y}_j$ we may write $\mathbf{F}_j\mathbf{u}_j = \mathbf{F}_j\mathbf{T}\mathbf{S}\mathbf{u}_j = \mathbf{G}_j\mathbf{y}_j$. The equation $\mathbf{u}_j = \mathbf{T}\mathbf{y}_j$ gives the reparameterization. **qed**.

The reparametrization in the Proposition gives the μ_j's and d_j's in terms of the category quantifications, but from the proof it is clear that we also have

$$y_{+j} = \mu_j + d_j, \tag{12}$$

and

$$y_{-j} = \mu_j - d_j. \tag{13}$$

If $p_{+j} = p_{-j}$, we must obtain $\hat{\mu_j} = 0$ by (10), and in that case the category quantifications only reflect discrimination. In fact, the diagnostic quantities called discrimination measures in Gifi (1990) are defined as

$$\eta_j^2 = p_{+j} y_{+j}^2 + p_{-j} y_{-j}^2.$$

Using (12) and (13), and simplifying with the aid of (10) we can express the discrimination measures as

$$\eta_j^2 = 4p_{+j}(1 - p_{+j})d_j^2.$$

(Essentially the same result is given in Yamada & Nishisato, 1993, p. 60). With equal marginals, we find $\eta_j^2 = d_j^2$. Otherwise, since $4p_{+j}(1 - p_{+j}) \leq 1$, it will always be true that $d_j^2 \geq \eta_j^2$.

Using (10), we can write (7) with the third set of parameters partialled out, i.e.,

$$\tau(\boldsymbol{\theta}, \mathbf{d}_j, *) = m^{-1} \sum_j \left\| d_j(\mathbf{z}_j - \bar{z}_j \mathbf{e}) - \boldsymbol{\theta} \right\|^2,$$

which in turn can be minimized over $\mathbf{d}_j$ to obtain $\tau(\boldsymbol{\theta}, *, *)$ as

$$\tau(\boldsymbol{\theta}, *, *) = \|\boldsymbol{\theta}\|^2 - m^{-1} \sum_j \frac{\left[\boldsymbol{\theta}'(\mathbf{z}_j - \bar{z}_j \mathbf{e})\right]^2}{\left\|\mathbf{z}_j - \bar{z}_j \mathbf{e}\right\|^2}. \tag{14}$$

In (14) we have used the fact that the minimum of $\tau(\boldsymbol{\theta}, \mathbf{d}_j, *)$ over d_j is attained for

$$\hat{d}_j = \frac{\boldsymbol{\theta}'(\mathbf{z}_j - \bar{z}_j \mathbf{e})}{\left\|\mathbf{z}_j - \bar{z}_j \mathbf{e}\right\|^2}. \tag{15}$$

In (15), $\boldsymbol{\theta}$ is standardized. From this and (14) we may conclude that the loss function (7) is minimized when $\boldsymbol{\theta}$ is chosen in such a way that the sum of squared correlations with the observed item responses is maximized.

5. The linear approximation to IRT

The loss function proposed in the previous section can be considered a linear approximation of nonlinear IRT models. In this section it is demonstrated that the method is closely related but not equivalent to the linear IRT approach described by McDonald (1982). The nonlinear logistic IRF in (4) can be approximated linearly as

$$P_j(\theta) \approx a_j^*(\theta - b_j^*) + 1/2. \tag{16}$$

The linear approximation (16) is a special case of the model of congeneric test models (Jöreskog, 1971). To be able to make a comparison with (7), the loss function corresponding to (16) is defined as the average sum of squared residuals

$$\nu\left(\boldsymbol{\theta}, \mathbf{a}^{**}, \mathbf{b}^{*}\right) = m^{-1} \sum_{j} \left\| \mathbf{z}_j - a_j^{**}\left(\boldsymbol{\theta} - b_j^{*}\mathbf{e}\right)\right\|^2, \tag{17}$$

where m is the number of variables and $\mathbf{z}_j$ is the vector with scores $z_{ij} = 1$ for a correct response and $z_{ij} = -1$ for an incorrect response, under the restrictions that $\boldsymbol{\theta}$ has zero mean and unit variance, and $a_j^{**} = 2a_j^{*}$. Formulation (17) can be written as

$$\nu\left(\boldsymbol{\theta}, \mathbf{a}^{**}, \mathbf{c}\right) = m^{-1} \sum_{j} \left\| \mathbf{z}_j - a_j^{**}\boldsymbol{\theta} - c_j\mathbf{e}\right\|^2, \tag{18}$$

where $c_j = -a_j^{**}b_j^{*}$, and the other parameters are the same as above.

Stationary equations. The minimum of (18) provides the parameter estimates:

$$\hat{c}_j = \bar{z}_j = p_{+j} - p_{-j}, \tag{19}$$

$$\hat{a}_j^{**} = p_{+j}\bar{\theta}_{+j} - p_{-j}\bar{\theta}_{-j}, \tag{20}$$

where $\bar{\theta}_{+j}$ and $\bar{\theta}_{-j}$ are the average abilities of the correct group and the incorrect group respectively,

$$\hat{\theta}_i = m^{-1}\frac{\sum_j a_j^{**}(z_{ij} - \bar{z}_j)}{\sum_j a_j^{**}},$$

$$\hat{b}_j^{*} = -\frac{\hat{c}_j}{\hat{a}_j^{**}}.$$

Using (19) we may write (18) as

$$\nu\left(\boldsymbol{\theta}, \mathbf{a}^{**}, *\right) = m^{-1}\sum_j \|\mathbf{z}_j - \bar{z}_j\mathbf{e}\|^2 + m^{-1}\sum_j (a_j^{**})^2\,\|\boldsymbol{\theta}\|^2 - 2m^{-1}\sum_j a_j^{**}(\mathbf{z}_j - \bar{z}_j\mathbf{e})'\boldsymbol{\theta}, \tag{21}$$

where we have used the notation $\nu\left(\boldsymbol{\theta}, \mathbf{a}^{**}, *\right)$ to indicate that the third set of parameters is eliminated by inserting the estimates. Minimizing $\nu\left(\boldsymbol{\theta}, \mathbf{a}^{**}, *\right)$ over $\mathbf{a}^{**}$ we obtain $\nu\left(\boldsymbol{\theta}, *, *\right)$ as

$$\nu\left(\boldsymbol{\theta}, *, *\right) = m^{-1}\sum_j \|\mathbf{z}_j - \bar{z}_j\mathbf{e}\|^2 - nm^{-1}\sum_j (a_j^{**})^2.$$

Since $\hat{a}_j^{**}$ defined in (20) can also be written as

$$\hat{a}_j^{**} = n^{-1}\sum_j (\theta_i - \bar{\theta})(z_{ij} - \bar{z}_j),$$

it is a covariance, and hence (21) shows that this approach determines $\boldsymbol{\theta}$ in such a way that the sum of squared covariances with the observed item responses is maximized. Hence, (7) and (18) are related, yet different, whenever the p_{+j}'s differ. Apart from that they are different, it is not intuitively clear which approach is to be preferred.

6. Some demonstrations

In this section numerical illustrations are provided to illustrate the method proposed in section four. But before this is done, some relevant results by de Gruijter (1984) should be pointed out.

Lord (1958) showed that the optimal scaling weights from a HA maximize coefficient alpha. For dichotomous scores the item weight which maximizes alpha is given by d_j. The homogeneity item weight was compared to the discrimination parameter from model (4) in a study by de Gruijter. Under the the assumption that the latent variable is normally distributed it was shown that d_j is related to the point-biserial correlation of the item with the latent variable. Also, with simulated data de Gruijter showed that d_j is different from the IRT discrimination parameter: the increase in d_j diminishes as a_j becomes larger.

Considering the results by de Gruijter, it is expected that (7) has the clearest interpretation in terms of IRT if it is applied to models with equal discrimination parameters, a special case of (5). For the class in (5) the proposed method and the IRT approach are likely to give different outcomes with respect to discrimination and difficulty parameters.

In the remainder of this section two models are used as gauges or benchmarks. Artificial datasets were generated from both the logistic 2-PM in (4) and from the Rasch model, i.e. (4) with $a_j = a$. Both datasets consisted of responses of 1000 simulated persons on 50 items. The subjects were sampled from a standard normal distribution. With respect to the 2-PM data, the 50 items consisted of 5 sets of 10 items, each with a different discrimination parameter (0.5, 1.0, 1.5, 2.0, 2.5). For each set, the 10 location parameters were the same and ranged from -1.8 to 1.8 with step size .4, and the value 0 omitted. With respect to the Rasch data, all discrimination parameters were set at 1; the difficulty parameters ranged from -1.96 to 1.96 with stepsize .08.

For both datasets the IRT item parameter estimates were obtained using marginal maximum likelihood (abbreviated MML, Multilog; Thissen, Chen, & Bock, 2003); HA item parameter estimates were obtained with the reparametrization of the quantifications from the Proposition. Subject IRT parameter estimates were obtained using the Bayesian method maximum a posteriori (abbreviated MAP, Multilog; Thissen, Chen, & Bock, 2003); the HA person estimates are the subject scores. The subject estimates of both IRT and HA of the 2-PM dataset are plotted in Figure 1.

Close inspection of Figure 1 shows a sigar-shaped relationship between the sets of estimates. The relationship is not strong, but an increase in the latent variable using one estimate is approximately the same as an increase using the other estimate. The discrimination estimates for the 2-PM dataset are plotted in Figure 2.

The relationship between both sets of discrimination estimates is clearly not linear. In fact, Figure 2 is a clear visualization of a result derived by de Gruijter (1984): the increase in d_j diminishes as a_j becomes larger. The difficulty estimates for the 2-PM dataset are plotted in Figure 3.

Remember that the 2-PM data consisted of 5 sets of 10 items, each with a different discrimination parameter. Close inspection of Figure 3 reveals that within each set the estimates for the difficulty parameter are approximately proportional. However, this is

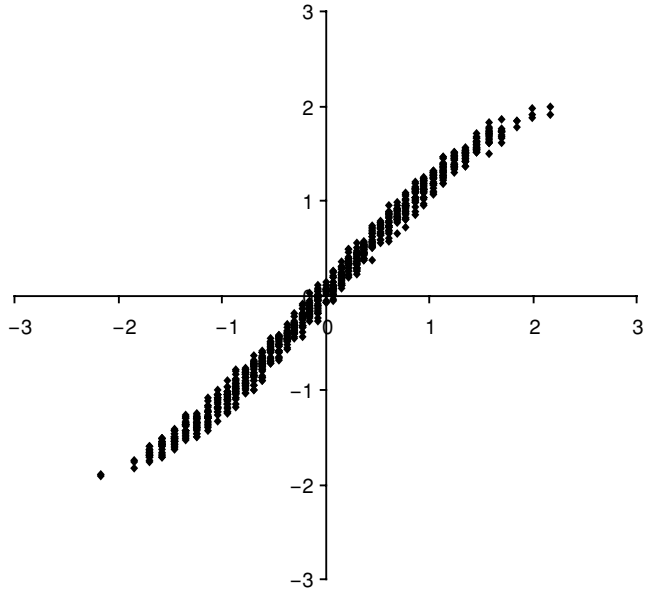

Figure 1: Plot of subject estimates for the 2-PM dataset; MAP (horizontal) versus HA (vertical).

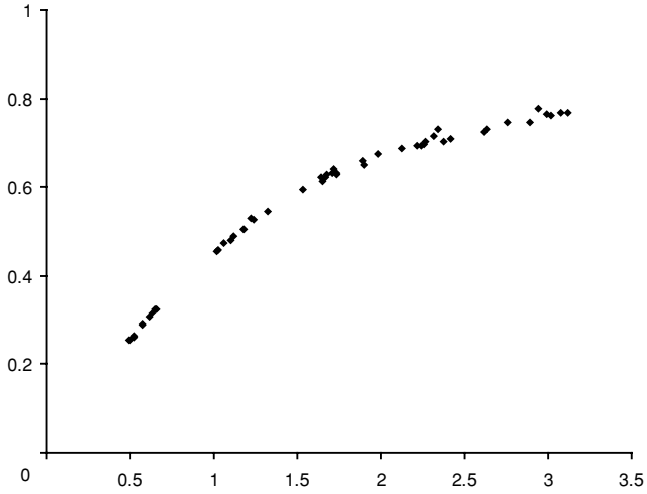

Figure 2: Plot of discrimination parameter estimates for the 2-PM dataset; MML (horizontal) versus HA (vertical).

clearly not the case for the total group. Furthermore, the HA difficulty estimates are on quite a different scale than the IRT estimates.

Note that HA standardization is on the subject scores, which are approximately on the same scale as the IRT estimates (see Figures 1 and 4). However, the values of the item category quantifications, and hence the HA estimates for difficulty and discrimination, depend on the score patterns in the data. For IRFs with steep slopes the subjects are further apart, which will result in higher values for the quantifications. With flat IRFs

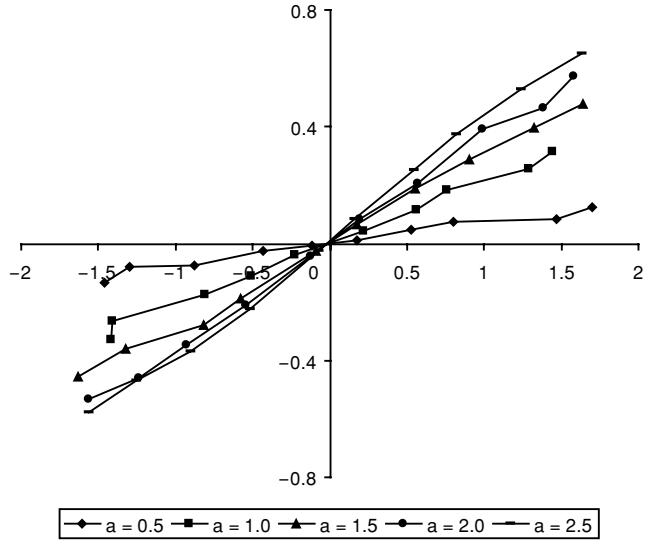

Figure 3: Plot of difficulty parameter estimates for the 2-PM dataset; MML (horizontal) versus HA (vertical).

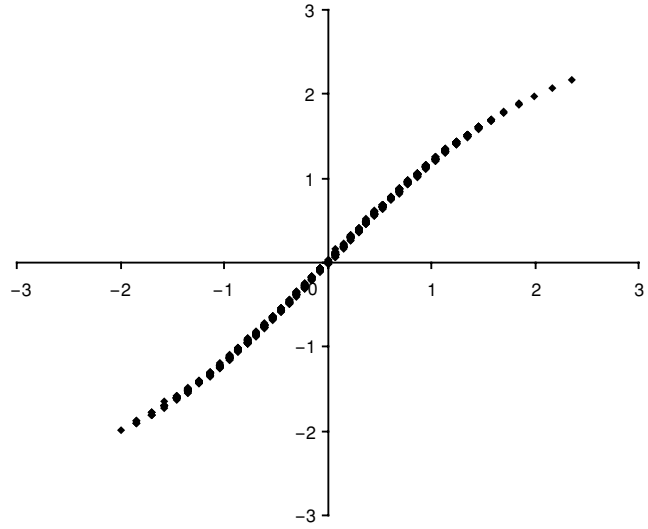

Figure 4: Plot of subject estimates for the Rasch dataset; MAP (horizontal) versus HA (vertical).

the item quantifications are very close together, because of (2). The subject estimates for the Rasch dataset are plotted in Figure 4.

The relationship plotted in Figure 4 is much stronger compared to the one in Figure 1. Close inspection reveals that there are $m + 1 = 51$ clusters of scores. Under the Rasch model the sum score is a sufficient statistic for the subject estimate and there is an unique IRT score for each sum score. With HA, for each different score pattern there is a possible unique HA score, irrespective of the corresponding sum score. Nevertheless, the HA

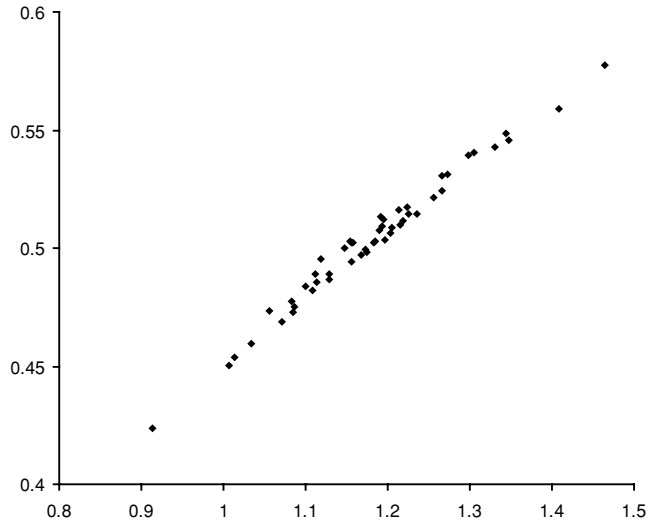

Figure 5: Plot of discrimination parameter estimates for the Rasch dataset; MML (horizontal) versus HA (vertical).

subject scores corresponding to the same sum score must be very close together under the Rasch model, in order to obtain the clusters as observed in Figure 4.

Although all items have the same discrimination parameter under the Rasch model, the stochastic process used for generating the data introduces a slight form of variance. Because the HA approach is essentially a 2-PM, the data is analyzed with the logistic 2-PM. The discrimination estimates from both approaches for the Rasch dataset are plotted in Figure 5.

Figure 5 can be interpreted as a visualization of a selection of Figure 2. When the discrimination parameters are closer together, as in the Rasch dataset, the relationship between the HA and IRT estimates is approximately linear. Finally, the difficulty parameters for the Rasch dataset are plotted in Figure 6.

In Figure 6 a relatively strong relationship between the IRT and HA estimates can be observed. The relationship is quiet an improvement compared to the relationship plotted in Figure 3.

7. Discussion

In this manuscript a distinction was made between two test theoretical approaches to item analysis, namely HA and IRT. The psychometric literature on the relationships between HA and IRT had only a few contributions. For the case that the scores are dichotomous and a single latent variable is assumed to underlie the data a contribution to this subject was made in this manuscript. Homogeneity based analyses of test results were compared to IRT analyses, both theoretically as with applications to simulated data. A loss function was proposed to accommodate monotonically increasing IRFs. It was demonstrated that a solution for the loss function can be obtained by a simple transformation

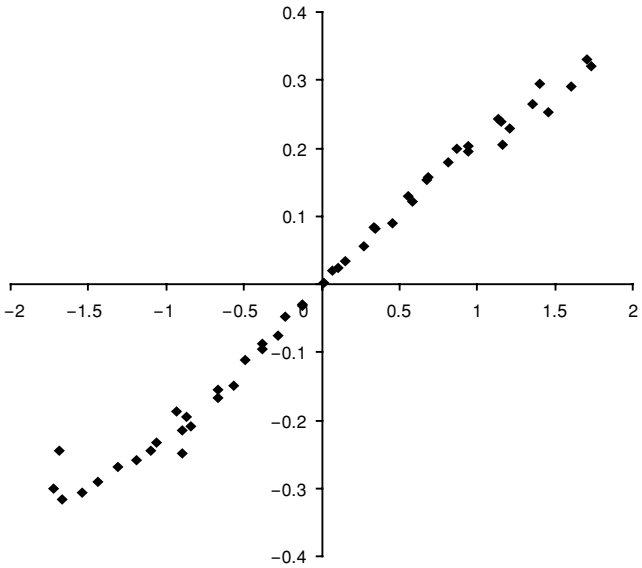

Figure 6: Plot of difficulty parameter estimates for the Rasch dataset; MML (horizontal) versus HA (vertical).

of the HA solution.

The end of section four and section five were used to show the relation of the proposed method to the linear IRT approach. In the latter approach θ can be estimated in such a way that the sum of squared covariances with the observed item responses is maximized. On the other hand, the method described in section four estimates θ is such a way that the sum of squared correlations with the observed item responses is maximized. Thus, the approaches have related, but different solutions.

The discrimination parameter used by the proposed method, turned out to be the maxalpha weight for dichotomous scores (Lord, 1958). Using simulated data de Gruijter (1984) showed that d_j is different from the IRT discrimination parameter: the increase in d_j diminishes as a_j becomes larger. This result was confirmed in section six (see Figure 2), which contains several illustrations. Although the HA method is essentially formulated as a 2-PM, it is different from the logistic IRT 2-PM. However, for the case that the discrimination parameters of a model are close together (e.g. the Rasch model), the two methods seem to provide very similar results.

In this paper it was made clear, for the unidimensional, dichotomous item case, what the relationship is between HA and IRT modelling. The properties of HA described in this manuscript give a new insight in an old method. It turns out, that HA, or similar methods, i.e. Hayashi's third method of quantification or dual scaling, can be used to model aspects of IRFs to a certain degree. Since aspects of IRFs can only be modelled to a certain degree, HA should not be considered an estimation heuristic for IRT item parameters.

References

Andrich, D. (1978). A rating formulation for ordered response categories, *Psychometrika*, **43**, 561–573.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, In F. M. Lord & M. R. Novick (Eds), *Statistical theories and mental test scores*, Reading: Addinson-Wesley.

Cheung, K.C., & Mooi, L.C. (1994). A comparison between the rating scale model and dual scaling for Likert scales. *Applied Psychological Measurement*, **18**, 1–13.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chicester: Wiley.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*, New York: Academic Press.

Gruijter, D.N.M. de (1984). Homogeneity analysis of test score data: A confrontation with the latent trait approach, *Applied Psychological Measurement*, **8**, 385–390.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction, In P. Horst (Ed), *The prediction of personal adjustment*. New York: SSRC.

Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, **3**, 69–98.

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

Linden, W. van der, & Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York: Springer Verlag.

Lord, F.M. (1952). A theory of mental test scores. *Psychometrika Monograph No. 7*.

Lord, F.M. (1958). Some relations between Guttman's principal components analysis and other psychometric tests. *Psychometrika*, **36**, 109–133.

McDonald, R.P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, **6**, 379–396.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests. Studies in Mathematical Psychology*. Copenhagen: Danish Institute for Educational Research.

Serlin, R.C. & Kaiser, H.F. (1978). A method for increasing the reliability of a short multiple-choice test. *Educational and Psychological Measurement*, **38**, 337–340.

Takane, Y., & Oshima-Takane, Y. (2003). Relationships between two methods of dealing with missing data in principal components analysis. *Behaviormetrika*, **30**, 145–154.

Thissen, D., Chen, W.H., & Bock, D. (2003). *Multilog 7: Analysis of multiple-category response data*. Scientific Software International.

Yamada, F., & Nishisato, S. (1993). Several mathematical properties of dual scaling as applied to dichotomous item-category data. *Japanese Journal of Behaviormetrics*, **20**, 56–63.

(Received October 6 2004, Revised May 23 2005)