Measurement and Research Department Reports 95-1

A Heuristic Procedure for Suggesting an Appropriate Scoring Function for Polytomous Items with Ordered Response Categories

Niels H. Veldhuijzen

Measurement and Research Department Reports 95-1

A Heuristic Procedure for Suggesting an Appropriate Scoring Function for Polytomous Items with Ordered Response Categories

Niels H. Veldhuijzen

Cito
Arnhem, 1995

Abstract

Polytomous items with two or more *ordered* response categories, called ordered polytomous items, are considered. A *scoring function* assigns a real number to each response category. Each answer to an ordered polytomous item is scored as the number assigned to the response category the answer belongs to.

A useful model for ordered polytomous items is the Generalized Partial Credit Model. To use this model, a scoring function for each item in a test must be available. In this report a heuristic procedure is developed which may be used to supply tentative scoring functions to be used in an application of the Generalized Partial Credit Model. Two examples and an algorithm are presented.

Key words: IRT, weighted least squares, logistic models, optimal scaling, scoring functions.

Introduction

In this report the problem of finding an appropriate *scoring function* for ordered polytomous items is studied. Ordered polytomous items are items with an *ordered* set of response categories. All possible answers to the item can be classified as belonging to just one of the response categories. A *scoring function* assigns a *category score* to each response category; each answer given to the item is scored as the score of the category the answer belongs to. For example, in a test for music education you may be asked to identify the instruments you hear in a piece of music. Usually, the number of instruments identified correctly is taken as the item score.

The problem with polytomous items having more than two ordered response categories is that both the number and the order of the response categories are not always easily established. For example, in the music item presented above, it may be that one pair of instruments can be identified more easily than some other pair. If this is the case, identifying the one pair should be scored differently from identifying the other pair. Or, to give another example, some item writers may include the response category "don't know" in one or more items. Where to place this category compared to categories as "correct" and "incorrect" will not always be obvious.

In classical item analysis, the problem of finding an appropriate scoring function has been solved long ago by Guttman (1941; see also Nishisato, 1980; Gifi, 1990). The procedure known as *homogeneity analysis*, *multiple correspondence analysis* and *optimal scaling* assigns scores to response categories such that Cronbach's coefficient alpha is maximized. Homogeneity analysis treats polytomous items as nominal variables; order restrictions can be incorporated, however (Gifi, *o.c.*) to accommodate ordered polytomous items.

In item response theory, a counterpart of homogeneity analysis may be found in the Nominal Item Response Model by Bock (1972). As the name suggests, this model also treats items as nominal variables. No incorporation of order restrictions in Bock's model has been presented in the psychometric literature. The heuristic procedure for finding appropriate scoring functions for ordered polytomous items, presented in this report, will be based on Bock's model. This procedure should provide suggestions about the right number and order of response categories, to be used in a formal analysis of ordered polytomous items with a dedicated model. The model dedicated to ordered polytomous items that will be the focus of this report is the *Generalized Partial Credit Model*.

In the next sections, the Partial Credit Model (Masters, 1982) and a useful generalization will be presented. It will be shown that Bock's model is a relaxation of the Generalized Partial Credit Model. A quick least squares estimation procedure for Bock's model is

developed. A detailed description of the algorithm can be found in the Appendix. The results of the heuristic procedure, applied to a contrived and a real data set, are discussed in the last few sections. Some comments on relating the results of the procedure to the parameters of the Generalized Partial Credit Model are given.

The Partial Credit Model

Suppose that all possible answers to a certain test item j can be classified as belonging to just one of k_j+1 ordered categories. These ordered categories will be labelled from 0 to k_j. That the categories are ordered means that a response belonging to category m is assumed to indicate a higher ability level than a response belonging to category m' where $m' < m$.

Let U_{ijm} be an indicator variable defined as follows:

$$U_{ijm} = \begin{cases} 1 \text{ if the answer of subject } i \text{ to item } j \text{ belongs to category } m; \\ 0 \text{ otherwise.} \end{cases}$$

Let S_{ij} be a scoring function defined as follows:

$$S_{ij} = \psi_j(m) \Leftrightarrow U_{ijm} = 1$$

where $\psi_j(m)$ is some increasing function of the category label m, reflecting the order of the response categories for item j.

The Partial Credit Model (Masters, *o.c.*) is an item response model where the functions $\{ \psi_j(m) \}$ are chosen to be the identity function. The model postulates the following probability distribution function for the indicator variable U_{ijm}:

$$Prob(U_{ijm}=1;\vartheta_i,\underline{\beta}_j)=\exp(z_{ijm})/[\Sigma_{h=0}^{k_j}\exp(z_{ijh})]$$

where ϑ_i is the ability parameter of subject i, $\underline{\beta}_j \equiv (\beta_{j0},\cdots,\beta_{jk_j})$ is a vector with category parameters, and $z_{ijm}=m\vartheta_i-\Sigma_{c=0}^{m}\beta_{jc}$. Some insight into the meaning of the category parameters may be gained by looking at the following conditional probability of a response belonging to category m rather than category m-1:

$$Prob(U_{ijm}=1 \mid U_{ij,m-1}=1 \text{ or } U_{ijm}=1;\vartheta_i,\underline{\beta}_j)=\exp(z_{ijm})/[\exp(z_{ij,m-1})+\exp(z_{ijm})]=$$

$$=\exp(z_{ijm}-z_{ij,m-1})/[1+\exp(z_{ijm}-z_{ij,m-1})]=$$

$$=\exp(\vartheta_i-\beta_{jm})/[1+\exp(\vartheta_i-\beta_{jm})].$$

Now, this probability, which may be recognized as the Rasch model (Rasch, 1960), is equal to its complement if $\exp(\vartheta_i-\beta_{jm})$ is equal to 1, so if ϑ_i equals β_{jm}. So, for each point β_{jm} on the ability scale, known as a *category bound*, the following condition holds:

$$Prob(U_{ijm}=1;\vartheta_i,\underline{\beta}_j) \begin{Bmatrix} > \\ \leq \end{Bmatrix} Prob(U_{ij,m-1}=1;\vartheta_i,\underline{\beta}_j) \Leftrightarrow \vartheta_i \begin{Bmatrix} > \\ \leq \end{Bmatrix} \beta_{jm}.$$

The Generalized Partial Credit Model

The Partial Credit Model may lead to some undesired effects if the items in a test differ in the number of response categories. Suppose that there are two test items, one having two categories scored as 0 and 1, and one having three, scored as 0, 1 and 2. A perfect answer to the first item results in an item score of 1; a perfect answer to the second item results in an item score of 2. If the item scores are added and their sum is taken as the test score, the second item weights more than the first item. This may be undesirable, for a perfect answer to the first item may indicate the same ability level as a perfect answer to the second item. That the second item has more response categories than the first has nothing to do with it.

One way to overcome this problem is to insert a weight parameter a_j into the Partial Credit Model. The resulting Generalized Partial Credit Model can be written as follows:

$$Prob(U_{ijm}=1;\vartheta_i,a_j,\underline{\beta}_j)=\exp(z^*_{ijm})/[\textstyle\sum_{h=0}^{k_j}\exp(z^*_{ijh})],$$

where $z^*_{ijm}=a_j(m\vartheta_i-\sum_{c=0}^{m}\beta_{jc})$. The weight parameter a_j is usually called the discrimination parameter of item j, because it monitors the rate with which $Prob(U_{ijm}=1;\vartheta_i,a_j,\underline{\beta}_j)$ changes as ϑ_i changes.

The Generalized Partial Credit Model has been described by Muraki (1992). A special case has been in use for several years at Cito. This special case, called the One Parameter Logistic Model (Verhelst, Glas and Verstralen, 1994; Verhelst and Glas, 1995) treats the discrimination parameter a_j not as a parameter but as a known constant, called a *discrimination index*. If the discrimination indices are known, weighted sumscores with these indices as weights are sufficient statistics for the parameters of the model, and conditional maximum likelihood estimates (Andersen, 1973) of the category boundaries are available. The discrimination indices, of course, must be specified beforehand. A computer program for the One Parameter Logistic Model (Verhelst, Glas and Verstralen, *o.c.*) incorporates test statistics that are sensitive to badly specified discrimination indices.

On Scoring Items

The results of the Generalized Partial Credit Model depend on the scoring functions used for the items. These scoring functions imply that the number and the order of the response categories of the items are known. This will not always be the case. Badness of fit may result

from an injudicious ordering of the possible item responses. Consider, for example, a dichotomous item. The possible responses to this item are classified as "correct" and "incorrect". But now, the item writer adds another response category to the item: the "don't know" category. Obviously, "don't know" is not a perfect answer. But it is not sure how to relate it to the "incorrect" category. Is it "more incorrect" than "incorrect", or does it show an ability level high enough to be able to rule out incorrect responses?

Some item writers specify a relatively large number of response categories for their items, hoping that the items will lead to a reliable measurement instrument. Item analysis may show that, without loss of information, categories can be collapsed.

Even if it is possible to specify an ordered set of response categories for an item, it may be very difficult to have them "equidistant", as reflected in the scoring function of the Generalized Partial Credit Model.

To solve these problems, procedures for evaluating a tentative response classification and a tentative scoring function before formal item analysis takes place are required. One such procedure, based on ideas from the field of optimal scaling (Nishisato, *o.c.*; Gifi, *o.c.*), is presented hereafter. Another procedure, based on category information functions, has been formulated by Muraki (1993).

The key idea of the procedure is to abandon, for the moment, the idea of *ordered* categories and treating items as nominal variables. Consider the following scoring function:

$S_{ij}=\delta_j(m) \Leftrightarrow U_{ijm}=1$

where the function $\delta_j(m)$ has the value δ_{jm} if $U_{ijm}=1$. Since, for each j and m, the function $\delta_j(m)$ can only have the value δ_{jm}, it is appropriate to write $\delta_j(m) \equiv \delta_{jm}$. Note that the numbers $\{\delta_{jm}\}$ are unknown constants, to be determined by minimizing an appropriate loss function. Implementing this scoring function in the Generalized Partial Credit Model, the following model results:

$$Prob(U_{ijm}=1;\vartheta_i,a_j,\underline{\beta}_j)=\exp(z_{ijm}^{**})/[\Sigma_{h=0}^{k_j}\exp(z_{ijh}^{**})] \qquad (1)$$

where $z_{ijm}^{**}=a_j(\vartheta_i\delta_{jm}-\Sigma_{c=0}^{m}\beta_{jc})$. Now, z_{ijm}^{**} can be rewritten as follows:

$$z_{ijm}^{**}=a_j(\vartheta_i\delta_{jm}-\Sigma_{c=0}^{m}\beta_{jc})=\vartheta_i(a_j\delta_{jm})-a_j\Sigma_{c=0}^{m}\beta_{jc}=\vartheta_i\delta_{jm}^{*}-\gamma_{jm}$$

where $\delta_{jm}^{*}=a_j\delta_{jm}$ and $\gamma_{jm}=a_j\Sigma_{c=0}^{m}\beta_{jc}$. With this reparametrization, the model in Equation (1) may be recognized as Bock's Nominal Item Response Model (Bock, *o.c.*; Thissen and Steinberg, 1984).

Therefore, a reasonable way to have the data suggest an appropriate scoring function is to begin with Bock's model. The results may be used to revise the original scoring functions, or to supply initial estimates for the parameters of the Generalized Partial Credit Model.

Of course, one may stop after fitting Bock's model and forget about the Generalized Partial Credit Model. But one has to pay a price: one ends up with a lot of parameters. The view taken in this report is that, because of parsimony, the Generalized Partial Credit Model should be used if possible. Bock's model should only be used as a data analytic tool.

Fitting Bock's model to data may be more expensive than fitting the Generalized Partial Credit Model. As the model only plays its part as a data analytic tool, it would be nice to have a quick but not dirty estimation procedure for that model. Before such a procedure will be described, some invariance properties of Bock's model will be stated.

Bock's Nominal Item Response Model

Using the reparametrization given in the last section, Bock's model can be written as follows:

$$Prob(U_{ijm}=1;\vartheta_i,\underline{\delta}_j^*,\underline{\gamma}_j)=\exp(\vartheta_i\delta_{jm}^*-\gamma_{jm})/[\Sigma_{h=0}^{k_j}\exp(\vartheta_i\delta_{jh}^*-\gamma_{jh})]$$

where $\underline{\delta}_j^*=(\delta_{j0}^*,\cdots,\delta_{jk_j}^*)$ and $\underline{\gamma}_j=(\gamma_{j0},\cdots,\gamma_{jk_j})$. Now, let $\vartheta_i=\alpha\vartheta_i^*+\eta$ where η and α are arbitrary constants, but $\alpha\neq 0$. Then:

$\vartheta_i\delta_{jm}^*-\gamma_{jm}=(\alpha\vartheta_i^*+\eta)\delta_{jm}^*-\gamma_{jm}=\vartheta_i^*(\alpha\delta_{jm}^*)-(\gamma_{jm}-\eta\delta_{jm}^*)=\vartheta_i^*\delta_{jm}^{**}-\gamma_{jm}^*$, where $\delta_{jm}^{**}=\alpha\delta_{jm}^*$ and $\gamma_{jm}^*=\gamma_{jm}-\eta\delta_{jm}^*$. So, the origin and the unit of the ability scale can be chosen freely. For convenience, let $\eta=\Sigma_i\vartheta_i$ and $\alpha^2=\Sigma_i\vartheta_i^2$. Then Bock's model may be written as:

$$Prob(U_{ijm}=1;\vartheta_i^*,\underline{\delta}_j^{**},\underline{\gamma}_j^*)=\exp(\vartheta_i^*\delta_{jm}^{**}-\gamma_{jm}^*)/[\Sigma_{h=0}^{k_j}\exp(\vartheta_i^*\delta_{jh}^{**}-\gamma_{jh}^*)] \qquad (2)$$

where $\underline{\delta}_j^{**}=(\delta_{j0}^{**},\cdots,\delta_{jk_j}^{**})$, $\underline{\gamma}_j^*=(\gamma_{j0}^*,\cdots,\gamma_{jk_j}^*)$, $\Sigma_i\vartheta_i^*=0$ and $\Sigma_i\vartheta_i^{*2}=1$.

Dividing both numerator and denominator of Equation (2) by the *number* $\exp(\vartheta_i^*\delta_{js}^{**}-\gamma_{jt}^*)$, where both s and t are chosen arbitrarily from the set $\{0,\cdots,k_j\}$, does not change the value of the equation, of course. The right hand side of Equation (2) now becomes:

$$\frac{\exp(\vartheta_i^*\delta_{jm}^{**}-\gamma_{jm}^*)/\exp(\vartheta_i^*\delta_{js}^{**}-\gamma_{jt}^*)}{\left[\Sigma_{h=0}^{k_j}\exp(\vartheta_i^*\delta_{jh}^{**}-\gamma_{jh}^*)\right]/\exp(\vartheta_i^*\delta_{js}^{**}-\gamma_{jt}^*)}=\frac{\exp[\vartheta_i^*(\delta_{jm}^{**}-\delta_{js}^{**})-(\gamma_{jm}^*-\gamma_{jt}^*)]}{\Sigma_{h=0}^{k_j}\exp[\vartheta_i^*(\delta_{jh}^{**}-\delta_{js}^{**})-(\gamma_{jh}^*-\gamma_{jt}^*)]}.$$

From this expression it may be seen that the origins of the δ_{jm}^{**}'s and the γ_{jm}^*'s can be chosen freely. Now, let $s=t=0$ and redefine ϑ_i, δ_{jm} and γ_{jm} as follows: $\vartheta_i\equiv\vartheta_i^*$, $\delta_{jm}\equiv\delta_{jm}^{**}-\delta_{j0}^{**}$ and $\gamma_{jm}\equiv\gamma_{jm}^*-\gamma_{j0}^*$. Then Bock's model can be written as follows:

$$Prob(U_{ijm}=1;\vartheta_i,\underline{\delta}_j,\underline{\gamma}_j)=\frac{\exp(\vartheta_i\delta_{jm}-\gamma_{jm})}{1+\sum_{h=1}^{k_j}\exp(\vartheta_i\delta_{jh}-\gamma_{jh})} \tag{3}$$

with $\sum_i^N \vartheta_i=0$, $\sum_i^N \vartheta_i^2=1$ where N is the number of testees, and where $d_{j0}=0$ and $\gamma_{j0}=0$ for all items $j=1,\cdots,K$ with K the number of items in the test.

A Least Squares Procedure

Define $p_{ijmm'} \equiv Prob(U_{ijm}=1 \mid U_{ijm}=1 \text{ or } U_{ijm'}=1;\vartheta_i,\underline{\delta}_j,\underline{\gamma}_j)$, $m \neq m'$. Using Equation (3), it follows that:

$$\text{logit } p_{ijmm'} \equiv \ln[p_{ijmm'}/(1-p_{ijmm'})]=\vartheta_i(\delta_{jm}-\delta_{jm'})-(\gamma_{jm}-\gamma_{jm'}). \tag{4}$$

Logit $p_{ijmm'}$, however, can not be estimated because $U_{ijm}=0$ or $U_{ijm}=1$. This is a pity, for logit $p_{ijmm'}$ is linear in the parameters of the model and would make estimation easy. Suppose, however, that for each subject i a surrogate score t_i is available, which is believed to be positively associated with the latent ability. An obvious choice for a surrogate score is a sumscore or weighted sumscore, using the current scoring function. Divide the subjects into G groups which are homogeneous with respect to the surrogate score. Let N_{gjm} be the number of subjects in group g with $U_{ijm}=1$. Define: $p_{gjmm'}=(\frac{1}{2}+N_{gjm})/(1+N_{gjm}+N_{gjm'})$, where the correction terms will guarantee that logit $p_{gjmm'}$ will exist (Cox, 1970). Moving from individual subjects to groups of subjects, the following restatement of Equation (4) seems reasonable:

$$\text{logit } p_{gjmm'} \equiv \ln[p_{gjmm'}/(1-p_{gjmm'})]=\vartheta_g(\delta_{jm}-\delta_{jm'})-(\gamma_{jm}-\gamma_{jm'}),$$

where ϑ_g denotes a group ability level. Then an appropriate least squares loss function can be formulated:

$$F=\sum_g^G\sum_j^K\sum_{\substack{h}}^{k_j}\sum_{h\neq h'}^{k_j} w_{gjhh'}\{[\vartheta_g(\delta_{jh}-\delta_{jh'})-(\gamma_{jh}-\gamma_{jh'})]-\text{logit } p_{gjhh'}\}^2$$

where $w_{gjhh'}$ is a suitable weight. Note that in the loss function, all response categories within an item are paired with each other.

Least squares estimates have minimum variance if the weights $\{w_{gjhh'}\}$ are elements of the inverted dispersion matrix of the logits. As $p_{gjhh'}$ and $p_{gjhh''}$ $(h \neq h'')$ have a term N_{gjh} in common, the dispersion matrix of the logits will not be diagonal. Nevertheless, since the estimates of the $\{\delta_{jh}\}$ do not have to be too precise, the off-diagonal elements of the dispersion matrix of the logits will be ignored. So, $w_{gjhh'}$ will be taken as the reciprocal of the variance of logit $p_{gjhh'}$.

Now, $p_{gjhh'}$ converges, at least weakly, to a number $\pi_{gjhh'}$. By the *delta method* (Rao, 1973) it can be established that $Var(\text{logit } p_{gjhh'})=[N_{gjh}\pi_{gjhh'}(1-\pi_{gjhh'})]^{-1}$. As $\pi_{gjhh'}$ is unknown, it is replaced by $p_{gjhh'}$. Thus, the weight $w_{gjhh'}$ is defined as: $w_{gjhh'}=N_{gjh}p_{gjhh'}(1-p_{gjhh'})$. Define: $l_{gjhh'}=\text{logit } p_{gjhh'}$, $\Delta_{jhh'}=\delta_{jh}-\delta_{jh'}$ and $\Gamma_{jhh'}=\gamma_{jh}-\gamma_{jh'}$. Then the loss function F can be written as follows:

$$F=\sum_g^G\sum_j^K\sum_{\substack{h\\h\neq h'}}^{k_j}\sum_{h'}^{k_j}w_{gjhh'}\{(\vartheta_g\Delta_{jhh'}-\Gamma_{jhh'})-l_{gjhh'}\}^2. \tag{5}$$

The estimation equations for the parameters can be obtained by equating to zero the partial derivatives of the loss function with respect to the parameters. These partial derivatives are:

$$\frac{\partial F}{\partial \vartheta_g}=2\sum_j^K\sum_{\substack{h\\h\neq h'}}^{k_j}\sum_{h'}^{k_j}w_{gjhh'}\Delta_{jhh'}\{(\vartheta_g\Delta_{jhh'}-\Gamma_{jhh'})-l_{gjhh'}\};$$

$$\frac{\partial F}{\partial \Delta_{jhh'}}=2\sum_g^G w_{gjhh'}\vartheta_g\{(\vartheta_g\Delta_{jhh'}-\Gamma_{jhh'})-l'_{gjhh'}\};$$

$$\frac{\partial F}{\partial \Gamma_{jhh'}}=(-2)\sum_g^G w_{gjhh'}\{(\vartheta_g\Delta_{jhh'}-\Gamma_{jhh'})-l_{gjhh'}\}.$$

Equating these partial derivatives to zero entails the following estimation equations:

$$\left.\begin{aligned}\vartheta_g&=\left[\sum_j^K\sum_{\substack{h\\h\neq h'}}^{k_j}\sum_{h'}^{k_j}w_{gjhh'}\Delta_{jhh'}(\Gamma_{jhh'}+l_{gjhh'})\right]\cdot\left(\sum_j^K\sum_{\substack{h\\h\neq h'}}^{k_j}\sum_{h'}^{k_j}w_{gjhh'}\Delta_{jhh'}^2\right)^{-1};\\ \Delta_{jhh'}&=\left[\sum_g^G w_{gjhh'}\vartheta_g(\Gamma_{jhh'}+l_{gjhh'})\right]\cdot\left(\sum_g^G w_{gjhh'}\vartheta_g^2\right)^{-1};\\ \Gamma_{jhh'}&=\left[\sum_g^G w_{gjhh'}(\vartheta_g\Delta_{jhh'}-l_{gjhh'})\right]\cdot\left(\sum_g^G w_{gjhh'}\right)^{-1}.\end{aligned}\right\} \tag{6}$$

Note that, for each set of three different categories within item j the following restriction holds: $\Delta_{jhh'}+\Delta_{jh'h''}=\Delta_{jhh''}$; similar restrictions hold for the Γ's. The estimation equations can be solved iteratively by an Alternating Least Squares Algorithm. A detailed description of such an algorithm is presented in the Appendix.

A Contrived Example

Data were generated for five thousand subjects and nine items in accordance with the Generalized Partial Credit Model. The first three items had two response categories each. Responses in category 1 were recoded as '3', and the responses in category 0 were randomly divided over the codes '0', '1' and '2'. The items four, five and six had three response

categories each. Responses in catcgory 0 were coded as '0', responses in category 2 were recoded as '3', and the responses in category 1 were randomly divided over the codes '1' and '2'. The last three items had four response categories each, coded as '0', '1', '2' and '3', respectively. The ability levels of the subjects were drawn from a standard normal distribution.

The discrimination indices and category boundaries used in generating the data are displayed in Table 1, together with the category scores obtained by the heuristic procedure. The pattern formed by these scores is displayed in Figure 1. The number of homogeneous groups of testees was set equal to 6. The way the number of groups to be used is specified, is explained in the Appendix.

Table 1

Discrimination indices, boundary parameters and category scores for nine contrived items

Item	a_j	β_j	δ_1	δ_2	δ_3
1	1	-0.5	-0.1023	-0.1087	0.9573
2	2	0.0	0.1093	0.0353	2.0988
3	3	0.5	-0.2875	-0.2177	2.2546
4	1	-0.5, 0.0	0.9367	0.9719	1.8142
5	2	0.0, 0.5	1.9784	2.2092	3.9509
6	3	0.5, 1.0	2.9283	3.0285	3.4811
7	1	-1.0, -0.5, 0.0	1.0712	1.9112	2.9253
8	2	-0.5, 0.0, 0.5	1.8193	3.2965	4.2973
9	3	0.0, 0.5, 1.0	2.7836	4.2389	4.5642

Note: the scores for categories '0' are equal to 0 by default.

In Figure 1, the results for each item are displayed in a strip. Within each strip, the category scores for the item are connected by a solid line. As for each item j, $\delta_{j0}=0$ by definition, a horizontal base line is drawn at level 0.

From Figure 1 it can be seen that the true number of response categories per item is recovered for all items except, perhaps, for items 6 an 9. For all items except items 6 an 9 categories are collapsed as they should be. Category scores for item 6 show a bit of the expected pattern, though: the middle scores are relatively close together. The category scores for item 9 show a slight deviation from the expected equidistance pattern.

Because in this example, data were generated in accordance with the Generalized Partial Credit Model, Equation (1), the relationship $\delta_{jm}=a_j \cdot m$, where a_j is the discrimination index for item j, should hold. One way of looking at the "fit" of the procedure is to estimate discrimination indices by minimizing the following loss functions: $\lambda_j=\sum_{h=1}^{k_j-1}\sum_{h'=h+1}^{k_j}[(\delta_{jh}-\delta_{jh'})-a_j\cdot(s_{jh}-s_{jh'})]^2$ with, for item j, the scores $\{s_{jh}\}$ taken at their

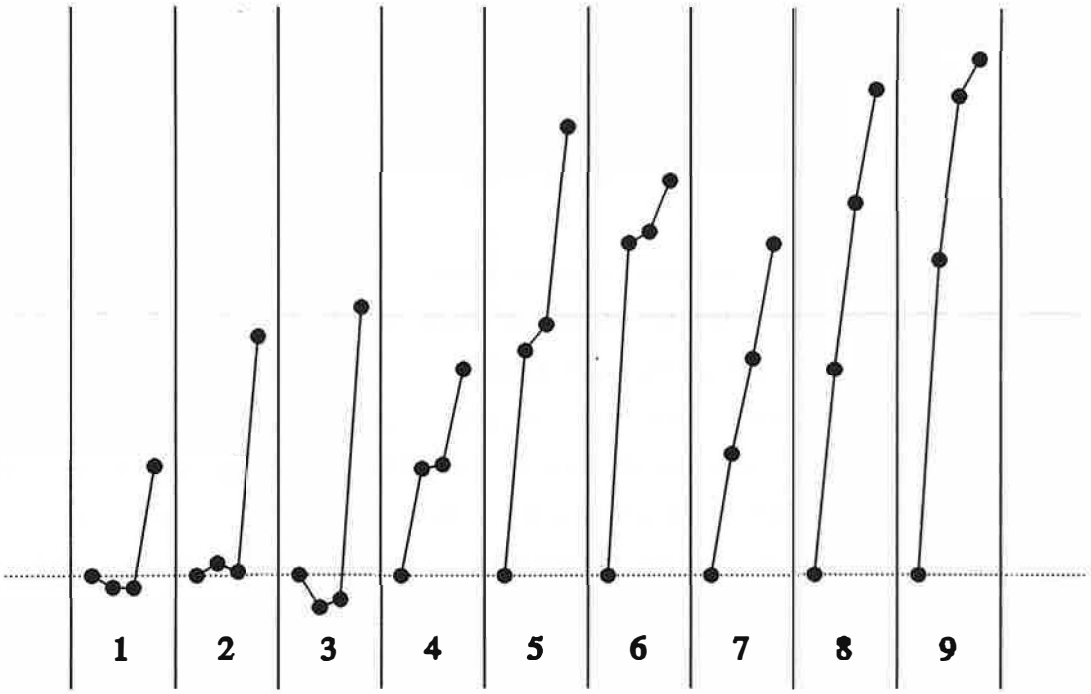

Figure 1 Category parameters for nine contrived items

original values. For example, as the first item is a dichotomous item where the score 0 was randomly divided over the labels '0', '1' and '2', the scores to be used in the loss function λ_1 are equal to 0, 0, 0 and 1 respectively. In the loss functions, differences of category scores are used in order to get rid of the arbitrary origins. By setting the derivative of the function λ_j with respect to a_j equal to zero, the following estimate of the discrimination index a_j is obtained: $\hat{a}_j = ([k_j+1]\Sigma_{h-0}^{k_j}\delta_{jh}\cdot s_{jh} - \Sigma_{h-0}^{k_j}s_{jh}\Sigma_{h-0}^{k_j}\delta_{jh})/([k_j+1]\Sigma_{h-0}^{k_j}s_{jh}^2 - [\Sigma_{h-0}^{k_j}s_{jh}]^2)$. The results are given in Table 2.

Table 2

True and estimated discrimination indices for contrived items

Item	true a_j	$\hat{a}_j$
1	1	1.0276
2	2	2.0507
3	3	2.5230
4	1	0.9071
5	2	1.9755
6	3	1.7406
7	1	0.9616
8	2	1.4369
9	3	1.5148

From Table 2 it is seen that not all discrimination indices are recovered equally well. But as the main purpose of the heuristic procedure is to suggest a plausible scoring function, the conclusion is that the results for this contrived data set are not too bad.

Real Data

In the Dutch National Assessment Program, music is one of the subjects tested. A small test from the program contained ten items, in each of which a tune is played by one instrument. The instrument should be identified. The possible answers to each item were classified into three categories: "correct", "incorrect" and "don't know". Data from one hundred eighty-eight twelve-year old children were used in the heuristic procedure. For each item, the response categories were labelled as follows: "correct"=2, "incorrect"=1 and "don't know"=0. The category scores obtained by the heuristic procedure are displayed in Table 3 and Figure 2. Table 3 also contains the estimated discrimination indices for the music items, using the category labels as scores for all items.

Table 3
Category scores and estimated discrimination indices for music items

Item	δ_1	δ_2	$\hat{a}_i$
1	1.1324	2.5470	1.2735
2	0.4721	0.9719	0.4860
3	0.4721	1.2452	0.6226
4	0.7151	1.2412	0.6206
5	0.6081	0.5148	0.2574
6	1.3364	1.4872	0.7436
7	0.7607	1.4778	0.7389
8	1.7521	2.8553	1.4277
9	1.8584	2.7264	1.3632
10	1.3756	1.5235	0.7618

Note: the scores for categories '0' are equal to 0 by default.

From Figure 2 it can be seen that for most items, a "don't know" response (category 0 for each item) should not only be scored differently from an "incorrect" response, but that such a response should get the lowest possible score. Interestingly, items 5, 6 and 10 seem to be dichotomous items with the "correct" and "incorrect" categories collapsed! If two different response categories m and m' get nearly the same category scores, this means that the conditional probability $Prob(U_{ijm}=1 \mid U_{ijm}=1 \text{ or } U_{ijm'}=1)$ is equal to about one half. Such categories do not discriminate between ability levels. For the music items, items 5, 6 and 10 just separate the "don't know" responses from the other two responses. These items should

be reconsidered before they are entered in a formal analysis. Perhaps the items should be discarded because the instruments used are totally unknown to the testees, or the recordings of the tunes were very bad.

If the category *labels* 0, 1 and 2 are used as category *scores*, it is possible to estimate discrimination parameters of the Generalized Partial Credit Model for these items, as shown in the first example. The results for the music items are displayed in Table 3. Note the relatively small estimated discrimination index for item 5, the item that does not discriminate between "correct" and "incorrect" responses. The estimated discrimination index for item 10 is relatively high. A possible explanation is that the zero category of this items stands more out with respect to the other two categories than is the case with item 5.

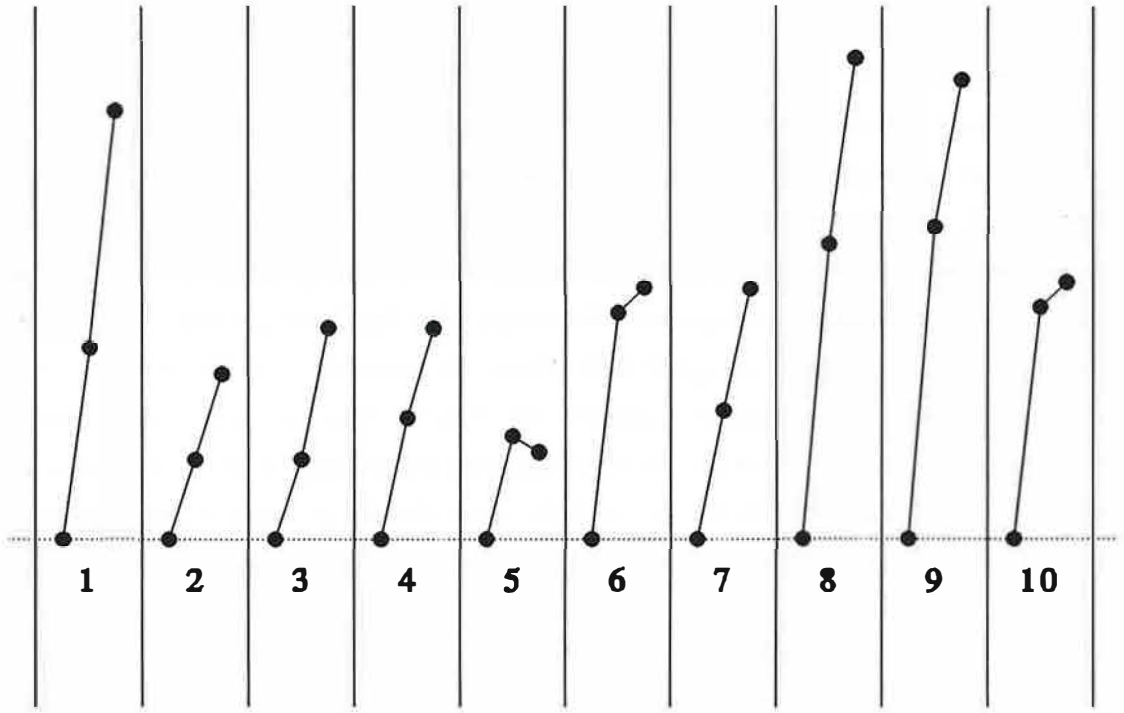

Figure 2 Category scores for ten music items

This small example shows the usefulness of the heuristic procedure. First, it suggests that some items may hamper the fit of the Generalized Partial Credit Model because some items are supplied with the wrong scoring function, as items 5 and 10. Second, it suggests a way to handle response categories whose order is not obvious. It is widely believed that a "don't know" response indicates a higher ability than an "incorrect" response, because the testee can at least eliminate incorrect answers. For these ten music items, this is not the case.

Discussion

In this report, the focus has been on the Generalized Partial Credit Model and its special version, the One Parameter Logistic Model. The model is being used for the analysis of ordered polytomous items; that is, polytomous items where the possible answers can be classified into a finite number of ordered response categories. The model pairs versatility to parsimony as it fits many data sets without too many parameters.

Although both versions of the model contain parameters for discriminatory power and category boundaries for items, their fit depends on the scoring function chosen. Both versions of the model assume that the response categories per item are ordered, and assign integer scores ranging from 0 to m for an item with $m+1$ response categories.

The heuristic procedure presented in this report looks for a scoring function such that the Generalized Partial Credit Model or the One Parameter Logistic Model will fit a data set well. While Bock's Nominal Item Response Model is the core of this procedure, it has been argued that this model *per se* is less attractive because of its many parameters. The heuristic procedure ends up with the same number of parameters, of course; the point is that the parameters are used for revising the scoring functions. For this purpose, the quick heuristic procedure is to be preferred to expensive maximum likelihood estimation.

The heuristic procedure does quite well. From the example with the contrived items it can be deduced that the procedure captures the true scoring function. As a bonus, the procedure may give initial values for the discrimination parameters in the Generalized Partial Credit Model. The example with the music items show that the procedure can pinpoint items that behave contrary to what is expected. Such items, or their scoring functions, are up for revision.

Muraki (1993) also addresses the problem of reordering and collapsing response categories of ordered polytomous items. His procedure involves computation of mean total scores for those testees responding in each category m of item j. The vector of mean scores suggests a reordering or collapsing of response categories for one item at a time. The procedure described in this report gets the same suggestions with much less computations, and for all items simultaneously.

Another way to find an appropriate scoring function is by applying multiple correspondence analysis to a data set. It has been noted before (Schriever, 1986; Gifi, *o.c.*) that multiple correspondence analysis gives a good linear approximation to certain item response models in the case of dichotomous items. From experience with the heuristic procedure described in this report it appears that multiple correspondence analysis may give a good approximation in the case of polytomous items as well.

Appendix

The estimation equations for the heuristic procedure, Equations (6), can be solved by an Alternating Least Squares Algorithm. Such an algorithm may be used when several sets of parameters must be estimated. The idea is to keep all sets of parameters but one fixed to certain values, and to solve for the one set of free parameters. Then, holding these estimates, another set of parameters is freed and estimated. This procedure goes on until the estimates have converged.

From Equations (6) it will be clear that all estimates will depend on the number of groups. As it is not known beforehand what a good choice for the number of groups will be, the number of groups will be determined interactively. So the algorithm to be described contains an outer loop, in which a new number of groups may be specified, and an inner loop in which the parameters and category scores will be estimated. In the outer loop, a frequency distribution of the estimated group ability parameters $\{\vartheta_g\}$ will be displayed. From this display it may be found that some groups contain very few subjects; the number of groups should be set lower in that case. If the specified number of groups seems appropriate, it may be set fixed and no further frequency distribution for the $\{\vartheta_g\}$ will be displayed.

As the latent ability levels of the subjects are unknown, the weighted sum score $t_i = \sum_j^K \sum_h^{k_j} \delta_{jm} U_{ijm}$ will be used. Since this surrogate score is not available at the very beginning of the algorithm, optimal subject scores obtained by a multiple correspondence analysis will be used. These subject scores are normalized. The category scores are initialized at zero.

Let $\underline{\vartheta}^{(o,i)}$ and $\underline{t}^{(o)}$ denote the estimates of the group ability levels and the individual surrogate scores respectively, at the end of outer loop o and inner loop i. Let $\underline{\delta}^{(i)}$ and $\underline{\gamma}^{(i)}$ denote the estimates of the category scores and item parameters at the end of inner loop i. Let $G^{(o)}$ denote the number of groups used in outer iteration o. Then the algorithm performs the following steps, which describe outer loop $(o+1)$ and inner loop $(i+1)$:

i) Specify $G^{(o+1)}$. (If it is decided not to change the number of groups any more, $G^{(o+1)}$ will be set equal to $G^{(o)}$.)

ii) Compute $\underline{\vartheta}^{(o,i+1)}$ from the surrogate scores $\underline{t}^{(o)}$.

iii) Given $\underline{\vartheta}^{(o,i+1)}$ and $\underline{\delta}^{(i)}$, solve the estimation equations for $\underline{\gamma}^{(i+1)}$.

iv) Given $\underline{\vartheta}^{(o,i+1)}$ and $\underline{\gamma}^{(i+1)}$, solve the estimation equations for $\underline{\delta}^{(i+1)}$.

v) Given $\underline{\delta}^{(i+1)}$ and $\underline{\gamma}^{(i+1)}$, solve the estimation equations for $\underline{\vartheta}^{(o,i+2)}$.

vi) If no convergence: set $i := i+1$, $\underline{\vartheta}^{(o,i+1)} := \underline{\vartheta}^{(o,i)}$, $\underline{\delta}^{(i)} := \underline{\delta}^{(i-1)}$, $\underline{\gamma}^{(i)} := \underline{\gamma}^{(i-1)}$ and return to step *iii*.

vii) Compute $\underline{t}^{(o+1)}$ as the weighted-sumscore, using $\underline{\delta}^{(i)}$. Normalize $\underline{t}^{(o+1)}$.

viii) If $\underline{t}^{(o)} \neq \underline{t}^{(o-1)}$ using some convergence criterion, set $o := o+1$ and return to step *i*.

The grouping in step *ii* is done as follows. Order and normalize the surrogate scores and divide them into $G^{(o)}$ fractiles. Then take the mean surrogate score of fractile g as the estimate for ϑ_g.

As measure of convergence in step *vi* the maximum absolute elementwise difference between two consecutive estimates of each parameter vector is computed. As soon as the maximum of these maxima is smaller than a positive constant ε, convergence is reached. A similar criterion is used in step *viii*. The constant ε is set to 10^{-5}.

Within each item all pairs of categories are used, as can be seen from the loss function, Equation (5). This may lead to a prohibitive computing time. Using the fact that in most applications, the order in the *labels* of the categories reflect the order in the categories themselves, the loss function actually implemented is Equation (5), but restrained to using only two "adjacent" categories at a time:

$$F_{implemented}=\sum_g^G\sum_j^K\sum_{h=2}^{k_j} w_{gjh,h-1}\{[\vartheta_g(\delta_{jh}-\delta_{j,h-1})-(\gamma_{jh}-\gamma_{j,h-1})]-\text{logit } p_{gjh,h-1}\}^2.$$

The example with the contrived items was run with six groups of subjects in each iteration. It needed fourteen outer iterations; the number of inner iterations varied between eighteen and twerty.

References

Andersen, E.B. (1973). *Conditional Inference and Models for Measuring*. Mentalhygiejnisk Forskningsinstitut, Copenhagen.

Bock, D.R. (1972). Estimating Item Parameters and Latent Ability when Responses are Scored in Two or More Nominal Categories. *Psychometrika*, **37**, 29-51.

Cox, D.R. (1970). *The Analysis of Binary Data*. Methuen, London.

Gifi, A. (1990). *Non-Linear Multivariate Analysis*. Wiley, Chichester.

Guttman, L.A. (1941). *The Quantification of a Class of Attributes*: *A Theory and Method of Scale Construction*. In: The Committee on Social Adjustment (Ed.), *The Prediction of Personal Adjustment*. Social Science Research Council, New York.

Masters, G.N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, **47**, 149-174.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, **16**, 159-176

Muraki, E. (1993). Information Functions of the Generalized Partial Credit Model. *Applied Psychological Measurement*, **17**, 351-36376

Nishisato, S. (1980). *Analysis of Categorical Data*: *Dual Scaling and its Applications*. University of Toronto Press, Toronto.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications*. Wiley, New York.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. The Danish Institute for Educational Measurement, Copenhagen.

Schriever, B.F. (1986). Multiple Correspondence Analysis and Ordered Latent Structure Models. *Kwantitatieve Methoden*, **21**, 117-132.

Thissen, D. and Steinberg, L. (1984). A Response Model for Multiple Choice Items. *Psychometrika*, **49**, 501-519.

Verhelst, N.D., Glas, C.A.W. and Verstralen, H.H.F.M. (1994). *OPLM*: *One Parameter Logistic Model*. Cito, Arnhem.

Verhelst, N.D. and Glas, C.A.W. (1995). *The Generalized One Parameter Model: OPLM*. In: G.H. Fischer and I.W. Molenaar (Eds.): *Rasch Models*: *Their Foundations, Recent Developments and Applications*. [To appear].