KM 42 (1993)
pg.73-92



# A STOCHASTIC UNFOLDING MODEL DERIVED FROM THE PARTIAL CREDIT MODEL

**N.D.VERHELST AND H.H.F.M. VERSTRALEN**

**National Institute of Educational Measurement (CITO)**
**Arnhem, The Netherlands**

**Abstract**

Binary preference data can be considered as a special case of incomplete data. Rejection of a item is conceived as an observed response, originating from one of two possible latent responses: rejection because the item is too far to the 'left' from the respondent's ideal, or too far to the 'right'. Latent and observed endorsements, however, coincide. The latent responses are modeled through the 3-category Partial Credit Model (PCM). By a reparametrization of the PCM, each item is characterized by a location parameter and a width parameter. Marginal Maximum Likelihood estimators are derived, using the EM-algorithm. A class of statistical tests is derived, which can be used for diagnostic purposes. The 'nuclear energy' data and the 'traffic' data are analyzed. A discussion of the ML-estimator of the subject parameter exemplifies the difficulty of the estimation problem.

Key words: Unfolding, EM-algorithm, partial credit model, missing data, preference analysis.

Requests for reprints should be sent to N.D. Verhelst, P.O. Box 1034, 6801 MG Arnhem.

## Introduction

In psychological scaling theory and in item response theory, dichotomously scored responses have got far more attention than polytomies. In fact, models for polytomously scored responses - the different response categories being considered as ordered or unordered - tend to appear some years later in the literature than their dichotomous counterparts. A nice example is the Partial Credit Model (PCM) of Masters (1982) - also presented in another parametrization by Andersen (1977) (see Glas, 1989) - which is an elegant generalization of the Rasch model (Rasch, 1960). It seems as if sufficient experience has to be collected on the easier dichotomous case before the more general (and usually more difficult) polytomous case can be attacked. One might be tempted to conclude that scaling models should be developed, starting with the 'easy' dichotomous case, with a possible but in any case postponed attempt to generalization to the polytomous case. In the present report an attempt will be made to do the converse, the dichotomous case being considered as a kind of mutilated polytomous case.

As early as 1953, Coombs gave a nice characterization of the distinction between what is nowadays known as (unidimensional) item response models on the one hand and unfolding models on the other hand. However, he did so by means of dichotomously scored responses to a stimulus (an item in an aptitude test or a statement in an attitude questionnaire): "If an individual responds positively to a monotone item, he will respond positively to any item whose scale value is lower. A monotone item with two alternatives (e.g., correct or incorrect) then dichotomizes a continuum; individuals below some particular point all respond one way and individuals above that point respond the other way. If an individual responds positively to a non-monotone item, he will not necessarily respond positively to items on one side of it. A non-monotone item with two alternatives (e.g., endorse or reject a statement) trichotomizes a continuum. An intermediate segment of the continuum contains those individuals who respond positively, and the end segments contain those individuals who respond negatively." (Coombs, 1964, p. 562-563, in a reformulation of his 1953 paper.) Although formulated as if a deterministic mechanism governs the behavior, the idea is clear and up to date: in IRT, items are treated as monotone items, and the behavior is modeled through a monotone item characteristic curve, while in unfolding theory, item characteristic functions are invariably single peaked, a characteristic firmly founded on the theoretical work of Coombs & Avrunin (1977) and Aschenbrenner (1981)

The second part of the above quotation suggests that it might be wise either to collect trichotomously scored responses, or to consider the dichotomous data as a particular reduction from an underlying trichotomy. The approach followed in the present report is to consider a rejection (usually scored as a zero) as an ambivalent response, which represents one of two extreme latent

responses. More formally, let $Y_{vi}$ be the overt (observed) response of subject v on item i, and let $X_{vi}$ be the latent response, coded as '0' if the subject's position is 'far' to the left of the items position, or more precisely, if the item is judged to have too much of the latent attribute in comparison to the subject's ideal, coded as '1' if the item is endorsed, and coded as '2' if the items position is too far to the left of the subjects ideal. So, the observed response $Y_{vi}$ may be considered as a function of the latent response $X_{vi}$ onto $\{0,1\}$, defined by

$$Y_{vi}(X_{vi}) = \begin{cases} 1 \text{ if } X_{vi} = 1 \\ 0 \text{ if } X_{vi} = 0 \text{ or } X_{vi} = 2. \end{cases} \tag{1}$$

As the above paraphrasing suggests, the latent responses may be considered as ordered categories, and the more to the right an individuals ideal is located, the greater the probability that he will choose a higher ordered category. But this means that a non-monotone (dichotomous) item at the level of the overt response coincides with a monotone three-category item at the latent response level (a monotone item can be more generally defined as an item whose expected score is monotonically increasing in the latent variable). From this, the approach to develop a model is evident: Any model which describes adequately the latent trichotomy can in principle be used. All one has to do is to investigate the implications of the partially latent responses.

**The Model and the Interpretation of its Parameters**

At the level of the latent responses $X_{vi}$ the proposed model coincides with the three-category PCM of Masters. The PCM is defined by the so called category response functions which gives the probability of each response as a function of the latent variable $\vartheta$:

$$f_{ji}(\vartheta_v) \equiv P(X_{vi}=j \mid \vartheta_v) = \frac{\exp[\alpha(j\vartheta_v - \sum_{g=0}^{j} \beta_{ig})]}{\sum_{h=0}^{2} \exp[\alpha(h\vartheta_v - \sum_{g=0}^{h} \beta_{ig})]}, \tag{2}$$

$(j=0,1,2;\ i=1,\ldots,k;\ v=1,\ldots,N)$, where k is the number of items, and v denotes a respondent. A graph of these functions is depicted in Figure 1. From (2) it is clear that for all i, $\beta_{i0}$ can be chosen as an arbitrary constant. For reasons of interpretability and in line with the common use in IRT literature, $\beta_{i0}$ is put to zero for all i, leaving two free parameters per item.

As to the interpretation of the latent response $X_{vi}$, it should be clear that the graded response does not express the strength of approval of item i by individual v; it rather gives an indication of the position of the subject point on the latent continuum in comparison to the items position, as

described in the introduction section. Now, this description can be refined somewhat. The graphs of $f_{ji}$ and $f_{j-1,i}$ intersect at $\beta_{ij}$ (j=1,2). Referring to Figure 1, one sees that to the left of $\beta_{i1}$, the 'zero' response is the most probable one, and that the 'two'-response is modal to the right of $\beta_{i2}$. The 'one' response is modal for $\beta_{i1} < \vartheta < \beta_{i2}$. However, the inequality $\beta_{i1} < \beta_{i2}$ is not required by the PCM; if $\beta_{i1} > \beta_{i2}$, this means that responses '1' is never modal. The graph of such a case is given in figure 2.

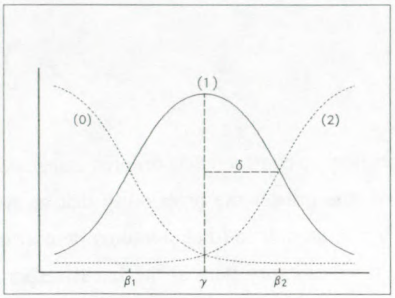



Figure 1
Category response functions
with $\beta_2 > \beta_1$

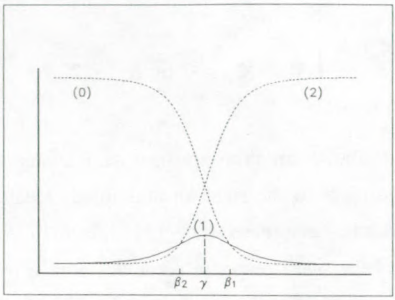



Figure 2
Category response functions
with $\beta_1 > \beta_2$

At the level of the observed responses, $Y_{vi}$, the item characteristic function is directly derived from (1) and (2):

$$F_i(\vartheta_v) \equiv P(Y_{vi}=1 \mid \vartheta) = f_{1i}(\vartheta_v). \tag{3a}$$

It immediately follows that

$$1 - F_i(\vartheta_v) = f_{0i}(\vartheta_v) + f_{2i}(\vartheta_v). \tag{3b}$$

The graph of $F_i(\vartheta)$ is the single peaked curve in Figures 1 and 2.

In order to have a nice interpretation of the parameters in the resulting unfolding model, the underlying PCM is reparametrized as follows. Let for all i

$$\gamma_i = \frac{\beta_{i2} + \beta_{i1}}{2}, \quad \delta_i = \frac{\beta_{i2} - \beta_{i1}}{2}. \tag{4}$$

With this parametrization, $F_i(\vartheta)$ can be written as

$$F_i(\vartheta_v) = \frac{\exp[\alpha(\vartheta_v + \delta_i - \gamma_i)]}{1 + \exp[\alpha(\vartheta_v + \delta_i - \gamma_i)] + \exp[2\alpha(\vartheta_v - \gamma_i)]}. \tag{5}$$

It is easily verified that the derivative $F'_i$ (with respect to $\vartheta$) is given by

$$F_i'(\vartheta) = \alpha f_{1i}(\vartheta)\,[f_{0i}(\vartheta) - f_{2i}(\vartheta)], \tag{6}$$

meaning that $F_i(\vartheta)$ reaches its maximum at the point where the graphs of $f_{i0}$ and $f_{i2}$ intersect:

$$F_i'(\vartheta) = 0 \;\Leftrightarrow\; \vartheta = \frac{\beta_{i2}+\beta_{i1}}{2} = \gamma_i. \tag{7}$$

So $\gamma_i$ can be interpreted as a location parameter of the item on the latent continuum: it gives the value of the latent variable where the probability of endorsement is maximal. This maximal value is given by

$$F_i(\gamma_i) = \frac{\exp(\alpha\delta_i)}{2+\exp(\alpha\delta_i)}. \tag{8}$$

Of course the location parameters are not uniqely determined by (5). Adding an arbitrary constant to $\gamma_i$ and to $\vartheta_v$ leaves (5) invariant. The origin of the scale may be fixed by fixing one of the $\gamma$-parameters or their sum at zero.

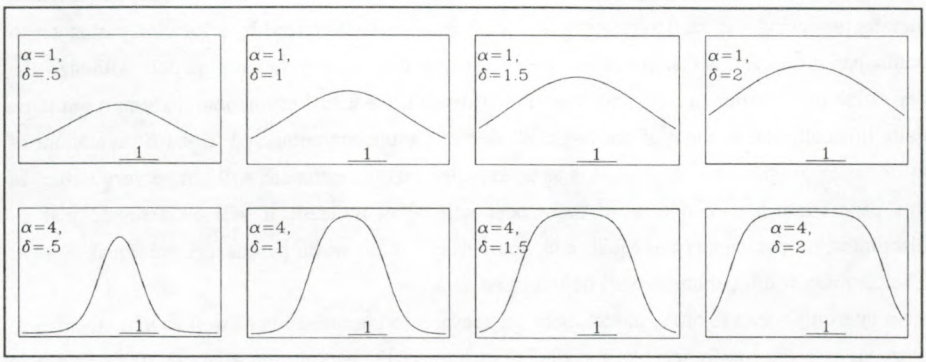



Figure 3 Item characteristic functions for different values of $\alpha$ and $\delta$

From (4) it is clear that $\delta_i$ can be interpreted as a width parameter. The first row of diagrams in Figure 3 shows clearly the interdependence of $F_i(\gamma_i)$ and the width parameter: the higher the maximal probability of acceptance, the broader the range on the latent continuum where the item will be endorsed with high probability. This may seem an unattractive characteristic of the model: a item that is very attractive at $\vartheta = \gamma_i$ is also attractive for a broad range of $\vartheta$ around $\gamma_i$, implying that high popularity at a given point goes together with weak discrimination in the neighborhood. However, this disadvantage can easily be taken away by careful inspection of the meaning of the scale parameter $\alpha$. Although $\alpha$ is arbitrary, fixing it at a specific value has implications for the

interpretation of the latent variable $\vartheta$. Considering a collection of $\vartheta$-values as a distribution, then doubling the value of $\alpha$ necessitates halving the $\vartheta$-values and thus halving the standard deviation of the $\vartheta$-distribution. So if we are to study the behavior of individuals belonging to a specified population, we might choose a scale unit related to the distribution of $\vartheta$, for instance, by fixing its standard deviation to one. By doing so, the $\vartheta$-values no longer can be multiplied by an arbitrary constant, and $\alpha$ becomes a parameter to be estimated. The effect of variation in $\alpha$ in combination with variation in the $\delta$-parameter is depicted in Figure 3: in a single column (with constant $\delta$), the curves become more peaked as $\alpha$ increases and so the item discriminates better in the neighborhood of $\gamma_i$. So, given a scale unit, $\alpha$ can be interpreted as a discrimination parameter. In the limit, as $\alpha \to \infty$, $F_i(\vartheta)$ is a step function:

$$\lim_{\alpha \to \infty} F_i(\vartheta) = \begin{cases} 1 \text{ if } \gamma_i - \delta_i < \vartheta < \gamma_i + \delta_i, \\ 0 \text{ elsewhere}, \quad (\delta_i > 0), \end{cases} \tag{9}$$

producing the trichotomization of the latent continuum as described in the quotation of Coombs, given above.

Of course, one might be tempted to allow variation in the discrimination parameter across items, thereby producing a three parameter model: each item is characterized by a location parameter, a width parameter and a discrimination parameter, yielding a very flexible model. Although the derivation of the estimation equations is straightforward, the actual estimation might turn out to be quite difficult, and in view of the large number of parameters compared to the little amount of information provided by the data, it is to be expected that the estimates will not be very stable. In this paper only a special case of the model will be considered: it will be assumed that all discrimination parameters are equal, and, moreover, that all width parameters are equal as well. The common width parameter will be indicated as $\delta$.

An interesting feature of the model, both with respect to interpretation as well as to mathematical elegance is given by (8) and (9): for all finite values of the parameters, $F_i(\vartheta) \in (0,1)$, implying that its logarithm is well defined everywhere, and from (5) it is easily seen that $\log F_i(\vartheta)$ is arbitrarily many times differentiable, assuring that most of the regularity conditions for constructing statistical tests, based on the ML-estimates, are fulfilled. With respect to the interpretation, the model implies that the probability of endorsement at the ideal point, that is, when $\vartheta = \gamma_i$, is strictly less than one, and can take all values in (0,1). This maximal probability is a simple estimable function of the parameters - given by (8) - in contrast to for instance, the PARELLA model of Hoijtink (1991), where this maximal probability is one per definition, or Andrich's (1988) model, where it is bound to be not larger than ½.

As a final remark, it should be noticed that (8) does not imply that $\delta_i > 0$. A negative width parameter means that $\beta_{i2} < \beta_{i1}$, implying that $F_i(\vartheta) < 1/3$. Although theoretically possible, such situations are likely to be uninteresting for real life applications; so no further special attention will be paid to this case.

**Parameter Estimation**

In IRT models, three methods of parameter estimation are commonly used: the first (and easiest) way is to consider all $\vartheta_v$, as well as $\alpha$, $\gamma_i$ and $\delta$ as parameters to estimate. ML-estimates are those values which maximize the likelihood function jointly with respect to all parameters. This procedure is known as the joint ML method (JML). For the Rasch model it is known that with a finite number of items, JML yields inconsistent estimators for the item parameters. It is likely that an JML procedure will yield inconsistent estimates of the parameters in the present model as well. For this reason, the JML procedure will not be discussed.

A second method, which is rather popular among users of the Rasch model, is the conditional ML method (CML). This procedure is applicable for any model where nontrivial minimal sufficient statistics for $\vartheta$ exist. For the PCM, these sufficient statistics are given by the score S, defined as

$$S_v = \sum_i X_{vi}. \tag{10}$$

However, in the present model, $X_{vi}$ is not observed, but $Y_{vi}$ is. From definition (1), it follows readily that there can exist response patterns $X_v$ and $X_w$ such that $S_v \neq S_w$ and $Y(X_v) = Y(X_w)$. For example, $X_v = (1\ 0)$ and $X_w = (1\ 2)$. Therefore there cannot exist a partition on the sample space $\{Y\}$ which coincides with the partition induced by (10). So CML is not feasible as an estimation procedure.

The third method - which is strictly speaking not just a method, but an extension of the measurement model - is known as Marginal Maximum Likelihood (MML). There, $\vartheta_v$ is no longer considered as a fixed parameter but as an (unobserved) realization of a random variable $\Theta$. The measurement model defined by (2) only defines the conditional probability of the overt response $Y_{vi} = 1$, given $\Theta = \vartheta_v$. By extending this measurement model with a structural model describing the distribution of $\Theta$, and by assuming that the realized sample is a simple random sample from this distribution, the likelihood of a specific response pattern $Y = (Y_1,..,Y_k)$ is given by

$$L(\eta; Y) = \int \prod_i [F_i(\vartheta)]^{Y_i} [1-F_i(\vartheta)]^{1-Y_i} \, dG_\varphi(\vartheta), \tag{11}$$

where $\eta = (\alpha, \delta, \gamma_1,\ldots,\gamma_k, \varphi)$ and $G_\varphi(\vartheta)$ is the distribution function of $\vartheta$, possibly indexed by one or more parameters, denoted $\varphi$. It will be assumed, as is common in IRT modeling, that the

parameters of the $\vartheta$-distribution are independent of the parameters of the measurement model. Maximizing (11) with respect to all parameters is known as MML, and by a result of Kiefer & Wolfowitz (1956), the estimates are consistent under very mild regularity conditions. It is clear that in order to maximize (11), something should be known or assumed about $G(\vartheta)$, and the stronger the assumptions, the more vulnerable the model becomes. In some cases, $G(\vartheta)$ is assumed to be a member of a parametric family of distribution functions - such as a normal distribution - or a more general family of distributions is assumed. (For applications in the Rasch model, see De Leeuw & Verhelst (1986), Folman (1989) and Engelen (1989)). In the present report the normal distribution will be considered, and also a special family of non-parametrized distributions.

First, the normal case will be discussed. Since unit and origin of the scale may be chosen arbitrarily, one may choose the mean of the distribution as origin and its standard deviation as unit; so in the normal case the distribution is completely fixed ($\varphi$ is empty), leaving $r=k+2$ free parameters to be estimated: k location parameters $\gamma_i$, the width parameter $\delta$ and the discrimination parameter $\alpha$.

In IRT modeling where the MML procedure is used, a common device is to consider the $\vartheta$-values as missing observations and to apply the EM-algorithm (Bock & Aitkin (1981), Glas (1989)). Using this approach here makes that data are missing at two levels: the $\vartheta$-value of the sampled individual is missing completely and, as explained above, the responses are only partially observed, because the distinction between an assumed '0' and '2' response is lost at the level of the observations. So one can approach the problem with Y as observed data and the pair $(X,\vartheta)$ as full data. Although the application of the EM-algorithm is straightforward, the computational burden is high, because expectations at two levels have to be computed. Since a large amount of data is modeled as being missing, it can be expected that the EM-algorithm will converge very slowly.

Another possible approach is to consider the pair $(Y,\vartheta)$ as the full data, ignoring (1), and to consider (5) as the complete specification of the measurement model. Although the formulae become more complicated, the amount of modeled missing information is less than in the former case, and it can be expected that the algorithm will converge faster. Therefore, the latter approach is chosen.

Let Y denote the Nxk data matrix, and define $\ell(.) \equiv \log L(.)$. Then, from (5), it is clear that

$$\ell(\eta;Y) = \sum_{v}^{N} \ln \int_{D} f(Y_v|\vartheta)\, g(\vartheta)\, d\vartheta, \qquad (12a)$$

where

free weights, one of them was estimated at zero, suggesting that an equal fit can be obtained with four nodes. A more flexible approach where the nodes are not fixed but have to be estimated (a full non-parametric model) might repair the lack of fit. This will be discussed further in the discussion section.

Table 1
Results of the analyses of the 'nuclear energy' data*)

| | | all items weights estimated | all items weights fixed | items 1-4 weights fixed |
|---|---|---|---|---|
| | $\alpha$ | 1.940 (.155) | 1.287 (.101) | 0.907 (.152) |
| | $\delta$ | 1.783 (.118) | 2.483 (.187) | 3.075 (.497) |
| NOALT | $\gamma_1$ | -1.798 (.118) | -3.243 (.241) | -4.036 (.622) |
| DIFFDE | $\gamma_2$ | -1.339 (.102) | -2.595 (.220) | -3.234 (.532) |
| PROBSOL | $\gamma_3$ | 0.091 (.116) | -0.716 (.181) | -0.859 (.342) |
| SAFEPRO | $\gamma_4$ | 1.180 (.117) | 0.846 (.174) | 0.863 (.423) |
| CLOSFOR | $\gamma_5$ | 2.275 (.195) | 2.484 (.206) | |
| | $G^2$ | 88.62 | 121.1 | 34.77 |
| | df | 20 | 24 | 9 |
| | p | < .0001 | < .0001 | .0001 |

*) standard errors in parentheses.

Table 2
W-statistics for two analyses of the 'nuclear energy' data (weights fixed)

| set | all items | items 1-4 | set | all items | items 1-4 |
|---|---|---|---|---|---|
| { } | -1.186 | -0.735 | {1,2,3} | 2.048 | 0.785 |
| {1} | 0.535 | -0.479 | {1,2,4} | 0.394 | -1.021 |
| {2} | -0.013 | -0.735 | {1,2,5} | -1.788 | |
| {3} | -1.186 | -0.735 | {1,3,4} | -1.180 | -2.240 |
| {4} | -2.676 | -1.962 | {1,3,5} | -0.571 | |
| {5} | -3.533 | | {1,4,5} | -2.236 | |
| | | | {2,3,4} | -0.553 | -1.787 |
| | | | {2,3,5} | 1.257 | |
| | | | {2,4,5} | -2.963 | |
| | | | {3,4,5} | -4.642 | |

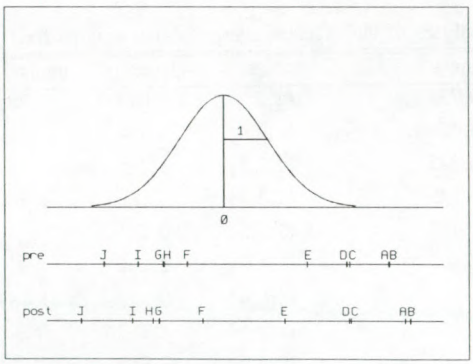
Of course, the lack of fit might be caused because the set of items lacks homogeneity with respect to the model. In Table 2 the W-statistics are displayed. In the analysis with 5 items, it appears clearly that items 4 and 5 in conjunction cause problems. Therefore a reanalysis was done with one of these two eliminated. The last column in the Tables 1 and 2 reports this analysis. Although the

$G^2$ statistic remains significant, the W-statistics are quite acceptable. An analysis with item 4 eliminated gave similar results. An analysis with four items and free weights failed to converge.

A negative W-statistic indicates that illegal triples occur less often than predicted by the model. Since the majority of them is negative, this might point to a lack of stochastic independence: persons answering to the items might have a more or less clear idea of the dimension involved and avoid illegal triples.

Table 3
Summary of the analyses of the 'traffic' data

| | pre | post | joint |
|---|---|---|---|
| $\alpha$ | 1.502 | 1.432 | 1.463 |
| $\delta$ | 2.367 | 2.533 | 2.428 |
| average $\gamma$ | 0.706 | 0.652 | 0.678 |
| $G^2$ (df=1011) | 444.15 | 394.15 | 544.52 |
| p | > .999 | > .999 | > .999 |
| log-lik ($\ell$) | $\ell_1$=-1587.1 | $\ell_2$=-1521.6 | $\ell_{12}$=-3123.2 |




| symbol | label |
|---|---|
| A | NOCONC |
| B | MAXTAX |
| C | HIGHTAX |
| D | PAYTAX |
| E | DECREA |
| F | CARSTA |
| G | OTHPOL |
| H | SELFDE |
| I | NOTHRE |
| J | EXTROA |

Figure 4
Solution of the pre- and post-analysis
of the 'traffic' data

The results of the second example, the 'traffic' data, (two independent samples each of size 300, one representing a pretest and the other a posttest relative to an information campaign) are summarized in Table 3. The three analyses (with the standard normal distribution as assumption) show a remarkable good fit. Nevertheless, a likelihood ratio test (with test statistic

$\lambda = 2(\ell_1 + \ell_2 - \ell_{12}) = 29.0$; df=12, $p < 0.01$; see Table 3) shows that the model parameters cannot be considered equal in the pre- and post-period. From the upper half of Table 3, it is seen that there is a slight tendency for the items to discriminate less at the post measurement, while at the same time the width parameter has increased. Both tendencies indicate that the region of acceptance for all items has increased somewhat, such that the measuring instrument has become less able to discriminate between people. It also means that with unchanged locations of the items, there would be a tendency to accept more items in the postmeasurement than in the premeasurement. Although this is the case in the sample (the average number of accepted items is 5.04 in the pre- and 5.14 in the post- measurement), this tendency is weakened by the changed locations. See Figure 4. The ordering at both measurement occasions is almost identical, and the interpretation of the continuum is easy: the positive direction indicates increasing concern with the environment. Besides, a nice pattern is visible when comparing the two scales: the most extreme items diverged still further, while the items in the middle came closer. At the postmeasurement people show less inclination to endorse extreme items. Since for both data sets, the itemparameters were estimated using a standard normal distribution of the latent attitudes, a decrease of the average location parameter is equivalent to an increase of the average attitude. So the estimate of the average increase is -(0.652-0.706) = 0.054. The standard error of the average location parameter estimate is 0.068 and 0.076 for pre- and postmeasurement respectively. The estimated increase does not differ significantly from zero (t = 0.53).

**Properties of Maximum Likelihood estimators of $\vartheta$.**

As explained above the present unfolding model can be viewed as a Partial Credit model with missing information on the two extremes of the three possible item scores. Therefore, it is to be expected that the data contain less information on the person parameter $\vartheta$ than the full data. Also the likelihood functions of the data will show some changes that might complicate the estimation procedures. The consequences of these two problems as well as the bias of the $\vartheta$-estimates will be explored with the aid of two examples.

The first example builds on the 'nuclear energy' data; the second example is chosen to reveal some properties in a more pronounced way. It consists of two groups of three items with gamma parameters resp. -1.50, -1.45, -1.40, and 1.40, 1.45, 1.50, two tight clusters, located symmetrically about the origin of the scale. The scale parameter $\alpha = 1.0$, and the width parameter $\delta = 2.0$. Choosing a smaller $\delta$ results in a greater loss of information, because the nondistinguishable item categories gain in likelihood at the cost of the middle category. The results

are given in the Figures 5 through 8. The left part of each Figure refers to the 'nuclear energy' data, the right part to the 'two cluster' example.

Figure 5 shows the loss of information of the unfolding model with respect to the full PCM. The solid line represents the information in the PCM and the dashed line the information in the present unfolding model. Especially from the artificial example it emerges that the loss of information is the most severe at a concentration of gamma's. This is a result of the fact that the item information function vanishes at its value of $\gamma$.

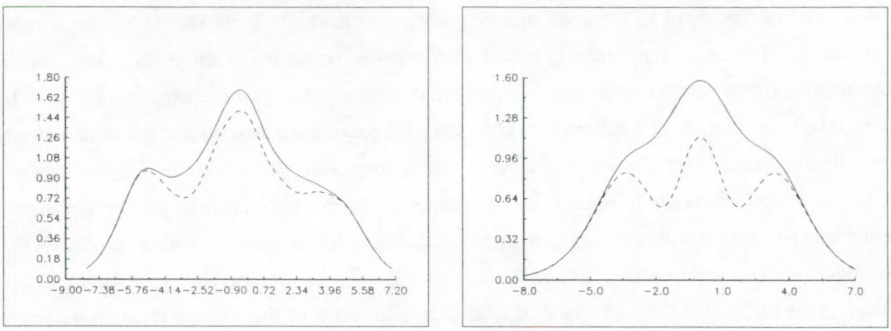



Figure 5. Information loss due to collapsing item scores 0 and 2.

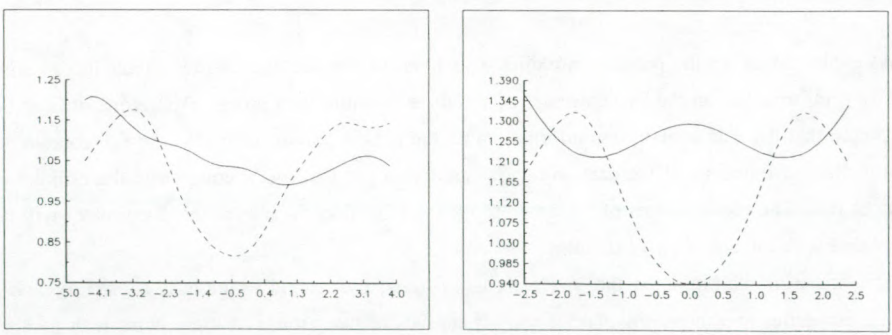



Figure 6 'Exact' error of estimation (solid line) and its approximation

The information function $I(\vartheta)$ is often used as measure of estimation accuracy, because the standard error of estimation can be approximated by $I(\vartheta)^{-\frac{1}{2}}$. Figure 6 shows the standard deviation

$$f(Y_v \mid \vartheta) = \prod_i [F_i(\vartheta)]^{y_{vi}} [1 - F_i(\vartheta)]^{1-y_{vi}}, \tag{12b}$$

$g(\vartheta)$ denotes the standard normal probability density function and D is the domain of $\vartheta$. Taking the partial derivative of $\ell$ with respect to an element of $\eta$ yields

$$\frac{\partial}{\partial \eta_j} \ell(\eta; Y) = \sum_v \int_D \frac{\partial \ln f(Y_v \mid \vartheta)}{\partial \eta_j} h(\vartheta \mid Y_v) \, d\vartheta \,, \tag{13}$$

where $h(\vartheta \mid Y_v)$ is the conditional probability density function of $\vartheta$ given the response pattern $Y_v$. The log-likelihood is maximized by equating (13) to zero, and solving for $\eta$. This may be complicated because both functions in the integrand of (13) are a function of $\eta$. In a single iteration of the EM-algorithm the conditional density is treated as a completely known function, $\eta$ being evaluated at the solution value of the previous iteration. Therefore only expressions for the first function in the integrand of (13) have to be derived. It proves useful to define the following auxiliary functions:

$$h_{1i}(\vartheta) \equiv h_{1i} = \frac{(f_{0i} - f_{2i})}{1 - f_{1i}}, \tag{14a}$$

$$h_{2i}(\vartheta) \equiv h_{2i} = \alpha[(\vartheta - \gamma_i) h_{1i} + \delta], \tag{14b}$$

Since the sign of $\alpha$ is irrelevant, it is convenient to restrict $\alpha$ to the positive reals. In order to keep $\alpha$ in its domain during iterations, all derivatives are taken with respect to $\ln \alpha$ rather than $\alpha$ itself. By straightforward differentiation it is found that

$$\frac{\partial \ln f(Y \mid \vartheta)}{\partial \ln \alpha} = \sum_i h_{2i} (y_i - f_{1i}) \,, \tag{15a}$$

$$\frac{\partial \ln f(Y \mid \vartheta)}{\partial \delta} = \alpha \sum_i (y_i - f_{1i}) \,, \tag{15b}$$

$$\frac{\partial \ln f(Y \mid \vartheta)}{\partial \gamma_i} = -\alpha \, h_{1i} (y_i - f_{1i}) \,. \tag{15c}$$

Substituting (15) into (13) and equating to zero gives the likelihood equations. In order to evaluate the integral, a numerical integration method can be used. A natural choice for this problem is to use Gauss-Hermite quadrature, with a suitable number Q of quadrature points. Loosely speaking, Gauss-Hermite quadrature can be described as replacing the original normal density function (which is continuous), by a discrete distribution, where all the probability mass is concentrated at the Q quadrature points. Quadrature points as well as the mass associated with each of them are fixed by the method itself, and are published (e.g., Abramowitz & Stegun, 1964).

In order to arrive in a smooth way to the non-parametric family of distributions, one can in a way reverse the above reasoning: the discrete distribution used in Gauss-Hermite quadrature can be considered as the model itself, and since Gauss-Hermite integration is a good approximation to integration with the normal density as part of the integrand, this discrete model can be interpreted as an approximation to the continuous normal one. This reasoning has a double advantage: in the first place the accuracy of the Gauss-Hermite integration is no longer of primary concern, since the basic model postulates a specific discrete distribution; in the second place, there is a natural way to generalize this model to a more general family. Fixing Q at a particular value, the above model may be relaxed in two respects: the weights or probability masses, $\omega_q$, $q=1,\ldots,Q$, attached to the quadrature points or nodes may be considered as variable (with the only restriction of being nonnegative and their sum being one) while the nodes remain fixed, or, still more general, nodes as well as weights may be considered as variable. The latter problem will not be considered in detail here. In the case where the weights have to be estimated, these very weights act as parameters of the model, although the distribution function does not belong to a parametric family, this being a reason to call it a non-parametric or semi-parametric distribution. Strictly speaking, the distribution is a simple multinomial distribution with index 1. By straightforward algebra, it is found that in each iteration of the EM-algorithm, the $\omega$-values are given by

$$\omega_j = \frac{\sum_v h(\vartheta_j | Y_v)}{\sum_q^Q \sum_v h(\vartheta_q | Y_v)}, \qquad (j=1,\ldots,Q). \tag{16}$$

Of course, changing the parameters $\omega$ will automatically change the mean and the variance of the distribution. In order to keep the mean at zero and the variance at 1, a simple linear transformation of the nodes can be carried out, together with the accompanying transformation of $\alpha$, $\delta$ and the $\gamma$-parameters described above.

Estimates of the standard errors of the parameter estimates through the inversion of the observed information matrix can be computed without evaluating the second partial derivatives of the log-likelihood function (12a), using an identity given by Louis (1982), which requires only evaluation of the expressions (15) and their partial derivatives. A more detailed discussion of this method is contained in Verhelst and Glas (1993).

**Testing the Model**

An overall test of the model is given by the well known $G^2$ test:

$$G^2 = 2N\sum_s p_s \ln\frac{p_s}{\hat{\pi}_s} \tag{17}$$

where the sum runs over all possible response patterns, $p_s$ denotes the observed proportion and $\pi_s$ the expected proportion of pattern s. $G^2$ is asymptotically chi squared distributed with $2^k$ - 1 - r degrees of freedom, where r is the number of parameters estimated. In order to test a restricted model against a more general one, the difference of the associated $G^2$ statistics is asymptotically chi squared distributed with the difference between the number of estimated parameters giving the number of degrees of freedom. This test is only possible if the parameter space of the restricted model is a subspace of the parameter space of the more general model. However, if the overall test yields a significant result, indicating that the model assumptions are not valid, one might need a more specific test to find out which of the assumptions fail. For example, one or a few items may violate the model assumptions, and eliminating them might result in a valid model for the remaining items.

If the parameters of the model are known, a class of one-degree-of-freedom tests is easily constructed as follows. Let $K = 2^k$ and let S be the index set $\{1,\ldots,K\}$ and let the response patterns be ranked in some arbitrary but fixed way. Define the K-variate random variable $\underline{Z} = (Z_1,\ldots,Z_K)$ as

$$Z_s = \begin{cases} 1 \text{ if response pattern s is observed,} \\ 0 \text{ otherwise,} \quad (s=1,\ldots,K). \end{cases} \tag{18}$$

Let $\pi_{s|\vartheta}$ be the conditional probability of observing pattern s given $\vartheta$, then $\underline{Z}$ is distributed multinomially with parameter $\underline{\pi} = (\pi_1,\ldots,\pi_K)$, where $\pi_s$ is given by

$$\pi_s = \int_\Omega \pi_{s|\vartheta} g(\vartheta)\,d\vartheta. \tag{19}$$

Now, let I be any strict and non-empty subset of S, and define $Z_I$ as

$$Z_I = \sum_{s\in I} Z_s, \tag{20}$$

then it follows easily that

$$\pi_I \equiv E(Z_I) = \sum_{s\in I} \pi_s, \quad \text{and} \quad \text{var}(Z_I) = \pi_I(1-\pi_I). \tag{21}$$

Since for all observations in the sample, $\underline{Z}$ is identically distributed, the average of the $Z_I$, i.e. the proportion of response patterns belonging to I and denoted $p_I$, is asymptotically normally distributed. Hence the statistic

$$W_I \equiv \sqrt{N}\,\frac{p_I - \pi_I}{\sqrt{\pi_I(1-\pi_I)}} \tag{22}$$

follows asymptotically the standard normal distribution.

In case the parameters are estimated from the data, it does in general not suffice to replace the theoretical proportions $\pi_I$ by their estimates in order to preserve the asymptotic normality of the test statistic; a suitable correction has to be applied to the the test statistic itself. This correction is easily derived for exponential family models, but not otherwise (Glas & Verhelst, 1990). Experience in the framework of a generalization of the Rasch model (Verhelst, 1992) shows that ignoring the correction does not lead to gross errors in many cases. Therefore, (28) with the ML-estimates substituted for the true theoretical probabilities will be used as test statistic.

As to the choice of I, the approach of van Schuur (1989) will be followed. In the deterministic model, a partial response pattern (1 0 1) on the items i, j and h with $\gamma_i < \gamma_j < \gamma_h$ is evidence against the model. In a probabilistic model, the proportion of this partial response pattern can be compared with its expected value using (22) and defining I as the subset of response patterns comprising this partial pattern. As a shorthand notation this subset will be denoted as {i,j,h}. From the above definition it is clear that

$$\pi_{\{i,j,h\}} = \int_D F_i(\vartheta)\,[1 - F_j(\vartheta)]\,F_h(\vartheta)\,g(\vartheta)\,d\vartheta. \tag{23}$$

In order to summarize the many test statistics having item i as the middle or as a lateral item in an illegal triple, one can define I as the subset of response patterns having at least one illegal triple comprising item i. This subset will be denoted as {i}. Finally, an overall test is constructed by choosing I as the subset of response patterns comprising at least one illegal triple. This subset will be denoted as { }. In contrast to (23), there are no compact expressions for $\pi_{\{i\}}$ and for $\pi_{\{\}}$. So, for each of the K possible response patterns, it has to be checked if they belong or not to {i} or { }. Therefore these tests are only useful if the number of items is not too large.

**Examples**

In Table 1 the results of some analyses of the 'nuclear energy' data (N=600) are summarized. Both solutions with five items yield a scale that is clearly interpretable as a bipolar scale as suggested by the initial ordering given in the introductory chapter. It is clear, however, that the solution with the normal distribution (weights fixed) does not fit the data very well. Although the analysis with estimated weights yields a significant drop of the $G^2$ statistic, the solution does not reproduce the data very well. Both analyses used five quadrature points, and in the analysis with

of the 'exact' conditional distribution of the Maximum Likelihood (ML) estimator given $\vartheta$ and its approximation $I(\vartheta)^{-½}$. 'Exact' is put between quotes because the ML estimator for the zero response pattern (no proposition is supported) does not exist. In the calculation of the conditional distribution of the ML estimator given $\vartheta$, the estimator for the zero pattern was chosen to be equal to $\vartheta$ itself,so as to prevent a distortion of the results by this problem. Moreover, the probability of a zero pattern is smaller than 1% for in the two plots for resp. $-3.65 < \theta < 2.13$, and $-2.10 < \theta < 2.10$, and smaller than 0.3% in the center of the plots. The two plots show that, the information function gives, in general too optimistic an impression of the accuracy of the ML estimator, especially round the mean of the $\gamma$ parameters of the scale. However, the differences are relatively small.

Besides the variance of the ML estimator also its bias $[\vartheta - E(\hat{\vartheta} \mid \vartheta)]$ is an important property. The bias function for the two examples is shown in Figure 7. For the zero response pattern, the above mentioned strategy is applied again. From the plots it emerges that, with respect to the Rasch model and the Partial Credit model the bias of the ML estimator is reversed. The estimator tends toward the center instead of away from it.

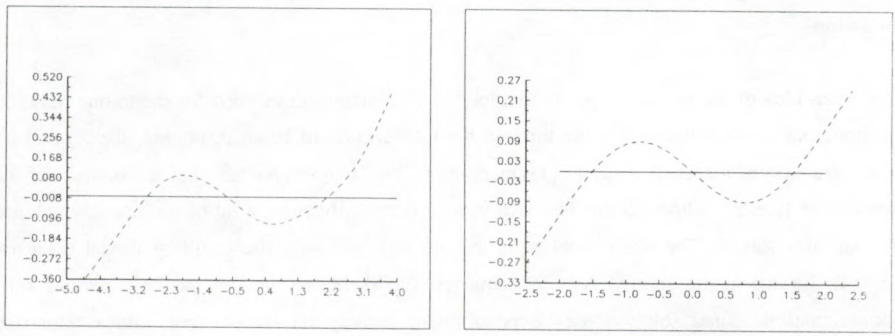



Figure 7 Bias $[\vartheta - E(\hat{\vartheta} \mid \vartheta)]$ of the ML estimator

There are many reponse patterns in the present two examples for which the likelihood shows more than two maxima. Fortunately most times the local maxima are located at a large enough distance from the global maximum and the initial estimator to cause serious problems, although straightforward application of the Newton algorithm in the computation of the estimate may cause problems. The secant method showed very efficient. However, the likelihood of pattern 10010 in the 'nuclear energy' example shows that problems cannot be excluded. The first plot in Figure 8 shows that two (almost) equal maxima are not imaginary. The second plot shows that the likelihood

for the pattern with only ones allows an ML estimator. One of the two extreme patterns in this model causes no problem in this respect.

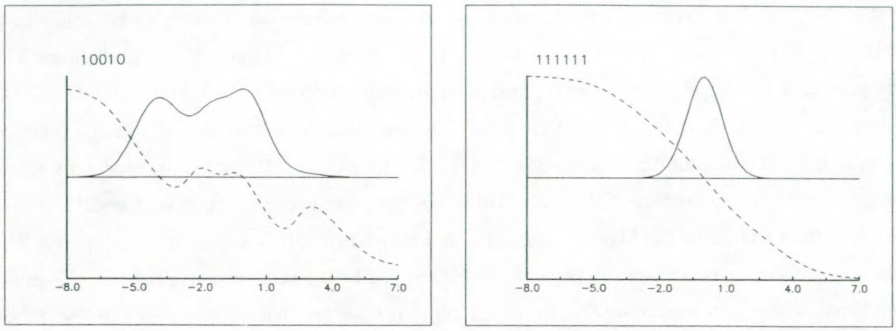



Figure 8 The log-likelihood function (solid line) and its derivative

**Discussion**

The main idea of the present paper is to apply an IRT model, developed for monotone items to non-monotone items. This is possible through the introduction of latent responses (the X) and by the construction of a special many-to-one mapping of the latent responses onto the domain of the observations (the Y). Although this key idea is very simple, it causes a lot of problems which are only partially solved. The main problem is that by this mapping, the resulting model does not belong to the exponential family. As a consequence, the likelihood function can have several local maxima, and the final solution may depend rather heavily on the starting values. Standard procedures, such as the Newton-Raphson algorithm may easily fail (see Figure 8), not only in the estimation of $\vartheta$, but also for the other parameters: it can easily be seen from (5) that all partial derivatives with respect to a $\gamma$ parameter and $\vartheta$ are similar. As a consequence the development of a reliable computer program to compute the estimates depends on many ad hoc rules and the one example reported where convergence was not reached, is but one example of the misery encountered in the analysis of a considerable number of data sets.

Adding parameters, such as allowing the $\delta$ or $\alpha$ parameters to vary or trying to estimate the nodes of the $\vartheta$-distribution, certainly will aggravate the problem. For the examples reported, the analyses with free weights required about 10 times as many iterations as the analyses with fixed weights.

Apart from the algorithmic problems, however, there is also a hard theoretical problem: There is no proof that the model presented here is identified, and the use of the model is justified through

the finding of plausible and interpretable solutions. But some work on an extension of the model with free $\delta$ parameters (not reported here) shows that in many cases no solution is found, or the computed solution is highly implausible. It is our conjecture that this extension yields an unidentified model.

From the point of view of interpretation, all these problems are caused because the data collection procedure is not able to distinguish between a latent '0' and '2' response. With regard to future investment of effort in unfolding models an important decision is at stake. On can continue to develop clever heuristics to face the computational problems mentioned and try to prove that the model and its extensions are identified (or not). But it is also possible to try to devise clever data collection procedures which allow for a distinction between a latent '0' and '2'. One might for example ask the respondent, after the collection of the binary responses, if he is able to classify the rejected items into two piles using the criterion of similarity with regard to the reasons of rejection, and to rankorder to two piles with repect to the subset of endorsed items. If successfull, the latent zeros and twos are distinguished, although not identified. The use of this information can be a challenge for future model development.

**References**


Abramowitz, M. & Stegun, I.A. (1964). Handbook of Mathematical Functions. Washington D.C., National Bureau of Standards.

Andersen, E.B. (1977). Sufficient Statistics and Latent Trait Models. Psychometrika, 42, 69-81.

Andrich, D. (1988). The Application of an Unfolding Model of the PIRT Type to the measurement of attitude. Applied Psychological Measurement, 12, 33-51.

Ashenbrenner, K.M. (1981). Efficient Sets, Decision Heuristics and Single-peaked Preferences. Journal of Mathematical Psychology, 23, 227-256.

Bock, R.D. & Aitkin, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: an Application of an EM Algorithm. Psychometrika, 46, 443-459.

Coombs, C.H. (1953). Theory and Methods in Social Measurement. In: L. Festinger & D. Katz (Eds), Research methods in the behavioral sciences. New York, The Dryden Press.

Coombs, C.H. (1964). A Theory of Data. New York, Wiley.

Coombs, C.H. & Avrunin, G.S. (1977). Single-peaked Functions and the Theory of Preference. Psychological Review, 84, 216-230.

De Leeuw, J. & Verhelst, N.D. (1986). Maximum Likelihood Estimation in Generalized Rasch Models. Journal of Educational Statistics, 11, 183-196.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-38.

Engelen, R.J.H. (1989). Parameter Estimation in the Logistic Item Response Model. Enschede, doctoral dissertation.

Follmann, D. (1988). Consistent Estimation in the Rasch Model Based on Nonparametric Margins. Psychometrika, 53, 553-562.

Formann, A.K, (1988). Latent Class Models for Non-monotone Dichotomous Items. Psychometrika, 53, 45-62.

Glas, C.A.W. (1989). Contributions to Estimating and Testing the Rasch Model. Arnhem, CITO.

Hoijtink, H. (1991). PARELLA: Measurement of Latent Traits by Proximity Items. Leiden, DSWO Press.

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. Annals of Mathematical Statistics, 27, 887-903.

Louis, T.A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. Journal of the Royal Statistical Society, Series B, 44, 226-233.

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

Molenaar, I.W. (1991). A heuristic to find the rank order of the stimuli in unfolding data. Groningen, Heijmans Bulletin.

Rasch, G. (1960). Probabilistic Models for some Intelligence and Attainment Tests. Copenhagen, Danish Institute for Educational Research.

Schuur, W.H. van (1989). Unfolding the german political parties: a description and application of multiple unidimensional unfolding. In: G. de Soete, H. Feger & K.C. Klauer (Eds), New developments in psychological choice modeling. Amsterdam, North Holland, pp. 259-290.

Verhelst, N.D. (1992). Het eenparameter logistisch model (OPLM). Een theoretische inleiding en een handleiding bij het computerprogramma. Arnhem, Cito, OPD Memorandum, 92-3.

Verhelst, N.D. & Glas, C.A.W. (1993). A dynamic generalization of the Rasch model. Psychometrika, 58, (in press).