Measurement and Research Department Reports 96-1

Estimating Integer Parameters in IRT Models for Polytomous Items

H.H.F.M. Verstralen

Cito

Measurement and Research Department Reports 96-1

Estimating Integer Parameters in IRT Models for Polytomous Items

H.H.F.M. Verstralen

Cito
Arnhem, 1996

Abstract

When the number of observations per item is moderate (< 3000) the number of item parameters in the Nominal Response Model of Bock (1972) is too large, especially with polytomous items, for accurate and sometimes even unique estimates to be expected. A solution is suggested by creating a link between this model and three simpler IRT models by estimating integer category scores, and possibly also integer discrimination parameters. Given the integer parameters as constants the remaining parameters can be estimated with a Restricted Maximum Likelihood procedure. In a simulation study four procedures to estimate integer category scores are investigated.

Key words: item response theory, polytomous items, category scores, integer estimation, optimal scores, multiple correspondence analysis.

Introduction

The Generalized Partial Credit Model (GPCM, Muraki, 1992), the One Parameter Logistic Model (OPLM, Verhelst, & Glas, 1995, Verhelst, a.o., 1995), and the PCM (Masters, 1982) can be viewed as progressive specializations of the Nominal Response Model (NRM) of Bock (1972). All these models are used to describe responses to items with two or more response categories. However, the NRM contains at least twice as much to be estimated item parameters as the other three models. Especially with moderately size data sets the NRM-parameters may be inaccurately estimated or even be undetermined, while the simpler models are more likely to yield unique estimates with acceptable accuracy. Moreover, the OPLM, and PCM offer a sufficient statistic for the person parameter, and, therefore, the opportunity to consistently estimate the item parameters by Conditional Maximum Likelihood (CML). A disadvantage is, of course, that the simpler models require appreciably stronger assumptions.

The approach suggested here can be viewed as a compromise, that almost keeps the advantages of the NRM, also with smaller data sets, while at the same time avoiding the assumptions of the (G)PCM and OPLM. Moreover, this approach may reduce the number of parameters even further. Some characteristics of the proposed procedure are investigated with a simulation study.

The Nominal Response Model and Three Specializations

Let $\boldsymbol{x} = (x_1,\dots,x_I)$ be a response pattern with $x_i = j$ the response on item i $(i = 1,\dots,I)$, $(j = 0,\dots,J_i)$, and let

$$z_{ij}(\theta) = \exp\left(\delta_{ij}\theta_v - \zeta_{ij}\right), \tag{1}$$

where δ_{ij} is the score parameter and ζ_{ij} the propensity parameter for category j of item i. θ_v denotes the ability of subject v $(v = 1,\dots,V)$. In the NRM the probability for subject v to score in category j of item i is given by

$$f_{ij}(\theta) \doteq f(x_{ij} \mid \theta) = \frac{z_{ij}(\theta)}{\sum_{h=0}^{J_i} z_{ih}(\theta)}. \tag{2}$$

It is easily checked that the elements of the vector $\boldsymbol{\delta}_i$ are determined by (2) up to an additive constant. Therefore, a linear restriction on $\boldsymbol{\delta}_i$ must be imposed, for instance $\delta_{i0} = 0$. For the same reason an arbitrary linear restriction has to be imposed on the elements of the vector ζ_i, often $\zeta_{i0} = 0$ is chosen. Moreover, the zero and scale of the ability distribution are also arbitrary and are usually fixed by assuming a standard normal ability distribution.
The 'sufficient statistic' for θ is denoted as

$$r_S = \sum_{i=1}^{I} \delta_{ix_i},$$

which, in the NRM is not a statistic, because it depends on the itemparameters.

In the GPCM and OPLM the category scores $\boldsymbol{\delta}_i$ are restricted to be a product of a discrimination parameter a_i and the index j of category $j : \delta_{ij} = ja_i$. In the GPCM the discrimination parameter a_i is estimated, whereas in OPLM it is considered a known constant. In the PCM (Masters, 1982) it is even assumed that all a_i are equal. In OPLM and PCM the scale of the ability distribution is fixed by the selected vector $\boldsymbol{a}$.

The OPLM and PCM, in contradistinction to the NRM and GPCM, have the advantage that r_s is really a statistic, independent of item parameters, and enables consistent CML estimation of the item parameters ζ_{ij}. When the parameters are estimated by CML the origin of the item parameter scale has to be fixed by a linear restriction on ζ. It is an obvious disadvantage of OPLM that the user has to acquire his prior knowledge about a_i from other sources. To compensate this disadvantage OPLM supplies the so called M-statistics (see Table 1) that indicate whether the discrimination was well chosen, or inform about the direction and size of a desired change.

However, experience shows that the assumption to identify the score of a category with its index ($\delta_{ij} = ja_i$) easily leads to erroneous results (see the korfball example below). The alternative is to drop the assumption that the category score of category j equals the index j, and instead to estimate an integer category score s_{ij}. With the OPLM also an integer discrimination parameter a_i has to be estimated. If

$$z_{ij} = \exp\left(a_i s_{ij}\theta - \zeta_{ij}\right), \tag{3}$$

is substituted in (2) one obtains an extended version of the GPCM or the OPLM, without the discussed assumptions. The propensity parameters ζ can, given the estimated integer parameters, be estimated by a restricted CML or MML procedure. The extended version may deviate from the original in two respects.

1 Different categories may share the same score. These categories are called 'collapsed'.

2 The difference between two adjacent category scores may be larger than one. The integer scores in between the two are called 'skipped'.

Suppose that two categories j and j' of an item share the same score, then with respect to the estimation of the ability θ these two categories can be considered identical. Consider the ratio of the two likelihoods as a function of θ given that a subject scores in category j, and given that the subject scores in category j'. It is easily checked that this ratio is a constant, independent of θ. This means that responding in either of the two categories j or j' gives the same information about θ. Therefore, unless there is some interest in estimating the separate category parameters, the two categories j and j' can be considered as one collapsed category j'' with parameter

$$\eta_{j''} = -\ln\left[\exp(-\eta_j) + \exp(-\eta_{j'})\right].$$

Jansen and Roskam (1986) discuss the case of collapsing categories and show that the PCM cannot fit both the collapsed and noncollapsed categories. If the data have enough power to distinguish between both cases for a pair of categories j and j' of item i, this can be considered a simple example of the estimation of s_{ij} and $s_{ij'}$.

Although in the original formulation of the PCM skipped category scores were not allowed, a note by Glas (1992) and, Wilson and Masters (1993) show that skipped scores can be handled in CML estimation by a simple adaptation of the combinatorial functions. The adaptation amounts to omitting all the response patterns that contain a skipped score from the collection of response patterns that yield the sufficient statistic r_S. The same procedure applies to OPLM. The GPCM does not allow for a CML procedure, but its parameters can be estimated by an MML procedure (Muraki, 1992), where skipped scores do not pose an estimation problem.

An item from a seven item test that measures proficiency in a korfball game can serve as an example. First some results of an OPLM-analysis with the category indices as scores are shown in Tables 1 and 2. Later the results with estimated integer scores will be presented. The S-test is a $\chi^2[df]$ distributed statistic that measures overall model fit. The M-tests embody three variations on a statistic that asymptotically follows the standard normal distribution. They measure model fit with special sensitivity to discrimination. The S- and M-tests of the category indices as scores are shown in Table 1.

Table 1

S- and M-Tests of Item 7 in a Korfball Game from an OPLM-Analysis with the Original Item Scores

Item		S-test	df	P	M	M2	M3
7 APass		7.751	5	.171	1.235	.913	-.984
	[:2]	5.083	6	.533	1.649	1.441	.589
	[:3]	13.546	7	.060	2.277	2.631	1.485
	[:4]	10.499	7	.162	1.914	2.180	.892
	[:5]	9.186	7	.240	-.047	1.113	.116
	[:6]	5.592	7	.588	.055	.489	-.893
	[:7]	13.101	6	.041	.071	-.800	-1.716
	[:8]	11.535	6	.073	-.366	-.616	-1.367
	[:9]	12.287	4	.015	99.999	-1.060	-1.743

The entries $[:j]$ in the first column indicate that the expected conditional probability given θ of scoring in categories $j,\ldots,J$, is compared with the conditional observed proportion of scoring in these categories, where $J = 9$ for this item. In the first row [:1] is omitted. The entry 99.999 in the last row indicates that the M-test could not be calculated for lack of enough observations.

Table 1 shows a tendency for the scores of categories 3 and 4 to be valued too high (3 and 4 are too high scores), and of categories 7 through 9 too low. The overall picture of the M-statistics does not necessitate a change in the 'known' integer discrimination parameter. The general test of model fit $R1c$ (asymptotically distributed as $\chi^2[df]$) and the distribution of p-values of the S-tests for all items are shown in Table 2.

Table 2

Some Overall Results of an OPLM-Analysis of the Original Korfball Data

Distribution of p-values for S-tests.

```
0.----/-----/-----.1-------.2-------.3-------.4-------.5-------.6-------.7-------.8-------.9-------1.
    9/   5/ 11      7        3        0        3        4        2        3        4        3
```

$R1c$ = 219.426; df = 158; p = .0007

Table 2 shows that the p-value distribution, which under the assumption of independent tests is uniform, is shifted toward zero. In actual fact, the S-tests are not independent under the model, certainly not within one item. The dependence of S-tests across items, however, is weak enough to judge the quality of an OPLM-analysis on the extent to which the p-values are 'uniformly' distributed. Moreover, the $R1c$ test indicates that the data significantly deviate from the model. In the sequel it is shown that estimated integer category scores, and discrimination parameters, improve the model fit.

General Approach

Assume that the data set X can be described by the GPCM or OPLM for some set of integer category scores (OPLM and PCM are identical in this approach). The problem of estimating integer category scores s_{ij} will be solved in two steps. The first step is to find a vector valued statistic $\boldsymbol{t}_i$ that is about linear in s_i per item, that is

$$t_{ij} \approx c_i s_{ij} + d_i . \tag{4}$$

In the second step integer score vectors $\hat{\boldsymbol{s}}_i$ are sought with an optimal linear relationship with $\boldsymbol{t}_i$, optimal in a sense to be made more precise below.

First the second step is treated. It is assumed that a statistic $\boldsymbol{t}$ is available. The following lines apply to each item separately, therefore, the item index is omitted. Loosely speaking the element $\hat{\boldsymbol{s}}$ from a set S of integer category score vectors $\boldsymbol{s} = (s_0, \ldots, s_J)$ is to be found that gives the best linear prediction of $\boldsymbol{t}$. More precisely this can be formulated as follows. Let W be a positive definite symmetric matrix of order $J + 1$, and $\mathbf{1}'W\mathbf{1} = 1$, with $\mathbf{1} = (1,\ldots,1)$. For the moment the reader may assume $W = 1/(J+1)\boldsymbol{I}$, with $\boldsymbol{I}$ the identity matrix. Let $\bar{x} = \mathbf{1}'W\boldsymbol{x}$, and $Cov(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{x}'W\boldsymbol{y} - \bar{x}\bar{y}$. Define a set S of all integer category score vectors $\boldsymbol{s}\,(s = s_0,\ldots,s_J)$ to be considered (see below). Given an element $\boldsymbol{s} \in S$ let $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{s},\boldsymbol{t})$ be a vector with elements $v_j = cs_j + d - t_j \; (j = 0,\ldots,J)$. Then the optimum integer score vector is found by minimizing over S the GLS loss function

$$L(\boldsymbol{s};\boldsymbol{t}) = \boldsymbol{v}'W\boldsymbol{v}, \tag{5}$$

where the values c and d are the GLS regression coefficients

$$c = \frac{Cov(\boldsymbol{s},\boldsymbol{t})}{Cov(\boldsymbol{s},\boldsymbol{s})}, \text{ and } d = \bar{t} - c\bar{s}. \tag{6}$$

Formula (5) shows that if s minimizes L then so does $-s$. This indeterminacy can be removed by restricting S to contain only elements s with $Cov(s,t) \geq 0$.

The following equivalent geometrical formulation facilitates some further developments. Let $<x,y>$ be a bilinear form on $\mathbb{R}^{J+1} \times \mathbb{R}^{J+1}$, defined by $<x,y> = x'Wy$. Let P be the projection on the orthocomplement of $\mathbf{1} \in \mathbb{R}^{J+1}$ (subtracting the mean): $P(x) = x - <\mathbf{1},x>\mathbf{1}$ (recall that $<\mathbf{1},\mathbf{1}> = 1$). Further let N normalize the length of a vector to $1: N(x) = x/|x|$, with $|x| = <x,x>^{1/2}$. Using the linearity properties of $<.,.>$ it is easily derived that

$$L(s;t) = |Pt|^2(1 - <NPs,NPt>^2).$$

It follows that minimization over S of $L(s;t)$ is equivalent with maximization over S of $<NPs,NPt> = WCos(Ps,Pt)$, the generalized cosine in the W-metric of the vectors Ps and Pt. Note that $<NPs,NPt>$ equals the ordinary inner product $(NPs,WNPt)$.

The Set S

The loss function L is to be minimized over S, the set of all integer category score vectors s to be considered. In principle the set S contains all vectors of integer scores with $J+1$ elements and at least one element equal to 0, further called 0-vectors. Now recall that models with integer scores were preferred over Bock's model, because they contain less parameters. This advantage tends to be more lost the more the set S is unrestricted. Indeed, with no restriction on S, the extended GPCM and OPLM can approximate the NRM indefinitely close, and the rationale for the current approach vanishes. An obvious way to restrict S is to bound the differences between successive scores, ordered from small to large. Some experience shows that this bound should be chosen as low as 1, 2 or 3. Consider s an ordered member of S, then this bound means that $s_j - s_{j-1} \leq$ 1, 2 or 3 for $j = 1,\ldots,J$. In the sequel we call these bounds B1, B2, B3, etc. Although the bounds B2 or B3 seem very low, and are a severe restriction on S, nevertheless very close linear relationships with any instance of a statistic t can be achieved with integer vectors thus restricted (for B2 a mean W-Cosine of about 0.995, see Table 3). However, for items with $J > 6$, this still leaves a set S too large for exhaustive search (for B2 see Table 3 under #Evaluations). Therefore, it would be of great help if the set S could be reduced further, without excluding the optimal vector

$\boldsymbol{v}$. The following lemma yields enough reduction of the size of S to enable exhaustive search up to $J = 9$, even under B3. Unfortunately, the lemma is restricted to a very limited class of matrices W, which is expressed in the following property.
A matrix W is called permutation-invariant, if $Q'WQ = W$ for any permutation-matrix Q.

Lemma: For permutation-invariant W S needs to contain only vectors $\boldsymbol{s}$ such that $\boldsymbol{s}$ and $\boldsymbol{WNPt}$ have the same weak order ($\boldsymbol{s}$ is called $\boldsymbol{Wt}$-ordered). Because if $\boldsymbol{s}$ is not $\boldsymbol{Wt}$-ordered there is a $\boldsymbol{Wt}$-ordered $\boldsymbol{s}'$ with $L(\boldsymbol{s}';\boldsymbol{t}) < L(\boldsymbol{s};\boldsymbol{t})$.
For example, let $W = k\boldsymbol{I}$, for some constant k, and let

$\boldsymbol{t}$ =	0.12	0.33	0.30	then
$\boldsymbol{s}$ =	1	0	1	is not $\boldsymbol{Wt}$-ordered and can be kept out of S
$\boldsymbol{s}'$ =	0	1	1	is $\boldsymbol{Wt}$-ordered and is a member of S

Proof: Recall that the geometrical reformulation shows that the optimal $\boldsymbol{s}$ maximizes the ordinary inner product $(\boldsymbol{NPs}, \boldsymbol{WNPt})$ over S. Further note that NP does not alter the order of the elements of a vector. Therefore $\boldsymbol{NPs}$ has the same order as $\boldsymbol{s}$. Let $\boldsymbol{u} = \boldsymbol{NPs}$ and $\boldsymbol{w} = \boldsymbol{WNPt}$. Now, suppose that the optimal s is not $\boldsymbol{Wt}$-ordered, then there exist two components u_h and u_l of u with $u_l > u_h$ while $w_l < w_h$ that contribute $x = u_h w_h + u_l w_l$ to the inner product. By exchanging the elements u_l and u_h in $\boldsymbol{u}$ to obtain the vector denoted by $\boldsymbol{u}'$ their contribution to the ordinary inner product becomes $x' = u_h w_l + u_l w_h$. Simple algebra shows that $x' > x$. Denote the integer vector that is obtained from $\boldsymbol{s}$ by exchanging elements l and h with $\boldsymbol{s}'$. Unfortunately, $\boldsymbol{NPs}'$ will not, in general, be equal to $\boldsymbol{u}'$. A sufficient condition for $\boldsymbol{NPs}' = \boldsymbol{u}'$ to hold is permutation-invariance of W. Therefore, if W is permutation-invariant then the $\boldsymbol{Wt}$-ordered $\boldsymbol{s}'$ results in a larger inner product and so in a smaller L. This contradicts the assumption that $\boldsymbol{s}$ is optimal for a permutation invariant W. ■

I did not succeed in finding a similar reduction of S for a broader class of matrices than the permutation-invariant matrices. This is unfortunate. In practice off-diagonal cells of W will generally be set to zero. Consequently, permutation-invariant W reduce to $W = k\boldsymbol{I}$. If one prefers another weight matrix than $k\boldsymbol{I}$, one could, in principle, find the minimum loss over the set S of all relevant permutations of the set of rank ordered vectors conforming to B1, B2, or B3. However, already from the moderate value of

$J = 7$ the size of S becomes prohibitively large (see Table 3 under #Evaluations E). To circumvent this problem, some investigations show that one of the following two heuristic procedures H1 or H2 quickly finds either the optimal 0-vector s, or a 0-vector s' with $1-\varepsilon < <NPs, NPs'> \leq 1$, with $\varepsilon > 0$ and small (Table 3: at most $\varepsilon \approx 0.03$ for H1 and $\varepsilon \approx 0.01$ for H2).

The basic idea of H1 is to start from the integer 0-vector $s = \mathbf{0}$, add a one to its first component to obtain s_0, and calculate $<NPs_0, NPt>$. Next add a 1 to the second component of s to obtain s_1, and calculate $<NPs_1, NPt>$. Repeat this for each of the components of s to find out which one has the largest *W*-Cosine with $\boldsymbol{t}$. Set s equal to this optimum, and repeat the procedure with the new s for all its descendents $s_i \in S$, until the highest possible category score is encountered. Thus traversing a small subset of S 'close in *W*-direction' to $\boldsymbol{t}$, keep the 0-vector s that showed the maximum *W*-Cosine during this process. In this way, while enlarging the length of s, the angle with $\boldsymbol{t}$ is kept almost as small as possible. The basic idea for H2 is similar to H1, but now at a certain s not only all points are checked that can be viewed from s along one of the axes in one step from $\boldsymbol{0}$, but now all points are checked that can be viewed from s as corners of the smallest integer hypercube of which s is the corner nearest to $\boldsymbol{0}$. The procedures are outlined in more detail in Appendix B. To compare H1 and H2 with exhaustive search a small simulation study was performed for items with $J = 2,\ldots,6$. For $J = 7,\ldots,9$ only H1 and H2 were executed. S contains all *0*-vectors under the restriction B2. The extrapolation of the number of evaluated *0*-vectors per item (Table 3, #Evaluations) shows that exhaustive search defeats commonly available computing capabilities already for moderate values of J. For each J 100 times a random diagonal matrix W was drawn and a random vector $\boldsymbol{t}$. Some results are shown in Table 3. The maximum difference between H1-optimal and optimal *W*-Cosines (Max) does not show an increase with J. The mean of these differences does show some positive relation, but this is due to an increase in the percentage of suboptimal solutions. As a matter of fact, the mean deviation restricted to suboptimal solutions, declines for $J = 2$ to 6 from 0.0083 to 0.0037. The agreement of H2 with E is too close to warrant a conclusion with respect to a trend of means or maximum differences. Because H2 clearly outperforms H1 and comes very close to E, it is the heuristic of choice. The increased computing time stays within reasonable bounds even for items with $J = 9$.

In some practical applications there are items, where one is certain about the order of the elements of s. For instance, if one discerns J categories in the amount of time it takes a person to run 100 meters, one wants to consider only s with a nonincreasing relation to the registered time. In such a case S contains only elements s that conform to the indicated order, which reduces S to a size that enables exhaustive search.

Table 3

Comparison (100 samples) of Exhaustive Search and two Heuristics to Find an $s \in S$ (under B2) in as Much the Same Direction as t

		Heuristic H1						
		Mean	*W*-Cos	Differences W-Cos E-H1			#Evaluations	
J	%Optimal	H1	*E*	Max	Mean	St. Dev.	H1	*E*
2	100	0.99679	0.99679	0.00000	0.00000	0.00000	12	27
3	97	0.99472	0.99497	0.02159	0.00025	0.00217	31	324
4	88	0.99451	0.99576	0.03110	0.00115	0.00431	65	4455
5	70	0.99411	0.99520	0.01714	0.00109	0.00272	118	71358
6	61	0.99402	0.99550	0.02035	0.00144	0.00334	196	1315545
7		0.99465					295	28 10^6
8		0.99483					428	680 10^6
9		0.99489					595	20000 10^6

		Heuristic H2					
		Mean	W-Cos	Differences W-Cos E-H2			#Evaluations
J	%Optimal	H2	*E*	Max	Mean	St. Dev.	H2
2	100	0.99679	0.99679	0.00000	0.00000	0.00000	14
3	100	0.99497	0.99497	0.00000	0.00000	0.00000	46
4	98	0.99563	0.99576	0.01096	0.00013	0.00110	127
5	93	0.99516	0.99520	0.00078	0.00004	0.00015	329
6	93	0.99529	0.99549	0.01249	0.00020	0.00134	793
7		0.99623					1816
8		0.99665					3986
9		0.99663					8738

Statistics Approximately Linear in *s*

To execute the minimization procedure over S outlined above, a category statistic is needed that can be expected to be approximately linear in the true category scores of an item. Below a simulation study to evaluate the procedure is described where four category statistics are investigated:

1 the mean raw score per category,

2 the mean optimal score per category,

3 category weights γ from multiple correspondence analysis (MCA),

4 category score parameters δ from the NRM.

The sequence of these statistics is with increasing accuracy of the linearity property. In the next sections the (approximate) linearity of the statistics in the score vector s is discussed.

The Mean Raw Score r_U per Category

Let B be a test, then the raw score on B can be denoted as

$$r_U = \sum_{i \varepsilon B} x_i .$$

When the used category scores of the items in B are more or less equal to the true category scores, and the item scores in a population of test takers correlate positively, it can be stated for a polytomous item k that

$$s_j > s_{j'} \Leftrightarrow \mathscr{E}(r_U \mid x_k = j) > \mathscr{E}(r_U \mid x_k = j').$$

If, for example, the assigned category scores are in reverse order of the true category scores for half of the items in the test, the statement does not apply. In many practical applications the assumption that the majority of the category indices are about appropriate as category scores seems reasonable. Moreover, whatever the relation between assigned category scores and their true values, the mean test scores for two response categories with the same true scores are equal:

If $j \neq j'$ and $s_j = s_{j'}$ then $f_j(\theta)/f_{j'}(\theta) = c$, with c a constant, and

$$\mathscr{E}(r \mid x_k = j) = \mathscr{E}(r \mid x_k = j') = \int \sum_{i=1}^{I} \sum_{h}^{J_i} h f_{ih}(\theta)\, g(\theta \mid x_k = j)\, d\theta,$$

where $g(\theta \mid x_k = j)$ denotes the posterior density given that $x_k = j$. The posterior density $g(.\mid.)$ only depends on s_{kj}, because the likelihood f_{kj} does. It follows that the mean score per category $(1/N_{kj}\, \Sigma (r_U \mid x_k = j))$ is expected to be successful at least in detecting categories that share the same score.

The idea that the mean raw score per category conveys information on the category scores was already put forward in Muraki (1992).

The Mean Optimal Score r_O per Category

The optimal score

$$r_O = \sum_i^I o_i x_i,$$

yields the largest reliability coefficient α among weighted scores. Note that the optimal weights are associated with items, not with individual categories within items. It is not difficult to derive that optimal weights are obtained by taking the dominant eigenvector $\boldsymbol{v}$ of the correlation matrix of the item scores, and dividing element v_i by the standard deviation of the scores of item i:

$$o_i = v_i / \sigma_i.$$

The optimal score is included because it is expected to yield better estimates of s than the raw score. If an item i has a large discrimination a_i, it will tend to have a large optimal weight o_i. Therefore, r_O is expected to show a higher correlation with the true category scores than r_U.

Multiple Correspondence Weights γ

The vector γ of category weights associated with the largest eigenvalue obtained from Multiple Correspondence Analysis (Greenacre (1984), Israëls (1987), Gifi (1990)), has the following property. Let

$$r_{MC}(\boldsymbol{x}) = \frac{1}{I}\sum_i \gamma_{ix_i}, \text{ and}$$

$$c_{ij} = \frac{1}{N_{ij}} \sum_{\{v:x_{vi}=j\}} r_{MC}(\boldsymbol{x}_v),$$

with N_{ij} the number of persons that scored in category j of item i, then $\gamma_{ij} = \xi c_{ij}$ $(i=1,\ldots,I)$, $(j=1,\ldots,J_i)$, for minimum $\xi > 1$. As explained in Gifi (1990, pp. 120-123) one could substitute the category-weights γ obtained from Multiple Correspondence Analysis for the category indices in the data-matrix X to obtain a new data matrix Y. The correlation matrix R_Y of Y can subsequently be analyzed by a Principle Components Analysis. The scores on the first component from this analysis are proportional to r_{MC}. Moreover, the first eigenvalue of R_Y is maximal over all possible category weights substituted for the category indices. With the categories represented by γ the rank one correlation matrix obtained from the implied one-dimensional linear model approximates best in a least squares sense the correlation matrix R_Y over all possible weighings of the categories.

The NRM also achieves the best one-dimensional representation, although with a non-linear model. However, in practical applications, the regression of the expected score on θ, is often strikingly well approximated in the relevant θ-range by only the linear component in its Taylor expansion. Let

$$\bar{s}_i(\theta) = \sum_{j=1}^{J_i} f_{ij}(\theta)\delta_{ij},$$

be the one-dimensional nonlinear NRM equivalent of the one-dimensional linear model of the MCA approach, then

$$\bar{s}_i(\theta) \approx \bar{s}_i(\theta_0) + Var(s_i \mid \theta_0)(\theta - \theta_0),$$

for some well-chosen θ_0.
To the extent that the NRM can be linearized, the expected δ-score in the NRM is approximately linear in the underlying ability like the expected γ-score in the MCA-approach. The ability dimension in the linear 'NRM' and in the MCA both achieve a best rank 1 approximation to the correlation matrix of the data, either with γ- or with δ-category scores. Therefore, both ability dimensions must be approximately the same. Because the γ-parameters are determined up to an arbitrary linear transformation, like the δ's, γ_i should be approximately linear in $\boldsymbol{\delta}_i$. Moreover, if the NRM is assumed to specialize to the extended GPCM or OPLM, γ_i is approximately linear in $\boldsymbol{s}_i$.

The Nominal Item Response Model
Because $\delta_{ij} = a_i s_{ij}$ by assumption $\boldsymbol{\delta}_i$ is linear in $\boldsymbol{s}_i$ for each i. Therefore, by estimating the vector $\boldsymbol{\delta}$ in the NRM one obtains a direct continuous estimate of $\boldsymbol{s}$. The relevance of the NRM for obtaining category scores was noted by Veldhuijzen (1995). In the mentioned report a quick estimation procedure for the estimation of $\boldsymbol{\delta}$ is introduced. This procedure is flawed because it yields estimates with an appreciable positive bias for the item with the largest δ-parameter. Therefore, the procedure is adapted such that this bias can be avoided, and, moreover, can be used with missing data or incomplete designs without extra complications. The estimation procedure is described in Appendix A.

With $\boldsymbol{\delta}$ as category statistics integer discrimination parameters are easily estimated. The NRM equivalent of equation (4) is $\boldsymbol{\delta}_i \approx c_i \boldsymbol{s}_i + d_i \mathbf{1}$. Because δ is determined up to a constant, one may arbitrarily set the minimum of $\boldsymbol{\delta}_i$ equal to zero. The category k

associated with the minimum of δ_i, will, according to equation (4) be associated with category score $s_k = 0$. Therefore, d equals zero, and it follows that $c_i \propto a_i$ (see Formulas (6) and (3)). Because $\boldsymbol{a}$ is a scale parameter it has to be centered by its geometric mean. Let

$$\bar{c} = \left(\prod_i c_i\right)^{\frac{1}{7}}.$$

Because in OPLM the scale is arbitrary the geometric mean μ has to be supplied by the user. Let b_i be a continuous approximation to a_i. Then using Formulas (1), and (5) and (6) it is easily verified that

$$b_i = \frac{\mu}{\bar{c}} c_i,$$

is the discrimination parameter with the required geometric mean. Because in OPLM a_i has to be an integer, the elements of $\boldsymbol{a}$ have to be rounded to the nearest integer in a geometrical sense. This is defined as follows. Let $t(a)$ be the largest integer $\leq a$ (= trunc(a)), and let $gm(i) = (i \times (i+1))^{\frac{1}{2}}$ (the geometric mean of i and $i+1$). Define the function $gi(x)$, the g-nearest integer of a as $t(a)$ if $a < gm(t(a))$ and as $t(a)+1$ otherwise. Then $\hat{a}_i = gi(\mu / \bar{\gamma}\gamma_i)$ is an integer estimate of a_i.

It may happen with a certain data set that the NRM-parameter estimates diverge for a particular item, for instance because unique estimates do not exist. In that case the procedure has to be restarted without the trouble causing item. The integer estimates for this item can subsequently be obtained by one of the previous three statistics. The simpler models are less likely to suffer the same problem with such a data set.

The Korfball Example Revisited

Returning to the previous example from the korfball game, first the data set was analyzed with the NRM. Next for all items the optimal integer category scores were estimated, with exhaustive search and diagonal weight matrix cN, with diagonal element $n_{ij,ij}$ equal to the number of observations in category j of item i. The optimal category scores for item 7 were 0121233444, with the next best candidate 012223344. So for item 7 instead of category scores 0,...,9 only 0,...,4 are kept, in a slightly different rank order.

The results are presented in Tables 4 and 5.

Table 4

S- and M-Tests of Item 7 in a Korfball Game from an OPLM-Analysis with the Optimal Integer Category Scores

Item		S-test	df	P	M	M2	M3
7 APass	2	4.371	4	.358	.758	.222	-.033
	[:2]	1.251	6	.974	.300	-.382	-.356
	[:3]	4.007	7	.779	.214	-.466	.676
	[:4]	11.813	7	.107	.617	.071	-.907

Table 5

Some Overall Results of an OPLM-Analysis of the Optimal Korfball Data

Distribution of *p*-values for S-tests.

```
0.----/-----/-----.1-------.2-------.3-------.4-------.5-------.6-------.7-------.8-------.9-------1.
   2/   3/  1       1       4       4       3       1       2       4       0       3
```

Total: 28
p-values out of range : 1

R1c = 83.286; df = 83; *p* = .4705

With the optimal integer scores, not only do the M-tests show a better fit, also the result of the OPLM-analysis as a whole shows better compliance with the model as shown by a more uniform distribution of the *p*-values for the S-tests, and a nonsignificant *R1c* test.

A Simulation Study

The aim of the simulation study is to evaluate the quality of the estimation of integer category scores on the basis of the four statistics. All data are generated under the OPLM model for a test with 15 items, each with observed categories $0,\ldots,3$. For all items the rank order of s and the category indices is kept equal. Let $a_i\eta_i = \zeta_i$ and $\beta_{ij} = \eta_{ij} - \eta_{ij-1}$. Parameter η_{i0} is set equal to 0. The simulations are conducted with a large and a small spread of item parameters. Therefore, let k take the values 1 or 3. The parameter of category 1 η_1 is distributed as $N(-0.25/k,(0.25/k)^2)$. The difference $\beta_{j+1} - \beta_j$ $(j > 1)$ has been given the same variance as η_1, but with mean

$0.25 / k$. Two categories with the same score share the same parameter. For instance for an item with s_i = 0113, $\eta_1 = \eta_2$. The discrimination indices of the items are all integer and uniformly distributed on $\{1, \ldots, 5\}$. The person parameters θ are distributed as $N(0, 0.3^2)$. This would compare to about $N(0, 1)$ in the Rasch model, which is often approximately found in practice. The cells of the simulation study vary on four dimensions:

1 small or large spread of item parameters,
2 the number of records 500 and 2500,
3 three levels of the amount of categories with the same score (called a collapse),
4 three levels of the amount of skipped scores.

The amount of collapsing and skipping was implemented as follows. Call the size of a collapse the number of categories that share their score with a preceding category. For instance with s_i = 0011 or 0111, the collapse size equals two. The size of skipping is the number of extra skipped categories. It is immaterial whether for instance two extra skipped categories are consecutive like in 0345, or separate, like in 0245. In both cases the size of skipping equals two. For collapsing and skipping as well a maximum size of two is allowed. The level of, for instance the amount of collapsing, is quantified by the number of randomly collapsed categories in the total set of 15 items. Therefore, a random item is drawn and if its collapse size is less than the maximum, randomly two neighboring category classes (one or more categories that share the same score) are given the score of the lower class, and the higher scores are decreased by 1. The same procedure applies to the random application of skipping. The three levels of collapsing and skipping were 1, 7, and 22. That is, at level 1 only 1 item was drawn and randomly collapsed, at level 2 items were drawn and randomly collapsed if possible until 7 times categories could be randomly collapsed, etc. The same procedure applies to the three levels of skipping.

Because the four conditions were crossed they yield $2 \times 2 \times 3 \times 3 = 36$ cells. For each of these 36 cells 400 times

1 a set of item parameters was drawn,
2 a data set was generated,
3 the four statistics were calculated, and,
4 per item for each of the four statistics the optimal vector of integer category scores was found. The NRM category statistics $\boldsymbol{\delta}$ were estimated by the procedure outlined in Appendix A.

Table 6

100 × Mean W-Cosines of Integer Estimates with True Category Scores CSType Indices Indicate the 3 Levels of resptively Collapse and Skip of the Item

#Records	500;	Small spread					
CSType	ScorU	ScorO	MCA	NRM	Marg	Index	#Items
00	99.53	99.50	99.34	99.34	99.43	100.00	14623
01	99.61	99.60	99.54	99.51	99.56	98.96	7053
02	99.61	99.61	99.53	99.54	99.57	98.78	6407
10	99.21	99.17	99.06	98.98	99.11	96.42	6871
11	99.45	99.44	99.49	99.43	99.45	95.28	3415
12	99.57	99.56	99.51	99.50	99.53	95.38	3148
20	97.65	97.63	97.73	97.54	97.63	87.39	6512
21	99.09	99.17	99.41	99.19	99.22	87.47	3120
22	99.61	99.65	99.78	99.61	99.66	87.51	2851
Marg	99.26	99.25	99.20	99.14	99.21	95.79	54000
#Records	2500;	Small spread					
CSType	ScorU	ScorO	MCA	NRM	Marg	Index	#Items
00	99.89	99.87	99.80	99.89	99.86	100.00	14571
01	99.84	99.84	99.87	99.95	99.88	98.97	6981
02	99.81	99.80	99.79	99.86	99.81	98.82	6512
10	99.73	99.76	99.82	99.83	99.79	96.43	7019
11	99.82	99.86	99.94	99.95	99.89	95.36	3348
12	99.81	99.82	99.81	99.83	99.82	95.42	3105
20	98.61	99.34	99.61	99.60	99.29	87.34	6460
21	99.63	99.85	99.95	99.94	99.84	87.49	3171
22	99.93	99.98	100.00	100.00	99.98	87.52	2833
Marg	99.68	99.78	99.82	99.86	99.78	95.81	54000
#Records	500;	Large spread					
CSType	ScorU	ScorO	MCA	NRM	Marg	Index	#Items
00	99.36	99.33	99.07	98.89	99.16	100.00	14508
01	99.31	99.29	99.15	99.21	99.24	98.66	7086
02	99.30	99.29	99.13	99.32	99.26	98.49	6439
10	99.02	98.96	98.81	98.65	98.86	95.79	7071
11	99.24	99.21	99.25	99.24	99.24	94.72	3340
12	99.42	99.40	99.34	99.42	99.40	94.97	3123
20	97.18	97.43	97.57	97.37	97.39	86.09	6481
21	98.92	99.07	99.36	99.23	99.14	86.82	3054
22	99.53	99.60	99.78	99.67	99.65	87.21	2898
Marg	99.02	99.04	98.96	98.88	98.98	95.38	54000
#Records	2500;	Large spread					
CSType	ScorU	ScorO	MCA	NRM	Marg	Index	#Items
00	99.69	99.68	99.51	99.79	99.67	100.00	14620
01	99.51	99.52	99.49	99.91	99.61	98.65	6921
02	99.50	99.49	99.39	99.82	99.55	98.50	6491
10	99.55	99.60	99.63	99.79	99.64	95.78	7012
11	99.60	99.65	99.78	99.93	99.74	94.69	3438
12	99.66	99.67	99.62	99.81	99.69	94.87	3086
20	98.12	99.16	99.57	99.54	99.10	85.83	6474
21	99.51	99.78	99.96	99.93	99.79	86.85	3029
22	99.88	99.95	100.00	100.00	99.95	86.96	2929
Marg	99.43	99.58	99.59	99.81	99.60	95.32	54000

To evaluate the performance of the four statistics the W-Cosines of the optimal vectors and the index-vector with the true score vectors and the percentage of items with correct estimates of s were recorded. The presented results are achieved with WLS (W = cN), with N a diagonal matrix with the number of responses in each category. Using the inverse of the covariance matrix of the *t*-vector as the W-matrix is problematical because it is of deficient rank. For Bock's NRM the information matrix of δ was tried with a somewhat inflated diagonal. However, the results did not improve with respect to those obtained with the diagonal matrix. Both exhaustive search and the heuristic H1 were applied under restriction B2. The results of Exhaustive search and H1 proved to be 99.8% identical.

The results on the W-Cosines are shown in Table 6 and Figure 1, the percentages correct in Table 7 and Figure 2. CSType refers to the sizes of Collapsing and Skipping within a single item. For instance CSType 12 means that the item has one pair of categories collapsed and two categories extra skipped. Two examples of items of CSType 12 are 0224, and 0034.

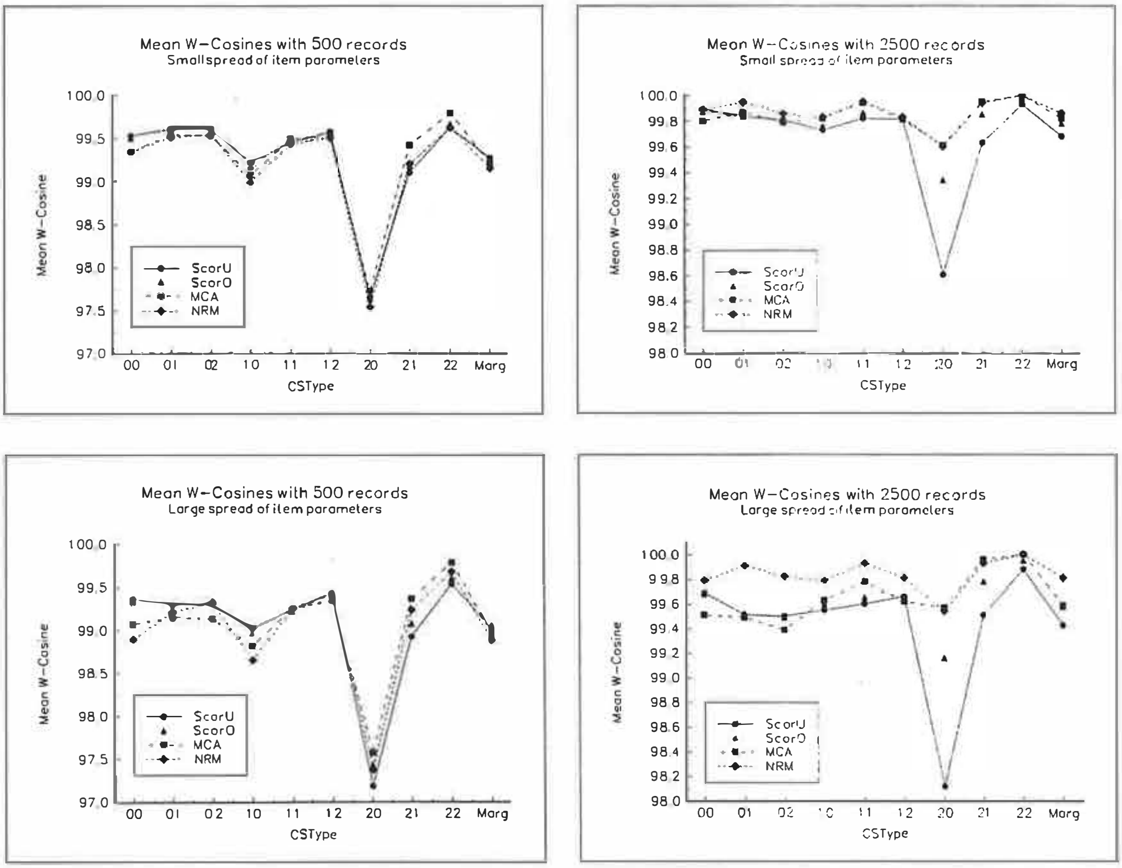

Figure 1

100 × Mean W-Cosines of Integer Estimates with True Category Scores

The percentages correct (Table 7, Figure 2) with 500 records may look a little disappointing. However, one has to consider that an incorrectly estimated score vector, in general, is not way out of target as can be verified by their Mean W-Cosines as shown in Table 6 and Figure 1. Closer inspection of the estimates shows that almost invariably one skip too much causes the estimate to deviate from its original value.

Table 7

Percentage of items with correctly estimated category scores

#Records	500;	Small spread				
CSType	ScorU	ScorO	MCA	NRM	Marg	#Items
00	51.25	50.70	40.29	42.71	46.24	14623
01	62.24	62.10	62.81	60.60	61.94	7053
02	38.11	38.22	36.09	37.13	37.39	6407
10	52.93	52.28	51.54	50.41	51.79	6871
11	67.38	67.53	73.50	69.72	69.53	3415
12	33.89	33.29	32.37	34.88	33.61	3148
20	47.76	51.97	59.24	55.04	53.50	6512
21	52.44	53.56	57.56	54.04	54.40	3120
22	59.21	59.91	62.26	59.00	60.09	2851
Marg	51.42	51.76	50.24	49.62	50.76	54000
#Records	2500;	Small spread				
CSType	ScorU	ScorO	MCA	NRM	Marg	#Items
00	85.90	84.44	76.02	87.58	83.49	14571
01	82.19	81.99	86.06	94.83	86.27	6981
02	55.36	54.81	54.91	63.53	57.15	6512
10	79.23	81.99	87.33	88.83	84.35	7019
11	87.37	89.01	94.89	96.92	92.05	3348
12	44.28	44.83	45.06	47.54	45.43	3105
20	64.02	81.67	89.71	89.33	81.18	6460
21	61.27	65.00	66.92	66.57	64.94	3171
22	67.67	68.51	68.90	68.80	68.47	2833
Marg	73.55	75.93	76.36	82.05	76.97	54000
#Records	500;	Large spread				
CSType	ScorU	ScorO	MCA	NRM	Marg	#Items
00	44.29	43.48	32.85	33.26	38.47	14508
01	53.20	53.39	51.69	55.38	53.42	7086
02	30.33	30.11	26.91	35.33	30.67	6439
10	46.83	45.96	44.08	43.80	45.17	7071
11	59.19	58.35	64.19	64.28	61.50	3340
12	30.68	30.36	28.24	34.36	30.91	3123
20	43.37	49.13	55.49	52.86	50.21	6481
21	50.46	52.98	57.43	54.49	53.84	3054
22	59.32	60.63	63.35	60.59	60.97	2898
Marg	45.31	45.81	43.50	44.79	44.85	54000
#Records	2500;	Large spread				
CSType	ScorU	ScorO	MCA	NRM	Marg	#Items
00	68.13	66.65	55.18	78.68	67.16	14620
01	66.16	66.49	67.79	92.07	73.13	6921
02	40.41	40.27	38.51	62.36	45.39	6491
10	68.41	72.08	75.36	86.08	75.48	7012
11	73.68	76.35	84.82	95.72	82.64	3438
12	37.46	37.46	34.96	48.48	39.59	3086
20	55.50	76.32	88.25	87.44	76.88	6474
21	58.77	63.45	66.19	65.70	63.53	3029
22	65.28	66.17	66.75	66.75	66.23	2929
Marg	60.99	64.07	63.36	78.43	66.71	54000

The 'performance' of the index is included in Table 6 to translate the deviance from the index vector of the different item CSTypes to the W-Cosine scale. If the order of the true category scores would have been independent of the index vector, the W-Cosines of the index with the true scores would have been around zero. It is not relevant to include the 'performance' of the index with the percentage correct data. They are 100% correct for the items with CSType = 00, and 0% otherwise.

Because the estimation of the discrimination index is to a large extent dependent on the estimates of the category scores, and the user supplied geometric mean, it is not a simple matter to evaluate its accuracy in a simulation study of this kind. However, practical application of the NRM-procedure in connection with OPLM, generally shows excellent model fit with respect to the discrimination parameters.

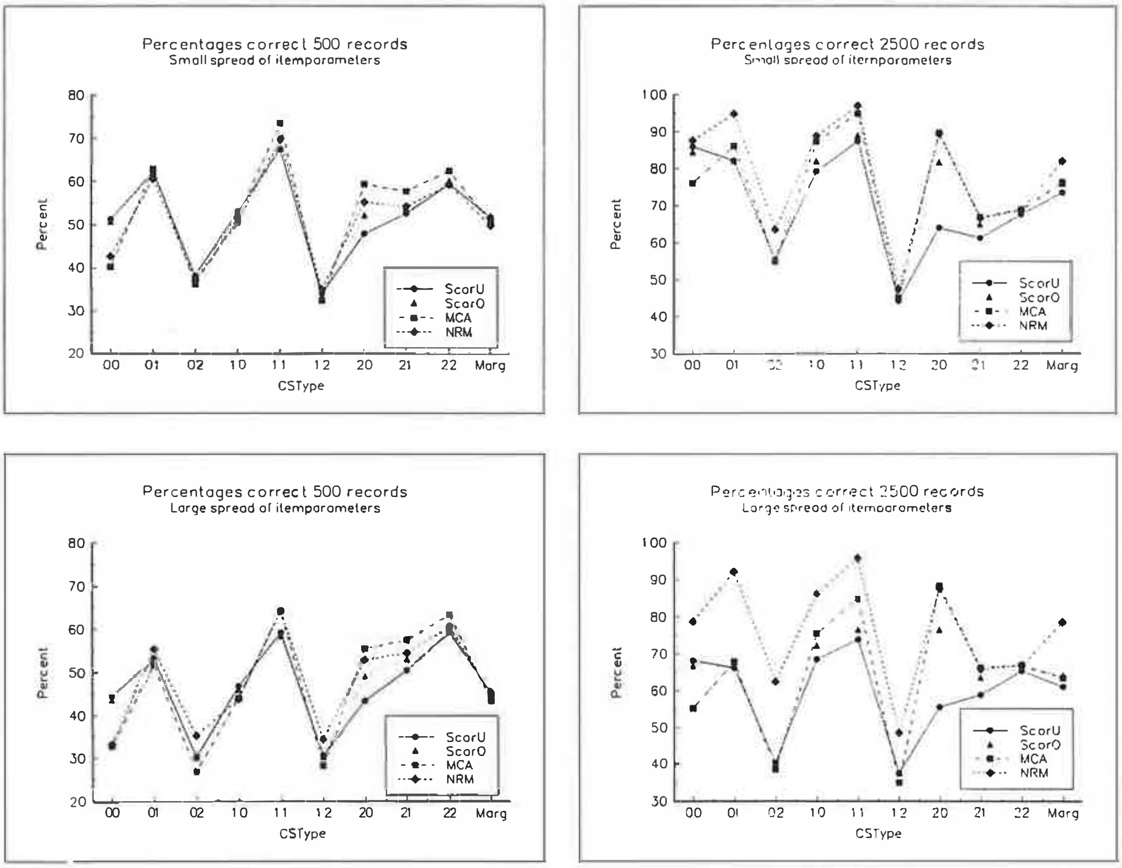

Figure 2

Percentages of Items Correct

Conclusion

Inspection of Tables 6 and 7 and Figures 1 and 2 shows that the four statistics perform very similarly with the 500 record data sets, although with the larger spread of item parameters a slightly larger differentiation can be perceived. With 2500 record data sets the NRM estimation clearly outperforms the other statistics, certainly with the larger item parameter spread. The effect of small versus large spread of item parameters can be summarized as follows. All statistics perform worse under the large spread condition. However, with 2500 record data sets NRM clearly shows the smallest drop in performance ($\pm$ 3.5%). The other statistics loose about 12%. MCA shows the largest difference between the small and large spread conditions with 2500 records (13%), probably because the linear approximation is less accurate with larger spread of item parameters. Looking at the Raw scores (ScorU) it is especially striking that ScorU performs worst on CSType 20, because in the section on the raw score it was deduced that this statistic would at least correctly identify category collapses. Comparing the W-Cosines with the percentages correct plots, it is noteworthy that, for the 500 record data sets, and less conspicuous for the 2500, the W-Cosines show a dip at CSType 20, and the percentages correct do not. The reverse holds for CSType 12. I have no explanation for these reversals. It is also remarkable that the more sophisticated statistics MCA and NRM do not perform as well as the simpler statistics ScorU and ScorO with the 500 record data.

Assuming that data were generated that can be considered representative or at least relevant for real psychometric data sets the simulation study shows that the true integer category scores can be estimated with a high degree of accuracy. Of course, the size of the data set is decisive in this respect. The mean W-Cosines with 500 records are about 0.991 and with 2500 records 0.997, which must be considered highly accurate.

Appendix A

A Quick Algorithm for Parameter Estimation in the NRM

The following procedure is partly adopted from Veldhuijzen (1995) with weights from Verhelst (1996), but adapted in such a way that dominance by the item with the largest δ-parameter cannot occur, moreover, it can be used with missing data or incomplete designs without extra effort. Assume that the abilities of all persons can be grouped into $h = 1,\ldots,H$ homogeneous ability groups. Let θ_h be the ability of the members of group h. Denote the number of persons in ability group h that scores in category j of item i with N_{hij}, and let

$$p_{hij} = \frac{N_{hij} + 0.5}{N_{hij} + N_{hij-1} + 1} \; (j = 1,\ldots,J_i), \; \Delta_{ij} = \delta_{ij} - \delta_{ij-1}, \text{ and } Z_{ij} = \zeta_{ij} - \zeta_{ij-1}.$$

Then equations (1) and (2) show that

$$l_{hij} \doteq logit\, p_{hij} = ln\, p_{hij} / (1 - p_{hij}) \approx \theta_h \Delta_{ij} - Z_{ij}. \tag{7}$$

Formula (7) shows that a least squares estimate of the item parameters is obtained by minimizing the following loss function

$$F = \sum_{hij} w_{hij} (\theta_h \Delta_{ij} - Z_{ij} - logit\, p_{hij})^2, \tag{8}$$

with w_{hij} an appropriate weight. Verhelst (1996) derives that

$$w_{hij} = \frac{N_{hij} N_{hij-1}}{N_{hij} + N_{hij-1}} \text{ is an adequate choice.}$$

Formula (8) is minimized by iteratively solving the following estimation equations

$$\Delta_{ij} = \frac{\sum_h w_{hij} \theta_h (\Gamma_{ij} + l_{hij})}{\sum_h w_{hij} \theta_h^2}$$

$$\Gamma_{ij} = \frac{\sum_h w_{hij} (\theta_h \Delta_{ij} - l_{hij})}{\sum_h w_{hij}}. \tag{9}$$

In Veldhuijzen (1995) formula (9) also entails estimation equations for θ_h. Because the estimation of θ_h and calculation of N_{hij} are approached differently here, these are absent. For the calculation of the numbers N_{hij} we need for each person v and for each item i an estimate of the group membership of v independent of his score on i.

Let $g(\theta) = N(0,1)$ be the apriori ability density. Define $H+2$ ability groups by selecting $H+1$ edges $\zeta_h (h = 1,\ldots,H+1)$ on the latent continuum. In the present algorithm the C-scale boundaries for the standard normal are used $\zeta_h = -2.25\,(0.50)\,2.25$. Persons are classified for each item i separately into one of these ability groups on the basis of their EAP from their response pattern omitting their response x_i on item i. EAP's outside the range [-2.25, 2.25] are neglected. The midpoint $\theta_h = (\zeta_h + \zeta_{h+1})/2$ of the interval $[\zeta_h, \zeta_{h+1}]$ is taken as the common ability estimate for the members of class h.

For each item i and each ability group h one obtains the number N_{hij} of respondents in group h that responded in category j of item i as follows. The a posteriori density of θ given a response vector $\boldsymbol{x}$ is given by

$$g(\theta \mid \boldsymbol{x}) = \frac{\prod_{i=1}^{I} f(x_i \mid \theta)\, g(\theta)}{\int \prod_{i=1}^{I} f(x_i \mid \theta) g(\theta)\, d\theta}.$$

The integral in the denominator is approximated with the Gauss-Hermite sum with Q points

$$\int_{-\infty}^{\infty} \prod_i f(x_i \mid \theta)\, g(\theta)\, d\theta = \sum_{q=1}^{Q} w_q \prod_i f(x_i \mid \theta_q), \tag{10}$$

where w_q are the Gauss-Hermite weights associated with the Q Gauss-Hermite values θ_q. First calculate the Q summands

$$y_q = w_q \prod_i f(x_i, \theta_q),$$

of (10), and denote the response vector $\boldsymbol{x}$ omitting the response from item i with $\boldsymbol{x}^{(i)}$. Then the a posteriori distribution of θ at θ_q given the response pattern $\boldsymbol{x}$ omitting item i is given by

$g\left(\theta_q \mid \boldsymbol{x}^{(i)}\right) \propto z_{iq} = y_q / f(x_i \mid \theta_q)$, and the EAP_i given $\boldsymbol{x}^{(i)}$ is found with

$$EAP(\boldsymbol{x}^{(i)}) = \frac{\sum_q \theta_q z_{iq}}{\sum_q z_{iq}}.$$

Suppose $x_{vi} = j$, and $\zeta_h \leq EAP_{vi} < \zeta_{h+1}$ then v contributes 1 to N_{hij}. It is conceivable that for $i' \neq i$ $EAP_{vi'}$ is contained in another interval $h' \neq h$ than EAP_{vi}. In that case v contributes 1 to $N_{h'i'j'}$, and not to $N_{hi'j'}$.

Now, the estimation procedure proceeds as follows.

Select the category indices as initial score vector $\delta^{(0)}$ for the categories of items $1,\dots,I$. Select as initial vector $\zeta^{(0)}$ the vector $\boldsymbol{0}$, and initiate F at ∞.

1 Calculate N_{hij} for all h, i, and j, and loss function F (Formula (8)). If F does not decrease with respect to its previous value then stop.

2 Iteratively solve Formula (8) until convergence.

3 If the results of 2 deviate more than a criterion from the parameters used in 1, then return to 1, otherwise stop.

The stop criterion given in 1 is included because the algorithm does not always converge in the sense of 3. This is due to the fact that after some iterations the ability grouping for some records tends to swap from one iteration to the next. However, the parameter estimates obtained from this procedure did not result in a worse performance of *NRM* in the simulation study, than estimates obtained when this least squares estimation was followed with 10.

GEM estimation iterations (Verstralen, 1996). It is perhaps noteworthy that the percentages correct decreased by about 1.5% if only one *EAP* per record was estimated (the response to item i was not omitted in the calculation of N_{hij}). Moreover, the dominance of the item with the highest δ-parameter could not be reproduced when this algorithm (one *EAP* per record) was tried on the example in Veldhuijzen (1995). The reason for this improvement is not entirely clear.

Appendix B

Heuristics H1 and H2 for Finding a Near Optimal Integer Vector in S

H1

Let $\boldsymbol{e}_i$ be the binary vector $(0,\ldots,1,\ldots,0)$ with component $e_{ii} = 1, (i = 0,\ldots,J)$ and the other components equal to zero. Let E denote the set $\{\boldsymbol{e}_i, i = 0,\ldots,J\}$. Select a bound $Bb = B1, B2$, or $B2$ (higher bounds are not recommended).

Start with $\boldsymbol{s} = \boldsymbol{0}, i = 0, j = 0, C' = 0, C'' = 0, k = bJ$.

1 Let $\boldsymbol{u} = \boldsymbol{s} + \boldsymbol{e}_i$.
 If at least one component of $\boldsymbol{u}$ equals 0
 continue with 2, otherwise
 If $i < J$ repeat 1 with $i = i + 1$, otherwise
 Continue with 4.

2. If $\boldsymbol{u} \in B$
 If the largest component of $\boldsymbol{u}$ equals k set $j = 1$
 Continue with 3,
 otherwise
 Continue with 1.

3 Calculate $C = \langle NP(\boldsymbol{s} + \boldsymbol{e}_i), NP(\boldsymbol{t}) \rangle$.
 If $C \geq C'$ then let $C' = C$ and $\boldsymbol{s}' = \boldsymbol{s} + \boldsymbol{e}_i$.
 Continue with 1.

4 If $C' > C''$ let $C'' = C'$, and $\boldsymbol{s}'' = \boldsymbol{s}'$,
 Let $\boldsymbol{s} = \boldsymbol{s}'$.
 If $j = 1$ stop, otherwise
 Set $i = 0$, and proceed with 1.
 $\boldsymbol{s}''$ is the H1-optimal integer score vector.

H2

H2 follows the same scheme as H1, only i runs from 1 to $2^{J+1} - 2$, and $\boldsymbol{e}_i$ is the binary vector with $J + 1$ components that reflects the binary representation of i. For instance if $i = 5$ and $J = 3$ $\boldsymbol{e}_i = (0, 1, 0, 1)$.

References

Bock, D.B. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 1, 29-51.

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

Glas, K. (1992). *Unobserved categories in the partial credit model.* Unpublished note. Arnhem: Cito.

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

Israëls, A.Z. (1987). *Eigenvalue techniques for qualitative data*. Leiden: DSWO Press.

Jansen, P.G.W., & Roskam, E.E. (1986). Latent trait models and dichotomization of graded responses. *Psychometrika, 51*, 69-91.

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*, 351-363.

Thissen, D. (1988). *Multilog™, (version 5.1) Multiple, categorical item analysis and test scoring using item response theory*. Mooresville, IN, USA: Scientific Software.

Veldhuijzen N.H. (1995). *A heuristic procedure for suggesting an appropriate scoring function for polytomous items with ordered response categories.* Measurement and Research Department Reports, 95-1 (2nd version). Arnhem: Cito.

Verhelst, N.D. (1996). *SDA, opnieuw gewogen.* Internal note. Cito: Arnhem.

Verhelst, N.D., & Glas, C.A.W. (1995). The one parameter logistic model. In G.H.Fischer, & I.W. Molenaar (Eds.). *Rasch models: Foundations, recent developments, and applications.* New York: Springer-Verlag.

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). *OPLM: One Parameter Logistic Model.* Arnhem: Cito.

Verstralen, H.H.F.M. (1996). OPCAT, Estimating integer category weights in OPLM. User manual. Arnhem: Cito.

Wilson, M., & Masters, G.N. (1993). The partial credit model and null categories. *Psychometrika, 58*, 87-99.