Measurement and Research Department Reports 91-10

Finding Starting Values for the Item Parameters and Suitable Discrimination Indices in the One-Parameter Logistic Model

N.D. Verhelst
H.H.F.M. Verstralen
Th.J.H.M. Eggen

Cito

Measurement and Research Department Reports 91-10

Finding Starting Values for the Item Parameters and Suitable Discrimination Indices in the One-Parameter Logistic Model

N.D. Verhelst
H.H.F.M. Verstralen
Th. J.H.M. Eggen

Cito
Arnhem, 1991

Abstract

In the one parameter logistic model (OPLM), the discrimination indices of the items are considered known constants which are part of the model hypothesis. Suitable values for these indices can be estimated from (part) of the data by a weighted least squares algorithm. It turns out that the same algorithm can also be used with a slight modification to obtain good initial estimates of the item parameters. The algorithm is applicable for binary as well as for polytomous data, in complete as well as in incomplete designs. In case of complicated designs a two-step procedure can be used, which needs far less computer storage than the original algorithm. Two examples are given.

Key words: IRT, Weighted least squares, logistic models.

Introduction

In educational testing, the Rasch model is a much celebrated model because of its mathematical elegance and because of the possibility to yield consistent estimates of its parameters independently of any assumption on the distribution of the latent ability or the way the sample of respondents is drawn. On the other hand it lacks flexibility, mainly because of the very stringent condition that all items in a test have equal discriminating power. A natural generalization of the Rasch model, which repairs for this rigidity is the two-parameter logistic model or Birnbaum model (Lord & Novick, 1968), where a discrimination parameter as well as a difficulty parameter is associated with each item. However, the latter model does not belong to the exponential family, and therefore does not share the mathematical elegance of the Rasch model. In an attempt to combine the mathematical advantages of the Rasch model and the flexibility of the Birnbaum model, the one-parameter logistic model (OPLM) and a powerful computer program with the same name were developed (Verhelst, Glas, & Verstralen, 1991). In case of binary items, OPLM is formally identical to the Birnbaum model, but the discrimination parameters are treated as fixed constants, and are called discrimination indices. In case of polytomous items, OPLM is a generalization of the partial credit model (Masters, 1982). The model is defined by its so-called category response functions, which give the probability of the score j ($j \in \{0,1,\ldots,m_i\}$, $m_i > 0$) on an item i ($i=1,\ldots,k$), conditional on the value of a so-called latent variable, denoted ϑ, and by the axiom of local independence which states that any two item responses are independent given ϑ. Letting X_i denote the random variable 'item score', the category response function of OPLM is given by

$$\text{Prob}(X_i=j\,|\,\vartheta) = \frac{\exp\left[a_i\left(j\vartheta - \sum_{g=0}^{j}\beta_{ig}\right)\right]}{\sum_{h=0}^{m_i}\exp\left[a_i\left(h\vartheta - \sum_{g=0}^{h}\beta_{ig}\right)\right]}, \qquad (j=0,\ldots,m_i;\ a_i>0), \quad (1)$$

where β_{i0} is defined to be zero. The quantities a_i are the discrimination indices. Although OPLM is - in the dichotomous case - formally identical to the two-parameter logistic model (Birnbaum, 1968), it is quite different from a statistical point of view. In OPLM the discrimination indices are treated as known quantities and not to be estimated, thus making the estimation problem in OPLM much easier than in the two parameter model. Moreover, since the sufficient statistic for ϑ, the test score, depends on these indices, the test score is a mere statistic and not a function of unknown parameters. This means that the category parameters β can be estimated by the method of conditional maximum likelihood (CML), which amounts to maximizing the likelihood conditional upon the observed frequencies of the test scores. Andersen (1973) showed that under rather mild conditions this method gives consistent estimates.

Inspection of (1) shows that the model as stated is unidentified. Addition of an arbitrary value to ϑ and to the β's (for $j > 0$) does not change the function value. We should therefore fix the zero of the scale, for instance, by fixing an arbitrarily chosen β at some value, or by making the sum of the β's equal to zero. Moreover, multiplying the discrimination indices with an arbitrary positive constant and dividing ϑ and the β's by the same constant does not alter the function value either. Restricting the a's to the positive rationals (which is not of any practical meaning) makes it possible to choose the multiplicative constant in such a way that the discrimination indices are all integer valued for a given collection of k items. In the sequel we will assume integer valued discrimination indices. The details of the estimation procedure are described in Verhelst, Glas, and Verstralen (1991).

The present report deals with two problems which at first seem to be unrelated, but which turn out to be solvable in an elegant way by the same technique. The first problem is concerned with the initial estimates of the β-parameters for the iterative procedure that solves the estimation equations. The second problem is to find appropriate values of the discrimination indices used in (1). Treating these indices as known

constants is easy, but choosing an adequate value for them is not. Considering (1) as a statistical hypothesis, tested by goodness of fit tests, it will be clear that since the discrimination indices are fixed, their specified value is part of the hypothesis. Verhelst et al. (1991) developed statistical tests which are especially sensitive against misspecification of the indices, and also give an indication on how to alter them. However, this indication is useful only in case the discrimination indices are not grossly inadequate. The reason for this is that, although the statistical tests are item oriented, they are not independent of each other. This means that every statistical test on the adequacy of a particular discrimination index is influenced by the adequacy of the others, and if they are all inadequate, the tests become useless. It is therefore important to have a good approximation to the discrimination indices. This report offers a technique to extract a fairly good hunch from the data. The technique is first described in its generality. A variant, which is actually implemented in the program OPLM to estimate the a's is discussed in a separate section. An example is presented next and in the final section the whole procedure is discussed.

A Least Squares Estimation Procedure in OPLM

Consider a collection of k items and the index set I={1,...,k}. If k is large, it becomes very impractical to administer each item to each respondent in the sample. In many practical applications each respondent answers only to a subset of the items. These subsets will be called booklets, and they can formally be indicated as subsets of I. Assume there are B booklets, each subset being described by $I_b \subseteq I$. As will be immediately clear, two numerically equal scores obtained on different booklets are not comparable. So in incomplete designs, every score has meaning only if it is associated with a booklet I_b. In OPLM the test score is defined as follows: Let $\underline{X}$ be a k-dimensional random variable with elements X_i (i=1,...,k), taking values from $\{0,1,...,m_i\}$

if the item is responded to, and taking a not specified value * if the item is not presented. The test score obtained on booklet b is given by

$$S_b = \sum_{i \in I_b} a_i X_i . \qquad (2)$$

Since the item scores are weighted with the discrimination indices, this score will be called the weighted score. It will prove useful to work also with unweighted test scores U_b defined as

$$U_b = \sum_{i \in I_b} X_i . \qquad (3)$$

Now consider the conditional probability of obtaining score j on item i ($j=1,\ldots,m_i$) conditional on ϑ and on the item scores j and j-1:

$$\text{Prob}(X_i=j \mid \vartheta, j-1 \leq X_i \leq j) = \frac{\text{Prob}(X_i=j \mid \vartheta)}{\text{Prob}(X_i=j-1 \mid \vartheta) + \text{Prob}(X_i=j \mid \vartheta)}, \quad (j=1,\ldots,m_i) . \qquad (4)$$

Substituting (1) in (4), and taking the logit transformation gives

$$\text{logit Prob}(X_i=j \mid \vartheta, j-1 \leq X_i \leq j) = a_i(\vartheta-\beta_{ij}), \quad (j=1,\ldots,m_i) . \qquad (5)$$

By replacing the probability in the left-hand side of (5) by the corresponding observed proportion, an approximate equality which can be used to estimate the β's is obtained. However, since (5) depends on the value of the latent variable ϑ, it is impossible to group respondents into classes of equal ϑ, and therefore the corresponding proportions cannot be computed. However, we can use an approximation, based on the following result: If the number of items responded to is not too small, then $\text{var}(\vartheta \mid s)$ becomes rather small and the regression of s on ϑ is well approximated by a linear function. Using this approximation and neglecting the conditional variance we get

$$\text{logit Prob}(X_i=j \mid s, j-1 \leq X_i \leq j) \approx a_i[(Bs+A)-\beta_{ij}], \quad (j=1,\ldots,m_i) . \qquad (6)$$

Considering (6) as a strict equality, one can in principle replace the probabilities by proportions and use these to estimate the parameters, where B and A are now also to be considered as parameters. However, there are two objections to the use of (6), and neglecting them may cause gross distortions in the solutions. First, the linear transformation Bs + A is used as an approximation of the regression of ϑ on s. Even if the regression of s on ϑ is approximately linear, the regression of ϑ on s may be very nonlinear. While the regression of s on ϑ is completely specified by the measurement model (1), the regression of ϑ on s also depends on the distribution of ϑ, so that the adequacy of (6) may differ grossly from application to application. Therefore, it might be safer not to assume anything about this regression. So we replace (6) by

$$\text{logit Prob}(X_i=j|s,j-1 \leq X_i \leq j) \approx a_i[\vartheta_s-\beta_{ij}], \quad (j=1,\ldots,m_i), \qquad (7)$$

where the ϑ_s are considered as parameters. But using different discrimination indices, many of them being larger than 1, causes the score range to be much larger than the number of items, so that the number of parameters ϑ_s may become quite large. Moreover, and this is the second problem with (6) as well as with (7), many score frequencies will inevitably be low, causing the conditional proportions to be unstable and in many cases to be zero or one, requiring ad hoc corrections for the logit transformation to be defined. These corrections though relatively harmless if used with large samples, will influence the solutions in an inadmissible way if the score frequencies are low. It seems therefore useful not to condition on separate scores but on homogeneous classes of scores. For a single test (booklet), let all scores from 0 to the maximum score be partitioned into G subsets S_g (g=1,...,G) such that

$$s \in S_g \text{ and } s+2 \in S_g \Rightarrow s+1 \in S_g, \qquad (8)$$

and (7) can be replaced by

$$\text{logit Prob}(X_i=j \,|\, s \in S_g, j-1 \le X_i \le j) \approx a_i[\vartheta_g-\beta_{ij}], \quad (j=1,\ldots,m_i;\ g=1,\ldots,G). \tag{9}$$

Letting n_{bgij} denote the number of respondents answering booklet b, belonging to group g (within booklet b) and having score j on item i, and let n_{bg} be the number of respondents in group g of booklet b and define

$$\pi_{bgij} \doteq \text{Prob}(X_i=j \,|\, i \in I_b, s \in S_g, j-1 \le X_i \le j), \quad \begin{array}{l}(j=1,\ldots,m_i;\\ g=1,\ldots,G_b;\\ b=1,\ldots,B),\end{array} \tag{10}$$

then it is clear that the corresponding proportion is given by

$$p^{*}_{bgij} = \frac{n_{bgij}}{n_{bgij}+n_{bgi,j-1}}. \tag{11}$$

For the logit transformation to be defined in any case, we apply a common correction, and define

$$p_{bgij} = \frac{\frac{1}{2}+n_{bgij}}{1+n_{bgij}+n_{bgi,j-1}}. \tag{12}$$

So (12) can be considered as the observed proportion corresponding to (10), and its logit transformation can be used as an estimate of the right-hand part of (9). It will be obvious that a suitable loss function is given by

$$F = \sum_{b=1}^{B} \sum_{g=1}^{G_b} \sum_{i \in I_b} \sum_{j=1}^{m_i} w_{bgij}[a_i(\vartheta_g-\beta_{ij}) - \text{logit}(p_{bgij})]^2, \tag{13}$$

where w_{bgij} is some suitable weight. Before discussing the weights, it should be noted that the parameter estimates given by minimizing F are not consistent. Equation (9) is almost never strictly true, not even if k goes to infinity if we hold G_b fixed, because (9) implies that within a homogeneous score group the variance of ϑ vanishes, which holds only if the distribution of ϑ is degenerated. So in general, (9) is not equivalent with (1), and it seems a little bit pointless to look for the 'best' weights,

which might be taken here to mean variance minimizing weights. On the other hand, it seems natural to give less importance to proportions which are based on a small number of observations than on a large sample. Least squares estimates have minimum variance if the inverse of the variance-covariance matrix of the random variable is used as a weight matrix. Here this means that we have to determine the variance-covariance matrix of all $\text{logit}(p_{bgij})$. Referring to (12), it is easily seen that this matrix is not diagonal, because n_{bgij} $(0<j<m_i)$ is used in two proportions. Because the estimates need not be very accurate it seemed appropriate to neglect the covariances and to take the weights proportional to the inverse of the (estimated) variance of $\text{logit}(p_{bgij})$. By the univariate delta method (Bishop, Fienberg & Holland, 1975) it is easy to show that

$$\text{Var}[\text{logit}(p_{bgij})] = [n_{bg}\,\pi_{bgij}(1-\pi_{bgij})]^{-1}. \qquad (14)$$

Replacing π_{bgij} by p_{bgij} in (14) gives as weights

$$w_{bgij} = n_{bg}\,p_{bgij}(1-p_{bgij}). \qquad (15)$$

When initial estimates of the β's are needed, and the discrimination indices are fixed, one minimizes (13) with respect to the β's and the ϑ's, where the construction of homogeneous score groups is based on the weighted scores s, given by (2). The estimated ϑ's then simply are not used. As the correlation between weighted and unweigted scores is usually very close to one, a grouping based on the unweighted scores may give useful results as well. So, when one needs an indication of suitable discrimination indices, one may minimize (13) with respect to the a's, the β's and the ϑ's with the grouping based on the unweighted scores, and neglect the estimates of the β's and the ϑ's. The technique used to minimize (13) is a simple alternating least squares procedure. If two sets of parameters are to be determined, for example, the ϑ's

and the β's, then in each iteration F is minimized twice: the first time with respect to the first set, the β's say, while the other set is fixed at its current value. The second time F is minimized with respect to the second set, while the first set is fixed at the value just found.

A Two Stage Estimation Procedure

The computer program OPLM is designed to analyze data in an incomplete design for rather large numbers of items. At Cito, an analysis on 250 items distributed over as many as 25 booklets is a common enterprise. As may be seen from the quadruple subscripts of the w's and the p's in (13), the total number of weights and conditional proportions, which are all needed within a single iteration, may be too large to store them all in the available RAM of the machine. So, of necessity, part of them will be stored on an external memory device, a disk say, with the consequence that in every iteration a substantive number of disk accesses will be necessary, making the total processing time a multiple of the actual computing time. Moreover, it appears that when the booklets are poorly linked, that is, the number of common items in any pair of booklets is small, the alternating least squares algorithm converges rather slowly, so that the number of iterations may grow quite substantially. In order to avoid these problems a two stage procedure was devised, which will be discussed for the case of the a-estimates. In the first phase, the a's are estimated per booklet. But as has been argued earlier, the a's are determined up to an arbitrary (positive) multiplicative constant. This means that in every booklet the unit of measurement is arbitrary, so that two estimates for the same item are not on a common scale, and thus cannot be compared with each other. So the second phase is meant to bring the a-estimates from different booklets on a common scale.

Let a_{bi} denote the estimate of the discrimination index for the i-th item in booklet b, then the rationale says that for any two booklets b and b' where item i appears, there should be a positive constant $c_{bb'}$ such that

$$a_{bi} \approx a_{b'i} c_{bb'}, \tag{16}$$

where the equality is not strict because of the sampling error in the estimates. Assuming strict equality, it follows that

$$c_{bb'} = \frac{c_{bb''}}{c_{b'b''}}. \tag{17}$$

So, if for all b one knows $c_{bb''}$ for some b", all c's are known. Because $c_{b''b''} = 1$, only B-1 free parameters are to be estimated. For the sake of notational convenience, assume b" = B.

Estimating a common unit for the a's corresponds to estimating a common zero for their logarithms, which seems more natural when applying least squares procedures. Therefore, define

$$\begin{aligned} \gamma_b &= \ln c_{bB}, \quad (b=1,\ldots,B-1), \\ \alpha_{bi} &= \ln a_{bi}, \quad (i \in I_b;\ b=1,\ldots,B), \\ d_{bi} &= \begin{cases} 1 & \text{if } i \in I_b, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{18}$$

The loss function to be minimized with respect to γ_b, (b=1,...,B-1), is given by

$$F_2 = \sum_{i=1}^{k} \frac{\sum_b w_b d_{bi} [\alpha_{bi} + \gamma_b - \tilde{\alpha}_i]^2}{\sum_b w_b d_{bi}}, \tag{19}$$

where

$$\tilde{\alpha}_i = \frac{\sum_b w_b d_{bi} (\alpha_{bi} + \gamma_b)}{\sum_b w_b d_{bi}}, \tag{20}$$

and w_b is some suitable weight. It seems reasonable to take the weights proportional to the number of respondents answering to booklet b. The interpretation of (19) is straightforward: (20) gives the average α_i after adjustment, and the numerator in the fraction of (19) gives the

(weighted) squared distance of the individual adjusted α_{bi}'s to their adjusted average. So the fraction in (19) is the within item variance of the adjusted α_{bi}'s.

The partial derivative of F_2 with respect to the γ-parameters is given by

$$\frac{\partial F_2}{\partial \gamma_j} = 2 \sum_{i=1}^{k} \frac{w_j d_{ji} (\alpha_{ji} + \gamma_j - \tilde{\alpha}_i)}{\sum_b w_b d_{bi}}, \qquad (j=1,\ldots,B-1). \tag{21}$$

Defining

$$\bar{\alpha}_i \doteq \frac{\sum_b w_b d_{bi} \alpha_{bi}}{\sum_b w_b d_{bi}}, \tag{22}$$

and using this in (20) yields

$$\tilde{\alpha}_i = \bar{\alpha}_i + \frac{\sum_b w_b d_{bi} \gamma_b}{\sum_b w_b d_{bi}}, \qquad (i=1,\ldots,k). \tag{23}$$

Substituting (23) into (21), and equating to zero gives a system of linear equations

$$\sum_{b=1}^{B-1} \gamma_b \sum_{i=1}^{k} \frac{w_b w_j d_{bi} d_{ji}}{\left(\sum_g w_g d_{gi}\right)^2} - \gamma_j \frac{\sum_i w_j d_{ji}}{\sum_g w_g d_{gi}} =$$

$$\sum_i \frac{w_j d_{ji} (\alpha_{ji} - \bar{\alpha}_i)}{\sum_g w_g d_{gi}}, \qquad (j=1,\ldots,B-1). \tag{24}$$

Letting $\underline{\gamma} = (\gamma_1,\ldots,\gamma_{B-1})$, $\underline{z}$ a (B-1) vector with elements

$$z_j = \sum_i \frac{w_j d_{ji} (\alpha_{ji} - \bar{\alpha}_i)}{\sum_g w_g d_{gi}}, \qquad (j=1,\ldots,B-1), \tag{25}$$

and H a square matrix of order B-1 with elements

$$h_{jb} = \sum_i \frac{w_b w_j d_{bi} d_{ji}}{\left(\sum_g w_g d_{gi}\right)^2} - \delta_{bj} \sum_i \frac{w_j d_{ji}}{\sum_g w_g d_{gi}}, \qquad (j,b=1,\ldots,B-1), \tag{26}$$

where δ_{bj} is the Kronecker delta, then (24) can be equivalently written as

$$H\underline{\gamma} = \underline{z}, \tag{27}$$

and the solution is given by

$$\underline{\gamma} = H^{-1}\underline{z}. \tag{28}$$

The final solutions for the discrimination indices are given by

$$a_i^* = \exp(\tilde{\alpha}_i). \tag{29}$$

These a^*'s will in general not be integer values, they have to be rounded to the nearest integer. Since only the ratios of the a^*'s do really matter, they could be multiplied by a large constant before rounding, so that the ratios of the rounded a^*'s are almost the same as the ratios before rounding. However, this will in general result in large discrimination indices, resulting in turn in very large score ranges and thus increasing the risk that CML estimates of the category parameters do not exist. On the other hand, choosing a small multiplicative constant will map really different discriminations on the same integer. In practice, the multiplicative constant is chosen in such a way that the geometric mean of the transformed a^* equals some specified constant J. If G represents the geometric mean of the a*'s, then the final values of the discrimination indices are given by

$$a_i = \max[1, \text{int}(0.5 + \frac{J a_i^*}{G})]. \tag{30}$$

The only problem that remains is to choose a suitable value for J. Practice showed that acceptable values mostly range from 1.5 to 5.

Examples

In this section two examples of the estimation of the discrimination indices will be discussed. The problem with the

procedure, however, is that there are no objective criteria to evaluate it. In an analysis of the same data set with two different sets of discrimination indices, the likelihood function values at the solution cannot be compared by a formal statistical test because the number of degrees of freedom is equal in both cases. So we should have recourse to informal criteria, or to artificial data.
Both will be discussed.

An analysis of artificial data is discussed as a first example. The item collection consists of k=15 items, each allowing a score of 0, 1 or 2; so $m_i = 2$ for all i. The items are distributed over two booklets, booklet one consisting of the items 1 to 10 and booklet two of the items 6 to 15. The category parameters β_{i1} and β_{i2} are -1 respectively 1 for items 1 to 5; -.5 respectively .5 for the items 6 to 10 and both zero for the items 11 to 15. The 'true' discrimination indices are 1 for odd numbered items and 2 for the even numbered ones. Responses for thousand artificial subjects were generated according to (1), where the ϑ-values were randomly and independently drawn from a standard normal distribution. Five hundred subjects responded to booklet 1 and five hundred to booklet 2.

In estimating the discrimination indices with the two stage algorithm described above, the only variable which is under the control of the user is the geometric mean J of the estimated indices (before rounding). The indices were estimated three times with J equal to 1.5, 3 and 6 respectively. The rounded estimates are displayed in Table 1. The geometric mean of the true values is 1.38. Setting J to 1.5 estimates the indices correctly, whereas for J = 3 the index for item 7 is somewhat overestimated while the index for item 10 is underestimated. For J = 6, the general pattern is still there, the estimates for the even items being systematically larger than the estimates for the odd items, but there is a remarkable deviation from the true ratio of 2. The ratio of the geometric mean of the even numbered estimates to the geometric mean of the odd numbered estimates is only 1.58, displaying some sort of regression towards the mean. The ratio before rounding is 1.65, so the rounding with J=6

strengthens the regression, whereas in the other two cases the rounding tends to neutralize the regression effect. Although the precise cause of this regression is not quite clear, it may well be due to the coarse grouping (G=6 in both booklets), which implies ignoring a substantial part of the variance of ϑ. Further research on this topic may be needed.

TABLE 1
Estimates of the discrimination indices
in the artificial example

item	J=1.5	J=3	J=6	before rounding
1	1	2	5	0.7128
2	2	4	8	1.3065
3	1	2	5	0.7667
4	2	4	7	1.1429
5	1	2	5	0.7500
6	2	4	9	1.3656
7	1	3	5	0.8058
8	2	4	8	1.2729
9	1	2	5	0.7340
10	2	3	7	1.0735
11	1	2	4	0.6587
12	2	4	8	1.1936
13	1	2	5	0.7572
14	2	4	7	1.1520
15	1	2	5	0.6983

Of course, one cannot expect that the ratio of the estimates will reflect the true ratio, since the sample is finite, and the estimates will reflect the peculiarities of the sample. Therefore, the β-parameters were estimated with the CML-procedure, holding the a's fixed at the values given in Table 1. Since the program yields a number of statistical tests per item and the log-likelihood for the total data set, the three analyses may be compared by counting for instance the number of significant test statistics at a predetermined significance level. The results are summarized in Table 2. The computer program computes m_i general test statistics per item which are asymptotically

chi-squared distributed. These tests are not sensitive against specific alternatives. Besides $3xm_i$, so-called M-tests are computed, which are sensitive to misspecification of the discrimination index. However, these tests are not independent and their degree of dependence varies depending on the two tests being concerned with the same item or with different items.

TABLE 2
Number of significant tests and log-likelihood in the artificial example

		J=1.5	J=3	J=6
general test	α=.05	1	0	1
	α=.01	0	0	0
M-tests	α=.05	0	1	4
	α=.01	2	2	0
log-likelihood		-5584	-4952	-4324

Two tests on the same item are highly dependent, while two tests concerned with different items are quasi independent. (For a detailed discussion, see Verhelst et al., 1991). Therefore we counted the items for which at least one of the tests (the general test or one of the M-tests) was significant. Note that an entry in a row labelled α=.05 refers to the number of items for which at least one of the tests yielded a result significant at the 5% level, and none was significant at the 1% level.

It can be concluded that all three sets of estimates do an excellent job in view of the purpose they were devised for: giving a good hunch at the true discrimination indices such that only minor adaptations, based on the statistical tests are required to arrive at an acceptable model. The very clear trend in the last line of Table 2 will be commented upon in the discussion section.

A pilot study for the Dutch national assessment program in geography will be taken as a second example. Four groups of items were arranged in a

balanced incomplete block design with two groups per block yielding six different booklets. The design is presented in Table 3.

TABLE 3
Design of the geography pilot study

group		A	B	C	D
number of items	25	25	24	25	
booklets					
1		x	x		
2	x		x		
3	x			x	
4		x	x		
5		x		x	
6			x	x	

The six subsamples, one for each booklet, consisted of approximately 250 students, so that each item was presented to approximately 750 respondents. Fifteen items consisted of a short introduction story, followed by four to nine true-false questions. In order to avoid or at least to lessen the difficulties caused by interdependencies between these questions, it was decided to treat these multiple question items as single polytomous items. The analysis started with values of the discrimination indices computed by a heuristic which did not do a good job with polytomous items, and which will not be discussed here. The important point is that seven analyses were required, each with a modification of the discrimination indices of the preceding one, in order to get an acceptable solution. Since with real data it is impossible to know what the 'true' solution is, and since acceptance of the model is the result of a quite complicate decision process involving content oriented and statistical considerations, these criteria will not be described here in great detail. In general the iterative reanalysis of the data is decided upon by clear cut indications of the M-tests that one or more item fits may be improved by altering the discrimination index in the direction indicated by these M-tests.

The method of estimating the discrimination indices described in the present paper became available after the analyses on the geography data were carried out. In order to test the quality of our algorithm, the discrimination indices were estimated with taking for J the value of the geometric mean of the a's used in the final analysis. Out a total of 99 indices, the algorithm found 77 identical values and the remaining 22 differed one unit from the ones in the final analysis. An analysis with these estimated indices was carried out, and of the 22 items with deviant discrimination indices, the associated M-tests of 8 of them pointed clearly to a change in the direction of the final solution, while the others yielded acceptable values. Although it is quite common that by changing a few discrimination indices, the M-tests for some formerly accepted items may become significant as the result of the mutual dependence of the statistical tests, the analysis with the estimated a's certainly yielded a very good approximation, as only one of the M-statistics, which are to be interpreted as standard normal deviates, exceeded 3.

Discussion

Equation (9), where the probability in the left-hand member can be replaced by the corresponding proportion, is the basic equation used in the algorithm. A (weighted) least squares loss function then is readily constructed, see (13). If initial estimates of the β's are needed, (13) is minimized with respect to the β's and the ϑ's, while the a's are kept fixed. The grouping of respondents is based on the weighted scores. If the a's have to be estimated, (13) is minimized jointly with respect to the β's, the ϑ's and the a's and the grouping is based as the unweighted scores. A practical constraint with respect to computer memory and computing time arises with this latter procedure if the number of booklets is too large. In such a case recourse can be taken to a two-step procedure: the first step consists in minimizing (13) separately per booklet, and in the second step the estimated a's are brought to the same scale by minimizing (19). However, it should be kept in mind that considering (9) as

a strict equality implies a specification error. Equation (9) is not equivalent to (1), and using the minimization of the derived loss function (13) as a genuine parameter estimation method will inevitably yield inconsistent parameter estimates. As has been pointed out, the specification error is that the conditional variance of ϑ given the score is considered as being zero, while in (1) it is definitely non-zero for all nondegenerate ϑ-distributions. Therefore, it is not hoped that some clever refinement of the least squares method will yield estimates with acceptable statistical properties.

Using the β's resulting from the minimization of (13) as start values for the algorithm which yields the maximum likelihood estimates is, of course, only a numerical recipe whose merits are to be judged on empirical results. In fact, it turns out that in the large number of analyses carried out thus far at Cito, these starting values do a good job in that they closely approximate the CML estimates, requiring a small to moderate number of iterations to find the maximum likelihood estimates. An exception to this rule is formed by poorly linked designs, where usually a larger number of iterations is required.

As to the use of the estimates of the discrimination indices, either by minimizing (13) directly with regard to the ϑ's, the β's and the a's or by using the two stage procedure, the question on the statistical status of these indices remains. In the model, as defined by (1), the a's are treated as known constants being an integral part of the null hypothesis, while in the least squares procedure the a's are treated as unknown parameters whose value are estimated from the data, usually the same data that are subsequently analyzed under the hypothesis of known a's. Formally speaking this implies a contradiction which invalidates to an unknown degree the statistical tests used in evaluating the goodness of fit of the model. However, the estimation of the discrimination indices is not part of the model and the estimation procedures it implies. Using the same data to estimate the a's and then to analyze them with the a's given the status of known constants is

chance capitalization. This can clearly be seen in the bottom row of Table 2: with a large value of J, the rounded a's better approximate the estimated ones and hence should generally give a better description of the data. On the other hand, there are arguments which may make the present procedure acceptable to a certain level. First, extracting hypothetical values for certain parameters is common use. Take a simple model such as factor analysis. The number of factors used is a discrete parameter which in many cases is estimated from the data which are being analyzed, but is nevertheless treated as a fixed constant. (See, for example, Schönemann (1981) for a discussion). Yet those who make a 'sharp' distinction between exploratory and confirmatory factor analysis or some broader class of structural models usually derive their 'final' hypothesis from one or more analyses on the same data (see, for example, Jöreskog, 1974). An analogous way of working is found in loglinear analysis (e.g., Fienberg, 1977) where a 'best fitting' model is chosen from among a number of tested models on the same data, and where the test statistic used to select the preferred model is also presented as evidence for the goodness of fit of the chosen model. Although such practice is to be rejected from an orthodox statistical point of view, a good alternative is mostly not available, since it would imply a new and independent data collection for every new hypothesis which is not entirely independent of the former one(s). And this gives us a second argument: it is by no means necessary for the analysis with the model described by (1) to extract the values of the a's and the estimates of the β's from the same data set. Any values provided by the user are acceptable as a hypothesis. This could be values dictated by content oriented theory, or by a least squares procedure as described above from some independent data set. Ultimately the user decides where his hypotheses stem from, and it is the user's risk to take a chance capitalization by extracting his hypotheses from the very data to be analyzed. As a third argument, one should not exaggerate the risk of chance capitalization. Since the acceptance or rejection of any statistical model is never free of risk, a decision based on statistical criteria if it has

serious theoretical or practical implications, will always be cross validated; so the hypothesized invariance of the discrimination indices will almost automatically be investigated in a cross validation study. In summary then it can be concluded that if the discrimination indices, as estimated with the weighted least squares algorithm, lose by definition their status of hypothesis, they cannot be denied to yield at least valuable suggestions.

References

Andersen, E.B. (1973). *Conditional inference and models for measuring.* Unpublished dissertation. Copenhagen: Mentalhygienisk Forskningsinstitut.

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examiner's Ability. In F.M. Lord, & M.R. Novick, *Statistical theory of mental test scores.* Reading (Mass.): Addison-Wesley.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis: theory and practice.* Cambridge (Mass.): MIT Press.

Fienberg, E.S. (1977). *The analysis of cross-classified categorical data.* Cambridge (Mass.): MIT Press.

Jöreskog, K.G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D.H. Krantz, R.C. Atkinson, R.D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol 1). San Francisco: Freeman.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Schönemann, P.H. (1981). Factorial definitions of intelligence: dubious legacy of dogma in data analysis. In I. Borg (Ed.), *Multidimensional data representations: when and why.* Ann Arbor: Mathesis Press.

Verhelst, N.D., Glas, C.A.W., & Verstralen H.H.F.M. (1991). *OPLM: a one parameter logistic model for dichotomous and polytomous data.* Arnhem: Cito.

Recent Measurement and Research Department Reports:

91-1 N.D. Verhelst & N.H. Veldhuijzen. A New Algorithm For Computing Elementary Symmetric Functions And Their First And Second Derivatives.

91-2 C.A.W. Glas. Testing Rasch Models For Polytomous Items: With An Example Concerning Detection Of Item Bias.

91-3 C.A.W. Glas & N.D. Verhelst. Using The Rasch Model For Dichotomous Data For Analyzing Polytomous Responses.

91-4 N.D. Verhelst & C.A.W. Glas. A Dynamic Generalization Of The Rasch Model.

91-5 N.D. Verhelst & H.H.F.M. Verstralen. The Partial Credit Model With Non-Sequential Solution Strategies.

91-6 H.H.F.M. Verstralen & N.D. Verhelst. The Sample Strategy Of A Test Information Function In Computerized Test Design.

91-7 H.H.F.M. Verstralen & N.D. Verhelst. Decision Accuracy In IRT Models.

91-8 P.F. Sanders & T.J.H.M. Eggen. The Optimum Allocation Of Measurements In Balanced Random Effects Models.

91-9 P.F. Sanders. Alternative Solutions For Optimization Problems In Generalizability Theory.