DOCUMENT RESUME

ED 308 238 TM 013 630

AUTHOR	Kelderman, Henk
TITLE	Loglinear Multidimensional IRT Models for Polytomously Scored Items. Project Psychometric Aspects of Item Banking No. 41. Research Report 88-17.
INSTITUTION	Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE	Nov 88
NOTE	39p.
AVAILABLE FROM	Bibliotheek, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Early Childhood Education; Equations (Mathematics); Estimation (Mathematics); Foreign Countries; Goodness of Fit; *Latent Trait Theory; Learning Problems; Mathematical Models; Maximum Likelihood Statistics; *Multidimensional Scaling; *Scoring; *Test Items; Young Children
IDENTIFIERS	*Log Linear Models; Multidimensional Models; Netherlands; Partial Credit Model; *Polytomous Scoring; Rasch Model; Response Function Discrepancies

ABSTRACT

A loglinear item response theory (IRT) model is proposed that relates polytomously scored item responses to a multidimensional latent space. Each item may have a different response function where each item response may be explained by one or more latent traits. Item response functions may follow a partial credit model (D. Andrich, 1978; and G. N. Masters, 1982), a multidimensional Rasch model (G. Rasch, 1961; and E. B. Andersen, 1973, 1983), or other forms of response functions to be defined by the user. Conditional maximum likelihood estimates are derived, and the models may be tested generally or against alternative loglinear models. The latter tests are sensitive to deviations from local independence subgroup invariance or assumptions about the form of the operating characteristic curves. The model was illustrated through application to data from a test to identify learning problems in Dutch children from 4 to 6.5 years of age. Fifteen items were administered to 66 children aged 4 to 5 years, 132 children aged 5 to 5.5 years, and 65 children aged 5.5 to 6 years. Three appendices illustrate the dichotomous Rasch model, the partial credit model, and Rasch's multidimensional model. Two tables contain data from the illustrative study and three figures give examples to support the text. (SLD)

Loglinear Multidimensional IRT Models for Polytomously Scored Items

Research Report
88-17

Henk Kelderman

department of
EDUCATION

Division of Educational Measurement and Data Analysis

University of Twente

Project Psychometric Aspects of Item Banking No.41

Colofon:
Typing: Lorette Padberg
Cover design: Audiovisuele Sectie TOLAB Toegepaste Onderwijskunde
Printed by: Centrale Reproductie-afdeling

Loglinear Multidimensional IRT models

for Polytomously Scored Items

Henk Kelderman

Loglinear multidimensional IRT models for polytomously scored items / Henk Kelderman – Enschede : University of Twente, Department of Education, November, 1988. – 32 pages

Abstract

A loglinear IRT model is proposed that relates polytomously scored item responses to a multidimensional latent space. Each item may have a different response function where each item response may be explained by one or more latent traits. Item response functions may follow a partial credit model (Andrich, 1978; Masters, 1982) a multidimensional Rasch model (Rasch, 1961; Andersen, 1973, 1983) or other forms of response functions to be defined by the user. Conditional maximum likelihood estimates are derived and the models may be tested generally or against alternative loglinear models. The latter tests are sensitive to deviations from local independence subgroup invariance or assumptions about the form of the operating characteristic curves.

Key words: Multidimensional item response theory, Loglinear model, polytomous responses, graded response models, goodness of fit testing.

Loglinear Multidimensional IRT models for Polytomously Scored Items

Loglinear models have been used for the estimation and testing of IRT models (Cressie and Holland, 1983; Duncan, 1984; Kelderman, 1984; Tjur, 1982). They have proved useful in the solution of practical psychometric problems such as item bias detection (Kelderman, 1985) and equating (Kelderman, 1986). All loglinear IRT models considered so far are dealing with a one dimensional latent trait and dichotomously scored items. In many testing situations, however, different subjects may give responses that differ in psychologically meaningful ways. To produce one answer may require quite another ability from the examinee than to produce another answer. And even if both responses are related to the same latent trait, one response may be more difficult to give than another or one response may require the repeated application of the same ability whereas another response may require only a single application of that ability.

In the present paper, a loglinear IRT model is proposed that applies to the situation of polytomously scored test items that may be explained by a multidimensional latent space. The flexibility and generality of loglinear IRT modeling enables the analyst to formulate a model that is precisely tailored to the particular items in the test. In the proposed model each item may have a different number of response categories each of which may be related in a different way to the latent traits. Item response functions may be specified according to a multidimensional Rasch model (Rasch, 1961; Andersen, 1973, 1983) or to a graded response model (Andrich, 1978; Masters, 1982) or another response function to be defined by the user. The usual assumption of local independence of the item responses given the latent traits is made. The parameters are estimated by the conditional maximum likelihood method and the models are tested either generally or against special alternative models that contain parameters describing deviations of local independence, subgroup invariance and the specified operating characteristic curves.

The Model

Suppose that each of N subjects respond to k test items where the answers of subject i to item j may be any of r_j+1 responses x_{ij} ($x_{ij}=0,...,r_j$). The response pattern of subject i on all k test items is denoted by the vector $x_i = (x_{i1}, x_{i2},...,x_{ik}$. The corresponding random variables are denoted by capital letters X_{ij} and X_i. Let θ_{iq} be a value of subject i on a latent ability continuum q=1,...,s and let $\theta_i = (\theta_{i1}, \theta_{i2},...,\theta_{is})$ be the vector of ability values.

To produce a score x_{ij} on item j, subject i must perform certain operations where each operation depends on a certain proficiency on a latent trait. For example to produce a correct answer on the item "What is the square root of fifteen minus six?", involves three operations. First, the expression $\sqrt{(15-6)}$ must be obtained from the verbal formulation. Second, the subject must make the subtraction 15–6=9. And finally, the square root $\sqrt{9}=3$ must be taken. It may be hypothesized that to perform the first operation successfully, the subject must have a certain level on a verbal ability trait θ_{i1}, and to perform the second and the third operation, a latent numerical ability θ_{i2} is needed. Producing the correct answer '3' requires ability θ_{i1} once and ability θ_{i2} twice. Producing the partially correct response '9' requires ability θ_{i1} and θ_{i2} each once. Producing the formula '$\sqrt{(15-6)}$' requires only θ_{i1}.

Let $B_{jq}(x)$ be the number of times that a person has to apply latent trait q to produce a response x on item j and let $\phi_j(x)$ be a parameter describing the easiness of response x of item j. The probability that subject i has a response x on item j can now be written as:

$$(1) \qquad P(X_{ij}=x) = \frac{\exp\{ \sum_{q=1}^{s} \theta_{iq} B_{jq}(x) + \phi_j(x)\}}{\sum_{y=0}^{r_j} \exp\{ \sum_{q=1}^{s} \theta_{iq} B_{jq}(y) + \phi_j(y)\}}$$

By choosing the functions $B_{jq}(.)$, different models can be chosen for the dependence of item responses on latent traits. In the sequel, the function $B_{jq}(x)$ for $x=0,\ldots,r_j$ and $q=1,\ldots,s$ will be called the item scoring function (ISF) of item j.

Figure 1, gives some examples of item scoring functions that might be employed in model (1). In figure 2 each of these ISF's are showed in a diagram, where a latent variable is depicted in a circular box and an observed response in a square box. An arrow is drawn from a latent trait to an item response if that response depends on the latent trait. The value next to the arrow denotes corresponding value of $B_{jq}(x)$.

ISF (a) describes the dichotomous Rasch model. A wrong response (x=0) is scored $B_{j1}(0)=0$ and a correct response (x=1) is scored $B_{j1}(1)=1$ indicating that a successful application of trait 1 is required. It is easily shown that with ISF (a) model (1) reduces to the dichotomous Rasch model (see Appendix I)

In ISF (b) three responses are scored where response x=0 corresponds to a wrong response and the responses x=1 and x=2 are both right responses involving one successful application of the latent trait. Both responses, however, may not be equally likely as the parameters $\phi_j(1)$ and $\phi_j(2)$ may differ.

ISF (c) describes the partial credit model (Andrich, 1978; Masters, 1982)(see Appendix II). The response x=2 has the score 2 indicating that the latent trait has to be applied twice to obtain the correct answer. The response x=1 corresponds to a partial answer, for which the latent trait has to be applied only once.

In ISF (d) through (h) two latent traits are involved. In ISF (d) the response x=1 and x=2 each depend on a different latent trait. This is the multidimensional Rasch model described by Rasch (1961) and Andersen (1973)(see Appendix III). In ISF (e) each response depends on its own latent trait but the item scores are not all equal to one. This model corresponds to Andersen's (1983) generalized Rasch model.

ISF (f) describes a multidimensional partial credit model where each operation depends on a different latent trait. The correct response $x=2$ requires two operations depending on latent trait θ_{i1} and θ_{i2} respectively. The partial response $x=1$ requires only one operation depending on the latent trait θ_{i1}. This might be an alternative ISF for the item √(15–6) where the first latent ability is subtraction and the second latent trait is taking the square root.

Obviously a combination of (c) and (f), where there are different latent traits but some operations depend on the same latent trait, is also possible. ISF (g), for example, may model the item √(20–5–6) where there are two subtractions and one square root.

ISF (h) is a two dimensional model for a dichotomous item. Finally, ISF (i) is the null function. The variable does not measure an underlying latent trait. This ISF may be used to add background variables (e.g. sex) to the model. Adding background variables may be useful to study subgroup invariance of the psychometric model. This will be discussed in the section on goodness-of-fit testing.

We have discussed some examples of ISF's. There may be many more ISF's than are shown here that make sense in a particular application. Also in one test each item may have its own ISF.

Estimation

Assuming local independence, the simultaneous distribution of $\mathbf{x}_i = (X_{i1}, \ldots, X_{ik})$ given θ_i is

$$(2) \qquad P(X_i = x_i | \theta_i) = \prod_{j=1}^{k} P(X_{ij} = x_{ij} \mid \theta_i) =$$

$$\exp\{\sum_{q=1}^{s} \theta_{iq} t_{iq} + \sum_{j=1}^{k} \phi_j(x_{ij})\} \cdot \prod_{j=1}^{k} c^{-1}(\theta_i, \phi_j)$$

where $\phi_j = (\phi_j(0), \ldots, \phi_j(r_j))$,

$$c(\theta_i, \phi_j) = \sum_{y=0}^{r_j} \exp\{\sum_{q=1}^{s} \theta_{iq} B_{jq}(y) + \phi_j(y)\}$$

and

$$t_{iq} = \sum_{j=1}^{k} B_{jq}(x_{ij}) \qquad q = 1, \ldots, s$$

Note that (2) is an exponential family model and the sums of scores $t_{i1}, \ldots, t_{is}$ are sufficient statistics for the latent ability parameters $\theta_{i1}, \ldots, \theta_{is}$. That is, all information about those latent ability parameters is contained in the sums of scores. The simultaneous distribution of the sum score variables taking values $t_i = (t_{i1}, \ldots t_{is})$ is

(3) $$f(t_i|\theta_i) = \sum_{x_i|t_i} f(x_i|\theta_i) =$$

$$\gamma(t_i, \phi) \cdot \exp(\sum_{q=1}^{s} \theta_{iq} t_{iq}) \prod_{j=1}^{k} c^{-1}(\theta_i, \phi_j)$$

with

(4) $$\gamma(t_i, \phi) = \sum_{x_i|t_i} \exp(\sum_{j=1}^{k} \phi_j(x_{ij})).$$

where $\phi = (\phi_1, \ldots, \phi_k)$ and where $\sum_{x_i|t_i}$ means summation over those values of x_i for which $(\sum_j B_{j1}(x_i), \ldots, \sum_j B_{js}(x_i))$ is equal to t_i. The conditional distribution of $X_{i1}, \ldots, X_{ik}$ given $t_{i1}, \ldots, t_{is}$ is

(5) $$P(X_i = x_i|t_i) = P(X_i = x_i|\theta_i) \;/\; f(t_i|\theta_i)$$

$$= \exp\{\sum_{j=1}^{k} \phi_j(x_{ij})\} \;/\; \gamma(t_i, \phi),$$

which does no longer depend on θ_i.

If it is assumed that individuals respond independently of one another, model (5) describes a multinomial distribution of the response patterns x for each score pattern t, where the subject index i is dropped.

Let

$$(6) \qquad m_{xt} = N_t P(X=x|t)$$

$$x_1 = 0, \ldots, r_1; \ldots; \quad x_k = 0, \ldots, r_k; \quad x = (x_1, \ldots, x_k)$$
$$t_1 = \sum_{j=1}^{k} B_{j1}(x_j); \ldots; t_s = \sum_{j=1}^{k} B_{js}(x_j); \quad t = (t_1, \ldots, t_s)$$

be the conditional expected frequency of the response pattern x given score vector t and let f_{xt} be the corresponding observed frequency. Taking logarithms we have the loglinear model:

$$(7) \qquad \log m_{xt} = \sigma_t + \sum_{j=1}^{k} \phi_j(x_j),$$

where $\sigma_t = -\log(\gamma(t, \phi)/Nt)$ is a fixed normalizing constant and the parameters $\phi_j(x_j)$ $(j = 1, \ldots, k; \; x_j = 0, \ldots, r_j)$ are considered random parameters.

Model (7) is a quasi-loglinear model for an incomplete item 1 x item 2 x ... x item k x score 1 x ... x score s contingency table. The table is incomplete because the cells for which $t_q \neq B_{1q}(x_1) + B_{2q}(x_2) + \ldots + B_{kq}(x_k)$ are structurally zero (Bishop, Fienberg and Holland, 1975; see 5.4; Haberman, 1979, see 7.3).

Unless further restrictions will be set on the ϕ parameters they will not be identifiable. To formulate identifiability conditions for model (7) first rewrite it as

(8) $\log m_t = \sigma_t + Dt\phi$,

where log is an element-wise operator and m_t is the vector of expected counts with sumscores t, $\phi = (\phi_1(0), \ldots, \phi_1(r_1), \ldots, \phi_k(0), \ldots, \phi_k(r_k))$ and D_t is the design matrix with zero's and ones in the appropriate places. The identifiability criteria are then (Imrey, Koch and Stokes, 1981)

(9) $\text{rank } [1\ D_t] = 1 + \text{rank } [D_t]$

and

(10) $\text{rank } [D] = a$

where $D' = [D'_0, \ldots, D'_{t_{max}}]$ and a is the number of columns of D, i.e. the number of parameters.

Condition (10) ensures that the ϕ parameters are not linearly dependent upon each other and condition (9) ensures they are not linearly dependent on the sumscore parameters.

If D does not satisfy these identifiability conditions certain columns must be removed, which is equivalent to setting the corresponding ϕ parameters to zero. To derive the likelihood equations, it is assumed that the identifiability conditions are met.

The conditional likelihood of model (7) is

$$(11) \quad L_C = \log \Pi_t \; \Pi_{x|t} [(N_t! \; P(X = x|t)^{f_{xt}}) \;/\; f_{xt}!])$$
$$= \Sigma_t \; \Sigma_{x|t} \; (f_{xt} (\sum_{j=1}^{k} \phi_j (x_j) - \gamma(t, \phi)) + \text{constant}$$
$$= \sum_{j=1}^{k} \phi_j(x_j) \; f_{x_j}^{X_j} - \Sigma_t \; N_t \; \gamma(t, \phi) + \text{constant},$$

where $f_{x_j}^{X_j}$ is the marginal observed frequency of response x_j on item j.

Model (11) is an exponential family model with sufficient statistics $f_{x_j}^{X_j}$. Maximum likelihood equations can be obtained by taking the derivatives of the loglikelihood for ϕ and setting them to zero. This yields the equations (Andersen, 1980; Haberman, 1975)

$$(12) \quad f_{x_j}^{X_j} = m_{x_j}^{X_j} , \qquad x_j = 0,\ldots,r_j; \; j = 1,\ldots,k,$$
$$f_t^T = N_t , \qquad \forall \; t .$$

Solving (12) for the parameters $\phi_j(x_j)$ yields maximum likelihood estimates of the parameters.

The system (12) can be solved iteratively by a Newton Raphson Algorithm (Adbey and Dempster, 1974) or iterative proportional fitting (IPF) (Goodman, 1968; Bishop et all. 1975, p.189). Kelderman (1987) uses a version of IPF that does not require to internal storage of the full table of observed and expected counts and uses a special algorithm that considerably reduces the amount of computation to obtain expected marginal tables. The method is implemented in the LOGIMO (Loglinear IRT modeling) program.

Standard errors for the parameter estimates can be obtained by taking the inverse of the Fischer information matrix I which has elements

$$(13) \qquad \frac{d^2 \log L_c}{d\,\phi_j(x)\; d\,\phi_l(y)}$$

$j = 1, \ldots, k$; $l = 1, \ldots, k$; $x = 0, \ldots, r_j$; $y = 1, \ldots, r_l$, where $j \neq l$ or $x \neq y$.

The Fischer information matrix I provides a practical method of checking the identifiability of the model. A possibly indeterminate solution can be obtained via the IPF algorithm, since IPF does not require the model to be of full rank. This solution can be used to calculate the information matrix I. If I is not of full rank, rows (and columns) that are linearly dependent on the preceding columns may be successively deleted from the matrix I until a full rank matrix is obtained. The parameters corresponding to the deleted columns are not identifiable and they are removed from the model by setting them to zero.

Testing

The overall goodness of fit of model (1) can be tested by the Pearson statistic

(14) $$X^2 = \sum_x \frac{(f_{xt} - m_{xt})^2}{m_{xt}}$$

or the likelihood ratio statistic

(15) $$G^2 = -2 \sum f_{xt} \ln(m_{xt} / f_{xt}).$$

Both statistics are asymptotically distributed as chi-square with degrees of freedom equal to the difference between the number of logically nonzero cells and the number of estimable parameters. If any of the expected counts become to small, however, the approximation of the distribution of X^2 and G^2 by a chi-square distribution becomes inappropriate (Lancaster, 1961; Koehler, 1977, 1986; Larnz, 1978). Although the distribution of X^2 is generally closer to chi-square than G^2 (Cox and Plackett, 1980; Larnz, 1978). The traditional criterion for the size of the expected counts is five but if the distribution is smooth the minimum expected count could be as small as one (Cochran, 1952).

This, however, would still mean that the overall statistics (14) and (15) are only useful with small numbers of items. Suppose for example we have ten trichotomous items. We would have 59049 non logically-zero counts.

If the number of items is large, we choose to test the model against specific alternative loglinear models that contain model (7) as a special case and also contain parameters describing a particular deviation. Let m^*_{xt} denote the expected counts under a particular alternative model. The likelihood ratio test statistic is then

$$(16) \quad G^2(m,m^*) = -2(L_C - L^*_C)$$

and is asymptotically distributed as chi-square with degrees of freedom equal to difference in numbers of estimable parameters of both models. Any loglinear model (8) that arises by extending the design matrix D with columns and extending the parameter vector ϕ with parameters might be chosen as alternative model, provided that the identifiability conditions (9) and (10) are met.

Three types of alternative models may be used, models that are sensitive to deviations from the operating characteristic curves, models that are sensitive to violations of the local independence assumption and models that are sensitive to deviation of subgroup invariance.

In the first type of the model, parameters describing an interaction between one or more sum scores t_q and an item response x_j may be added to model (7). For example, the model

$$(17) \quad \text{Log } m_{xt} = \sigma_t + \sum_{j=1}^{k} \phi_j(x_j) + \phi_{qj}(t_q x_j)$$

contains a model term $\phi_{qj}(t_q x_j)$ describing the interaction of the score t_q and item response x_j comparing (17) against (7) using the likelihood-ratio test (16) yields a test for the hypothesis that this term is different from zero. It means that the operating characteristic curve of this item with respect to latent trait q deviates from that expected by the model. If one wants to look only at the operating characteristic of one answer category, say y, the parameter may be set to zero for all other item responses $x_j \neq y$, the model

$$\text{(18)} \qquad \log m_{xt} = \sigma_t + \sum_{j=1}^{k} \phi_j(x_j) + \phi_{jl}(x_j x_l)$$

may be compared with model (7). The parameter $\phi_{jl}(x_j x_l)$ describes an interaction between both items that cannot be explained by the original model. Terms describing the interactions between other item pairs of items and terms describing triplets of items may be added to the models.

If x_j is not an item but a subgroup variable the term $\phi_{jl}(x_j x_l)$ in model (18) describes deviations from subgroup invariance of item l. In that case the item scoring function of item x_j must be the null function (see Figure 1(i)). If $\phi_{jl}(x_j x_l)$ is not zero, the parameters differ from subgroup to subgroup. These subgroup-dependent item parameters may be used to study item bias with respect to, for example, ethnic group or sex.

Like the loglinear IRT-model alternative loglinear models may be estimated by solving likelihood equations of the form (12) where again the sufficient statistics of the extra parameters are set equal to their expectations, in terms of the parameters.

Likelihood-ratio tests can be used to compare the fit of two models if one model is nested within another model. If two models are not nested, Akaike's (1977) information criterion is used:

$$AIC = G^2 + 2\ (\#\ of\ parameters) + C,$$

where C is a constant that is the same for all models fitted to the same data. The model with the smallest AIC (or AIC-C) is chosen as the best fitting model.

An Example

Van Kuyk (1988) developed a test to identify learning problems in children from four to six-and-a-half years old. The tests measures the application of antonimic dimensional size concepts such as long-short, high-low, thick-thin (see Figure 3). Fifteen items are administered to 66 children of age 4 to 5, 132 children of age 5 to 5.5, and 65 children of 5.5 to 6 years old.

The figures were shown and pointed at by the test administrator, while (s)he said "here you see some shirts, they gradually become a bit ... (shorter)". the answer was rated correct if the right size concept (e.g. long-short) was given and was correctly applied (e.g. shorter rather than longer). Small children may be unable to produce the correct specific concept (e.g. long-short) but use the general size concept 'big-small' instead. If 'big-small' is correctly applied (e.g. the shirt is smaller) the answer is rated partially correct. All other answers are rated incorrect.

Three ISF are applied to this data: the one dimensional partial credit model (1PC, Figure 1c), the two dimensional partial credit model (2PC, Figure 1f) and the two dimensional Rasch model (2RM, Figure 1d). In the 1PC it is hypothesized that one spatial ability has to be applied twice to obtain the correct answer: once to know the extensiveness of the objects (e.g. small or large) and once to identify the correct dimension (e.g. short-long).

In the 2PC model both processes are supposed to depend on distinct abilities, an extensiveness ability and a specific concept ability. To produce the correct answer both abilities have to be applied once. To produce the partially correct answer (small) only the first ability is needed.

Finally in the 2RM model two latent abilities are conjectured: a general extensiveness and a specific extensiveness ability. The general extensiveness ability is applied to get the partially correct answer and the specific extensiveness ability is sufficient to produce the correct answer. The difference between the 2PC and the 2RM model is that 2RM does not require an application of a general extensiveness ability to produce the correct answer but the 2PC does.

For all three ISF's two models are fitted: model (7) where the item response parameters are invariant over subgroups:

$$\text{(19)} \quad \log m_{\underset{\sim\sim}{xt}} = \sigma_{\underset{\sim}{t}} + \phi_1(x_1) + \ldots + \phi_{15}(x_{15}) + \phi_{16}(x_{16})$$

with constraints $\phi_j(0) = 0$, $j = 1, \ldots, 16$, and model (18), where the item response parameter vary from subgroup to subgroup:

$$\text{(20)} \quad \log m_{\underset{\sim\sim}{xt}} = \sigma_{\underset{\sim}{t}} + \phi_1(x_1) + \ldots + \phi_{15}(x_{15}) + \phi_{16}(x_{16})$$

$$+ \phi_{1,16}(x_1 x_{16}) + \ldots + \phi_{15,16}(x_{15}, x_{16})$$

with constraints $\phi_j(0) = 0$, $j = 1, \ldots, 16$ and $\phi_{j,16}(0, x_{16}) = \phi_{j,16}(x_j, 1) = 0$, $j = 1, \ldots, 15$. Thus the parameters are normed such that $\phi_j(x_j)$ $(j = 1, \ldots, 15)$ are the item response parameters in the youngest age group $(x_{16}=1)$ and the $\phi_{j,16}(x_j x_{16})$ are deviations of those parameters for older age groups.

Table 1 presents the test results of the six models. It is seen that the subgroup dependent versions of the models have a significantly better fit than the subgroup invariant versions. The likelihood-chi-square statistics are all very large relative to the degrees of freedom.

Since the 1PC, 2PC and 2RM models are not nested their fit is compared with Akaike's information criterion. The AIC values are given plus a constant to produce a relative small number. The best fitting model according to the AIC criterion is the 2RM model closely followed by the 2PC model.

The maximum likelihood estimates of the age dependent 2PC model are given in Table 2. The 2RM model shows a similar picture. It is seen that the easiness of the partial response ard the correct response changes positively for some items and negatively for other items. For example item 13, showing bags with different amounts of candy, becomes relatively more easy as children get older.

APPENDIX I: The Dichotomous Rasch Model.

Using ISF a (Figure 1) with Model (1) we obtain:

$$P_j(x_{ij}=1|\theta_i) = \frac{\exp\{\theta_i + \phi_j(1)\}}{\exp\{\theta_i 0+\phi_j(0)\} + \exp\{\theta_i+\phi_j(1)\}}$$

$$= \frac{\exp\{\theta_i - \delta_j\}}{1 + \exp\{\theta_i - \delta_j\}}$$

where $\delta_j = (\phi_j(0) - \phi_j(1))$. This is identical to the dichotomous Rasch model.

APPENDIX II: The Partial Credit Model.

Since there is only one latent trait, the subscripts q are omitted. With ISF (c) model 1 becomes

$$P_j(X_{ij}=x|\theta_i) = \frac{\exp\{\theta_i x + \phi_j(x)\}}{\sum_{y=0}^{r_j} \exp\{\theta_i y + \phi_j(y)\}}$$

Dividing numerator and denominator by $\exp\{\phi_j(0)\}$ and writing $\psi_j(x) = (\phi_j(x) - \phi_j(0))$ we obtain:

$$P_j(X_{ij}=x|\theta_i) = \frac{\exp\{\theta_i x + \psi_j(x)\}}{\sum_{y=0}^{r_j} \exp\{\theta_i y + \psi_j(y)\}}$$

$$= \frac{\exp\{\sum_{l=1}^{x} (\theta_i - \delta_{jl})}{1 + \sum_{y=1}^{r_j} \exp\{\sum_{l=1}^{y} (\theta_i - \delta_{jl})\}}$$

where

$$\delta_{jx} = \sum_{l=1}^{x} \delta_{jl} - \sum_{l=1}^{x-1} \delta_{jl} = -(\psi_j(x) - \psi_j(x-1))$$

which is the Masters (1982) model.

APPENDIX III: Rasch's (1961) Multidimensional Model

Using ISF (d) (Figure 1) in (1) yields

$$P(X_{ij}=p|\theta_i) = \frac{\exp\{\theta_{ip} + \phi_j(p)\}}{\sum_{g=1}^{r_j} \exp\{\theta_{ig} + \phi_j(p)\}}$$

$$= \frac{\exp\{\theta_{ip} - \delta_{jp}\}}{\sum_{g=1}^{r_i} \exp\{\theta_{ig} - \delta_{jg}\}}$$

with $\theta_{i1} = 0$, where $\delta_{jp} = -\phi_j(p)$, which is equivalent to Andersen's (1973) Rasch model for polytomous responses.

References

Adbey, P.R., & Dempster, M.A.H. (1974). Introduction to optimization methods. London: Chapman and Hall.

Akaike, H. (1977). On entropy maximization principle. In Krisschnaiah, P.R. Applications of Statistics. North Holland, p. 27–41.

Andersen, E.B. (1973). Conditional inference and multiple choice questionnaires. British Journal of Mathematical and Statistical Psychology, 26, 31–44.

Andersen, E.B. (1983). A general latent structure model. In Wainer H. and Messick S. (Eds.) Principals of modern Psychological Measurement. Hillsdale: Lawrence Erlbaum Ass.

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

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

Cochran, W.G. (1952). The χ^2–Test of Goodness of Fit. Annals of Mathematical Statistics, 23, 315–345.

Cox, M.A.A., & Plackett, R.L. (1980). Small samples in contingency tables. Biometrika, 67, 1–13.

Cressie, N., & Holland, P.W. (1983). Characterizing the manifest probabilities of latent trait models. Psychometrika, 48, 129–142.

Duncan, O.D. (1984). Rasch measurement: Further examples and discussion. In C.F. Turner & E. Martin (Eds.), Surveying Subjective Phenomena, Vol. 2. (pp. 367–403). New York: Russell Sage Foundation.

Goodman, L.A. (1968). The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. Journal of the American Statistical Association, 63, 1091–1131.

Haberman, S.J. (1979). Analysis of qualitative data: New developments, Vol. 2. New York: Academic Press.

Imrey, P.B., Koch, G.C., & Stokes, M.E. (1981). Categorical data analysis: Some reflections on the loglinear model and logistic regression. Part I: Historical and methodological overview. International Statistical Overview, 49, 265–283.

Kelderman, H. (1984). Loglinear Rasch model tests. Psychometrika, 49, 223–245.

Kelderman, H. (1985). Item bias detection using the loglinear Rasch model. Paper presented at the Fiftieth Annual Meeting of the Psychometric Society, Nashville Tennessee, June 1–4.

Kelderman, H. (1986). Common item equating using the loglinear Rash model. Paper presented at the Annual Meeting of the Society for Multivariate Behavioral Research in the Behavioral Sciences. Denmark, Copenhagen, August 26–28.

Kelderman, H. (1987). Estimating quasi-loglinear for a Rasch table if the number of items is large. University of Twente, Department of Education, Research Report 87-5.

Koehler, K.J. (1977). Goodness of fit statistics for large sparse multinomials. Unpublished Ph.D. Dissertation, School of Statistics, University of Minnesota.

Lancaster, H.O. (1961). Significance tests in discrete distributions. Journal of the American Statistical Association, 56, 223-234.

Larnz, K. (1978). Small-sample comparisons of exact levels for chi-square statistics. Journal of the American Statistical Association, 73, 412-419.

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

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 5, 321-333.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. Scandinavian Journal of Statistics, 9, 23-30.

van Kuyk, J.J. (1988). Verwerven van grootte-begrippen [Acquiring size concepts.] Pedagogische Studiën, 65.

Table 1

Goodness of Fit Statistics for Six Models on the Kuyk Data

Item Scoring Function	Subgroup Invariant Param.			Subgroup Dependent Param.			Difference		
	LR	# of param.	AIC-C	LR	# of param.	AIC-C	LR	DF	p-value
1PC	3436.51	108	652.51	3280.01	165	610.01	156.50	57	.00
2PC	3298.51	152	602.51	3141.92	209	559.92	156.59	57	.00
2MR	3291.86	156	603.86	3132.94	213	558.94	158.92	57	.00

Table 2

Parameter Estimates of the Age Dependent Two Dimensional Partial Credit Model (2PC)

	Response Effect		Response X Age Effect					
			Partial Response			Correct Response		
j	Part.R.	Corr.R.	4–5	5–5.5	5.5–6	4–5	5–5.5	5.5–6
1	0.85	2.39	–	0.52	0.17	–	1.36	0.10
2	1.00	–1.00	–	–0.94	–1.50	–	–0.43	–1.14
3	2.35	2.78	–	0.21	0.26	–	1.00	–0.00
4	0.58	2.77	–	–0.96	0.42	–	–0.90	–0.03
5	2.36	3.42	–	–1.51	–2.19	–	–0.83	–0.54
6	–0.41	–0.27	–	1.25	2.30	–	0.17	1.39
7	–0.96	–2.15	–	1.35	1.82	–	1.80	2.75
8	–1.76	–0.43	–	–1.72	0.49	–	0.57	1.32
9	–0.42	0.47	–	1.25	2.10	–	1.06	2.82
10	0.63	–1.09	–	1.36	2.60	–	1.52	2.77
11	1.02	1.47	–	0.82	0.11	–	1.03	0.25
12	0.77	1.95	–	1.27	1.88	–	1.26	2.20
13	1.18	–1.69	–	1.70	2.53	–	2.96	4.86
14	0.16	1.53	–	0.64	0.52	–	0.92	1.27
15	2.42	–0.19	–	1.24	0.91	–	0.87	1.47

Figure Caption

<u>Figure 1.</u> Examples of item scoring functions.

<u>Figure 2</u>. Diagrams for the item scoring functions in Figure 1

<u>Figure 3</u>. An item from Kuyk's size concept test.

		q
		1
x	0	0
	1	1

(a)

		q
		1
	0	0
x	1	1
	2	1

(b)

		q
		1
	0	0
x	1	1
	2	2

(c)

		q	
		1	2
	0	0	0
x	1	1	0
	2	0	1

(d)

		q	
		1	2
	0	0	0
x	1	1	0
	2	0	2

(e)

		q	
		1	2
	0	0	0
x	1	1	0
	2	1	1

(f)

		q	
		1	2
	0	0	0
	1	1	0
x	2	2	0
	3	2	1

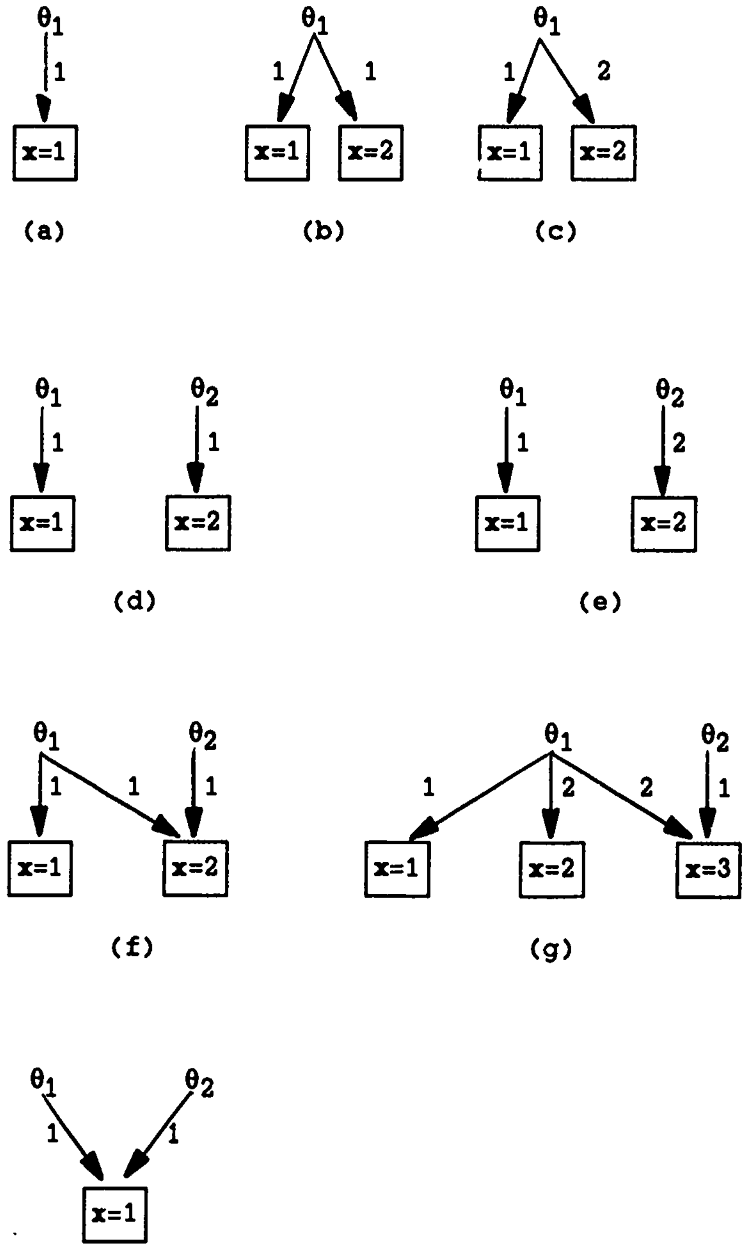
(g)

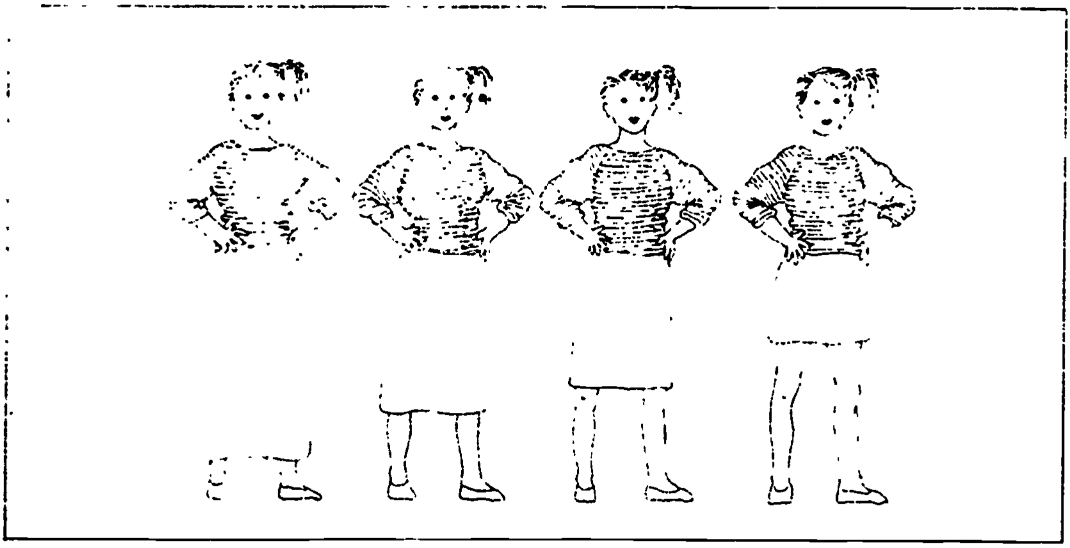
		q	
		1	2
x	0	0	0
	1	1	1

(h)

		q
		1
	0	0
x	1	0
	2	0

(i)


θ1
1
x=1
(a)
θ1
1
1
x=1
x=2
(b)
θ1
1
2
x=1
x=2
(c)
θ1
θ2
1
1
x=1
x=2
(d)
θ1
θ2
1
2
x=1
x=2
(e)
θ1
θ2
1
1
1
x=1
x=2
(f)
θ1
θ2
1
2
2
1
x=1
x=2
x=3
(g)
θ1
θ2
1
1
x=1
(h)



Titles of recent Research Reports from the Division of Educational Measurement and Data Analysis, University of Twente, Enschede, The Netherlands.

RR-87-1	R. Engelen, *Semiparametric estimation in the Rasch model*
RR-87-2	W.J. van der Linden (Ed.), *IRT-based test construction*
RR-87-3	R. Engelen, P. Thommassen, & W. Vervaat, *Ignatov's theorem: A new and short proof*
RR-87-4	E. van der Burg, & J. de Leeuw, *Use of the multinomial jackknife and bootstrap in generalized nonlinear canonical correlation analysis*
RR-87-5	H. Kelderman, *Estimating a quasi-loglinear models for the Rasch table if the number of items is large*
RR-87-6	R. Engelen, *A review of different estimation procedures in the Rasch model*
RR-87-7	D.L. Knol & J.M.F. ten Berge, *Least-squares approximation of an improper by a proper correlation matrix using a semi-infinite convex program*
RR-87-8	E. van der Burg & J. de Leeuw, *Nonlinear canonical correlation analysis with k sets of variables*
RR-87-9	W.J. van der Linden, *Applications of decision theory to test-based decision making*
RR-87-10	W.J. van der Linden & E. Boekkooi-Timminga, *A maximin model for test design with practical constraints*
RR-88-1	E. van der Burg & J. de Leeuw, *Nonlinear redundancy analysis*
RR-88-2	W.J. van der Linden & J.J. Adema, *Algorithmic test design using classical item parameters*
RR-88-3	E. Boekkooi-Timminga, *A cluster-based method for test construction*
RR-88-4	J.J. Adema, *A note on solving large-scale zero-one programming problems*
RR-88-5	W.J. van der Linden, *Optimizing incomplete sample designs for item response model parameters*
RR-88-6	H.J. Vos, *The use of decision theory in the Minnesota Adaptive Instructional System*

RR-88-7 J.H.A.N. Rikers, *Towards an authoring system for item construction*

RR-88-8 R.J.H. Engelen, W.J. van der Linden, & S.J. Oosterloo, *Item information in the Rasch model*

RR-88-9 W.J. van der Linden & T.J.H.M. Eggen, *The Rasch model as a model for paired comparisons with an individual tie parameter*

RR-88-10 H. Kelderman & G. Macready, *Loglinear-latent-class models for detecting item bias*

RR-88-11 D.L. Knol & M.P.F. Berger, *Empirical comparison between factor analysis and item response models*

RR-88-12 E. van der Burg & G. Dijksterhuis, *Nonlinear canonical correlation analysis of multiway data*

RR-88-13 J. Kogut, *Asymptotic distribution of an IRT person fit index*

RR-88-14 J.J. Adema, *The construction of two-stage tests*

RR-88-15 H.J. Vos, *Simultaneous optimization of decisions using a linear utility function*

RR-88-16 H. Kelderman, *An IRT model for item responses that are subject to omission and/or intrusion errors*

RR-88-17 H. Kelderman, *Loglinear multidimensional IRT models for polytomously scored items*

Research Reports can be obtained at costs from Bibliotheek, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

department of

EDUCATION

A publication by
the Department of Education
of the University of Twente
P.O. Box 217
[illegible]de
[illegible]nds