DOCUMENT RESUME

ED 320 956 TM 015 256

AUTHOR Westers, Paul; Kelderman, Henk
TITLE Differential Item Functioning in Multiple Choice Items. Project Psychometric Aspects of Item Banking No. 47. Research Report 90-1.
INSTITUTION Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE Apr 90
NOTE 37p.
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 Cognitive Processes; Estimation (Mathematics); Foreign Countries; Guessing (Tests); Higher Education; *Item Bias; Item Response Theory; Mathematical Models; *Multiple Choice Tests; Probability; Secondary Education; *Test Items
IDENTIFIERS Latent Class Models; Rasch Model; Second International Mathematics Study

ABSTRACT

In multiple-choice items the response probability on an item may be viewed as the result of two distinct latent processes--a cognitive process to solve the problem, and another random process that leads to the choice of a certain alternative (the process of giving the actual response). An incomplete latent class model is formulated that describes the first process by a Rasch model and the second process by a guessing model. Alternative models are specified that contain additional parameters describing differential item functioning (DIF) in the two processes. DIF with respect to either known or unknown subgroups can be tested by a likelihood ratio test that is asymptotically distributed as chi-square. As an example of the model, four five-choice items from the Second International Mathematics Study (1987) with a sample of 3,002 secondary students were considered. A 42-item list of references and 3 data tables are included. (Author/SLD)

Differential Item Functioning in Multiple Choice Items

Research Report 90-1

Paul Westers
Henk Kelderman

ment of EDUCATION

University of Twente

Division of Educational Measurement and Data Analysis

Project Psychometric Aspects of Item Banking No.47

Colofon:
Typing: L.A.M. Bosch-Padberg
Cover design: Audiovisuele Sectie TOLAB Toegepaste Onderwijskunde
Printed by: Centrale Reproductie-afdeling

Differential Item Functioning in Multiple Choice Items

Paul Westers

Henk Kelderman

University of Twente

Differential item functioning in multiple choice items , Paul Westers - Enschede : University of Twente, Department of Education, April. 1990. - 29 pages

Abstract

In multiple choice items the response probability on an item may be viewed as the result of two distinct latent processes. A cognitive process to solve the problem and another random process that leads to the choice of a certain alternative. An incomplete latent class model is formulated that describe the first process by a Rasch model and the second process by a guessing model.

Alternative models are be specified that contain additional parameters describing differential item functioning (DIF) in the two processes.

DIF with respect to either known or unknown subgroups can be tested by a likelihood ratio test that is asymptotically distributed as chi-square.

Key words: differential item functioning, multiple choice items, Rasch model, guessing model, incomplete latent class model, goodness of fit testing

Differential Item Functioning in Multiple Choice Items

Items in educational or psychological tests show differential item performance (DIF) if the probability of a correct response among equally able test takers is different between racial, ethnic, or other subgroups. DIF may lead to tests that are unfair for certain subgroups, and it is important to spot such items so that they can be improved or deleted from the test.

Many DIF detection methods have been proposed since Binet and Simon (1916, see also Jensen, 1980, p. 367) were the first to draw attention to this problem. Reviews of older DIF (also called item bias) detection methods are given by Osterlind (1983) and Shepard, Camilli and Averill (1981). Handbooks on item bias detection methods are provided by Berk (1982) and Jensen (1980).

In the last decade methods have been improved by giving better possibilities to match on ability. Various methods have used the number correct score of the test for this purpose (Camilli, 1979; Holland & Thayer, 1986; Kok, Mellenbergh, & van der Flier, 1985; Mellenbergh, 1982; Nungester, 1977 (see Ironson 1982); Scheuneman, 1979).

Recently, DIF detection methods have been proposed that are based on item response theory (IRT) (Durovic, 1975; Fischer & Formann, 1982; Lord, 1980; Mislevy, 1981; Muthén & Lehman, 1985; Wright, Mead & Draba, 1975). An IRT model explains the probability of an item response on the basis of

a person parameter and one or more item parameters. Differences between estimated item parameters across subgroups are considered as an indication of DIF. Thissen, Steinberg and Wainer (1989b) give an overview of IRT-based DIF detection methods and demonstrate their use. They also discuss DIF detection methods that can be used with multiple choice items.

The fact that in multiple choice items response alternatives are given introduces new potential sources of DIF. Green, Crone and Folk (1989) focus on differential popularity of the incorrect responses (or "distractors"). If a particular distractor is more attractive to subjects from one subgroup than for another, Green et al. conjecture that "...the item probably means something different to the different groups". They perform loglinear analysis of the subgroup x score group x incorrect response contingency table for each item, to detect distractors that are more popular in one subgroup than in another.

Another source of DIF in multiple choice items does not involve the popularity of the distractors, but concerns differential difficulty of the problem to be solved. Just as in other types of items, an item may pose a problem that is more difficult to some subjects than to others, even if they are equally able on the trait of interest. In this paper an item bias detection model is described that separates both sources of bias.

In the model it is assumed that the subject's response to a certain item depends on two distinct processes. The

first process determines whether an individual with a certain ability solves the problem that is presented by the item, the second process determines the actual response given.

Furthermore, we assume that if the subject solves the problem, (s)he will give the correct answer. Here the probability that the subject solves the problem is assumed to be governed by a Rasch (1960) model. If the subject cannot solve the problem the subject will guess the answer, where the guessing probabilities may be different for different alternatives.

The model differs from that of Thissen, Steinberg and Fitzpatrick (1989a), who distinguish between a "Don't know" state and a state in which the subject has partial or complete knowledge of the answer. In the "Don't know" state he guesses the answer as before, but in the "Partial knowledge" state the subject may answer a response alternative, where the response probabilities are governed by Bock's (1972) nominal response model.

The proposed model is simpler than the model by Thissen et al (1989a). This has two advantages. Firstly, it contains less parameters. For example, in a four choice item, our model has five item parameters while Thissen's model has fourteen. Obviously, if the sample is not very large the parameters in the latter model cannot be estimated reliable. So, in that case one may be inclined to "Buy information by assumption" and use the simpler model. Secondly, the proposed model can easily be formulated as a latent class analysis (LCA) model. LCA models have been introduced by Lazarsfeld

(1950; see also Lazarsfeld & Henry, 1968) and developed further by Goodman (1973), Haberman (1979), Clogg (1981) and others. LCA models have been used extensively for measurement in sociology, psychology and education. Formann (1985), Kelderman (1988, 1989), Kelderman and Macready (1988) and Mislevy and Verhelst (1987) and Yamamoto, (1987, 1988) integrated IRT models into LCA models. There is a well-developed theory for maximum-likelihood estimation and likelihood-ratio testing of LCA models. By comparing the fit of different latent class models, DIF in the guessing probabilities and DIF in the parameters of the Rasch model can be tested separately. Also, the model can be extended with latent classes, so that the subgroups for which the items exhibits DIF may be latent too.

In what follows the model for multiple choice items is developed and formulated as a LCA model. Different models for the detection of DIF are formulated. Also a model with latent subgroup variable is discussed. A computationally efficient estimation method is described and its use is illustrated using empirical data.

A Model for Multiple Choice Items

Suppose that each subject, randomly drawn from a population of N subjects, respond to k test items. where his/her answer to item j may be any of r_j responses y_j ($y_j=1,\ldots,r_j$). The response pattern of this subject on the test items is denoted

by the vector $\mathbf{y}=(y_1,\ldots,y_k)$. The corresponding random variables are denoted by capital letters Y_j $(j=1,\ldots,k)$ and $\mathbf{Y}$. Let x_j indicate the latent response of the subject, taking values $x_j=1$ if (s)he solved the problem or $x_j=0$ if (s)he did not solve the problem posed by item j. And let $\mathbf{x}=(x_1,\ldots,x_k)$ be the vector of these values. The corresponding random variables are denoted by X_j and $\mathbf{X}$.

The relationship between the latent responses x_j and the observed responses y_j is described by the conditional probability

$$(1) \qquad \Phi^{X_jY_j}_{x_jy_j} \equiv P(y_j|x_j)$$

where the superscripts are symbolic notation indicating that the random variables X_j and Y_j are involved in the conditional probability. For the sake of simplicity, the notations y_j, x_j, etc. in the probabilities are used for $Y_j=y_j$, $X_j=x_j$, etc.

It is assumed that if the subject can solve the problem, (s)he chooses the correct alternative, that is $\Phi^{X_jY_j}_{1\,y_j}$ must equal to 1 if y_j is the right alternative.

Assuming that y_j depends on x_j only, we have

$$(2) \qquad P(\mathbf{y}|\mathbf{x},\,\theta) = \prod_{j=1}^{k} P(y_j|x_j) = \prod_{j=1}^{k} \Phi^{X_jY_j}_{x_jy_j}$$

where θ is the latent ability value.

The latent responses are assumed to be governed by an one-parameter-logistic model (Rasch, 1960), where the probability of the latent response x_j given that the subject has ability θ is

$$(3) \qquad P(x_j|\theta) = \exp(x_j(\theta-\delta_j))/(1 + \exp(\theta-\delta_j))$$

and δ_j is the difficulty of item j.

Assuming that x_j depends only on the latent ability θ we have

$$(4) \qquad P(\mathbf{x}|\theta) = \prod_{j=1}^{k} P(x_j|\theta) = \exp(t\theta - \prod_{j=1}^{k} x_j\delta_j)\ C(\theta,\delta)^{-1}$$

with

$$C(\theta,\delta) = \prod_{j=1}^{k} (1 + \exp(\theta-\delta_j))$$

where $\delta=(\delta_1,\ldots,\delta_k)$, and $t=x_1+\ldots+x_k$ is the number correct score.

Let $F(\theta)$ be the continuous distribution function of the latent ability θ. Using (2), and (4) the marginal probability of the observed responses $\mathbf{y}$ then becomes

$$(5) \qquad P(\mathbf{y}) = \sum_{\mathbf{x}} \int_{-\infty}^{\infty} P(\mathbf{y}|\mathbf{x}, \theta)\, P(\mathbf{x}|\theta)\, dF(\theta)$$

$$= \sum_{\mathbf{x}} \left(\prod_{j=1}^{k} \Phi_{x_j y_j}^{x_j y_j} \right) \exp\left(- \sum_{j=1}^{k} x_j \delta_j\right) \int_{-\infty}^{\infty} \exp(t\theta) C(\theta, \delta)^{-1} dF(\theta).$$

In the next section we will formulate this model as an incomplete latent class model. The integral in model (5) will then be absorbed into a latent class parameter which depends only on the number correct score t. This means that it is not needed to specify the distribution function $F(\theta)$ any further.

To detect DIF in multiple choice items, model (5) has to be extended with subgroups. In order to keep the main idea of this section clear the subgroups have been ignored so far. In the third section we will extend the incomplete latent class model with the subgroups.

An Incomplete-Latent-Class Model

Kelderman (1988) has showed that model (5) is an incomplete latent-class model in the sense of Haberman (1979, ch. 10)

$$(6) \qquad P(\mathbf{y}) = \sum_{\mathbf{x}} \Phi_t^{\tau}\, \Phi_{x_1}^{x_1} \cdots \Phi_{x_k}^{x_k}\, \Phi_{x_1 y_1}^{x_1 y_1} \cdots \Phi_{x_k y_k}^{x_k y_k}$$

with

$$\Phi_t^T = \int_{-\infty}^{\infty} \exp(t\theta) C(\theta,\delta)^{-1} dF(\theta),$$

$$\Phi_{x_j}^{X_j} = \exp(-x_j\delta_j) \qquad j = 1, \ldots, k,$$

and where the Φ-parameters are subject to the restrictions

$$(7) \qquad \Phi_0^{X_j} = 1, \qquad j = 1, \ldots, k,$$

$$(8) \qquad \Phi_{x_{j1}}^{X_j Y_j} + \ldots + \Phi_{x_{jr_j}}^{X_j Y_j} = 1, \qquad j = 1, \ldots, k,$$

In this model each value of $\mathbf{x}$ represents a latent class. Model (6) is incomplete because for certain given values of $\mathbf{X}$ only a limited number of combinations $(Y_1, \ldots, Y_k)$ are possible. Because of the fact that Φ_t^T depends on an underlying latent trait distribution $F(\theta)$, these parameters are subject to the following complex inequality constraints (Cressie & Holland, 1983; Kelderman, 1984):

$$\det.(\| \Phi_{r+s}^T \|_{r,s=0}^{q_1}) \geq 0$$

and

$$\det.(\| \Phi_{r+s+1}^T \|_{r,s=0}^{q_2}) \geq 0$$

where

$$q_1 = \begin{cases} k/2 & \text{if } k \text{ is even,} \\ (k-1)/2 & \text{if } k \text{ is odd,} \end{cases}$$

$$q_2 = \begin{cases} (k-2)/2 & \text{if } k \text{ is even,} \\ (k-1)/2 & \text{if } k \text{ is odd,} \end{cases}$$

$\det.(\| - \|^{q}_{r,s=0})$ means the determinant of a matrix with row index r and column index s both running from zero to q.

Since it is not our goal to fit a model for the data, but to decide if a certain item exhibits DIF, we will follow Cressie and Holland and ignore these inequality constraints. This, the so called generalized Rasch model, provides an easy way to decide that an item exhibits DIF. The generalized Rasch model is also equivalent to the "conditional" Rasch model. That is, a Rasch model in which there is a conditioning on the number correct score (Kelderman, 1984).

Incomplete table methodology can be used to formulate several hypotheses about DIF by specifying alternative models that contain additional subgroup-dependent parameters.

Parameters describing DIF

An item can show DIF in two different ways. First, as indicated before, the item exhibits DIF if equally able

individuals from different subgroups have different probabilities of solving the problem that the item poses. This will be called DIF in the latent response.

It was assumed earlier that if the subject can solve the problem (s)he will choose the correct alternative. But if the subject can't solve the problem, (s)he would guess the most attractive alternative. Therefore, the item exhibits also DIF if the attractiveness of the alternatives varies from subgroup to subgroup. This will be called DIF in the guessing probabilities.

In most applications subgroup membership (e.g., sex) is known. In some situations, however, items are expected to exhibit DIF with respect to certain subgroups, but it is not known to which subgroup each of the individuals belongs.

In the following models are formulated for studying the two types of DIF, i.e, both for DIF in the latent response and DIF in the guessing probabilities. Further, the cases that the subgroup i (i=1,...,g) is observed or that it is not observed are considered.

DIF in the Latent Response

To detect DIF with respect to the process of solving the problem, an alternative model is formulated as

$$(9) \qquad P(\mathbf{y}|i) = \sum_{\mathbf{x}} \Phi_{it}^{IT} \Phi_{ix_1}^{IX_1} \Phi_{x_2}^{X_2} \ldots \Phi_{x_k}^{X_k} \Phi_{x_1y_1}^{X_1Y_1} \ldots \Phi_{x_ky_k}^{X_kY_k}$$

where $P(\mathbf{y}|i)$ is the conditional distribution of observed

response y given observed subgroup i, $\Phi_{1x_1}^{Ix_1} = \exp(-x_1\delta_{1i})$, δ_{1i} is the difficulty of item 1 in subgroup i, and

$$\Phi_{it}^{IT} = \int_{-\infty}^{\infty} \exp(t\theta) C(\theta,\delta)^{-1} dF_i(\theta)$$

where $F_i(\theta)$ is the distribution of the latent trait in subgroup i.

To test whether the interaction between subgroup i and the latent response to item 1 is zero, i.e., item 1 exhibits DIF in the latent response, this alternative model is compared with the model

$$(10) \quad P(\mathbf{y}|i) = \sum_{\mathbf{x}} \Phi_{it}^{IT} \Phi_{x_1}^{X_1} \ldots \Phi_{x_k}^{X_k} \Phi_{x_1y_1}^{X_1Y_1} \ldots \Phi_{x_ky_k}^{X_kY_k}$$

If the test is significant, it may be concluded that the difficulty of item 1 varies from subgroup to subgroup. In this case the subjects in one subgroup may find it more difficult to solve the problem than subjects from another subgroup.

DIF in the Guessing Probabilities

To test the null hypothesis that the interaction between the subgroup and the observed response to item 1 is zero, i.e. item 1 exhibits DIF in the guessing probabilities, the alternative model

$$(11) \quad P(\mathbf{y}|i) = \sum_{\mathbf{x}} \Phi_{it}^{IT} \Phi_{x_1}^{X_1} \ldots \Phi_{x_k}^{X_k} \Phi_{ix_1y_1}^{IX_1Y_1} \Phi_{x_2y_2}^{X_2Y_2} \ldots \Phi_{x_ky_k}^{X_kY_k}$$

where $P(\mathbf{y}|i)$ is the conditional distribution of observed response $\mathbf{y}$ given observed subgroup i and $\Phi_{ixy}^{IXY} = P(y|x,i)$ is the conditional probability of observed response y given latent response x and observed subgroup i, is compared with model (10). If the test is significant, it may be concluded that the attractiveness of the alternatives of item 1 varies from subgroup to subgroup.

In model (9) and model (11) the Φ-terms are specified to test DIF for only one item. Obviously, similar model terms can be specified for two or more items if necessary. It is also possible to analyse models in which one item exhibits DIF in the latent response and another (or the same) item DIF in the guessing probabilities.

Latent Subgroup Models

When subgroup membership is unobserved, the subgroup variable I becomes also a latent variable. And the models for the detection of DIF are still latent-class models. Models with unobserved subgroups are very useful in situations where grouping information is not available, or when it is not desirable to link the concept of DIF to any specific manifest variable.

Unlike the models in (9) to (11), the models with unobserved subgroups are not always identified. For example,

a model with unobserved subgroups in which only one item exhibits DIF in the latent response, is not identified. In order to overcome this problem, models can be specified to test DIF for v items ($2 \leq v < k$). The models (9) to (11) then become

$$(12) \quad P(\mathbf{y}) = \sum_i \sum_{\mathbf{x}} \Phi^{IT}_{it} \Phi^{IX_1}_{ix_1} \ldots \Phi^{IX_v}_{ix_v} \Phi^{X_{v+1}}_{x_{v+1}} \ldots \Phi^{X_k}_{x_k} \Phi^{X_1Y_1}_{x_1y_1} \ldots \Phi^{X_kY_k}_{x_ky_k},$$

$$(13) \quad P(\mathbf{y}) = \sum_i \sum_{\mathbf{x}} \Phi^{IT}_{it} \Phi^{X_1}_{x_1} \ldots \Phi^{X_k}_{x_k} \Phi^{X_1Y_1}_{x_1y_1} \ldots \Phi^{X_kY_k}_{x_ky_k},$$

and

$$(14) \quad P(\mathbf{y}) = \sum_i \sum_{\mathbf{x}} \Phi^{IT}_{it} \Phi^{X_1}_{x_1} \ldots \Phi^{X_k}_{x_k} \Phi^{IX_1Y_1}_{ix_1y_1} \ldots \Phi^{IX_vY_v}_{ix_vy_v} \Phi^{X_{v+1}Y_{v+1}}_{x_{v+1}y_{v+1}} \ldots \Phi^{X_kY_k}_{x_ky_k}$$

where $\Phi^{X_1I}_{x_1i} = \exp(-x_1\delta_{1i})$, δ_{1i} is the difficulty of item 1 in latent subgroup i, and $\Phi^{XYI}_{xyi} = P(y|x,i)$ is the conditional distribution of observed response $\mathbf{y}$ given latent response $\mathbf{x}$ and latent subgroup i.

Just as in the case of observed subgroups, it is also possible to analyse models in which some items exhibit DIF in

the latent response and other (or the same) items DIF in the guessing probabilities.

Parameter Estimation and Model Testing

Let n_{ixy} be the number of individuals in subgroup i with X=x and Y=y under a certain model and let $m_{ixy} = N\,P(i,x,y)$ be the expected value of n_{ixy}. Although n_{ixy} is not observed, it is possible to estimate the means m_{ixy} of n_{ixy}, and the Φ-parameters from the observed n_{iy} (or n_y if the subgroup is unobserved) by the method of maximum likelihood. To illustrate this, consider the model defined by

$$(15) \qquad m_{ixy} = N\,\Phi_{it}^{IT}\,\Phi_{ix_1}^{IX_1} \ldots \Phi_{ix_k}^{IX_k}\,\Phi_{ix_1y_1}^{IX_1Y_1} \ldots \Phi_{ix_ky_k}^{IX_kY_k} ,$$

The maximum likelihood equations for model (15) would be (Haberman, 1979):

$$\hat{m}_{it}^{IT} = \hat{n}_{it}^{IT} , \qquad \hat{m}_{ix_jy_j}^{IX_jY_j} = \hat{n}_{ix_jy_j}^{IX_jY_j} , \qquad j = 1 , \ldots , k$$

where

$$(16) \qquad \hat{n}_{ixy} = (\hat{m}_{ixy} / \hat{m}_{iy}^{IY})\, n_{iy}^{IY} ,$$

and where n_{it}^{IT} and $n_{ix_jy_j}^{IX_jY_j}$ are the numbers of individuals in

subgroup i with $T=t$, $X_j=x_j$ and $Y_j=y_j$, respectively. Further m_{it}^{IT} and $m_{ix_jy_j}^{IX_jY_j}$ are the expected values of n_{it}^{IT} and $n_{ix_jy_j}^{IX_jY_j}$. If the subgroup i is not observed, then n_{iy} and m_{iy} in (16) has to be replaced by n_y and m_y.

The equations can be solved by the iterative proportional fitting algorithm or the scoring algorithm (Goodman, 1978; Haberman, 1979). The iterative proportional fitting algorithm is to be preferred, since it is less sensitive to the choice of starting values.

In model (15) all items were considered to exhibit DIF in the latent response and DIF in the guessing probabilities. If some items exhibit no DIF in the latent response or DIF in the guessing probabilities, then the Φ-parameters for these items are restricted. For example, if in a certain model item 1 exhibits no DIF in the latent response, then the $\Phi_{ix_1}^{IX_1}$-parameter is restricted in the following manner

$$\Phi_{1x_1}^{IX_1} = \ldots = \Phi_{gx_1}^{IX_1}$$

Similar estimation equations can be formulated for restricted models.

The overall goodness of fit of an incomplete latent-class model can be tested by the Pearson statistic (Q) or the likelihood-ratio statistic (LR) (see Haberman, 1979). Both statistics are asymptotically distributed as chi-square with degrees of freedom equal to the difference between the number of count n_y (or n_{iy} if the subgroup is observed) and the

number of estimable parameters. The number of estimable parameters of a model should be equal to the rank of the information matrix (cf McHugh, 1956; Goodman, 1978).

By the difference in likelihood-ratio test statistics of both models (LR(a;b)) it can be tested whether the alternative model (b) yields a significant improvement in fit over the compact model (a), which is a special case of model (b). Under the assumption of model (a), LR(a;b) is asymptotically chi-square distributed with degrees of freedom equal to the difference in numbers of estimable parameters of both models (Bishop, Fienberg & Holland, 1975).

An Empirical Example

As an example four items from the Second International Mathematics Study in the Netherlands were considered. (Eggen, Pelgrum & Plomp, 1987). Each item was a five-choice item with only one correct alternative.

A sample of 3002 students from two schooltypes of lower secondary education in the Netherlands representing the whole ability range was drawn. To illustrate the use of quasi-loglinear models for detection of DIF, the students level of education was chosen as grouping variable: subgroup MAVO (intermediate general education) and subgroup HAVO/VWO (higher general education and pre-university education).

The models (9) and (11) were fitted to the data using the computer-program LCAG (Hagenaars & Luijkx, 1987). LCAG is

a program for estimating the parameters of loglinear models with latent variables. LCAG yields, besides the estimated latent conditional probabilities (i.e. the guessing probabilities), the estimated expected frequency distribution of the latent variables under the model. From this frequency distribution the difficulty parameters were estimated using LOGIMO (Kelderman & Steen, 1988). LOGIMO is a general computer program especially written to analyse loglinear IRT models.

DIF is tested by comparing model (9) (for DIF in the latent response) and model (11) (for DIF in the guessing probabilities) with model (10) (no DIF). In Table 1 for each item the values of the likelihood ratio test and the degrees of freedom are shown for models (9) and (11). In both cases the level of education was observed.

Insert Table 1 about here

From Table 1 it may be concluded that, except for item 2, the difficulty to solve the problems represented by the items does not vary significantly between the subgroups MAVO and HAVO/VWO. In Table 2 the difficulty parameters of the four items in the model, in which item 2 exhibit DIF in the latent response, are given.

Insert Table 2 about here

It can be seen from Table 2 that the difficulty of item 2 was substantially smaller for MAVO-students then for HAVO/VWO-students.

Table 1 also shows that the attractiveness of the alternatives of the items 1, 2, and 4 were significantly different in both subgroups. To give a more detailed interpretation of the attractiveness of the alternatives, the guessing probabilities of the alternatives for each item are presented in Table 3.

Insert Table 3 about here

For a HAVO/VWO-student the correct alternative of item 1 is more attractive then for a MAVO-student. So (s)he is more inclined to choose the correct alternative. On the other hand a MAVO-student would be more inclined to choose the correct alternative of item 2, because his/her guessing probability of the correct alternative is twice as big as the guessing probability for a HAVO/VWO-student. However, for both subgroups the correct alternative is not the most attractive alternative.

The guessing probabilities for the correct alternative

of item 4 are almost the same for both subgroups, but for the alternatives B and C there is a curious different between the two subgroups. A HAVO/VWO-student would guess alternative B with almost the same probability as a MAVO-student would guess alternative C and guessing alternative C with almost the same probability as a MAVO-student would guess alternative B.

Item 3 exhibits no DIF in the guessing probabilities. However, alternatives B and D of item 3 have a relatively large attractiveness.

In the foregoing the two types of DIF were studied separately from each other. Also only one item at the time was studied. As was indicated earlier, it is also possible to analyse models in which more than one item exhibits DIF. To illustrate this possibility model M, in which the items 1, 2, and 4 exhibits DIF in the guessing probabilities and where item 2 exhibit DIF in the latent response, was considered. Model M gives a considerably improvement in fit to the data over model (10) (LR(10;M) = 100.5; DF = 13). From Table 2 it also follows that model M fits the data better than the models discussed before. The parameters, however, do not differ much from the parameters of the previous models. Therefore they are not given.

In summary, the difficulty of the four items can be ordered in the following way $\delta_3 > \delta_1 > \delta_4 > \delta_2$. That is, item 2 is the easiest item and item 3 is the most difficult one. The attractiveness of alternatives 1, 2, and 4 as well as the difficulty of solving item 2 is not the same for the two

subgroups. Item 3 exhibits no DIF in the latent response or DIF in the guessing probabilities.

Discussion

In the present paper a model for multiple choice items is proposed, which views the observed response of a subject to a certain item as a result of two distinct processes. The first process consists of solving the problem and the second process of giving the actual response. This model is extended with subgroups (observed or latent) in order to study DIF in the two processes. The model was illustrated with an example.

In this paper all tests of DIF are two-sided. This means, that it is not possible to test directional hypothesis about DIF. The estimated difficulty parameters and the estimated guessing probabilities provides only an indication for the direction of DIF.

Because of the fact that LCAG claims much memory-space, it was not possible to consider more than four five-choice items. A line of further research will be to find an estimation method that overcomes this problem. Further research should also give an answer to the question if a certain model is identified or not.

References

Berk, R.A. (1982). Handbook of methods for detecting test bias. Baltimore: The Johns Hopkins University Press.

Binet, A., & Simon, T. (1916). The development of Intelligence in Children. Baltimore: Williams & Wilkins.

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

Bock, R.D. (1972). Estimating item parameters and latent proficiency when the responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Camilli, G. (1979). A critique of the chi-square method for assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder.

Clogg, C.C. (1981). Latent structure models of mobility. American Journal of Sociology, 86, 836-868.

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

Durovic, J. (1975). Definitions of test bias: A taxonomy and an illustration of an alternative model. Unpublished doctoral dissertation, State University of New York at Albany.

Eggen, T.J.H.M., Pelgrum, W.J., & Plomp, Tj. (1987). The implemented and attained mathematics curriculum: Some results of the second international mathematics study in the Netherlands. Studies in Educational Evaluation, 13, 119-135.

Fischer, G.H., & Formann, A.F. (1982). Some applications of logistic latent trait models with linear constraints on parameters. Applied Psychological Measurement, 6, 397-416.

Formann, A.K. (1985). Constrained latent class models: Theory and applications. British Journal of Mathematical and Statistical Psychology, 38, 87-111.

Goodman, L.A. (1978). Analyzing qualitative/categorical data: Loglinear models and latent structure analysis. London: Addison Wesley.

Green, B.F., Crone, C.R., & Folk, V.G. (1989). A Method for studying differential distractor functioning. Journal of Educational Measurement. 26, p. 147-160.

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

Hagenaars, J., & Luijkx, R. (1987). LCAG: latent-class models and other loglinear models with latent variables. Tilburg: Tilburg University. Working paper 17.

Holland, P.W., & Thayer, D. (1986). Differential item performance and the Mantel-Haenszel statistic. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Ironson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R.A. Berk: Handbook of methods for detecting item bias. Baltimore: The Johns Hopkins University Press.

Jensen, A.R. (1980). Bias in mental testing. London: Methuen.

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

Kelderman, H. (1988). An IRT model for Item responses that are subject to Omission and/or Intrusion Errors. Enschede: University of Twente, Research Report 88-16.

Kelderman, H. (1989). Item bias detection using the loglinear Rasch model: observed and unobserved subgroups. Psychometrika, 54, 18-.

Kelderman, H., & Macready, G.B. (1988). Loglinear latent class models for detecting item bias. Paper presented at the Annual Meeting. Meeting of the American Educational Research Association, April 5-9.

Kelderman, H., & Steen, R. (1988). LOGIMO 1. Loglinear item response theory modeling. Computer manual, University of Twente, Department of Educational Technology.

Kok, F.G., Mellenbergh, G.J , & van der Flier, H. (1985). An iterative procedure for detecting biased items. Journal of Educational Measurement, 22, 295-303.

Lazarsfeld, P.F. (1950). The interpretation and computation of some latent structures. In S. A. Stouffer et al. (Eds.). Measurement and prediction in World War II. Vol. 4. Princeton: Princeton University Press.

Lazarsfeld, P.F., & Henry, N.W. (1968). Latent structure analysis, Boston: Houghton: Miffin.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, New Yersey: Lawrence Erlbaum.

McHugh, R.B. (1956). Efficient estimation and local identification in latent-class analysis. Psychometrika, 21, 331-347.

Mellenbergh, G.J. (1982). Contingency table methods for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Mislevy, R.J. (1981). A general linear model for the analysis of Rasch item threshold estimates. Unpublished doctoral dissertation University of Chicago, 1981.

Mislevy, R.J., & Verhelst N. (1987). Modeling item responses when different subjects employ different solution strategies. Research Bulletin. RR-87-47-ONR, Princeton NJ: Educational Testing Service.

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. Journal of Educational Statistics, 10, 133-142.

Nungester, R.J. (1977). An empirical examination of three models of item bias (Doctoral dissertation Florida State University, 1977). Dissertation Abstracts International, 38, 2726 A (University Microfilms No. 77-24, 289).

Osterlind, S.J. (1983). Test item bias. Beverly Hills: Sage.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press.

Scheuneman, J. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparisons of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-377.

Thissen, D., Steinberg, L., & Fitzpatrick, A.R. (1989a). Multiple choice models: The distractors are also part of the item. Journal of Educational Measurement, 26, 161-176.

Thissen, D., Steinberg, L., & Wainer, H. (1989b). Detection of Differential Item Functioning using the parameters of item response models. Kansas: University of Kansas.

Wright, B.D., Mead, R.J., & Draba, R. (1975). Detecting and correcting test item bias with a logistic response model (RM 22). Statistical Laboratory, Department of Education, University of Chicago.

Yamamoto, K. (1987). A hybrid model for item responses. Doctoral dissertation, University of Illinois.

Yamamoto, K. (1988). Hybrid model of IRT and latent class models. Princeton, NJ, Educational Testing Service.

Table 1

Likelihood ratio Tests for detecting DIF on the data of the Second International Mathematics Study

Item(s)	LR(10;9)	DF	LR(10;11)	DF
1	1.701	1	26.519*	4
2	4.720*	1	21.340*	4
3	1.747	1	6.033	4
4	.018	1	52.595*	4

Note. Tests marked with an asterisk are significant. (α = .05)

Table 2

Difficulty parameters of the items in the model of DIF in the latent response in item 2

	Item i			
Subgroup	1	2	3	4
HAVO/VWO	1.52	~ .82	3.54	-1.32
MAVO	1.52	-1.90	3.54	-1.32

Note. The difficulty parameters of items 1, 3 and 4 for MAVO are set equal to the difficulty parameters for HAVO/VWO.

Table 3

Guessing probabilities of the alternatives of item i

	Alternatives				
Item	A	B	C	D	E
	Subgroup HAVO/VWO				
1	.073	.033	.685	.174	.035
2	.743	.123	.061	.045	.028
3	.112	.327	.139	.323	.099
4	.110	.355	.235	.092	.208
	Subgroup MAVO				
1	.211	.024	.563	.193	.009
2	.662	.240	.068	.015	.015
3	.112	.327	.139	.323	.099
4	.068	.241	.341	.084	.266

Note 1. The correct alternatives are underlined.

Note 2. Item 3 was not significantly biased in the guessing probabilities.

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

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

RR-89-6 J.J. Adema, *Implementations of the Branch-and-Bound method for test construction problems*

RR-89-5 H.J. Vos, *A simultaneous approach to optimizing treatment assignments with mastery scores*

RR-89-4 M.P.F. Berger, *On the efficiency of IRT models when applied to different sampling designs*

RR-89-3 D.L. Knol, *Stepwise item selection procedures for Rasch scales using quasi-loglinear models*

RR-89-2 E. Boekkooi-Timminga, *The construction of parallel tests from IRT-based item banks*

RR-89-1 R.J.H. Engelen & R.J. Jannarone, *A connection between item/subtest regression and the Rasch model*

RR-88-18 H.J. Vos, *Applications of decision theory to computer based adaptive instructional systems*

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

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

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

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

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

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

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

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

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

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

RR-88-6 H.J. Vos, *The use of decision theory in the Minnesota Adaptive Instructional System*

RR-88-5 W.J. van der Linden, *Optimizing incomplete sample designs for item response model parameters*

RR-88-4 J.J. Adema, *A note on solving large-scale zero-one programming problems*

RR-88-3 E. Boekkooi-Timminga, *A cluster-based method for test construction*

RR-88-2 W.J. van der Linden & J.J. Adema, *Algorithmic test design using classical item parameters*

RR-88-1 E. van der Burg & J. de Leeuw, *Nonlinear redundancy analysis*

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

A publication by
the Department of Education
of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands