Statistica Sinica **8**(1998), 647-667

DETECTION OF DIFFERENTIAL ITEM FUNCTIONING USING LAGRANGE MULTIPLIER TESTS

C. A. W. Glas

University of Twente, Enschede, the Netherlands

Abstract: In the present paper it is shown that differential item functioning can be evaluated using the Lagrange multiplier test or Rao's efficient score test. The test is presented in the framework of a number of IRT models such as the Rasch model, the OPLM, the 2-parameter logistic model, the generalized partial credit model and the nominal response model. However, the paradigm for detection of differential item functioning presented here also applies to other IRT models. Two examples are given, one using simulated data and one using real data.

Key words and phrases: DIF, generalized partial credit model, item response theory, Lagrange multiplier test, model fit, nominal response model, OPLM, Rao's efficient score test, Rasch model, two-parameter logistic model.

1. Introduction

When a new test is constructed, it is important to find empirical evidence that contributes to the construct validity of the test (AERA, APA and NCME (1985)). Part of this process may be to show that the test fits a unidimensional item response theory (IRT) model, which means that the observed responses can be attributed to item and person parameters that are related to some unidimensional latent dimension. Construct validity is supported if the construct to be measured is also unidimensional and if the ordering of item difficulties imposed by the construct is reflected in the ordering of item parameters on the latent scale. Further, if it can be shown that the latent ability is unidimensional, a meaningful unidimensional variable for measuring the underlying construct can be created, and the respondent can be assigned a value on some scale. So the IRT model validates the scoring rule of the test. Construct validity implies that the construct to be measured is the same for all respondents of the population the test is aimed at. This is where the problem of differential item functioning (DIF) or item bias arises. For reasons of semantic clarity, many authors prefer the terminology "DIF" to the older term "item bias" (see, for instance, Angoff (1993) or Cole (1993)). In the present paper this practice is complied with. Studies of DIF deal with the question how item scores are affected by external variables that do not belong to the construct to be measured. Usually, the external variable imposes a division into a small number of sub-populations, where a sub-population refers to a set of persons that have the same value on the external variable. If the external variable is dichotomous, one usually speaks of the reference population, say the majority group or an advantaged group, and the focal population, say the minority or a disadvantaged group. In DIF studies, the null-hypothesis is that the external variable does not moderate the effect of ability on the item scores. So the responses to a dichotomous item are subject to DIF if, conditional on ability level, the probability of a correct response differs over the samples from the various sub-populations (Mellenbergh (1982, 1983)). The generalization to polytomous items is straightforward. The responses to a polytomous item are subject to DIF if the set of probabilities of scoring in the various response categories of the item, conditional on ability, differs between different sub-populations. Another, equivalent definition of DIF is that the expected scores on the item, conditional on ability, are different for the sub-populations under consideration (Chang and Mazzeo (1994)).

The essential problem in DIF studies is whether the response behavior of the samples of all sub-populations can be properly described by an IRT model. An additional problem is that the possible presence of DIF will influence the parameter estimates of all items, and this may confound model fitting. In the example section of this paper it will be shown that detection of DIF can be accomplished by an iterative process of model fitting, testing for DIF and modeling the responses to affected items, until a fitting model for all items and all samples of respondents is found.

Several techniques for detecting DIF have been proposed. Most of them are based on evaluating differences in response probabilities between groups, conditional on some measure of ability. The most generally used technique is based on the Mantel-Haenszel statistic (Holland and Thayer (1988)), others are based on log-linear models (Kok, Mellenbergh and van der Flier (1985)), on IRT models (Hambleton and Rogers (1989)), or on log-linear IRT models (Kelderman (1989)). In the Mantel-Haenszel, log-linear and log-linear IRT approaches, the difficulty level of the item is evaluated conditionally on the respondents' unweighted sum scores. However, adopting the assumption that the unweighted sum score is a sufficient statistic for ability (together with some technical assumptions, which will seldom be inappropriate) necessarily leads to the adoption of the Rasch model (Fischer (1974, 1993, 1995)). However, with the exception of the log-linear IRT approach, the validity of the Rasch model is rarely explicitly tested. Therefore, Glas and Verhelst (1995) suggested a procedure consisting of two steps:

(1) searching for an IRT model for fitting the data of the sample from the reference population, and, as far as possible, the sample from the focal population;

(2) evaluating the differences in response probabilities between the two samples in homogeneous ability groups.

In this paper, an approach is investigated that can be viewed as a generalization of the above method. In the first step, Glas and Verhelst (1995) use a generalized version of the Rasch model where discrimination indices are imputed for dealing with differences in discrimination between the items. This model, known as the one parameter logistic model (OPLM), will be considered below. These authors propose an iterative process of adjusting the discrimination indices using so-called generalized Pearson statistics, until an acceptable model fit is achieved. Evaluating the differences in response probabilities between the samples from the reference and focal population in homogeneous ability groups is also done using generalized Pearson statistics. The alternative approach of the present paper is not only applicable in the framework of the Rasch model and the OPLM, it can also be used in the context of the two-parameter logistic model and the nominal response model. These last two models are more flexible than the former models, but the tests for evaluating the fit to these models developed thus far are less sophisticated, in the sense that the asymptotic distribution of the statistics for these tests is unknown (see, for instance, Mislevy and Bock (1990)). On the other hand, the generalized Pearson tests for the Rasch model and the OPLM completely rely on the existence of sufficient statistics (see Glas and Verhelst (1995)), so these tests cannot be used for performing the second step of the above approach for the two-parameter logistic and nominal response model. Therefore, in the present paper it will be shown that the second step can be performed using Lagrange multiplier (LM) tests.

The remainder of this paper is organized as follows: (1) the relevant IRT models will be discussed, (2) an estimation procedure will be described, (3) the LM tests will be presented, and (4) two examples will be given, one using simulated data and one using real data.

2. Choosing an IRT Model

In IRT models, the influence of items and persons on the observed responses are modelled by different sets of parameters. Since DIF is defined as the occurrence of differences in expected scores conditional on ability, IRT modelling seems especially fit for dealing with this problem. However, first the question must be answered which IRT models are appropriate in this context. Before considering some significant models for studying DIF, the following definitions must be introduced. Consider items where the possible responses can be coded by the integers $0, 1, 2, 3, \ldots, m_i$. Let item i have $m_i + 1$ response categories, indexed $h = 0, 1, \ldots, m_i$. Notice that dichotomous items are the special case where $m_i = 1$. The response to an item will be represented by a vector

$(x_{il}, \ldots, x_{ih}, \ldots, x_{im_i})$, where x_{ih} is a realization of the random variable X_{ih} defined by

$$x_{ih} = \begin{cases} 1 & \text{if a response is given in categroy } h, \\ 0 & \text{if this is not the case.} \end{cases} \tag{1}$$

In this section, two classes of models will be considered. The first class comprises of exponential family IRT models, such as the unidimensional Rasch model (UPRM) by Rasch (1960, 1961), the partial credit model (PCM) by Masters (1982), the one-parameter logistic model (OPLM) by Verhelst and Glas (1995) and the generalized PCM (GPCM) by Wilson and Masters (1993). The second class comprises of generalizations of the first class of models outside the exponential family, such as the two-parameter logistic model (2-PL) by Birnbaum (1968) and the nominal response model by Bock (1972). The motivation for making this distinction is that there are many statistical testing procedures based on statistics with known (asymptotical) distributions for the first class of models and hardly any such procedures for the latter class of models. This point will be returned to below.

In the framework of polytomous items, Rasch (1960, 1961) (see also, Andersen (1972, 1973b, 1977) and Fischer (1974)) has introduced several exponential family IRT models. In the model most suited for ability measurement, the UPRM, the probability of scoring in category h of item i is given by

$$\Pr(X_{ih} = 1 | \theta_n, \beta_i) = \frac{\exp(h\theta_n - \beta_{ih})}{1 + \sum_{k=1}^{m_i} \exp(k\theta_n - \beta_{ik})}, \tag{2}$$

where θ_n is the unidimensional ability parameter of person n, and β_i is a vector with elements $\beta_{ih}, h = 1, \ldots, m_i$ which are the parameters of item i. For $m_i = 1$, equation (2) defines the item response function of the well-known Rasch model for dichotomous items. One of the reasons for considering this model is that it can be derived from a set of assumptions which will often apply in the context of ability measurement. Andersen (1977) has shown that the UPRM can be derived from the assumption that $R_n = \sum_{i,h} hX_{ih}$ is a minimal sufficient statistic for a unidimensional ability parameter θ, local stochastic independence and some technical assumptions. Masters (1982) develops a completely equivalent model from an entirely different perspective. Masters' version, the PCM, can be derived from the assumption that every category $h, h > 0$, can be seen as a step that is either passed or failed. The final score on the item is determined by the number of steps that the respondent has successfully taken. Further, it is assumed that the probability of scoring in category h, rather than in category $h-1$, is described by a Rasch model for a dichotomous item with item parameter η_{ih} . Glas and Verhelst (1989) have pointed out that the PCM is a reparametrization of the UPRM, that is, the parameters of the UPRM are obtained by the reparametrization $\beta_{ih} = \sum_{g=1}^{h} \eta_{ig}, h = 1, \ldots, m_i$.

One of the attractive features of the UPRM is the possibility of using a conditional maximum likelihood method (CML) for obtaining consistent estimates of the item parameters (see Fischer (1974), Molenaar (1995)). By conditioning on the minimal sufficient statistics R_n a likelihood function is obtained that does not depend on the person parameters. This has the important advantage that computation of CML estimates does not need any assumption concerning the distribution of ability in the population. Further, these consistent estimates can, in principle, be obtained using any arbitrary sample of persons where the model holds. The less attractive feature of the model is that the possible form of the item response curve is rather restricted, for instance, for the dichotomous case the item response curves must be parallel, in the sense that they are shifted along the latent continuum. Fortunately, many statistical tools are available for evaluating the fit of the Rasch model. The assumption that the unweighted sum score is a minimal sufficient statistic for the person parameter and the assumption concerning the form of the item response curves are the focus of Martin Löf's (1973) T-test, the R_1-test (Glas (1988), Glas and Verhelst (1989)), the U_i-test (Molenaar (1983)) and the S_i- and M-tests (Verhelst and Glas (1995), Verhelst, Glas and Verstralen (1993)). The property that the item parameters can be consistently estimated on every subgroup of the population is tested by Andersen's likelihood ratio test (Andersen (1973a)) and the Fischer-Scheiblechner test (Fischer (1974)). Finally, the assumption of unidimensionality and local stochastic independence are the focus of the likelihood ratio test of Martin Löf (1973, 1974) and the R_2-test of Glas (1988).

The combination of the axiomatic foundation of the model and the tradition in social research and educational measurement of working with unweighted sum scores makes the model an attractive starting point for statistical analyses. However, the restrictive character of the model will often obstruct model fit. There are several aspects of the Rasch model that may lead to rejection of the model. These violations can be accounted for by defining specific generalizations of the Rasch model. In this paper, the focus will be on models where the assumption of the form of the item response curves is relaxed. This can be done by introducing discrimination indices or discrimination parameters $\alpha_{ih}, h = 1, \ldots, m_i$, so that equation (2) generalizes to

$$\psi_{ih}(\theta_n) = \Pr(X_{ih} = 1|\theta_n, \alpha_i, \beta_i) = \frac{\exp(\alpha_{ih}\theta_n - \beta_{ih})}{1 + \sum_{k=1}^{m_i} \exp(\alpha_{ik}\theta_n - \beta_{ik})}, \tag{3}$$

were α_i and β_i are vectors of the elements α_{ih} and $\beta_{ih}(h = 1, \ldots, m_i)$, respectively. If the discrimination indices are viewed as known constants, this model can be derived from the assumption that $R_n = \sum_{i=1}^{k} \alpha_{ih} X_{nih}$ is a sufficient statistic for ability, local independence, and some technical assumptions (Andersen (1977)). In the framework of known discrimination indices, Verhelst and Glas (1995) have developed a CML estimation procedure and a procedure for evaluating model fit, for the so-called OPLM, where the item categories are assumed to have score weights $\alpha_{ih} = h\alpha_i$. Recently, Glas (1997) has generalized this procedure to the more general GPCM by Wilson and Masters (1993), where item categories are given scoring weights α_{ih}.

The discrimination indices can also be treated as unknown item parameters to be estimated. In the framework of dichotomous items this approach is known as the two-parameter logistic model (2-PL) by Birnbaum (1968). The nominal response model by Bock (1972) can be viewed as a generalization of the 2-PL to polytomous items. There are several considerations with respect to the choice between the two approaches. The OPLM and GPCM allow for CML estimation and have theoretically well-founded tools for testing model fit. In fact, most of the procedures mentioned above can easily be generalized to model (3) (Verhelst and Glas (1995), Glas (1997)). On the other hand, the nominal response model is more flexible with respect to possible item response curves. This flexibility is bought at the expense of needing an MML estimation procedure for obtaining consistent estimates of the item parameters. This introduces assumptions with respect to the distribution of ability, which, of course, introduce another source of possible model violations that needs to be accounted for. However, attempting to generalize the complete catalogue of tests of model fit for exponential family IRT to non-exponential family IRT is far beyond the scope of the present paper; here only an alternative for the DIF tests of exponential family IRT will be studied.

3. Estimation

In the present section, the well-known theory of MML estimation for IRT models will be re-iterated. In this presentation the formalism of Glas (1992) will be used, which, as will become apparent in the sequel, is especially suited for introducing LM tests for DIF. Consider the case of two sub-populations. A background variable will be defined by

$$y_n = \begin{cases} 1 & \text{if person } n \text{ belongs to the focal prpulation,} \\ 0 & \text{if person } n \text{ belongs to the reference population.} \end{cases} \tag{4}$$

The absence of DIF entails that respondents of equal ability of different sub-populations have the same expected item scores. This, of course, does not mean that the expected item scores in the different sub-populations are the same, because it may well be the case that the ability distributions of the sub-populations are different. Let $g(\theta_n; \lambda_{y(n)})$ be the density of the ability distribution of subpopulation y, with parameters $\lambda_{y_{(n)}}$, where $y(n) = y_n$ is the index of the subpopulation of person n. Further, if $\xi' = (\alpha', \beta', \lambda')$ is the vector of all item and population parameters, the log-likelihood can be written as

$$\ln L(\xi; X) = \sum_n \ln \Pr(x_n; \xi), \tag{5}$$

where x_n stands for the response pattern of person n and X stands for all data.

To derive the MML estimation equations, it proves convenient to introduce the vector of derivatives

$$b_n(\xi) = \frac{\partial}{\partial \xi} \ln \Pr(x_n, \theta_n; \xi) = \frac{\partial}{\partial \xi}[\ln \Pr(\boldsymbol{x}_n | \theta_n, \alpha, \beta) + \ln g(\theta_n | \lambda_{y(n)})]. \tag{6}$$

Glas (1992) adopts an identity due to Louis (1982) to write the first order derivatives of (5) with respect to ξ as

$$\frac{\partial}{\partial \xi} \ln L(\xi; \mathbf{X}) = \sum_n E(b_n(\xi) | \boldsymbol{x}_n, \xi). \tag{7}$$

This identity greatly simplifies the derivation of the likelihood equations. For instance, using the short-hand notation $\psi_{nih} = \psi_{ih}(\theta_n)$, it can be easily verified that

$$b_n(\alpha_{ih}) = \theta_n(x_{nih} - \psi_{nih}) \tag{8}$$

and

$$b_n(\beta_{ih}) = \psi_{nih} - x_{nih}, \tag{9}$$

so the likelihood equations are given by

$$\sum_n E(\theta_n x_{nih} | x_n, \xi) = \sum_n E(\theta_n \psi_{nih} | \boldsymbol{x}_n, \xi) \tag{10}$$

and

$$\sum_n x_{nih} = \sum_n E(\psi_{nih} | x_n, \xi). \tag{11}$$

The choice of a distribution of ability is not essential to the theory presented here; the test for DIF will both apply to the parametric MML framework (see Bock and Aitkin (1981)) and in the non-parametric MML framework (see De Leeuw and Verhelst (1986), Follmann (1988)). As an example of the parametric context, one might assume that the ability distribution is normal with parameters μ_y and σ_y. Then

$$b_n(\mu_{y(n)}) = (\theta_n - \mu_{y(n)})\sigma_{y(n)}^{-2} \tag{12}$$

and

$$b_n(\sigma_{y_{(n)}}) = -\sigma_{y(n)}^{-1} + (\theta_n - \mu_{y(n)})^2 \sigma_{y(n)}^{-3}, \tag{13}$$

so the likelihood equations are

$$\mu_y = \frac{1}{N_y} \sum_{n|y} E(\theta_n | x_n, \xi) \tag{14}$$

and

$$\sigma_y^2 = \frac{1}{N_y} \sum_{n|y} E(\theta_n^2 | x_n, \xi) - \mu_y^2, \tag{15}$$

where the right-hand summations are over the respondents in the sample from sub-population y, and N_y is the number of respondents in this sample. Below, this framework will be used for introducing an LM test for DIF. But first the principle of LM tests will be described.

4. Lagrange Multiplier Tests

Applications of LM tests to the framework of IRT have been described by Glas and Verhelst (1995). The principle of the LM test (Aitchison and Silvey (1958)), and the equivalent efficient-score test (Rao (1948)) can be summarized as follows. The arrangement of the LM test is the same as the arrangement of the likelihood-ratio test and the Wald test; all these three tests are used for testing a special model against a more general alternative. Consider a null-hypothesis about a model with parameters ϕ_0. This model is a special case of a general model with parameters ϕ. In the present case the special model is derived from the general model by fixing one or more parameters to known constants. To avoid problems beyond the scope of this paper, it will be assumed that these parameters are not fixed at points on a boundary of the parameter space. Let ϕ_0 be partitioned as $\phi_0' = (\phi_{01}', \phi_{02}') = (\phi_{01}', c)$, where c is a vector of postulated constants. Let $h(\phi)$ be the partial derivatives of the log-likelihood of the general model, so $h(\phi) = (\partial/\partial\phi) \ln L(\phi)$. This vector of partial derivatives gauges the change of the log-likelihood as a function of local changes in ϕ. Let $H(\phi, \phi)$ be defined as $-(\partial^2/\partial\phi\partial\phi') \ln L(\phi)$. Then the LM statistic is given by

$$LM = h(\phi_0)' \; H(\phi_0, \phi_0)^{-1} h(\phi_0). \tag{16}$$

If (16) is evaluated using the ML estimate of ϕ_{01} and the postulated values of c, it has an asymptotic chi-square distribution with degrees of freedom equal to the number of parameters fixed.

An important computational aspect of the procedure is that at the point of the ML estimates $\hat{\phi}_{01}$ the free parameters have a partial derivative equal to zero. Therefore, (16) can be computed as

$$LM(c) = h(c)' W^{-1} h(c) \tag{17}$$

with

$$W = H(c, c) - H(c, \hat{\phi}_{01})\ H(\hat{\phi}_{01}, \hat{\phi}_{01})^{-1}\ H(\hat{\phi}_{01}, c). \tag{18}$$

Note that $H(\hat{\phi}_{01}, \hat{\phi}_{01})$ also plays a role in the Newton-Raphson procedure for solving the estimation equations and in computation of the observed information matrix, so its inverse will generally by available at the end of the estimation procedure anyway. Further, if the validity of the model of the null-hypothesis is tested against various alternative models, the computational task is relieved because the inverse of $H(\hat{\phi}_{01}, \hat{\phi}_{01})$ is already available and the order of W is equal to the number of parameters fixed, which must be small to keep the interpretation of the outcome tractable.

The interpretation of the outcome of the test is supported by observing that the value of (17) depends on the magnitude of $h(c)$, that is, on the first order derivatives with respect to the parameters ϕ_{02} evaluated in c. If the absolute values of these derivatives are large, the fixed parameters are bound to change once they are set free, and the test is significant, that is, the special model is rejected. If the absolute values of these derivatives are small, the fixed parameters will probably show little change should they be set free, that is, the values at which these parameters are fixed in the special model are adequate and the test is not significant, that is, the special model is not rejected.

The rationale for using LM tests rather than likelihood ratio tests and Wald tests is based on the fact that LM tests only need ML estimates of the parameters of the special model. In many instances, the parameters of the general model will be quite complicated to estimate. But even if this is not the case, this procedure still has the advantage that many alternatives can be considered without needing repeated estimation of all these alternatives. In the sequel it will be shown that the hypothesis of DIF can be tested for one item at a time. If this was done using a Wald or likelihood ratio test, this would require computing new estimates for every tested item. Further, DIF is just one of the many possible violations that may be of interest. Scanning the whole spectrum of violations of a non-exponential family IRT model without repeated estimation presents a promising direction for further research, but this is beyond the scope of the present paper.

5. Lagrange Multiplier Tests for DIF

In section 3, the case of two sub-populations labeled $y = 0$ and $y = 1$, was considered. As a generalization of the model defined by (3) consider

$$\Pr(X_{ih} = 1 | y_n, \theta_n, \alpha_i, \beta_i, \gamma_i, \delta_i) = \frac{\exp(\alpha_{ih}\theta_n - \beta_{ih} + y_n(\gamma_{ih}\theta_n - \delta_{ih}))}{1 + \sum_{k=1}^{m_i} \exp(\alpha_{ik}\theta_n - \beta_{ik} + y_n(\gamma_{ik}\theta_n - \delta_{ik}))}. \tag{19}$$

This model implies that the responses of the reference population are properly described by (3), but that the responses of the focus population need additional location parameters δ_{ih}, additional discrimination parameters γ_{ih}, or both. In the dichotomous case, the first instance covers so-called uniform DIF, that is, a shift of the item response curve for the focal population, while the latter two cases are often labelled non-uniform DIF, that is, the item response curve for the focal population is not only shifted, but it also intersects the item response curve of the reference population (Mellenbergh (1982, 1983)). Application of the LM test boils down to postulating a special model where γ_{ih} and δ_{ih} are equal to zero and testing against the alternative that either $\gamma_{ih},\ h=1,\ldots,m_i, \delta_{ih},\ h=1,\ldots,m_i$, or both sets of parameters are non-zero.

The rest of this section will be devoted to the derivation and the interpretation of the expressions for the LM statistic. As with the derivation of the estimation equations, also for the derivation of the matrix of second order derivatives the theory by Louis (1982) can be used. Using Glas (1992), it follows that the matrix of second order derivatives for the special model,

$$H(\xi,\xi)=\frac{\partial^2 \ln L(\xi;X)}{\partial\xi\partial\xi'} \tag{20}$$

evaluated using MML estimates, is given by

$$H(\xi,\xi)=\sum_n[E(B_n(\xi,\xi)|x_n,\xi)+E(b_n(\xi)\boldsymbol{b}_n(\xi)|x_n,\xi)], \tag{21}$$

where

$$B_n(\xi,\xi)=\frac{\partial^2 \ln \Pr(x_n,\theta_n;\xi)}{\partial\xi\partial\xi'}. \tag{22}$$

Note that the expressions for the second of the two right-hand terms of (21) can be directly derived from (8), (9), (12) and (13). The expressions for evaluating $\boldsymbol{B}_n(\xi,\xi)$ for some item i are listed in Table 1. From (8) and (9) it is easily seen that the expressions for $B_n(\xi,\xi)$ involving two different items i and j are all equal to zero.

Table 1. Expressions for $B_n(\xi_1,\xi_2)$ for the paramters of item i

ξ_1,ξ_2	α_{ih}	α_{ig}	β_{ih}	β_{ig}
α_{ih}	$-\theta_n^2\psi_{nih}(1-\psi_{nih})$	$\theta_n^2\psi_{nih}\psi_{nig}$	$-\theta_n\psi_{nih}(1-\psi_{nih})$	$\theta_n\psi_{nih}\psi_{nig}$
α_{ig}	$\theta_n^2\psi_{nig}\psi_{nih}$	$-\theta_n^2\psi_{nig}(1-\psi_{nig})$	$\theta_n\psi_{nig}\psi_{nih}$	$-\theta_n\psi_{nig}(1-\psi_{nig})$
β_{ih}	$-\theta_n\psi_{nih}(1-\psi_{nih})$	$\theta_n\psi_{nih}\psi_{nig}$	$-\psi_{nih}(1-\psi_{nih})$	$\psi_{nih}\psi_{nig}$
β_{ig}	$\theta_n\psi_{nig}\psi_{nih}$	$-\theta_n\psi_{nig}(1-\psi_{nig})$	$\psi_{nig}\psi_{nih}$	$-\psi_{nig}(1-\psi_{nig})$

Inserting these structural zero's and the expressions of Table 1 into (21) gives the expression for $H(\xi,\xi)$ as far as the free item parameters are concerned. Further, from (6) it follows that for any population parameter λ_y,

$y=0,1, B_n(\alpha_{ih,\lambda_y})=B_n(\beta_{ih},\lambda_y)=0$. Continuing the example of a normal ability distribution with parameters μ_y and σ_y, it follows that $B_n(\mu_y,\mu_y)=-\sigma_y^{-2}$, $B_n(\sigma_y,\sigma_y)=\sigma_y^{-2}-3(\theta_n-\mu_y)^2\sigma_y^{-4}$, and $B_n(\mu_y,\sigma_y)=-2(\theta_n-\mu_y)\sigma_y^{-3}$. This concludes the derivation of the expressions for $H(\xi,\xi)$ for the free parameters in ξ.

The fixed parameters emerge from a general model, where it is assumed that for the focal population additional location δ_{ih} and discrimination parameters γ_{ih} have to be postulated. Under the null-hypothesis, these additional parameters are fixed at zero. For these fixed parameters, it can easily be shown that

$$b_n(\gamma_{ih}) = y_n\theta_n(x_{nih} - \psi_{nih}) \tag{23}$$

and

$$b_n(\delta_{ih}) = y_n(\psi_{nih} - x_{nih}), \tag{24}$$

so the entries of the vector $h(c)$ of the general LM statistic (17) are given by

$$h(\gamma_{ih}) = \sum_n y_n x_{nih} E(\theta_n|\boldsymbol{x}_n,\xi) - \sum_n y_n E(\theta_n\psi_{nih}|x_n,\xi) \tag{25}$$

and

$$h(\delta_{ih}) = \sum_n y_n E(\psi_{nih}|x_n,\xi) - \sum_n y_n x_{nih}. \tag{26}$$

Note that the right-hand side of (26) is the difference between the expected and observed number of persons in the focal group scoring in category h of item i. So for dichotomous items the right-hand side of (26) is the difference between the observed number correct in the focal group and its expectation, computed using parameter estimates obtained in both groups simultaneously. Since a test based on (26) is aimed at the hypothesis that there is no specific additional difficulty δ_{ih} present, it should be sensitive to uniform DIF, that is, a shift of the item response curve for the focal population. As a result of this shift, the observed number correct score for item i in the focal group will not be properly predicted if item parameter estimates obtained using both groups simultaneously will be used. This inconsistency between the observed and the predicted number correct score for item i in the focal group is exactly what is reflected in the difference at the right hand side of (26). If this difference is too large, the entry $h(\delta_{ih})$ of $h(c)$ will be large and the test will be significant. Also (25) is a difference between the expected and observed number of persons in the focal group scoring in category h of item i, but here the individual observations and expectations are weighted with θ. Therefore differences in the extremes of the ability range carry more weight than differences in the middle of the ability range. This is in accordance with the fact that the differences on the right-hand side of (25) arise when a test is derived for the hypothesis that the slope of the regression of the responses on θ is the same for all groups.

For computation of the LM statistic, the matrix of second order derivatives with respect to the fixed and free parameters must be evaluated. Using equation (19) the reader can easily verify that for the fixed parameters $B_n(\gamma_{ih}, \gamma_{ig}) = y_n B_n(\alpha_{ih}, \alpha_{ig})$, $B_n(\gamma_{ih}, \delta_{ig}) = y_n B_n(\alpha_{ih}, \beta_{ig})$ and $B_n(\delta_{ih}, \delta_{ig}) = y_n B_n(\beta_{ih}, \beta_{ig})$. In the same manner, it can also be derived that the second order derivatives with respect to fixed and free parameters are equal to $B_n(\gamma_{ih}, \alpha_{ig}) = y_n B_n(\alpha_{ih}, \alpha_{ig})$, $B_n(\gamma_{ih}, \beta_{ig}) = y_n B_n(\alpha_{ih}, \beta_{ig})$, $B_n(\delta_{ih}, \alpha_{ig}) = y_n B_n(\beta_{ih}, \alpha_{ig})$, and $B_n(\delta_{ih}, \beta_{ig}) = y_n B_n(\beta_{ih}, \beta_{ig})$. Again, inserting these expressions into (23) gives the desired expressions for the elements of $H(\xi, \xi)$.

6. Some Examples

In this section, various examples of LM tests for DIF will be presented. These examples must be viewed as an illustration of the technique, not as an exhaustive power study. The first example concerns data simulated with the Rasch model for dichotomous items. The second example concerns a data set that was recently analyzed using the OPLM in combination with CML estimates and generalized Pearson tests (Glas and Verhelst (1995)). This example will be re-analyzed here using MML estimates and LM tests, both for the OPLM and the nominal response model.

Table 2. A simulated example for the Rasch model for dichotomous items 100 replications

True Parameters					Estimated Parameters and LM Tests										
i	α_i	γ_i	β_i	δ_i	$\hat{\beta}_i$	$Se(\hat{\beta}_i)$	$LM(\gamma_i)$	Pr	Nr	$LM(\delta_i)$	Pr	Nr	$LM(\gamma_i, \delta_i)$	Pr	Nr
1	1.0		−1.0		−.95	.066	2.79	.32	20	2.61	.30	27	4.85	.27	27
2	1.0		−0.5		−.45	.063	1.94	.37	17	2.06	.38	19	3.70	.33	20
3	1.0		0.0		.05	.062	1.36	.45	9	1.95	.34	20	3.32	.36	21
4	1.0		0.5		.56	.062	1.59	.42	12	1.62	.38	11	3.82	.33	22
5	1.0		1.0		1.07	.064	1.38	.45	10	1.51	.42	12	3.87	.34	20
6	1.0		−1.0		−.96	.066	2.29	.36	21	1.66	.39	14	4.02	.31	25
7	1.0	0.0	0.0	0.5	.31	.062	3.05	.31	26	20.21	.00	100	21.83	.00	99
8	1.0	0.5	0.0	0.0	.04	.062	9.05	.02	87	2.86	.27	31	11.90	.02	91
9	1.0	0.5	0.0	0.5	.27	.062	7.58	.03	76	12.41	.00	97	23.31	.00	100
10	1.0		1.0		1.07	.064	1.39	.42	7	1.87	.36	16	4.19	.29	28
y	μ_y				$\hat{\mu}_y$	$Se(\hat{\mu}_y)$									
0	0.5				.60	.056									
1	0.0				.00	.000									
y	σ_y				$\hat{\sigma}_y$	$Se(\hat{\sigma}_y)$									
0	1.0				.99	.039									
1	1.0				1.06	.040									

To illustrate the possibilities of the technique, a number of simulation studies were carried out using data simulated for a test of 10 dichotomous Rasch items. The data for each replication consisted of 1000 response patterns for the reference group and 1000 response patterns for the focal group. The responses of the reference group were generated according to a Rasch model, the item parameters used are given in the second and fourth columns of Table 2. These columns are labeled "α_i" and "β_i", respectively. For the focal group, the items 1 through 6 and 10 were generated using the same Rasch model as for the reference group, but the responses for the items 7, 8 and 9 were generated using (19); the additional discrimination parameter γ_i and difficulty δ_i are given in the third and fifth columns of Table 2. The response patterns in the study were generated using normal ability distributions. To keep the illustration realistic, it was assumed that the means of the ability distribution of the reference group and the ability distribution of the focal group differed: the actual values used for generating the data are shown in the second column of the last four rows of Table 2. The remaining columns of this table give results of analyses averaged over 100 replications. For each replication, MML estimates and their standard errors were computed. The means of the estimates of the item parameters are shown in the sixth and seventh columns; these columns are labeled "$\hat{\beta}_i$" and "$Se(\hat{\beta}_i)$", respectively. The means of the estimates of the population parameters are shown in the last two columns of the four bottom lines of Table 2. In each replication, for each item three LM statistics were computed: $LM(\gamma_i)$ to test whether γ_i departed from zero, $LM(\delta_i)$ to test whether δ_i, departed from zero, $LM(\gamma_i, \delta_i)$ to test whether γ_i and δ_i simultaneously departed from zero. The results are given in the last nine columns of Table 2. The columns labeled "$LM(\gamma_i)$", "$LM(\delta_i)$" and "$LM(\gamma_i, \delta_i)$" contain the means of the test statistics, the columns labels "Pr" contain the mean probability levels of the statistics and the columns labeled "Nr" contain the number of times that the test was significant at the 5%-level. From the first columns of this table, it can be seen that the responses to item 7 are subject to uniform DIF only, that is, $\delta_i \neq 0$, item 8 is subject to non-uniform DIF only, that is, $\gamma_i \neq 0$, and item 9 both shows uniform and non-uniform DIF, so here both and $\delta_i \neq 0$ and $\gamma_i \neq 0$. The results show that the LM tests are indeed sensitive to the various forms of DIF imposed. For the items 8 and 9, the mean significance probabilities of $LM(\gamma_i)$ are below 0.02 and 0.03, respectively. Further, the test is significant at the 5%-level in 87 and 76 replications. The $LM(\delta_i)$ test for the items 7 and 9 has a probability level below 0.001 and 0.004 and the hypothesis of no uniform DIF is rejected at the 5%-level in 100 and 97 percent of the cases. Finally, for all three items, $LM(\gamma_i, \delta_i)$ is significant at the 5%-level in 99, 91 and 100 percent of the replications, the mean significance probabilities are below 0.003, 0.024 and 0.001, respectively. The DIF imposed on the three items does, of course, result in some bias in the parameter estimates of the other items, which, in turn, results in an augmentation of the number of erroneously significant LM tests. However, the consequences of this effect must not be exaggerated: it can be seen that the mean outcomes and probability levels of the tests for the items not affected by DIF are substantially different from the same indices for the items where the responses are subject to DIF. Therefore, it is advisable to adopt a procedure where the items with the most extreme outcomes are handled first, either by removing them or by modelling the responses to these items further; an example will be given below. For the present example, removing the items with DIF resulted in rejection rates of the hypothesis of no DIF for the other items at the proper chance level.

The second example entails a data set recently analyzed by Glas and Verhelst (1995) using the OPLM and generalized Pearson statistics. The objective of the present analysis is to investigate whether the DIF detected by these two authors will also be detected if LM tests are used, first in combination with the OPLM and then using the nominal response model. The example comprises a part of an examination of the business curriculum for the Dutch higher secondary education, the HAVO level. The example was part of a larger study of gender based DIF in examinations in secondary education. Since the objective, both here and in the Glas and Verhelst (1995) paper, is to illustrate the statistical procedures rather than to give an account of the findings with respect to gender based DIF, no actual examples of items with DIF will be shown. For a detailed report of the findings one is referred to Bügel and Glas (1991). The analyses were carried out using a sample of 1000 boys and 1000 girls from the complete examination population. For convenience of presentation the example is limited to 10 items. The items are open ended questions, the number of score points that could be obtained ranged from $m_i = 2$ to $m_i = 3$; the exact numbers of score points of the items can be seen in Table 3, in the column labeled "h".

In the first analysis, the OPLM was used. Glas and Verhelst (1995) have fitted an OPLM to the data used here. The discrimination indices that proved adequate are shown in Table 3 in the column labeled "a_i". These indices were also used in the present analyses. MML estimates were computed under the assumption of different normal ability distributions for the boys and the girls. The results of this MML estimation procedure are given in the columns labeled "$\hat{\beta}_{ih}$", "$Se(\hat{\beta}_{ih})$", "$\hat{\mu}_y$", "$Se(\hat{\mu}_y)$", "$\hat{\sigma}_y$" and "$Se(\hat{\sigma}_y)$" under the heading "Analysis 1". Glas and Verhelst (1995) have pointed out that the adequacy of the chosen scoring weights can be evaluated using an LM statistic for testing whether the value at which α_i is fixed is acceptable. This test, denoted by $LM(\alpha_{ih})$, was computed for every category within an item, that is, for every category h of item i it was tested whether the hypothesis $\alpha_{ih} = h\alpha_i$ had to be rejected. The results of this test are displayed in the columns labeled "$LM(\alpha_{ih})$" and "Prob". It can be seen that the items 3 and 9 do not fit the model. However, at this point it is unclear whether this lack of fit is due to DIF, since it might well be the case that the chosen discrimination index was inappropriate for boys and girls alike. Therefore, the LM statistics proposed in this paper were computed for testing whether non-zero shift parameters δ_{ih}, $h = 1, \ldots, m_i$, had to be added for the girls. The test was performed per item for all item category parameters simultaneously, therefore the test is labeled $LM(\delta_i)$. The results are shown in the columns labeled "$LM(\delta_i)$" and "Prob" of Table 4. It can be seen that the test is highly significant for the items 3 and 9.

Table 3. Parameter estimates and model fit for the OPLM

			Analysis 1				Analysis 4			
			no items splitted				items 3 and 9 splitted			
i	h	a_i	$\hat{\beta}_{ih}$	$Se(\hat{\beta}_{ih})$	$LM(\alpha_{ih})$	Prob	$\hat{\beta}_{ih}$	$Se(\hat{\beta}_{ih})$	$LM(\alpha_{ih})$	Prob
1	1	2	.27	.059	1.51	.219	.23	.060	.34	.544
	2		.49	.072	.00	.977	.43	.074	.00	.957
2	1	3	−1.25	.069	.09	.758	−1.34	.069	1.10	.293
	2		−.35	.098	2.01	.156	−.49	.100	2.11	.146
	3		.24	.121	1.91	.167	.10	.124	1.77	.183
3	1	4	−.70	.072	1.42	.232	−1.48	.103	.01	.892
	2		−.18	.105	6.96	.008	−1.14	.139	2.99	.083
4	1	2	.63	.066	2.48	.115	.59	.067	1.20	.273
	2		.39	.073	.14	.708	.32	.074	.66	.414
	3		1.86	.107	.00	.973	1.79	.109	.00	.977
5	1	2	−3.8	.073	.68	.406	−.44	.073	.00	.949
	2		.36	.101	.65	.418	.26	.102	.82	.363
	3		−1.12	.090	2.24	.134	−1.24	.092	1.64	.199
6	1	3	.00	.067	.02	.880	−.07	.068	.67	.412
	2		.08	.087	.03	.854	−.02	.090	.07	.791
7	1	3	.60	.066	.81	.366	.54	.067	2.11	.146
	2		.98	.089	1.96	.161	.90	.092	1.39	.237
8	1	3	−.58	.077	.41	.520	−.67	.078	.00	.976
	2		−1.01	.094	.44	.505	−1.16	.096	.44	.507
	3		−.55	.118	.19	.660	−.73	.122	.04	.839
9	1	4	.35	.074	8.47	.004	.04	.101	.73	.390
	2		.49	.104	5.94	.015	−.01	.134	.31	.577
10	1	4	.33	.094	.72	.394	.21	.095	.17	.679
	2		−.06	.121	.08	.771	−.27	.124	.18	.666
	3		−1.00	.143	.66	.415	−1.25	.149	.81	.367
3*	1	4					−.22	.093	2.11	.146
	2						.42	.129	1.33	.249
9*	1	2					.68	.088	2.92	.087
	2						.54	.092	3.16	.075
y			$\hat{\mu}_y$	$Se(\hat{\mu}_y)$	$\hat{\sigma}_y$	$Se(\hat{\sigma}_y)$	$\hat{\mu}_y$	$Se(\hat{\mu}_y)$	$\hat{\sigma}_y$	$Se(\hat{\sigma}_y)$
1			.25	.015	0.35	.011	−0.08	.017	.34	.011
1			.00	.000	0.34	.012	0.00	.000	.35	.011

Table 4. Testing DIF using the OPLM

	Analysis 1 no items splitted		Analysis 2 item 9 splitted		Analysis 3 item 3 splitted		Analysis 4 both splitted	
i	$LM(\delta_i)$	Prob	$LM(\delta_i)$	Prob	$LM(\delta_i)$	Prob	$LM(\delta_i)$	Prob
1	6.04	.048	4.93	.085	3.54	.170	2.49	.287
2	5.66	.129	5.19	.158	2.32	.508	2.92	.404
3	100.98	.000	108.65	.000				
4	7.04	.070	4.81	.186	4.79	.188	2.93	.401
5	4.80	.187	3.80	.283	3.44	.328	2.86	.414
6	3.89	.142	2.04	.360	2.34	.310	1.19	.551
7	4.57	.101	2.26	.323	1.79	.408	.21	.900
8	1.47	.687	.95	.813	3.24	.355	4.44	.218
9	15.15	.000			25.65	.000		
10	10.63	.013	8.13	.043	3.58	.310	1.44	.696

However, the test is also significant at a 5% level for the items 1 and 10. Interestingly, these results are similar to the results of the Glas and Verhelst (1995) analysis: also there the items 3 and 9 were highly significant and the items 1 and 10 moderately significant. As already noted above, the presence of DIF can bias the estimates of the parameters of items that are not influenced by DIF. Therefore, it is advisable to try to model DIF for the highly significant items before drawing conclusions for the other items. The following additional analyses were carried out. First, item 9 was entered into the analysis as a different item for boys and girls, that is, it was assumed that the item parameters β_{ih} were different for these two groups. However, from computation of the $LM(\alpha_{ih})$ statistics it had to be concluded that the scoring weights α_i also differed across the two groups; this result was also encountered in the Glas and Verhelst (1995) analysis. Changing this weight from 4 to 2 resulted in non-significant $LM(\alpha_{ih})$ tests. In this analysis, also the $LM(\delta_i)$ statistics were computed, the results are shown under the heading "Analysis 2" in Table 4. The $LM(\delta_i)$ statistic could not be computed for item 9 since it was split into two so-called conceptual items. Note that the test for item 1 is no longer significant at 5% level. Next, this procedure was repeated first with item 3 split up into two conceptual items and then with both the items 3 and 9 split up. The results are displayed under the heading "Analysis 3" and "Analysis 4" in Table 4. It can be seen that in the last analysis all $LM(\delta_i)$ statistics are non-significant. In Table 3, the parameter estimates and the $LM(\alpha_{ih})$ statistics for the last analysis are shown. Inspection shows that also these last statistics are no longer significant at the 5% level. So after splitting up the items 3 and 9 into different conceptual items for the two groups, an OPLM could be fitted to the data. This result is consistent with the results of the Glas and Verhelst (1995) analyses.

Table 5. Parameter estimates for the nominal response model

		Analysis 1 no items splitted				Analysis 4 items 3 and 9 splitted			
i	h	$\hat{\alpha}_{ih}$	$Se(\hat{\alpha}_{ih})$	$\hat{\beta}_{ih}$	$Se(\hat{\beta}_{ih})$	$\hat{\alpha}_{ih}$	$Se(\hat{\alpha}_{ih})$	$(\hat{\beta}_{ih})$	$Se(\hat{\beta}_{ih})$
1	1	1.99	.184	.26	.056	1.99	.177	.22	.056
	2	4.21	.201	.51	.060	4.10	.199	.44	.061
2	1	2.77	.191	−1.23	.065	2.79	.182	−1.31	.065
	2	6.27	.234	−.29	.082	6.28	.231	−.42	.082
	3	8.95	.235	.26	.088	8.83	.237	.10	.089
3	1	3.53	.205	−.66	.060	3.61	.267	−1.41	.093
	2	8.26	.263	−.06	.073	7.95	.384	−1.05	.109
4	1	1.99	.227	.63	.063	1.99	.218	.59	.063
	2	3.94	.190	.38	.060	3.87	.186	.31	.060
	3	6.23	.228	1.93	.087	6.16	.224	1.85	.088
5	1	1.71	.219	−.34	.070	1.77	.205	−.40	.071
	2	4.25	.326	.37	.093	4.28	.316	.27	.094
	3	6.19	.245	−1.10	.070	6.01	.242	−1.23	.070
6	1	2.96	.207	.00	.060	2.95	.200	−.06	.060
	2	5.96	.21	.08	.064	5.82	.243	−.02	.064
7	1	3.27	.216	.62	.060	3.22	.211	.55	.060
	2	5.82	.237	.93	.064	5.73	.240	.84	.065
8	1	2.91	.240	−.57	.073	2.96	.224	−.68	.073
	2	6.12	.214	−1.02	.074	6.12	.210	−1.17	.075
	3	9.11	.237	−.55	.082	8.99	.240	−.74	.083
9	1	3.43	.255	.39	.062	3.82	.352	.04	.090
	2	7.33	.297	.49	.066	8.32	.434	.02	.097
10	1	3.75	.388	.35	.087	3.77	.357	.24	.087
	2	8.10	.402	−.09	.089	8.16	.398	−.29	.090
	3	12.30	.405	−1.00	.080	12.15	.414	−1.26	.081
3*	1					3.70	.292	−.21	.084
	2					8.41	.362	.54	.104
9*	1					1.53	.274	.68	.084
	2					4.42	.278	.62	.085
y		$\hat{\mu}_y$	$Se(\hat{\mu}_y)$	$\hat{\sigma}_y$	$Se(\hat{\sigma}_y)$	$\hat{\mu}_y$	$Se(\hat{\mu}_y)$	$\hat{\sigma}_y$	$Se(\hat{\sigma}_y)$
0		.23	.015	0.32	.010	−0.10	.016	.33	.010
1		.00	.000	0.33	.010	0.00	.000	.35	.011

Finally, it was investigated how the procedure would perform if the nominal response model was used instead of the OPLM. From the previous analyses it is already apparent that the OPLM fits the data quite well, so the nominal response model should give results close to the previous results. In Table 5 the parameter estimates are shown for two analyses with the same arrangement as the analysis labeled “Analysis 1” and “Analysis 4” in Table 3. It can be seen that the estimates of the scoring weights α_{ih} are in accordance with the weights α_i postulated for the OPLM. Also the estimates of β_{ih} differ little between the two models.

In Table 6 the values of the $LM(\gamma_i, \delta_i)$ statistics are shown for four analyses comparable to the four analysis of Table 4. The $LM(\gamma_i, \delta_i)$ statistic is used to test the simultaneous hypotheses that the parameters γ_{ih} and δ_{ih}, $h = 1, \ldots, m_i$ are all equal to zero. It can be seen that also in the present case the items 3 and 9 show DIF. However, in this case the tests for the items 4 and 10 were also significant in the first analysis. As with the previous analyses, this significant result vanished when the items 3 and 9 were split into conceptual items for boys and girls. Again, this shows that it is important to investigate the items one at a time, starting with the items that seem to show the most serious DIF, because DIF in one item may affect the estimates of the parameters of the other items in such a way that the LM tests produce spurious results.

Table 6. Testing DIF using the nominal response model

	Analysis 1 no items splitted		Analysis 2 item 9 splited		Analysis 3 item 3 splitted		Analysis 4 both splitted	
i	$LM(\gamma_i, \delta_i)$	Prob	$LM(\gamma_i, \delta_i)$	Prob	$LM(\gamma_i, \delta_i)$	Prob	$LM(\gamma_i, \delta_i)$	Prob
1	5.47	.242	7.15	.128	5.62	.229	2.83	.587
2	7.44	.114	7.11	.130	2.93	.570	3.66	.453
3	130.10	.000	141.90	.000				
4	11.28	.024	6.28	.179	7.40	.116	4.37	.359
5	5.56	.234	4.48	.345	2.98	.560	2.35	.672
6	5.45	.244	1.61	.808	2.54	.638	.55	.969
7	5.72	.221	2.46	.652	2.78	.596	.32	.989
8	.37	.985	3.04	.552	3.17	.530	5.69	.224
9	18.93	.001			34.53	.00		
10	13.43	.009	10.10	.039	3.69	.449	1.92	.751

7. Discussion

In the present paper a method for detection of DIF is proposed that is based on a test statistic with a known asymptotical distribution. In the simulated example, it is shown that the method cannot only be used to detect DIF, it can also be used to distinguish between uniform and non-uniform DIF. The validity of the procedure is further supported with a real data example, where the results obtained are in agreement with the results obtained using the OPLM, in combination with CML estimates and generalized Pearson statistics. However, a choice between the two methods is not straightforward. The LM procedure can handle a wider array of IRT models than the procedure based on generalized

Pearson statistics, which can only be applied in the framework of exponential family IRT models. On the other hand, the latter procedure can be embedded in a procedure where various sources of model violations can be systematically evaluated, whereas evaluation methods of model fit for non-exponential family IRT are still rather unsophisticated. This is the more serious because estimation in non-exponential family IRT relies on assumptions about the ability distribution. These assumptions are an integral part of the model and should be tested also. In summary, there is no clear answer to the question which method is to be preferred.

In the present paper the LM method for detection of DIF is worked out for the OPLM and the nominal response model with normal ability distributions. However, the procedure does not only apply to these models, it also applies to other unidimensional IRT models, such as for instance the models proposed by Samejima (1969, 1972) and to multidimensional models such as the model proposed by Glas (1992) and Adams and Wilson (1995). Further, the assumption of one or more normal ability distributions can be replaced with the assumption of the non-parametric MML method that the distribution of ability can be represented by one or more step-functions (De Leeuw and Verhelst (1986), Follmann (1988)). These applications remain a topic for further research.

References

Adams, R. J. and Wilson, M. R. (1995). Formulating the Rasch model as a mixed coefficients multinomial logit model. In *Objective Measurement: Theory into Practice, Vol.* 3 (Edited by G. Engelhard and M. Wilson), Nordwood, NJ: Ablex Publishing Corporation.

AERA, APA and NCME. (1985). *Standards for Educational and Psychological Tests.* Washington DC: American Psychological Association, American Educational research Association, National Council on Measurement in Education.

Aitchison, J. and Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29**, 813-828.

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *J. Roy. Statist. Soc. Ser. B* **34**, 42-54.

Andersen, E. B. (1973a). A goodness of fit test for the Rasch model. *Psychometrika* **38,** 123-140.

Andersen, E. B. (1973b). Conditional inference for multiple-choice questionnaires. *British J. Math. Statist. Psych.* **26**, 31-44.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika* **42**, 69-81.

Angoff, W. H. (1993). Perspective on differential item functioning methodology. In *Differential Item Functioning* (Edited by P. W. Holland and H. Wainer), 3-23. Hillsdale, N. J., Erlbaum.

Birnbaum, A. (1968). Some latent trait models. (hoofdstuk 17 in:) In *Statistical Theories of Mental Test Scores* (Edited by F. M. Lord and M. R. Novick), Addison-Wesley: Reading (Mass.).

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

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM-algorithm. *Psychometrika* **46**, 443-459.

Bügel, K. and Glas, C. A. W. (1991). Item specifieke verschillen in prestaties tussen jongens en meisjes bijtekstbegrip examens moderne vreemde talen. *Tijdschrift voor Onderwijsresearch* **16**, 337-351.

Chang, H. and Mazzeo, J. (1994). The unique correspondence of the item response functions and item category response functions in polytomously scored item response models. *Psychometrika* **59**, 391-404.

Cole, N. S. (1993). History and development of DIF. In *Differential Item Functioning* (Edited by P. W. Holland and H. Wainer), 25-29. Hillsdale, N. J., Erlbaum.

De Leeuw, J. and Verhelst, N. D. (1986). Maximum likelihood estimation in generalized Rasch models. *J. Educational Statistics* **11**, 183-196.

Fischer, G. H. (1974). *Einführung in Die Theorie Psychologischer Tests* (Introduction to the theory of psychological tests). Bern: Huber.

Fischer, G. H. (1993). Notes on the Mantel-Haenszel procedure and another chi-square test for the assessment of DIF. *Methodika* **7**, 88-100.

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika* **60**, 459-487.

Follmann, D. (1988). Consistent estimation in the Rasch model based on non-parametric margins. *Psychometrika* **53**, 553-562.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika* **53**, 525-546.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In *Objective Measurement: Theory into Practice, Vol.* 1 (Edited by M. Wilson). New Jersey: Ablex Publishing Corporation.

Glas, C. A. W. (1997). Testing the generalized partial credit model. In *Objective Measurement: Theory into Practice*, Vol. 4 (Edited by M. Wilson, G. Engelhard, Jr., and K. Draney), New Jersey: Ablex Publishing Corporation.

Glas, C. A. W. and Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika* **54**, 635-659.

Glas, C. A. W. and Verhelst, N. D. (1995). Tests of fit for polytomous Rasch models. In *Rasch Models. Their Foundation, Recent Developments and Applications* (Edited by G. H. Fischer and I. W. Molenaar). Springer, New York.

Hambleton, R. K. and Rogers, H. J. (1989). Detecting potentially biased test items: comparison of IRT area and Mantel-Haenszel methods. *Appl. Measurement in Education* **2**, 313-334.

Holland, P. W. and Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In *Test Validity* (Edited by H. Wainer and H. I. Braun), 129-145. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika* **54**, 681-697.

Kok, F. G., Mellenbergh, G. J. and van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *J. Educational Measurement* **22,** 295-303.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44**, 226-233.

Martin Löf, P. (1973). *Statistika Modeller. Anteckningar från seminarier Lasåret 1969-1970, utarbetade av Rolf Sunberg. Obetydligt ändrat nytryck, oktober* 1973. Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistik vid Stockholms Universitet.

Martin Löf, P. (1974). The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data. *Scand. J. Statist.* **1**, 3-18.

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

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *J. Educational Statist.* **7**, 105-118.

Mellenbergh, G. J. (1983). Conditional item bias methods. In *Human Assessment and Cultural Factors* (Edited by S. H. Irvine and W. J. Berry), 293-302. Plenum Press, New York.

Mislevy, R. J. and Bock, R. D. (1990). PC-BILOG. *Item Analysis and Test Scoring with Binary Logistic Models.* Scientific Software: Mooresville.

Molenaar, I. W. (1983). Some improved diagnostics for failure in the Rasch model. *Psychometrika* **48**, 49-72.

Molenaar, I. W. (1995). Estimation of Item Parameters. In *Rasch Models: Their Foundations, Recent Developments and Applications* (Edited by G. H. Fischer and I. W. Molenaar). Springer, New York.

Rao, C. R. (1948). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proc. Cambridge Philosophical Society* **44**, 50-57.

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

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

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika, Monograph Supplement, No.* 17.

Samejima, F. (1972). A general model for free response data. *Psychometrika, Monograph Supplement, No. 18.*

Verhelst, N. D. and Glas, C. A. W. (1995). The generalized one parameter model: OPLM. In *Rasch Models: Their Foundations, Recent Developments and Applications* (Edited by G. H. Fischer and I. W. Molenaar). Springer, New York.

Verhelst, N. D., Glas, C. A. W. and Verstralen, H. H. F. M. (1993). OPLM: One Parameter Logistic model. Computer program and manual. Arnhem: Cito.

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

Department of Educational Measurement and Data Analysis, Faculty of Educational Science and Technology, University of Twente, P. O. Box 217, 7500 AE Enschede, the Netherlands.

E-mail: C.A.W.Glas@edte.utwente.nl

(Received April 1996; accepted October 1997)