PSYCHOMETRIKA—VOL. 54, NO. 4, 681–697
DECEMBER 1989

ITEM BIAS DETECTION USING LOGLINEAR IRT

HENK KELDERMAN

UNIVERSITY OF TWENTE

A method is proposed for the detection of item bias with respect to observed or unobserved subgroups. The method uses quasi-loglinear models for the incomplete subgroup × test score × Item 1 × ··· × item k contingency table. If subgroup membership is unknown the models are Haberman's incomplete-latent-class models.

The (conditional) Rasch model is formulated as a quasi-loglinear model. The parameters in this loglinear model, that correspond to the main effects of the item responses, are the conditional estimates of the parameters in the Rasch model. Item bias can then be tested by comparing the quasi-loglinear-Rasch model with models that contain parameters for the interaction of item responses and the subgroups.

Key words: loglinear models, Rasch model, item bias, differential item performance, latent-class models, IRT

Educational or psychological tests are biased if the test scores of equally able test takers are systematically different between racial, ethnic, cultural, and other similar subgroups. Biased test scores may lead to unfair decisions or erroneous conclusions about individuals from particular subgroups. A test score is biased only if one or more of the test items are biased. A test item is biased if individuals with the same ability level from different subgroups have a different probability of a right response, that is, the item has different difficulties in different subgroups. A test can be made fairer by deleting or improving the biased items.

Binet and Simon (1916; see also Jensen 1980, p. 367) were already concerned with bias when they applied their test of general intelligence that was standardized on working class children to children of higher social status. Since then, many methods for detecting item bias have been developed. Reviews are given by Osterlind (1983), and Shepard, Camilli and Averill (1981). Handbooks on item bias detection methods and research are provided by Berk (1982) and Jensen (1980).

Over the years methods have been improved by better control for ability. This is done either by using the number-correct score of test items to control for ability (Camilli, 1979; Holland & Thayer, 1986; Kok, Mellenbergh, & van der Flier, 1985; Mellenbergh, 1982; Nungester, 1977 [see also Ironson, 1982]; Scheuneman, 1979) or by studying item bias under an IRT model (Durovic, 1975; Fischer & Forman, 1982; Lord, 1980; Muthén & Lehman, 1985; Mislevy, 1981; Wright, Mead, & Draba, 1975).

Using IRT models, item bias is detected as differences of item parameters across subgroups. Since IRT models provide a clear separation of person ability and item difficulty, they are ideally suited to detect item bias. In this paper this advantage of the Rasch model is combined with the evaluative power of loglinear models.

The Rasch model describes the probability $P(X_j = x_j|\alpha)$ that an individual with parameter α gives a response X_j to item j ($j = 1, \ldots, k$), where the random variable X_j can take values $x_j = 0, 1$ for a wrong (0) or a right (1) response:

The author thanks Wim J. van der Linden and Gideon J. Mellenbergh for comments and suggestions and Frank Kok for empirical data.

Requests for reprints should be sent to Henk Kelderman, University of Twente, PO Box 217, 7500 AE Enschede, THE NETHERLANDS.

0033-3123/89/1200-8068$00.75/0

© 1989 The Psychometric Society

$$P(X_j = x_j|\alpha) = \frac{\exp(x_j(\alpha - \delta_j))}{1 + \exp(\alpha - \delta_j)}, \tag{1}$$

where $\delta_j (j = 1, \ldots, k)$ is a single item parameter describing the difficulty of item j. If this item parameter varies from subgroup to subgroup, the item is considered biased. Although the Rasch model is a rather simple model, its parsimony yields several virtues in using it to detect item bias.

The Rasch model is an exponential family model wherein the simple number right score $T = X_1 + \cdots + X_k$ is a sufficient statistic for the person parameter α. Assuming local independence of the item responses for a given value of α and after conditioning on the number right score, the joint probability $P(X_1 = x_1, \ldots, X_k = x_k|T = t)$ of the item responses $X_1, \ldots, X_k$ for a given score $T = t$ becomes:

$$P(X_1 = x_1, \ldots, X_k = x_k|T = t) = \frac{\exp(-x_1\delta_1 \cdots - x_k\delta_k)}{\sum_{x_1} \cdots \sum_{\substack{x_k \\ t = x_1 + \cdots + x_k}} \exp(-x_1\delta_1 \cdots - x_k\delta_k)}. \tag{2}$$

By conditioning on the score, the nuisance parameter α has vanished (Rasch, 1966). In this paper the invariance over subgroups i $(i = 1, \ldots, m)$ of the joint item response distributions for given values of T

$$P_i(X_1 = x_1, \ldots, X_k = x_k|T = t) = P(X_1 = x_1, \ldots, X_k = x_k|T = t) \tag{3}$$

is tested to study item bias. According to Model (2) any deviation of this invariance must be explained by differences in item difficulty between the subgroups. Note from (2) that the use of the Rasch model to study item bias is both an observed score method and a IRT method.

In this paper, item bias detection methods are described using a loglinear IRT model assuming a Rasch model for ability and difficulty. Quasi-loglinear models are formulated for test data and the Rasch model is formulated as one of them. Alternative models are described to test various aspects of item bias. The use of these tests is illustrated on a set of test data from Kok (1982), where item bias was introduced experimentally. Finally, further developments of the basic model are described where the subgroups are unknown.

Quasi-Loglinear Models for the Incomplete Subgroup × Score × Item 1 × ··· × Item k Table

Let $f_{itx_1 \cdots x_k}$ be the number of individuals from subgroup i $(i = 1, \ldots, m)$ with number right score $T = t$ $(t = 0, 1, \ldots, k)$ and item scores $X_1 = x_1, \ldots, X_k = x_k$. Since it is logically impossible to have a test score that is unequal to the number of correct item responses (excluding counting errors) the counts $f_{itx_1 \cdots x_k}$ are zero for $t \neq \Sigma_j x_j$. Contingency tables with structurally zero cells are called incomplete contingency tables.

Fienberg (1972; see also Bishop, Fienberg & Holland, 1975) presents a general theory for the statistical analysis of incomplete multiway contingency tables by quasi-loglinear models. We apply Fienberg's theory to the analysis of the subgroup × score × item 1 × ··· × item k contingency table to detect item bias.

Let $m_{itx_1 \cdots x_k}$ be the expected counts for the table under some model. If $t \neq x_1 + \cdots + x_k$, the expected counts are again structurally zero. If $t = x_1 + \cdots +$

x_k, the expected counts are structurally nonzero and these counts are explained by a quasi-loglinear model. The saturated or fully specified model for the table is:

$$\begin{aligned}\ln m_{itx_1 \cdots x_k} &= u + u_1(i) + u_2(t) + u_3(x_1) + \cdots + u_{(k+2)}(x_k) \\ &\quad + u_{12}(it) + u_{13}(ix_1) + \cdots + u_{(k+1)(k+2)}(x_{k-1}x_k) \\ &\quad + u_{123}(itx_1) + \cdots + u_{123\cdots(k+2)}(itx_1 \cdots x_k) \end{aligned} \tag{4}$$

for $i = 1, \ldots, m; x_1 = 0, 1; \ldots; x_k = 0, 1; t = x_1 + \cdots + x_k$, where ln is the natural logarithm. Model (4) has constraints:

$$\begin{aligned} u_1(+) = u_2(+) &= \cdots = u_{(k+2)}(+) = u_{12}(+t) = u_{12}(i+) \\ &= u_{13}(+x_1) = u_{13}(i+) = \cdots = u_{(k+1)(k+2)}(+x_k) \\ &= u_{(k+1)(k+2)}(x_{k-1}+) = \cdots = u_{123}(+tx_1) = u_{123}(i+x_1) \\ &= u_{123}(it+) = \cdots = u_{123\cdots(k+2)}(+tx_1 \cdots x_k) \\ &= u_{123\cdots(k+2)}(i+x_1 \cdots x_k) = \cdots \\ &= u_{123\cdots(k+2)}(itx_1 \cdots x_{k-1}+) = 0. \end{aligned} \tag{5}$$

The u-terms in Model (4) describe main effects and interaction effects of subgroup i, score t and item responses $x_1, \ldots, x_k$. The u-terms in (5) denote sums of parameters that occur in (4) where a plus sign replacing an index indicates that the summation is over the replaced index. The constraints (5), however, are not sufficient to ensure that all parameters in (4) are estimable. Additional constraints must be imposed to obtain a unique solution of the model parameters. These constraints will be discussed later.

Restrictive quasi-loglinear models are defined by setting u-terms in (4) equal to zero. The only models considered here will be hierarchical, that is, whenever a particular u-term is set to zero, all its higher order relatives must also be set to zero.

The Rasch Model as a Quasi-Loglinear Model

A restrictive quasi-loglinear model is

$$\ln m_{itx_1 \cdots x_k} = u + u_1(i) + u_2(t) + u_{12}(it) + u_3(x_1) + \cdots + u_{k+2}(x_k), \tag{6}$$

with the constraints

$$u_1(+) = u_2(+) = u_{12}(+t) = u_{12}(i+) = u_3(+) = \cdots = u_{k+2}(+) = 0 \tag{7}$$

Model (6) can be obtained from the saturated quasi-loglinear model (4) by setting all interactions with and between item responses equal to zero.

If the subgroup and score are taken as fixed variables and the item responses are considered as random variables, Model (6) is equivalent to the conditional Rasch model. In that case $m_{itx_1 \cdots x_k}$ is the conditional expected frequency of the response $X_1 = x_1, \ldots, X_k = x_k$ for given subgroup i and score t. The conditional probability of response $X_1 = x_1, \ldots, X_k = x_k$ for i and t can then be obtained from (6) by

$$P_i(X_1 = x_1, \ldots, X_k = x_k \mid T = t) = \frac{m_{itx_1 \cdots x_k}}{\sum_{x_1} \cdots \sum_{x_k} m_{itx_1 \cdots x_k}}$$

$$x_1 + \cdots + x_k = t$$

$$= \frac{\exp\,(u_3(x_1) + \cdots + u_{k+2}(x_k))}{\sum_{x_1} \cdots \sum_{x_k} \exp\,(u_3(x_1) + \cdots + u_{k+2}(x_k))}. \qquad (8)$$
$$x_1 + \cdots + x_k = t$$

Except for a reparameterization, (8) is equivalent to (2). In (2) the effect $-x_j\delta_j$ of a response $X_j = x_j$ on item j is $-\delta_j$ for a correct response ($X_j = 1$) and zero for an incorrect response ($X_j = 0$), whereas in (8) the effect of a correct response is $u_{j+2}(1)$ and the effect of an incorrect response is $u_{j+2}(0)$, where $u_{j+2}(0) = -u_{j+2}(1)$ by the constraints (7). Model (8) can be parameterized in the same way as (2) if $u_{j+2}(1)$ is added to each parameter $u_{j+2}(x_j)$ so that the new parameter $u_{j+2}(x_j) + u_{j+2}(1)$ becomes $u_{j+2}(0) + u_{j+2}(1) = 0$ with an incorrect response and $2u_{j+2}(1)$ with a correct response. This can be done by multiplying both numerator and denominator by

$$\exp\,(u_3(1) + \cdots + u_{k+2}(1)),$$

so that (8) becomes (2) with

$$-x_j\delta_j = u_{j+2}(x_j) + u_{j+2}(1) = x_j(2u_{j+2}(1)),$$

for all $j = 1, \ldots, k$; that is, $\delta_j = 2u_{j+2}(0)$. This shows that the conditional Rasch model is equivalent to the quasi-loglinear model (6). Other derivations of this fact are given by Cressie and Holland (1983), Duncan (1984), Kelderman (1984), and Tjur (1982).

In (6) there is an obvious overparameterization because of the linear dependence of the item responses and the score: adding a constant c to each of the item parameters $u_{j+2}(1)$ ($j = 1, \ldots, k$) and subtracting c from $u_{j+2}(0)$ ($j = 1, \ldots, k$) to satisfy the constraints (7) is equivalent to adding $t \cdot c - (k - t) \cdot c = (2t - k) \cdot c$ to $u_2(t)$. This indeterminacy can be removed from (2) by putting one linear constraint on the item parameters, for example, by setting $u_{k+2}(x_k)$ equal to zero. This also fixes the metric of the latent trait.

We now describe some less restrictive quasi-loglinear models that can be used to detect item bias.

Quasi-Loglinear Models to Detect Item Bias

To study item bias in a particular set of data, quasi-loglinear models may be set up that contain subgroup-dependent item parameters in addition to the parameters of the Rasch model (Rasch, 1960). The fit of these models can be compared by a likelihood ratio test with the fit of more restrictive models to test the significance of each of the subgroup-dependent item parameters. If a test yields a significant result, the item is considered biased. The subgroup-dependent item parameters each describe a particular type of item bias.

To detect the simplest type of bias, for example in item one, the model

$$\ln m_{itx_1 \cdots x_k} = u + u_1(i) + u_2(t) + u_{12}(it) + u_3(x_1) + \cdots + u_{k+2}(x_k) + u_{13}(ix_1), \qquad (9)$$

with the usual constraints (5), is compared with the loglinear Rasch model (6) to test the null hypothesis that the interaction between the subgroup and the response to item one, $u_{13}(ix_1)$ is zero. If the test is significant, it may be concluded that $u_{13}(ix_1)$ is not zero so that the difficulty of item one varies from subgroup to subgroup. The parameter $2u_{13}(i0)$ is the change in item difficulty in subgroup i. Note that loglinear Rasch models with subgroup $\times$ item interactions (such as (9)) can be viewed as loglinear Rasch models with item difficulties equal to the sum of the item-main effect and the item interaction-

effect parameter. In (9), $2(u_3(0) + u_{13}(i0))$ is the difficulty of item one in subgroup i. The unbiased items have the same difficulty $2u_{j+2}(0)\,(j = 2, \ldots, k)$ in all subgroups. In this way all item difficulties are put on the same scale.

In (9) a u-term is specified to test item bias for only one item. Obviously similar u-terms can be specified for two or more items if necessary. For example, comparing the loglinear Rasch model with the model:

$$\ln m_{itx_1 \cdots x_k} = u + u_1(i) + u_2(t) + u_{12}(it) + u_3(x_1) + \cdots + u_{k+2}(x_k) + u_{13}(ix_1) + u_{14}(ix_2), \tag{10}$$

yields a simultaneous test for bias in both item one and item two.

An item may be more difficult in one subgroup than another, because the item introduces some specific difficulty, e.g. reading ability, in which the members of one subgroup are generally more proficient than the members of another. If the ability to solve this difficulty varies from individual to individual within each of the subgroups and if there are two items in the test that both introduce the same difficulty we may expect these items to show an interaction that is not explained by the original latent trait.

This interaction may be investigated using the model:

$$\ln m_{itx_1 \cdots x_k} = u + u_1(i) + u_2(t) + u_{12}(it) + u_3(x_1) + \cdots + u_{k+2}(x_k) + u_{13}(ix_1) + u_{14}(ix_2) + u_{34}(x_1x_2) + u_{134}(ix_1x_2), \tag{11}$$

which contains two u-terms, $u_{34}(x_1x_2)$ and $u_{134}(ix_1x_2)$ describing an interaction between item one and two. If $u_{134}(ix_1x_2)$ is zero but $u_{34}(x_1x_2)$ is not zero, there is a simple interaction between both items that is the same in all subgroups. If $u_{134}(ix_1x_2)$ is not zero, the interaction is different from subgroup to subgroup. This may, for example, be the case if reading ability does introduce common variance in one subgroup but does not introduce any variance in another subgroup, because the individuals in that subgroup are all of relatively superior reading ability.

Comparing Model (11) with the loglinear Rasch model (6) yields a test for the hypothesis that all subgroup-dependent item parameters in (11) are simultaneously zero. If the test is significant, it may be concluded that one or more of these parameters are not zero. Comparing (11) with (10) yields a test for the item-interaction terms alone. To test both item-interaction terms $u_{34}(x_1x_2)$ and $u_{134}(ix_1x_2)$ separately, an intermediate submodel must be defined that contains $u_{34}(x_1x_2)$ but not $u_{134}(ix_1x_2)$.

If an item-interaction parameter is included in the model, it is no longer a conditional Rasch model (2) for ICC (1). Therefore, the model should not be considered as a "Rasch model with correlated errors". The model is merely meant to test whether the data deviate from the Rasch model in this respect.

Table 1 lists all relevant models (a through e) containing subgroup-dependent item parameters for the case of two items. Table 2 summarizes which models in Table 1 must be compared to test specific subgroup-dependent item parameters. Hypothesis 3 and 4 shows which models must be compared to test $u_{34}(x_1x_2)$ and $u_{134}(ix_1x_2)$ respectively.

Hypotheses 1 through 4 in Table 2 refer to what Mellenbergh (1982) has called "uniform" item bias. It means that the bias is constant within each subgroup. With "nonuniform" item bias (Mellenbergh) the bias of in each subgroup is dependent on the individual's ability level. Nonuniform bias may be studied with quasi-loglinear models containing item parameters that depend both on the subgroup and the score.

TABLE 1

Quasi-Loglinear Models for Detecting Item Bias.

Models with Subgroup-Dependent Item Parameters
a. Rasch + $u_{13}(ix_1)$
b. Rasch + $u_{14}(ix_2)$
c. Rasch + $u_{13}(ix_1) + u_{14}(ix_2)$
d. Rasch + $u_{13}(ix_1) + u_{14}(ix_2) + u_{34}(x_1x_2)$
e. Rasch + $u_{13}(ix_1) + u_{14}(ix_2) + u_{34}(x_1x_2) + u_{134}(ix_1x_2)$
Models with Subgroup and Score-Dependent Item Parameters
f. (a) + $u_{23}(tx_1)$
g. (a) + $u_{23}(tx_1) + u_{123}(itx_1)$
h. (b) + $u_{24}(tx_2)$
i. (b) + $u_{24}(tx_2) + u_{124}(itx_2)$
j. (c) + $u_{23}(tx_1) + u_{24}(tx_2)$
k. (c) + $u_{23}(tx_1) + u_{24}(tx_2) + u_{123}(itx_1) + u_{124}(itx_2)$
l. (d) + $u_{23}(tx_1) + u_{24}(tx_2) + u_{123}(itx_1) + u_{124}(itx_2) +$ $+ u_{234}(tx_1x_2)$
m. (e) + $u_{23}(tx_1) + u_{24}(tx_2) + u_{123}(itx_1) + u_{124}(itx_2) +$ $+ u_{234}(tx_1x_2) + u_{1234}(itx_1x_2)$

TABLE 2

Comparison of Quasi-loglinear Models to Test u-terms for Item Bias Hypothesis.

	Hypothesis	Model Terms	Comparison of Models
	Uniform Bias		
1.	One item uniformly biased	$u_{13}(ix_1)$	Rasch - a
2.	Two items uniformly biased	$u_{13}(ix_1), u_{14}(ix_2)$	Rasch - c
3.	Two items with common uniform bias:	$u_{34}(x_1x_2)$	c - d
4.	Two items with common uniform bias: subgroup dependent interaction	$u_{134}(ix_1x_2)$	d - e
	Non-uniform Bias		
5.	One item non-uniformly biased	$u_{123}(itx_1)$	f - g
6.	Two items non-uniformly biased	$u_{123}(itx_1), u_{124}(itx_2)$	j - k
7.	Two items with common non-uniform bias	$u_{234}(tx_1x_2)$	k - l
8.	Two items with common non-uniform bias: subgroup dependent interaction	$u_{1234}(itx_1x_2)$	l - m

Table 1 shows a series of models (f through m) with subgroup- or score-dependent item parameters. Since quasi-loglinear models are hierarchical, each model with a subgroup × score × item(s) interaction term must contain the corresponding subgroup × item(s) interaction term. In Table 1 all models f through m contain a submodel identical to one of the models a through e, which is indicated by its letter for brevity. Table 2 shows which of these models must be compared to obtain a statistical test that is sensitive to a specific type of nonuniform item bias. Note that these tests concentrate only on the nonuniformity of the bias and not on the uniform part of the bias. Therefore, if these tests are not significant, items may still be uniformly biased.

Hypothesis 5 in Table 2 concerns the simplest type of nonuniformity in item bias. If model g and f (Table 1) differ significantly, it can be concluded that the subgroup × score × item interaction $u_{123}(itx_1)$ is not zero. This nonuniformity in item bias may be expected, for example, if the difficulty of an item varies from subgroup to subgroup for low ability individuals only, which is the case if an item involves a specific skill that is not mastered by the low ability individuals of only one of the subgroups.

Hypothesis 6 (Table 2) concerns this hypothesis for two items simultaneously, whereas Hypotheses 7 and 8 address the question whether item interaction is nonuniform ($u_{234}(tx_1x_2) \neq 0$) or whether subgroup differences in item interaction are nonuniform ($u_{1234}(itx_1x_2) \neq 0$). This may be called *nonuniform common item bias*, where the amount of item bias that two items have in common depends on ability level. This type of item bias may occur, for example, if in only one subgroup two items introduce a common difficulty for low ability individuals but do not introduce a common difficulty for high ability subjects.

In most of the models in Table 1, the constraints are not sufficient to ensure identifiability of the model parameters. The same indeterminacy between the item main effect parameters and the sum score parameters that existed in the Rasch model are also present in the models of Table 2. This indeterminacy can again be removed by fixing one item-main-effect parameter to be zero. To interpret likelihood-ratio tests and interaction parameters, this need, however, not be the same item-main-effect parameter.

If the model is complex, other indeterminacies in the parameter estimates may be present. For example, the parameter $u_{23}(tx_1)$ with $t = 0$ and $x_1 = 1$ or $t = k$ and $x_1 = 0$ cannot be estimated because it corresponds to structurally zero cells only. A convenient and reliable way to determine the number of estimable parameters is to determine the rank of the information matrix, which should be equal to the number of estimable parameters for a given set of data (Goodman, 1974; McHugh, 1956). Baker and Nelder (1978, sec. 4.3) describe a weighted least-squares algorithm for the analysis of contingency tables, which estimates the parameters in a sequential fashion. If a parameter is linearly dependent on the preceding parameters, or if there are no observations to estimate it from, the parameter is removed from the model, thus the information matrix is of full rank.

Estimation and Testing

The kernel of the log likelihood is

$$l = \ln \prod_i \prod_t \prod_{x_1} \cdots \prod_{x_k} (m_{itx_1 \cdots x_k})^{f_{itx_1 \cdots x_k}}$$

$$= \sum_i \sum_t \sum_{x_1} \cdots \sum_{x_k} f_{itx_1 \cdots x_k} \ln m_{itx_1 \cdots x_k}. \tag{12}$$

Inserting a loglinear model for $\ln m_{it x_1 \ldots x_k}$ this log likelihood yields a sum of products of model parameters (e.g., $u_3(x_1)$) with the corresponding sufficient marginal counts (e.g., $f_{++x_1+\ldots+}$). For example, using the loglinear Rasch model (6) in (12) gives

$$\begin{aligned} l(\text{Rasch}) = {} & f_{+\ldots+}u + \sum_i f_{i+\ldots+}u_1(i) + \sum_t f_{+t+\ldots+}u_2(t) \\ & + \sum_i \sum_t f_{it+\ldots+}u_{12}(it) + \sum_{x_1} f_{++x_1+\ldots+}u_3(x_1)\cdots \\ & + \sum_{x_k} f_{+\ldots+x_k}u_{k+2}(x_k), \end{aligned} \tag{13}$$

where a plus sign replacing an index denotes summation over that index. Log likelihoods of larger models (e.g., Model (9)) may be obtained by adding terms (e.g., $\Sigma\Sigma f_{i+x+\ldots+}u_{13}(ix_1)$) to (13). If one model—say model M—is a special case of another model—say model M*—model M* may be tested against model M by -2 times the natural logarithm of the likelihood ratio of both models, or equivalently, by -2 times the difference in log likelihood of both models

$$G^2(\text{M};\ \text{M}^*) = -2(l(\text{M}) - l(\text{M}^*)). \tag{14}$$

Under the assumption of model M, G^2 is asymptotically distributed as chi-square with degrees of freedom equal to the number of estimable parameters of both models (Bishop, Fienberg & Holland, 1975, p. 525; Rao, 1965, p. 351).

An overall goodness of fit test for model M is obtained by testing it against the saturated model M* where the expected cell counts (m) in (12) are set equal to the observed cell counts (f).

For example the Rasch model (6) is a special case of (9). Model (9) has all parameters of the Rasch model but adds the term $u_{13}(ix_1)$. Testing (6) against (9) is a test for the hypothesis $u_{13}(ix_1) = 0$. If the parameter estimates of both (6) and (9) are known, the likelihood-ratio statistic $G^2(\text{M};\ \text{M}^*)$ can be calculated easily from the sufficient marginal sums corresponding to the parameters.

Maximum-likelihood estimates of the model parameters can be obtained by setting the observed marginal counts corresponding to each of the parameters equal to the corresponding expected marginal counts and solving the resulting system of equations for the parameters (Haberman, 1979, p. 448). For example, for the Rasch model the maximum-likelihood equations are

$$f_{it+\ldots+} = m_{it+\ldots+},$$

and (15)

$$f_{+\ldots+x_j+\ldots+} = m_{+\ldots+x_j+\ldots+},$$

for $i = 1, \ldots, m$; $t = 0, \ldots, k$ and $x_j = 0, 1$; $j = 1, \ldots, k$.

In general, for quasi-loglinear models, the maximum-likelihood equations yield no direct solution of the model parameters. The equations must be solved iteratively. Algorithms to solve the maximum-likelihood equations for quasi-loglinear models have been described by Goodman and Fay (1974; ECTA) and Baker and Nelder (1978; GLIM). These programs require the internal storage of the full observed and expected tables of counts which is virtually impossible if the number of items is larger than, say, 10. To deal with larger numbers of items a new computer program LOGIMO (loglinear IRT modeling) has been written (Kelderman & Steen, 1988). The program uses a very

efficient algorithm to calculate expected sufficient marginals in (15), the Marginalization-by-Variable (MBV) algorithm (Kelderman, 1987). Furthermore, it calculates the parameter estimates from the observed and expected marginals using an iterative proportional fitting procedure. This means that it is not necessary to store the full observed and expected contingency table. The program can be used to estimate the parameters in ordinary loglinear models or quasi-loglinear Rasch models with one or more subgroup variables, one or more sum score variables, items of any number of response categories and loglinear models for relatively large numbers (say 40) of variables.

An Example

Kok (1982) studied item bias in multiplication items by experimentally varying the test takers skill in bias factors that can be expected to be operating in differently formulated test items. In this section, fifteen multiplication items are reanalyzed to illustrate the use of quasi-loglinear models for the detection of item bias. In Item 1 through 12 the numbers are written out in Dutch and in Item 13 through 15 Roman numerals are used. The subjects were 286 Dutch undergraduates of which 144 randomly selected individuals received a short training in Roman numerals. It can be expected that the Roman numerals items are biased.

In Table 3 for each item the values of the likelihood-ratio test and the degrees of freedom are shown for both uniform (Hypothesis 1, Table 2) and nonuniform bias (Hypothesis 5, Table 2).

From Table 3 it is seen that for all Roman numerals items the likelihood-ratio chi-square value of the Rasch model against a model with one item uniformly biased (Model a, Table 1) yields a significant value. Table 3, also shows that none of these items is nonuniformly biased. Furthermore, two Dutch items, Items 6, 9 and 10, are identified as biased. The effect $u_{1,j+2}(11)$ of a correct response from the trained groups is given. Note that each of these parameters is from a different Model a (Table 1) and the other interaction parameters can be obtained from the constraints $u_{1,j+2}(11) = u_{1,j+2}(20) = -u_{1,j+2}(10) = -u_{1,j+2}(21)$.

It is seen that the Roman numerals items are all less difficult for the trained group than for the untrained group. The biased Dutch Items, 6, 9 and 10, however, are more difficult for the trained group than for the untrained group. This might indicate that the system of deciphering Roman numerals interferes with the method of obtaining the number from the Dutch in these items.

It was hypothesized that the Roman numerals Items 13, 14 and 15 are biased by a common cause and the Dutch Items 6, 9 and 10 are biased by another common cause. Likelihood-ratio statistics of Test 3 of Table 2 and the corresponding interaction parameters $u_{j+2,l+2}(11)$ in each Model d (Table 1) are computed for each pair of biased items. It is found that the Roman numeral, Items 13 and 14 have significant interaction with Item 15 ($u_{13+2,15+2}(11) = 0.18$, $G_1^2 = 7.04$, $p = 0.01$; and $u_{14+2,15+2}(11) = 0.16$, $G^2 = 4.06$, $p = 0.04$). A simultaneous test of the three interactions between the Roman numerals items, is also significant ($G_1^2 = 10.96$, $p = 0.01$).

To test whether the interactions between the items are different for the trained group than the untrained group, Test 4 of Table 2 is performed for each pair of biased items. For all biased items computed are the likelihood-ratio statistics and the values of the parameter $u_{1,j+1,l+2}(111)$, that is, the effect for the trained group with both item j and l correct. The results show that the Roman numerals Items 13 and 14 are less correlated in the trained group than in the untrained group. ($u_{1,13+2,14+2}(111) = -0.18$, $G_1^2 = 5.72$, $p = 0.02$).

TABLE 3

Likelihood-Ratio Tests for Uniform and Non-uniform Bias (Test 1 and Test 5).

Item	Uniform	Non-uniform	Difficulty (Rasch)	Bias (Model a)
G_1^2(Rasch;a)	G_{14}^2(f;g)	$u_j(0)$	$u_{1,j+2}(10)$	
1	0.67		0.00	0.06
2	0.10		0.65	-0.03
3	0.02		0.39	-0.00
4	3.06		1.27	0.14
5	0.16		0.38	0.03
6	4.03*	11.04	0.02	0.15
7	0.30		0.04	0.04
8	3.64	0.07	0.14	
9	4.41*	17.40	0.18	0.15
10	14.95***	9.08	0.89	0.28
11	2.03		0.88	0.10
12	1.04		1.63	-0.10
13	5.21*	9.19	0.34	-0.16
14	55.41***	9.22	0.44	-0.54
15	12.39***	11.35	0.37	-0.25

* $p < .05$, ** $p < .01$, *** $p < .001$

On the other hand, the two Dutch Items 6 and 9 have a significantly larger interaction in the trained group ($u_{1,6+2,9+2}(111) = 0.12$, $G_I^2 = 4.78$, $p = 0.03$).

This example shows that loglinear models can give us very useful information on which hypotheses about the causes of bias are confirmed by the data.

Further Developments

In some practical situations, items may be expected to be biased for certain subgroups of individuals, but it is not known a priori to which subgroup each of the individuals belongs. For example, for an item in an examination the probability of a correct response may be larger for a group of individuals with certain educational experiences than for individuals without that experience, or for an item in a mastery test the probability of a correct response may be larger for a subgroup of individuals having a different study strategy or for a subgroup of individuals having a different cognitive strategy to solve the item, and so forth. In these examples, information on the

individuals subgroup membership may be difficult to observe or, as in the last example, the test behavior itself may be the natural indicator of subgroup membership.

Within the theory of contingency table analysis a straightforward extension of the range of item bias detection methods is the inclusion of unobserved subgroups.

When subgroup membership is unobserved the subgroup variable becomes a latent variable. The models to detect item bias then become latent-class models. For example, if the latent classes are denoted by $\omega(\omega = 1, \ldots, m)$, a latent-class item-bias-detection model becomes,

$$\ln m_{\omega t x_1 \cdots x_k} = u + u_1(\omega) + u_2(t) + u_{12}(\omega t) + u_3(x_1) + \cdots + u_{k+2}(x_k) + u_{13}(\omega x_1) + \cdots + u_{1,r+2}(\omega x_r), \tag{16}$$

$\omega = 1, \ldots, m; x_1 = 0, 1; \ldots; x_k = 0, 1; t = x_1 + \cdots + x_k$; with the usual constraints (5).

Model (16) describes a Rasch model in each latent class ω, where the difficulty of Item 1 through r may be different in each latent class. The parameters $u_{13}(\omega x_1), \cdots, u_{1,r+2}(\omega x_r)$ describe the differences in item difficulty between the latent classes. If such a parameter is not zero, the corresponding item is biased with respect to the latent classes.

Latent-class models have been introduced by Lazarsfeld (1950; Lazarsfeld & Henri, 1968; Goodman, 1978). At first, latent-class models assumed local independence within each latent class. Goodman (1975) introduced latent-class models where the observed variables form an incomplete-contingency table assuming quasi independence within each latent class. Haberman (1979, chap. 10) formulates a latent-class model for an incomplete table where the model is not necessarily an independence model. The model can be any identifiable loglinear model containing unobserved categorical variables. Model (16) is a special case of Haberman's general latent class model where Items 1 through 4 may have a different difficulty in each of m latent classes, where the number m of latent classes is specified by the investigator.

Methods for the estimation and testing of latent-class-quasi-loglinear models differ from those for ordinary quasi-loglinear models. Since latent-class membership is unobserved, the frequencies $f_{\omega t x_1 \cdots x_k}$ are not known. Consequently, the maximum-likelihood equations (e.g., $f_{\omega + x_1 +} = m\hat{\omega}_{+x_1+}$) for parameters involving latent classes ω (e.g., $u_{13}(\omega x_1)$) cannot be solved because the frequencies are unknown. Haberman (1979, chap. 10), however, gives a rule for the derivation of maximum likelihood estimates in latent-class models from the known frequencies $f_{+t x_1 \cdots x_k}$. Haberman (1979) states, "The same maximum-likelihood equations apply as in the ordinary case in which all frequency counts are directly observed, except that the unobserved counts are replaced by their estimated conditional expected values given the observed marginal totals" (p. 543).

Under some loglinear model M (e.g., (16)), these estimates are

$$\tilde{f}_{\omega t x_1 \cdots x_k} = E_M(\tilde{f}_{\omega t x_1 \cdots x_k} | f_{+t x_1 \cdots x_k}) = \left(\frac{\tilde{m}_{\omega t x_1 \cdots x_k}}{\tilde{m}_{+t x_1 \cdots x_k}}\right) f_{+t x_1 \cdots x_k} \tag{17}$$

$t = x_1 + \cdots + x_k$; $x_1 = 0, 1; \ldots ; x_k = 0, 1$. For (16) the likelihood equations would then become

$$\tilde{f}_{\omega t+\cdots+} = \tilde{m}_{\omega t+\cdots+}, f_{++x_1+\cdots+} = \tilde{m}_{++x_1+\cdots+}$$

$$f_{+\cdots+x_k} = \tilde{m}_{+\cdots+x_k}, \tilde{f}_{\omega+x_1+\cdots+} = \tilde{m}_{\omega+x_1+\cdots+}, \ldots ,$$

$$\tilde{f}_{\omega+\cdots+x_r+\cdots+} = \tilde{m}_{\omega+\cdots+x_r+\cdots+}. \tag{18}$$

The estimated counts $\tilde{f}$ are obtained from (17) where the $\tilde{m}$ are described by (16). A scoring algorithm to solve these equations has been described by Haberman (1979, p. 556). An alternative way to solve these equations, is by using the EM algorithm (Dempster, Laird & Rubin, 1977) with (17) as the expectation step and solving (18) as the maximization step.

Unfortunately if the number of items is larger than say 10, these algorithms can no longer be used because the number of expected counts become too large. On the other hand, with a small number of items convergence is so slow that no final solution could be reached. An extension of the Marginalization-by-Variable algorithm as used in the LOGIMO program (Kelderman & Steen, 1988) for the case of latent class analysis might be made to estimate the parameters of the loglinear Rasch model with latent classes.

Discussion

In this paper an item bias detection method is proposed that uses a latent trait as an internal criterion for ability.

Latent trait parameters are removed from the model by conditioning on the number right score. The quasi-loglinear formulation of the model is extended with parameters that describe different types of item bias. The general theory of (quasi-) loglinear models is used to obtain maximum-likelihood parameter estimates and likelihood-ratio tests.

The models presented in this paper have two parts: one part contains parameters describing item bias, the other part contains parameters for the Rasch measurement model. The latter implies that the method assumes that all nonbiased items conform to the Rasch model. It may be asked how robust the item-bias-detection method is with respect to violations of this assumption. To check this, a simulation study was performed. Two hundred data sets were generated. Each data set contained the item responses of 1000 individuals on 11 items. In the first 100 data sets the data where generated from a Rasch model and in the last 100 data sets the data were generated from a two-parameter-logistic model. In all data sets the first item is a biased item and the last ten items are unbiased. Each data set contains two subgroups of 500 individuals each. In the first subgroup the biased item has a difficulty parameter of 0.5 and the ability parameters are drawn from a normal distribution with mean 0.5 and variance 1.0. In the second subgroup the biased item has a difficulty parameter 0.0 and the ability parameters are drawn from $N(0, 1)$. In all data sets the biased item has a slope parameter of one. Furthermore, in all data sets the ten remaining item difficulty parameters are drawn from a normal distribution with mean zero and variance one.

In the 100 data sets conforming to the two-parameter-logistic (2PL) model, slope parameters of the 10 unbiased items are chosen as follows. The angle of the item-characteristic curve is drawn from a uniform distribution from $\pi/8$ to $3\pi/8$. The slope parameter is the tangent of that angle so that they are between 0.5 and 2.0.

The slopes are not sampled uniformly between 0.5 and 2.0 because in that case

TABLE 4

Statistics for Simulated Data Sets with Item One Biased

Data Set			Difference in Likelihood Ratio						
Simulated	Mean	SD	0-5	5-10	10-15	15-20	20-25	25-30	30>
1PL	12.50	7.20	16	26	24	18	10	4	2
2PL	12.33	7.91	20	25	24	18	8	3	2

there would be about twice as much ICC's with slopes larger than the slopes 1 of the Rasch ICC than there would be ICC's with slopes smaller than the Rasch ICC. In that case, the mean slope parameter of the unbiased items in the 2PL data sets would be greater than that one of the Rasch data sets. We do not want to simulate mean differences in slope but variation or no variation in slope between the two data sets.

For each data set the loglinear Rasch model (6) and the item bias model with item one uniformly biased (9) where fitted, and the difference in likelihood-ratio statistics (14) of both models calculated. Table 4 gives the observed numbers of datasets with the difference in likelihood-ratio statistic, in each of several categories generated under the Rasch (or 1PL) model and generated under the two-parameter logistic (2PL) model. It is seen that the difference between the frequency distributions for the Rasch and the 2PL model is small (Pearson chi-square = 0.83, DF = 6). Table 11 also gives the means and standard deviations of the difference in likelihood-ratio statistics. It is seen that the means are about equal (t = .15, DF = 198). This simulation study suggests that the test for bias in item one is rather robust for deviations of the slope of the item characteristic curves in the remaining items. Item bias in tests following a two-parameter-logistic model is detected just as good as item bias in tests following a Rasch model.

In the preceding data sets there was a considerable bias effect in item one. It might be suspected that the method may erroneously reveal item bias in the 2PL data if the

TABLE 5

Statistics for Simulated Data Sets with Item One Unbiased

Data Set			Difference in Likelihood Ratio						
Simulated	Mean	SD	0-5	5-10	10-15	15-20	20-25	25-30	30>
1PL	1.04	1.33	48	21	7	8	7	1	8
2PL	1.28	1.55	38	19	16	7	6	5	9

real bias in item one is small or absent. To assess this, the same simulation study was performed except that no bias was introduced in item one. The item difficulty parameter of item one was set equal to 0.5 for all subjects. The results in Table 5 suggest that for the data generated under the Rasch model the distribution of the difference in likelihood-ratio statistics is located more to the left than for data generated under the two-parameter logistic model. The differences, however, are small and do not reach significance (Pearson chi-square = 7.65, DF = 6, $t = -1.19$, DF = 198). Thus both simulation studies indicate that the item bias detection method is robust with respect to deviations of the assumption of parallel item characteristic curves. Therefore, the assumption of a Rasch measurement model does not seem to be critical for the applicability of the method when the items follow a two-parameter logistic model.

Item bias detection methods using an internal ability criterion, assume that a good measure of this criterion is available, that is, the item used to measure this criterion fit the measurement model. If that is not the case, particularly if one or more of these items are biased themselves, the results may be erroneous. Marco (Lord, 1980, p. 228) proposed a procedure to purify a test of biased items. The total test is analyzed, items that appear to be biased are removed and the remaining items are used as an internal ability criterion to test the bias of all the test items one by one. Although this procedure does not escape the inherent circularity of the problem it should suffice if not too many items are biased. This procedure can also be used with the test presented in this paper where in the first phase only one-item-uniform bias is tested and in the second cycle the set of unbiased items is combined with pairs of possibly biased items to use the diagnostic tests presented in this paper.

If one or more items is uniformly biased and the uniform-item-bias model fits the data, it is not really necessary to remove the items from the test if one is willing to use different item difficulties in each subgroup. The uniform-item-bias model specifies a Rasch model within each subgroup. So latent trait values can be calculated for each subject provided that the item difficulties belonging to his or her subgroups are used in their calculation. It is doubtful, however, whether a scoring rule that makes use of subgroup membership would be acceptable to the testees.

Finally it should be remarked that the item bias part of the models may be more elaborate. The models in this paper contain parameters that indicate deviations due to item bias. Kok and Mellenbergh (1985) go further and formulate models that describe the processes involved in the genesis of item bias more precisely. Our models may be used to give directions as to which of Kok's models may be appropriate.

References

Baker, R. J., & Nelder, J. A. (1978). *The GLIM system: Generalized linear interactive modeling*. Oxford: The Numerical Algorithms Group.

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, MA: MIT Press.

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

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

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

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.

Fienberg, S. E. (1972). The analysis of incomplete multi-way contingency tables. *Biometrics, 28,* 177–202. (Corrig. 1972, *29,* 829)

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

Goodman, L. A. (1974). Exploratory latent structure analysis. *Biometrika, 61,* 215–231.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association, 70,* 755–768.

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

Goodman, L. A., & Fay, R. (1974). *ECTA program, description for users.* Chicago: University of Chicago, Department of Statistics.

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

Holland, P. W. (1985). *On the study of differential item performance without IRT.* Paper presented at the Annual Meeting of the Military Testing Association, San Diego.

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

Kelderman, H., & Steen, R. (1988). *LOGIMO: A program for loglinear IRT modeling.* Enschede: University of Twente, Department of Education.

Kok, F. G. (1982). *Het partijdige item. [The biased item]* Psychologisch Laboratorium, University of Amsterdam.

Kok, F. G., & Mellenbergh, G. J. (1985, July). *A mathematical model for item bias and a definition of bias effect size.* Paper presented at the Fourth Meeting of the Psychometric Society, Cambridge, Great Britain.

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.

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

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, pp. 413–472). 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 Jersey: 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.

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. *Dissertation Abstracts International, 38,* 2726 A. (University Microfilms No. 77-24, 289, Doctoral dissertation Florida State University, 1977)

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

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement,* 3–29.

Rao, C. R. (1965). *Linear statistical inference and its applications.* New York: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Paedagogiske Institut.

Rasch, G. (1966). An item analysis that takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19,* 49–57.

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). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317–377.

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

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

Manuscript received 1/24/86
Final version received 8/29/88