KM37(1991)
pag 5 - 17

COMPARISON OF TWO ITEM BIAS DETECTION PROGRAMS

(A simulation study)

H.J. Adèr*

Abstract

In 1984 van der Flier proposed an algorithm to identify test items that are differently responded to by equally able test takers in distinct (cultural) groups. Iterativily, biased items are removed from the test set. The method means a significant improvement over previous (non-iterative) approaches.

The present paper reports an extensive simulation study that has been recently conducted to analyse the behaviour of the algorithm and track down differences in performance between two well-known implementations.

Results show that the algorithm should be used with care. The programs show an overall performance of ± 85 % correct classifications, but performance is dependent on the characteristic curves of the items and on the ability distribution of the groups: easy items that are biased risk to remain undiscovered. Both programs tend to indicate unbiased items as biased more easily than the other way around. Therefore, one shouldn't use the technique in a 'diagnostic' setting, in which the biasedness of the items is explained by characteristics of the groups.

Finally, recommendations on the proper use of the technique are formulated, and some adaptations to the programs are proposed.

*Department of Psychology, Methodology Division, de Boelelaan 1111 Vrije Universiteit, Amsterdam.tel.: 020-5484404.

1 Introduction

Commonly, the historical development of Item Response Theory is started with Thurstone (1925), who tried to quantify the concept of 'mental age', on which the intelligence test of Binet and Simon is based. Thurstone assumed the (intellectual) *ability* of children of a fixed age to have a normal distribution. The present article concerns a special application of Item Response Theory: Item Bias Detection.

An item in a test is characterized as biased if *equally able test takers are systematically different between subgroups.* These subgroups may be racial, ethnic, or cultural subgroups but also bias in sex groups has been studied (Lucassen and Evers, 1984). The concept of item bias is relevant for tests in any domain: it may involve items in the educational or the psychological fields, but one could very well define bias in other test situations, for instance, market research.

In van der Flier, Mellenbergh, Adèr and Wijn (1984) an iterative algorithm was proposed to detect item bias. The comparison of two implementations of this algorithm by means of Monte Carlo methods is the subject of the present paper. The first program by Adèr (1982) is the 'prototype' implementation. It is written in Algol 68, a programming language that is very useful for the precise description of an algorithm, but which, unfortunately, produces rather slow object code. The second program, written in Pascal by Kok (1986) overcomes the practical drawbacks of the prototype. However, it also uses a slightly different method to assign observations to ability classes.

Experiments are carried out to test this difference and to get a general impression of the overall behaviour of the algorithm. It is expected that the influence of the shape of the item characteristic curves is not at all negligible. Finally, recommendations are given for the design of an 'optimal' item detection program along the lines of the original algorithm. The literature on item bias detection has steadily been growing in the past few years. Introductionary reviews have appeared in several articles and books. In his introduction, Kelderman (1986) gives such an overview. Also worthwhile in this respect is van der Flier *et al.* (1984). Handbooks by Berk (1982) and Jensen (1980) give broader information. Recently Kok's thesis (1989) appeared, which contains an account of earlier and recent developments.

Van der Flier distinguishes between *unconditional* and *conditional* methods, i.e. methods that are (or are not) conditioned on the ability level. Historically, unconditional methods come first. Cardall and Coffman (1964) used analysis of covariance to analyse p-values[1]in a group $\times$ items design. To satisfy the analysis of variance requirement of equal cell-variances they applied an arcsine transformation to their data. Echternacht (1974) and Angoff and Ford (1973) proposed other transformations, using standard normal deviates. Early research in this field concentrates on this 'equal-variances' problem and on problems related to the distribution of the (transformed) p-values.

Several authors have criticized unconditional methods. Hunter (1975) and Shepard, Camilli, and Averill (1981) show by example that these methods can lead to what will be called in paragraph 4 'type I unreliability' : indicating bias while items are not biased.

Scheuneman (1979) used a modified χ^2 as a test statistic for testing item bias. Her approach was improved by Camilli (1979) and Nungester (1977) and fitted into a more general loglinear framework by Mellenbergh (1982), who formulated a logit model to measure item bias. Van der Flier uses Mellenbergh's logit model in an iterative procedure to detect item bias.

[1] p-value $= \frac{\#\text{of correct responses}}{\text{total score}}$

2 Iterative Item bias detection

Let a test be administered to g groups of subjects. The test consists of m items. Item responses are scored correct (1) or incorrect (2). The total score t of a subject is the number of correctly answered items in the test. These total scores are divided into s score categories. The data for each item can be summarized in an $s \times g \times 2$, Score category $\times$ Group $\times$ Response, contingency table. Consider the three-dimensional contingency table for item $\mathcal{I}$, in which a cell contains the sample frequency f_{ijk} of score category i, group j and response category k — $k = 1$ for a correct response and $k = 2$ for an incorrect response —. Let F_{ijk} be the expected frequency. The logit is defined as the natural logarithm of the ratio of correct and incorrect responses. The saturated logit model may be formulated as (Fienberg, 1980, chap. 6):

$$ln(\frac{F_{ij1}}{F_{ij2}}) = C + S_i + G_j + SG_{ij} \tag{1}$$

with the usual constraints:

$$\sum_{i=1}^{s} S_i = 0; \quad \sum_{j=1}^{g} G_j = 0; \quad \sum_{i=1}^{s} SG_{ij} = \sum_{j=1}^{g} SG_{ij} = 0$$

C is the overall item difficulty parameter, S_i the main score category effect, G_j the main group effect and SG_{ij} the score category $\times$ group interaction effect parameter. When the following model is valid:

$$ln(\frac{F_{ij1}}{F_{ij2}}) = C + S_i \tag{2}$$

the item is said to be *unbiased*. If

$$ln(\frac{F_{ij1}}{F_{ij2}}) = C + S_i + G_j \tag{3}$$

represents the valid model, the item is called *uniformly biased* (Mellenbergh, 1982). If the SG_{ij} -term cannot be dropped from the model, i.e. the model is described by (1) but not by (3) then the item is said to be *nonuniformly biased*: in addition to the main group effect, in this model the interaction between score category and group is not neglectable. The expected frequencies for the unbiased item model of formula (2) are estimated by

$$\hat{f}_{ijk} = \left(\sum_{j=1}^{g} f_{ijk}\right)\left(\sum_{k=1}^{2} f_{ijk}\right) \Big/ \left(\sum_{j=1}^{g}\sum_{k=1}^{2} f_{ijk}\right) \tag{4}$$

As a test statistic the likelihood ratio statistic is used:

$$G^2 = 2 \sum_{i=1}^{s}\sum_{j=1}^{g}\sum_{k=1}^{2} f_{ijk} \ln(f_{ijk}/\hat{f}_{ijk}) \tag{5}$$

which is asymptotically χ^2 distributed with $s(g-1)$ degrees of freedom.

2.1 The iterative Algorithm

The algorithm given in van der Flier *et al.* (1984) is thus formulated:

Each iteration step T proceeds as follows:

I. First, for each item $\mathcal{I}$ four steps are taken:

1. Per subject a rest score is computed, i.e. the total score over all items except item $\mathcal{I}$ and those items that were found biased in the preceding iteration.
2. The overall frequency distribution of the rest scores is computed. Subjects are equally split up over the specified score categories. If a borderline between categories coincides with a tie, subjects are assigned randomly to the adjacent categories. In this way a uniform distribution over the score categories is obtained.
3. A Score × Group × Response table is constructed. If in the table a zero frequency is found, all frequencies are raised by .5.
4. Using Formulae (4) and (5) the likelihood ratio chi-square of item $\mathcal{I}$ is computed.

II. Next, the T items with the highest LRχ^2 values are considered biased, the rest of the items is included in a set considered unbiased for this iteration.

III. The Algorithm terminates if at the end of the iteration one of the following conditions arise:

- the prescribed number of iterations has been performed, or
- the maximal LRχ^2 of the set of unbiased items does not exceed the critical value.

Note that in each iteration G^2's are computed for all items of the test, including items found biased in the preceding iteration. In each iteration one more item than in the previous iteration is eliminated. The items excluded in a given iteration are not necessarily excluded in subsequent iterations: if an item that was considered biased before has an acceptable G^2 in the actual iteration, it is included again. In this way the score is iterativily freed from biased items and all items are tested using an unbiased rest score as ability indicator.

In step I.2 the distribution is chosen to be uniform over the categories to minimize the number of zero frequencies. This is where the two implementations differ.

2.2 BIASIT and BIAKOK

BIASIT is a direct implementation of the iterative algorithm. During each iteration rest scores are computed for each item. The data are classified accordingly into a fixed number of ability classes, each with an equal number of observations. Especially, in later iterations when the maximal rest score decreases, many subjects will have equal rest scores: since the program tries to keep the class frequencies equal, many random assignments to classes may occur. Recomputation of the rest scores and the splitting of the ties turn out to be very time-consuming.

The output of the program is straightforward: for each iteration a G^2-table is printed (see formula (5)). The output ends with an overview of the detected biasedness of items in subsequent iterations.

As the documentation states: "BIAKOK has been included in the Zielery-library since it is much faster than BIASIT" (Kok, 1986). Kok's handling of rest score class assignment deviates from the original algorithm: in BIAKOK, subjects with the same total score always belong to

the same class. Conceptionally this seems more logical than the original approach: subjects with the same rest score cannot be discerned in ability.

In this paper we concentrate on the different methods of assigning cases to score classes. BIASIT adheres closely to the original algorithm which indicates that cases should be assigned to subclasses in a way as to keep classes of equal size. BIAKOK assigns cases with the same rest score to the same rest score class possibly creating unequal classes. In the following we call the first method *random assignment,* the second strategy *fixed assignment.*

3 Design of the experiments

One could try to analyse analytically the way in which the class assignment strategy influences the item bias detection outcome. This appears to be a complicated theoretical exercise.

Here a Monte Carlo approach to the problem is reported. Before I describe the way test sets were generated, it should be remarked that a simulation experiment, in contrast to a complete algebraic treatment of the subject, is never conclusive, and rather gives an impression of the behaviour of the programs.

The test sets were generated using item parameters borrowed from van der Flier *et al.* (1984). Items were assumed to be scored on a five point scale. A number of three scoregroups was fixed throughout the experiment and a fixed number of 500 cases per group was generated. The characteristics of the groups were chosen after Kok (1982) in a way that often may be found in social science research. He argues that usually it is not realistic to assume the means of the ability distributions of both groups to be equal.

Data sets were generated that vary only in the number of items: for each test length, hundred (100) data sets were produced. A procedure was fixed to determine the ultimate iteration. As a measure for comparison the number of misclassifications of each program was taken.

To generate the item characteristic curves a three parameter normal ogive model has been used:

$$P(\theta) = c + \mu(1-c)\int_{-\infty}^{a(\theta-b)} f(t)dt \tag{6}$$

in which θ is the *latent variable* corresponding to the ability, a is the *item discrimination power*, b corresponds to the *difficulty* of the item, c is the *guessing*-parameter and f(t) is the standard normal density function (Anderson, 1980). Since we assume the items in the test to be five point items, the guessing parameter c is taken .20 for both groups. For a and b , itemparameters have been used, which were chosen by van der Flier *et al.* (1984) to generate a 29-item test set. μ is included in the model to be able to induce bias in the itemresponse. For the first group the μ has been taken 1.0 for all items, for the second group μ is taken .75 for half of the items, 1.0 for the other half, as was done in van der Flier's experiments. (It may seem that this is a rather high percentage of biased items. Observe that this will only put some extra strain on the technique, since ability estimators become less reliable if the number of biased items increases. Furthermore, in real life it is usually impossible to fix the amount of bias involved, due to the influence of confounding error sources (f.i. distribution). Therefore it is not easy to determine what a realistic amount of bias would be.)

In contrast to van der Flier, the normal distributions of the supposed ability distribution of the drawn subjects has been taken differently for both groups: group 1 is supposed to be $\mathcal{N}(0,1)$ distributed, group 2 is $\mathcal{N}(-.5,1)$. In this way the influence of differences in the location

parameter may be studied. The test length has been varied. The test lengths used are:

$$2, 3, 6, 9, 10, 11, 12, 15, 20, 24$$

It would have been attractive to have test length greater than 24. Only practical reasons prevented this: the amount of computer time required would be too great. The smaller, even degenerate test lengths have been included since it was expected that here the differences in class assignment strategy of the programs would appear most clearly.

For each test length, a subset of van der Flier's items have been used, so that the items in the different sets are comparable. For instance, for test set 6, items 1 to 6 correspond to items 24 to 29 of van der Flier. Hereafter, we will use van der Flier's numbering. For all test lengths hundred (100) data sets of two groups were drawn, each consisting of 500 cases. Both programs were run on all data sets. The scoring procedure was as follows: (a) The one iteration in which all bias induced items could have been discovered, was considered ultimate. Since the odd numbered items were bias induced, the number of iterations in which all biased items could have been detected is:

$$\text{number of iterations} = \begin{cases} \frac{1}{2} * \text{test length} & \text{if test length is even} \\ \lfloor \frac{1}{2} * \text{test length} \rfloor + 1 & \text{if test length is odd} \end{cases}$$

(b) Per test set misclassifications of each item for the ultimate iteration were scored.

4 Analysis

Two kinds of misclassification may be thought of in item bias detection:

(1) *unbiased* items are classified as *biased*

(2) *biased* items are classified as *unbiased*

(1) will be called 'type I unreliability'; (2) will be called 'type II unreliability' after 'type I error' and 'type II error' in hypothesis testing. It depends on the particular application what kind of misclassification is judged more undesirable: if item selection is the aim of the study, a technique that is type I unreliable is less harmful, since it keeps items that can be trusted to be unbiased: in other words, it rejects easily. On the other hand, if the group differences play an important role in the investigations and the emphasis is on the meaning of bias of the items for the make-up of the groups, type I unreliability may face the researcher with interpretative problems. He will tend to ascribe the bias of the item to certain characteristics of the groups, although it is due to dysfunctioning of the used technique.

In our analysis, the interaction between induced bias of the items and the way both programs classify them is of main interest: one would expect Kok's program to be type I unreliable in contrast to BIASIT which may be expected to produce conservative classifications errors (type II unreliability) since the randomization of the class boundaries could lead to loss of information. Test length may also play a role in the detection of bias.

A preliminary question is, whether the items in the test sets generated as described in the last section contain the induced bias, indeed. As we mentioned in section 3, the groups are taken to have different location parameter values. This may have some influence on the induced bias of the items.

Since our data consist of frequencies, loglinear analysis is the most attractive way to analyse them. Fienberg(1980) and Bishop, Fienberg and Holland (1975) are well known references for

this technique. As usual, we will only use hierarchical models in which together with any interaction term all lower order interaction terms or main effects are present. We adhere to the notation given in BMDP (1985) to indicate a model with its highest order interaction terms and we drop indices if no confusion may arise.

As an example, the model TIB stands for

$$F = T + I + B + TI + TB + IB + TIB \tag{7}$$

F : Classification effect (1 = misclassified, 2 = well classified)

T : Test length effect ($t = 2, 3, 6, 9, 10, 11, 12, 15, 20, 24$)

I : Item-effect (i = 1,2 ... 29)

B : Induced Bias effect (1 biased, 2 unbiased)

M : Method-effect (m = 1 (BIAKOK) or 2 (BIASIT))

Our design has fixed marginals if summation is over the classification score index. (In all sets the number right and the number wrong classifications sum to 100). Therefore the interaction term over the other factors is fixed and should be included in all models.

Corresponding to the two questions formulated above, two sets of analyses are conducted.

A. Do the generated test sets conform to expectations about induced bias? The factors included are: I(item), F(classification), M(method), B(bias induced) and T(test length).

The term $IMBT$ has to be included in all models. Our main interest with this question will be in the term IFB; we test models to find out if this term can or cannot be missed in the model. If this term appears to be needed the observed frequencies of the marginal table for IFB indicates in what way induced bias and item characteristics interact.

B. Do the programs differ in their detection performance ? The same model and data as mentioned in A may be used to give an impression of reliability of the programs.

Our main interest is now in the term FMB, indicating the interaction between method and induced bias of the item. However, there may be some interaction with test length, in which case the terms FMT and FBMT are of interest.

5 Results

The following model is used to test the first question (see Table 1 for an overview of other models):

$$IMBT, IFMB, IFBT, FMT \quad G^2(43) = 33.75 \quad (Prob. = .8429)$$

It turns out that the factor M(ethod) cannot be missed, so the term IFMB has to be included instead of only IFB and FMB. Inspection of the marginal frequency tables shows that for several items more than 40% misclassifications are made. This is true for item 23, 26, 27, 28 and item 29. Several considerations are justified here: the item characteristic curves of the two groups for these items may not differ due to the shift in location ability distributions of the groups. The items concerned are items with extreme difficulty parameters: ($b = 1.096, 1.516, 1.600, 2.954$),

Table 1: Fit of Various Models containing IFB

Model	df	G^2	Prob.
IMBT, IFMT, IFBT, FMBT	34	47.10	.0669
IMBT, IFMB, IFBT, FMT	43	33.75	.8429
IMBT, IFMB, IFBT	46	43.33	.5846
IMBT, FMB, IFM, IFBT, FMT	57	59.25	.3934
IMBT, IFMB, FBT, IFT, FMT	63	136.28	.0000

respectively. Test sets 2 and 3 are degenerate in that the ability estimates are based on 1 or 2 items only. Consequently, one can expect to find unreliable results for item 28 and 29. For the rest of our analysis we leave out these items and test lengths, although it should be stressed that in a real life situation it may not at all be clear that the group ability means differ: their estimation is based on total scores that contain biased information. Furthermore, it will be usually impossible to get reliable information on the interaction between biasedness and distribution. We conclude that due to these factors, both programs may completely misclassify certain items.

5.1 Performance analysis

From Table 2 we conclude, that BIAKOK performs better for test length 6,12 and 15; that the differences for test length 9,10,11 are slightly in the advantage of BIAKOK. BIASIT performs better for test length 20 and 24. Considering the terms in the model, we conclude that items are differently responded to (in the sense of misclassifications) for different test length. Percentages misclassifications for the different test length are indicated in Table 3. To investigate the connection with induced bias an analysis is done in which bias induced items are contrasted to unbiased items. We find as best fitting model (see Table 4):

$$BMT, BFT, FMT \quad \chi^2(4) = 5.64 \quad Prob. = .2280 \tag{8}$$

Table 5 shows that unbiased items are much more easily misclassified than biased items: both programs are Type I unreliable.

6 Discussion and conclusions

Simulation experiments, in contrast to a complete algebraic treatment of a subject, are never conclusive, and rather give an impression. In this case, generalizibility is further restrained by (a) a specific choice of test lengths and (b) an arbitrary (although realistical) shift of the location parameter of group 2. On the other hand, experiments have been extensive, so that some valuable information may be obtained about the behaviour of the programs and the iterative item bias detection technique in general.

Results in the previous section show that the algorithm as proposed by van der Flier should be used with care. The programs show an overall performance of $\pm$ 85 % correct classifications, but performance is dependent on the characteristic curves of the items and on the ability distribution of the groups: easy items that are biased risk to remain undiscovered. The same may be true for difficult items ($b < -1.0$). In practice, one should check on this: item

Table 2: Observed frequencies for effect FMT

Test length	Method	Classification		
		Incorrect	Correct	Total
6	BIAKOK	61	439	500
	BIASIT	78	422	500
	Total	139	861	1000
9	BIAKOK	93	707	800
	BIASIT	95	705	800
	Total	188	1412	1600
10	BIAKOK	102	798	900
	BIASIT	105	795	900
	Total	207	1593	1800
11	BIAKOK	102	898	1000
	BIASIT	107	893	1000
	Total	209	1791	2000
12	BIAKOK	113	987	1100
	BIASIT	124	976	1100
	Total	237	1963	2200
15	BIAKOK	179	1221	1400
	BIASIT	287	1113	1400
	Total	466	2334	2800
20	BIAKOK	308	1572	1880
	BIASIT	296	1604	1900
	Total	604	3176	3780
24	BIAKOK	427	1573	2000
	BIASIT	418	1582	2000
	Total	845	3155	4000

Table 3: Misclassifications (in percentages) for different test lengths

Test length:	6	9	10	11	12	15	20	24
Percentage:	14	12	12	10	11	17	16	21

Table 4: Fit of Various Models contrasting Biased and Unbiased Items

Model	df	G^2	Prob.
BMT, BFM, BFT, FMT	3	4.15	.2456
BMT, BFT, FMT	4	5.64	.2280
BMT, BF, FMT	7	175.91	0.000
BMT, BFT, FM	7	22.87	.0018

Table 5: Observed frequencies for effect BFT

Test length	Classification	Biasedness		
		Biased	Unbiased	Total
6	Incorrect	6	133	139
	Correct	394	467	861
	Total	400	600	1000
9	Incorrect	16	172	188
	Correct	584	828	1412
	Total	600	1000	1600
10	Incorrect	33	174	207
	Correct	767	826	1593
	Total	800	1000	1800
11	Incorrect	34	175	209
	Correct	766	1025	1791
	Total	800	1200	2000
12	Incorrect	53	184	237
	Correct	947	1016	1963
	Total	1000	1200	2200
15	Incorrect	167	299	466
	Correct	1033	1301	2334
	Total	1200	1600	2800
20	Incorrect	230	374	604
	Correct	1560	1616	3176
	Total	1790	1990	3780
24	Incorrect	276	569	845
	Correct	1524	1631	3155
	Total	1800	2200	4000

parameters should first be computed for both groups. If an item has a difficulty parameter with absolute value above 1.0, it should be discarded from further analysis. Estimation of the item parameters poses some problems, of course: since some of the individual items are biased, one may suspect the estimated itemparameters to be unreliable.

Both programs are type I unreliable: they tend to indicate unbiased items as biased more easily than the other way around. Therefore, one shouldn't use the technique in a 'diagnostic' setting, in which the biasedness of the items is explained by characteristics of the groups. BIAKOK and BIASIT may both be unrestrictedly used in the construction of unbiased test sets.

One may safely conclude that the iterative algorithm is also applicable with small test lengths, especially if the class attribution mechanism of Kok is used. BIASIT performs slightly better for the longer test lengths (20 and 24). The best approach would be to use BIAKOK for test lengths below 20 and BIASIT for longer test lengths. Generally, one would like an item bias detection procedure along the lines of van der Flier to take the following steps:

1. Determine mean and standard deviation of the (estimated) ability distribution (total score). Big differences between the groups will make any outcome unreliable. In the following steps one should check on this difference if items are left out from the total score.

2. Determine the difficulty parameters of the items. Remove those items that are very difficult or very easy and for which both groups perform more or less the same.

3. Test on nonuniform bias. If model (1) fits but (3) doesn't, ability and group membership are interacting and van der Flier's method is not applicable.

4. For test length below 20 use a fixed assignment method, for longer test length use a random assignment method.

5. Compute p-values for the chosen number of iterations.

Remark that in step 3 for test length below 20 one would wish to get G^2's for all iterations. We recommend the programs to be adapted in such a way that these steps may be taken more easily.

Acknowlegments

I gratefully acknowledge the support of Don Mellenbergh, who stimulated me to carefully describe the experiments. Frank Kok en Leo van der Weele suggested many improvements in the manuscript.

References

Adèr, H.J. (1982). BIASIT, Iterative program to select biased items. In: *Programma beschrijvingen Zielery.* Programma 4.1.6. Rekendienst Subfaculteit Psychologie, Vrije Universiteit, Amsterdam.

Andersen, E.B. (1980). *Discrete Statistical models with social science applications.* Amsterdam: North-Holland Publishing Company.

Angoff, W.H. & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. *Educational and Psychological Measurement*, **34**, 807-816.

BMDP Statistical Software Manual (1985). Multiway Tables - Analysis 4F. Berkeley: University of California Press

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

Bishop Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Mass.: MIT Press.

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

Cardall C. & Coffman, W.R.(1964). *A method for comparing performance of different groups on the items in a test* (RM 64-61). Princeton, N.J.: Educational Testing Service.

Echternacht, G. (1974). A quick method for determining item bias. *Educational and Psychological Measurement*, **34**, 271-280.

Fienberg, S.E (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge, MA: MIT press.

Van der Flier, H., Mellenbergh, G.J., Adèr, H.J. & Wijn, M. (1984). An Iterative Item Bias Detection Method. *Journal of Educational Measurement*, **21**, 131-145.

Hunter, J.E. (1975). *A critical analysis of the use of items means and item-test correlations to determine the presence of content bias in achievement test items.* Paper presented at the National Institute of Education conference on Test Bias. Annapolis, MD.

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

Kelderman, H. (1986). *Item Bias Detection using the Loglinear Rasch Model (Observed and Unobserved Subgroups)* (Research Report 86-2). University of Twente, Division of Educational Measurement and Data Analysis.

Kok, F. G. (1982). *Het partijdige item.* Amsterdam: University of Amsterdam, Department of Psychology.

Kok, F. G. (1986). BIAKOK, Iterative program to select biased items. In: *Programma beschrijvingen Zielery* (Programma 4.1.7.). Rekendienst Subfaculteit Psychologie, Vrije Universiteit, Amsterdam.

Kok, F.G. (1989). *Vraagpartijdigheid (Methodologische verkenningen).* Doctor's Thesis. Amsterdam: University of Amsterdam, Department of Psychology.

Lucassen, W. & Evers A. (1984). *Oorzaken en gevolgen van sexe-partijdigheid in de differentiële aanleg testserie Dat'83.* Paper held at the Dutch Psychologists Conference 1984, Ede.

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias *Journal of Educational Statistics*, **10**, 133-142.

Nungester, R.J. (1977). *An empirical examination of three models of item bias.* PhD Thesis, Florida State University.

Shepard, L., Camilli, G. & Averill, M. (1981). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. *Journal of Educational Statistics*, **6**, 317-375.

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

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, **16**, 433-449.

Ontvangen: 21-07-1989
Geaccepteerd: 25-02-1991