DOCUMENT RESUME

ED 383 739 TM 023 205

AUTHOR	van den Bergh, Huub; And Others
TITLE	Differential Item Functioning from a Multilevel Perspective.
PUB DATE	18 Apr 95
NOTE	30p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Cluster Analysis; Estimation (Mathematics); *Identification; *Item Bias; *Measurement Techniques; Models; Regression (Statistics); *Sex Differences; Test Bias; *Test Items
IDENTIFIERS	*Multilevel Analysis

ABSTRACT

The term differential item functioning (DIF) refers to whether or not the same psychological constructs are measured across different groups. If an item does not measure the same skills or subskills in different populations, it is said to function differentially or to display item bias. A multilevel approach to DIF is proposed. In such a model, the dependency between observations due to cluster effects is explicitly taken into account. Results of a multilevel logit model and of a multilevel logistic regression model are compared with results of analogous unilevel models. The procedure is illustrated with data from a national assessment of geography performed with respect to gender bias. Each of the 294 items was answered by an average of 2,161 respondents. Analysis supports the use of multilevel models, which have the advantage of accounting for cluster effects in data from a hierarchical population. DIF does seem to be more stable according to multilevel models than to unilevel models. Five tables and four figures present analysis results. Appendixes provide parameter estimates. (Contains 19 references.) (SLD)

Differential item functioning from a multilevel perspective

Huub van den Bergh[1], Hans Kuhlemeier[2], Johan Wijnstra[2]

[1] Utrecht University
Centre for Language and Communication
Trans 10
3512 JK Utrecht
The Netherlands

[2] National Institute for Educational Measurement [Cito])
Postbox 1023
6801 MG Arnhem
The Netherlands

1. Introduction

The term 'differential item functioning' refers to whether or not the same psychological constructs are measured across distinguished groups, for instance, males and females. If it can be shown that an item does not measure the same (sub)skill(s) in both populations, than such an item is said to function differentially, which is also sometimes referred to as item bias (Kok, 1988), or measurement bias (Millsap & Everson, 1993)[1]. Suppose, a person with ability A and group characteristic G has response R on an item. This item is considered to function differently if: $f\,(R \mid A, G) \neq f\,(R \mid A)$, that is an item functions differently if the response (R) is a function of both the ability (A) as well as the group characteristic (G), for at least one of the distinguished subpopulations, in stead of a function of ability (A) only.

To test whether an item functions differently, several types of model can be used, which all have specific advantages and disadvantages (Millsap & Everson, 1993). However, to detect differential item functioning (DIF) is one thing, to explain DIF is something quite different. It has been proven difficult to explain why some items function differently in certain subpopulations and others do not (Scheuneman & Steinhaus, 1987). Generally explanations have been put forward in terms of linguistic, cultural and school related factors (Taylor & Taylor, 1990; Uiterwijk, 1994).

Several suggestion have been made to explain the failure in pinning down the causes of DIF (cf, Schmitt, Holland & Dorans, 1992). One possible course of failure concerns an alleged lack of stability of DIF. Shaggs and Lissitz (1988), for instance, comparing several DIF indices in a simulation study with 92 items, found that across 33 replications not one item functioned differently in all cases. Only seven items --out of 92-- functioned differently in at least 20 of 33 replications, whereas only 13 items were never flagged as functioning differently. Therefore, one of the reasons for these deficiencies in the explanation of DIF might be that --at least for some items-- there is nothing to explain: some items that do not function differently actually might be incorrectly

[1] Note that a limited definition of item bias is presented, as only conditional DIF is considered as such. Unconditional DIF is not treated in this paper.

flagged as DIF. That is, we can not completely rule out the possibility of type I errors (H_0 is true but rejected).

A possible explanation may be that the procedures used to detect DIF are to sensitive, because the sample of students in studies on DIF is rarely a simple random sample (and should not be considered as such in the analysis). In the fast majority of studies on DIF, first a sample of school's is drawn, and at a next stage students within schools are sampled. It is well known that, due to selection, education and the like, students from the same school or class are more alike then students from different schools/classes. Recent estimates of the proportion between class/school variance may range from .1 to .5 (Kuhlemeier & Van den Bergh, 1989; Tate & King, 1994). Therefore, 'students' cannot be considered as independent observations. Usually, in a unilevel analysis no correction is made for this type of design effect. By consequence the true standard errors are underestimated (Fienberg, 1977, p.32), and the testing statistics are inflated (see for instance, Holt, Scott & Ewings (1980) for the $\chi2$ statistic). To avoid this problem we propose a multilevel approach to DIF. In such a model the dependency between observations due to cluster effects is explicitly taken into account (Goldstein, 1987).

2. Two models

In order to detect DIF two multilevel models are compared with their unilevel counterparts. That is, the results of a multilevel logit model and a multilevel logistic regression model are compared with results of analogous unilevel models.

2.1 Logit models

In a unileve, iterative logit procedure (Van der Flier, Mellenbergh, Ader & Wijn, 1984) a crosstable is constructed per item with dimensions Group and Ability. Per cell of this crosstable the logit of the proportion correct is calculated. These logits can be written as:

$$Logit(p_{ag}) = C + ABILITY_a + GROUP_g + ABILITY * GROUP_{ag} \qquad (1)$$
$$a = 1,2, ..., A; \; g = 1, 2, ..., G$$

According to this model an item functions differentially if either the main effect of Group reaches significance or the interaction of Group and Ability. If only the main effect of Group reaches significance, the item functions uniform differentially. Whereas a significant interaction term represents nonuniform DIF.

A well known problem is the construction of ability levels (compare, Millsap & Everson, 1993). Generally speaking these levels are based on the sum of the item scores (Mellenbergh 1982; Van der Flier et al., 1984). But since this sum is made up of items which function different too, it cannot be considered as a unbiased ability indicator. Therefore, the following procedure has been put suggested (Van der Flier et al, 1984):

- the sum score of the test is calculated as the sum of all items minus the item analyzed, and minus the scores of items classified as DIF on a previous iteration;
- the distribution of sum scores is investigated and *A* ability levels are constructed in such a way that the number of students in each ability level is more or less the same;
- the likelihood ratio χ^2 is calculated for every item;

This procedure is repeated until the sum solely consists of items which do not function differently.

Note that students are nested within classes. This is just another way of saying that there is a variance component between classes as well as a variance compor t between students within classes. Suppose, index *j* (*j* = *1, 2, ..., J*) indicates the class, than the corresponding multilevel logit model can be written as:

$$Logit(p_{agj}) = C + ABILITY_a + GROUP_g + ABILITY * GROUP_{ag} + \mu_{0j} \qquad (2)$$

$$a = 1, 2, ..., A \ ; \ g = 1, 2, ..., G \ ; \ j = 1, 2, ..., G.$$

In Equation 2 the random term μ_{0j} indicates the deviation for class j from the constant (C), representing the grand mean. It is assumed that μ_{0j} is normally distributed with mean zero and variance $\sigma^2_{\mu 0j}$. Not represented in Equation 2 are the level 1 --or within school-- residuals; we will return to these later on.

The model can also be written otherwise. The crosstable to be analyzed in the model 2 consists of A (ability levels) times G (groups) cells. Each cell of the crosstable can be indicated by a dummy-variable: X_{ag}. Then the model can be written as:

$$Logit(p_{agj}) = \sum_{a=1}^{a=A} \sum_{g=1}^{g=G} (\beta_a * X_{agj}) + \mu_{0j} \qquad (3)$$

$$h = 1, 2, ..., A \ X \ G; \ j = 1, 2, ..., J.$$

The variables X_{agj} are dummy variables which are turned on --X_{agj} = 1-- if a proportion is observed in the corresponding cell of the crosstable, and are turned off --X_{agj} = 0-- if otherwise. Hence, there are as many dummies as there are cells. Therefore, the fixed parameters --β_{ag}-- are the logits of the proportions correct in each cell. And the last term is a residual score for class j. These residuals are assumed to be normal distributed with E [μ_{0j} = 0].

For each cell a separate level 1 --within class-- variance term is estimated. Hence, a special pattern matrix is needed to indicate the level 1 residuals. Since the level 1 variances are dependent on the parameters in the fixed part of the model[2], the level 1 residuals are binomially distributed. To estimate these level 1 variances a weight matrix is constructed. This weight matrix, which is updated after each iteration, contains the ratio of unity and square root of the expected level 1 variance for each cell of the crosstable (i.e. 1

[2] Var (r_{ag} |j) = p_{agj} (1 - p_{agj})

/ $\sqrt{}\ s^2_{eag}$). Multiplication of this weight matrix with the pattern matrix, which indicates the level 1 variance, results in a known value of the level 1 variance (i.e. unity). That is, the level one variance is unity if, and only if, all cluster variation is accounted for. Extra binomial variation can be interpreted in terms of unmodelled cluster variation (Goldstein, 1991).

The main effects of group, ability as well as the interaction effect of group X ability can be tested using a contrast matrix[3]. This provides a testing statistic which is asymptotically χ^2 distributed.

2.2. Logistic regression models

A procedure related to the iterative logit model is the detection of DIF by means of a logistic regression model (Swaminathan & Rogers, 1990). The main difference between both methods pertains to the ability indicator. In the logit model ability levels are constructed whereas in the logistic regression model the ability is indicated by the sum of the unbiased items (for which the same procedure is followed as for the logit model above).

To explain the multilevel logistic regression procedure we must make a distinction between students and classes. Note that students are nested within classes. Suppose, Y_{ij} is the response of student *i* (*i* = *1, 2, ..., N_j*) in school *j* (*j* = *1, 2, ..., J*). The model to be analyzed can be written as:

$$Logit(Y_{ij}) = \beta_0\, X_0 + \beta_1 * AB_{ij} + \beta_2 * GR_{ij} + \beta_3 * AB_{ij} * GR_{ij} + \mu_{0j} \qquad (4)$$
$$i = 1, 2, \ldots ,N_j;\ j = 1, 2, \ldots, J.$$

The model in Equation 4 consists of four fixed parameters and a random term. The fixed parameters concern the constant (β_0), the main effect of ability (β_1), the main effect of group (β_2) and the interaction of ability and group (β_3). Of

[3] $\chi^2 = \beta\, C^T\, (C\ \Sigma\ C)^{-1}\, C\, \beta^T$

course, DIF is detected if either there is a main effect of group and/or an interaction between group and ability. The first parameter represents uniform DIF, whereas the second indicates nonuniform DIF.

The level 1 residuals --i.e. the deviation for student *i* in class *j*-- can be denoted as e_{ij}. These residuals are binomially distributed. Therefore, the same type of weight matrix can be used as is the case of the logit model. This results in an a priori known value (unity) of this variance component if all cluster variation is accounted for.

2.4. Some considerations

In Equations 3 and 4 a multilevel logit and a dito logistic regression model are specified. Both models have in common that there is one random term to represent the variance between classes. This tantamount to saying that the between class variance is homogeneous, that is, the between class variance does not depend on the ability of the students. In view of the results of studies on school effectiveness this seems a rather gross simplification. Therefore we can extent the random part of both models with variance terms. For instance, one for each ability level in the logit model, or specify that the regression from ability on the item score (β_1) is random over classes in the logistic regression model. These extensions of Equation 3 and 4 are represented in Equation 5 and 6 respectively:

$$Logit(p_{agj}) = \sum_{a=1}^{a=A} \sum_{g=1}^{g=G} \beta_{ag} * X_{agj} + \sum_{a=1}^{a=A} \mu_{aj} \qquad (5)$$

$$a = 1, 2, ..., A;\ g = 1, 2, ..., G;\ j = 1, 2, ..., J.$$

In Equation 5 there are as many variance terms for the differences between classes as there are ability levels.

$$Logit(Y_{ij}) = \beta_0 X_0 + \beta_1 * AB_{ij} + \beta_2 * GR_{ij} + \beta_3 * AB_{ij} * GR_{ij} + [\mu_{0j} + \mu_{1j} * AB_{ij}] \qquad (6)$$

$$i = 1, 2, ..., N_j; j = 1, 2, ..., J.$$

Note that Equation 6 results in a test of the assumption that the variance between classes is not homogenous[4].

Obviously, the multilevel logistic regression model has more power than the multilevel logit model, as the information of differences in ability is used --in stead of regarded as unordered categories as is done in the logit-analysis. The regression model has, however, an additional advantage. In the logit model the standard errors of the variance between classes are a function of the number of observations in each cell (Snijders & Bosker, 1990). If the number of observations per class per cell decreases, the standard errors increase. Hence, if two groups and three ability levels are distinguished, and class size varies from 20 to 30, one is left with three to five students per group per ability level to

[4] If we start from Equation 6, with the idea that the regression coefficient for the effect of ability varies between classes we get

$$Logit(Y_{ij}) = \beta_0 X_0 + \beta_{1j} * AB_{ij} + \beta_2 * GR_{ij} + \beta_3 * AB_{ij} * GR_{ij} + \mu_{0j} \qquad (6a)$$

Note that β_{1j} is indexed *j* in order to show that the coefficient may take different values for different classes. Now we can write β_{1j} as deviation from the population regression coefficient, say γ_{10}. This gives

$$\beta_{1j} = \gamma_{10} * AB_{ij} + \mu_{1j} \qquad (6b)$$

Substitution of Equation 6b in 6a leads to

$$Logit(Y_{ij}) = \beta_0 X_0 + \gamma_{10} * AB_{ij} + \beta_2 * GR_{ij} + \beta_3 * AB_{ij} * GR_{ij} + [\mu_{0j} + \mu_{1j} * AB_{ij}] \qquad (6c)$$

A result which is equal to Equation 6. At class level two random parameters are estimated: μ_{0j}, and μ_{1j}. As μ_{1j} is multiplied by AB_{ij}, the between class variance is a function of AB^2_{ij}.

estimate the between class, or level 2 variance (for each combination of ability and group). This limited number of observations leads to serious power problems. If the number of residual scores to be estimated diminishes the number of students per class per type of residual score increases. This can be accomplished in various ways. For instance, one may estimate only one variance term per ability level (which results in six to ten students per residual score in the example above), or estimate only different variance terms per group (which results in ten to fifteen students per residual score in the example above), in stead of a variance component for each combination of ability level and group. Note, that in view of the results of school effectiveness studies the former method is to be preferred over the latter.

In the logistic regression model such problems do not occur, as ability is considered to be a continuous variable. Therefore, only one parameter extra is needed to model differences in variance with ability level.

3. Data and design

Part of the data of a national assessment on geography (Kuhlemeier, Van den Bergh, Notté, Wagenaar, Verstralen, & Cappers, 1994) were analyzed with respect to gender bias; more than 13000 students (age ± 15) from 625 classes took at the start of the ninth grade and at the end of the ninth grade a core test with multiple choice items. For each school type or track this core tests consists of two (partly overlapping) subtests. In total 147 items were analyzed. Each item was answered (on average) by 2161 respondents (dependent on the school type the number of respondents vary from 1235 to 2633). In Table 1 the allocation of tests to students is presented.

--INSERT TABLE 1 ABOUT HERE--

Since every student took two tests --although half of the students took the same test twice-- in total 294 items were analyzed.

4. Results

For both the unilevel and multilevel logit model three ability levels were constructed for each item: low, medium and high achievers, each category containing about one third of the total number of students. Figure 1 presents an example of an unbiased item according to either the unilevel or the multilevel logit model.

--INSERT FIGURE 1 HERE--

As can be seen in Figure 1A through 1C the mean logits for males do not greatly considerably; the (logit of the) proportion correct only depends on the ability and not on either the gender or the interaction between gender and ability. Therefore, this item is classified as unbiased.

Note that there are slight differences between the estimated mean logits in the unilevel on the hand and both multilevel logit models on the other hand. This demonstrates that the mean of the class means per ability level (and gender) does not equal the mean of the students per ability level (and gender).

The second aspect to be noted in Figure 1B and 1C are the 80% confidence intervals. These are based on the estimated between class variance. In Figure 1B it is assumed that the variance ($\sigma^2_{\mu j}$) is the same for all three ability levels (see Equation 3), whereas in Figure 1C the between class variance ($\sigma^2_{a\mu j}$; *a* = *1, 2, 3*) is allowed to vary freely over the three ability levels (see Equation 5). As can be seen the differences in logits between classes are clearly larger for the low ability students than for the high ability students. The second multilevel logit model clearly fits better to the data than the first multilevel logit model ($\chi^2 = 14.1$; df = 5).

As items that function differently are more interesting, we will discuss one item that functions differently according all analyses with a logit model in more detail (in Appendix 1 the parameter estimates are presented).

--INSERT FIGURE 2 ABOUT HERE--

As can be seen in Figure 2 --which is based on the estimates in appendix 1-- males generally outperform females (χ^2 = 19.2; df = 1). Especially the differences in the low and high ability group are striking (the testing statistic χ^2 for the interaction affect equals: 21.7; df = 2). Hence, the DIF is nonuniform.

In Figure 3 and 4 an item is plotted which does not function differently and an item is plotted which does show DIF according to the multilevel logistic model. In Figure 3 the two lines for the (logits of the) probabilities for males and females differ only slightly and, therefore, the confidence intervals show overlap.

--INSERT FIGURE 3 & 4 ABOUT HERE--

Figure 4A and 4B plots an item which shows nonuniform DIF; low ability females outperform low ability males, whereas high ability males outperform high ability females. Both figures differ as to the confidence intervals (see Appendix 2 for parameter estimates). In Figure 4A just one random parameter is estimated (see Equation 4), whereas in Figure 4B, the between class variance is allowed to vary freely with the ability level of the students (see Equation 6). Obviously the latter model clearly fits the data batter than the former one (χ^2 = 23.6; df = 1). As can be seen in Figure 4B, compared to the middle of the ability scale, the between class variance is rather large at both extremes. (The same observations can be made in Appendix 2. Note that the between class, or level 2 variance is a function of ABILITY^2; see also note 4).

Note that, the differences between classes are, again, relatively large compared to the differences related to gender. Herefrom, one might pose the hypothesis that DIF is in some way related to the instruction the students received. We will return to this hypothesis later on.

Although the figures above give an impression of some of the results, a comparison of the unilevel versus the multilevel approaches cannot be based on examples. Therefore, we turn to the tables below, in which the number of items which do and which do not show DIF per method are presented.

We first compare the unilevel and multilevel logit model (in the last model differences in variance over ability levels were allowed; see Equation 5). In Table 2 the number of biased and unbiased items ($p < .01$) are cross classified.

--INSERT TABLE 2 ABOUT HERE--

As can be seen in Table 2, the majority of items is classified identically as either not showing DIF (217) or as showing DIF (24) according both models. Nevertheless a substantial number of items (53) is flagged DIF by only one of the models. As expected beforehand, the number of items functioning differently according the unilevel model clearly exceeds the number of DIF items according the multilevel model. That is, 44 items show DIF in the unilevel analysis but not in the multilevel one, whereas (only) nine items exhibit DIF in the multilevel analysis but not so in the unilevel analysis.

The comparison of both logistic regression models (unilevel versus multilevel with two random parameters at class level) provides the same results (see Table 3).

--INSERT TABLE 3 ABOUT HERE--

Again the majority of the items is classified identically in both types of analysis (280). Nevertheless, eleven items were shown to function differently in the unilevel analysis but not in the multilevel analysis, whereas three items proved to function differently only in the multilevel analysis. Again, as expected, the number of items exhibiting DIF in a unilevel model exceeds the number of items exhibiting DIF according a multilevel model.

The items flagged DIF in a unilevel model, but not in their multilevel counterparts, all seem to have one thing in common: the between class variance is relatively large (.15 or higher). Especially for these items there are --of course-- large differences between both type of models. Note, however, that a relatively large between class component indicated DIF by no means.

Remember that, earlier in this section --as well as in the introduction-- it was hypothesized that DIF might be a reflection of instructional practices. Since, there are two measurement occasions, we are able to develop this hypothesis a bit further, if we concentrate on which show DIF only at the start (and not at the end), and items which exhibit DIF only at the end of the ninth grade (and not at the start). These items might provide cues to causes of DIF. Therefore, in Table 4 the results per measurement occasion are presented.

--INSERT TABLE 4 ABOUT HERE--

From Table 4 it appears that in the unilevel logit model 34 items show DIF on both occasion, and 113 do not. This does not imply that the same 34 items function differently on both occasions. On the contrary, only 18 of these 34 function differently both at the beginning and at the end of the ninth grade. Therefore 16 items function differently only at the start or at the end of this grade.

As to the other three models (the multilevel logit model and both logistic regression models), the number of DIF items is clearly lower than for the unilevel logit model (as was already shown in Table 2). However, the percentage of items which show DIF on both occasions is somewhat higher then for the logit model. It also appears that the proportion of items which is biased at both occasions is somewhat higher for the multilevel models compared to the unilevel counterparts.

Analysis of the lower part of Table 4[5] shows a main effect of unilevel versus multilevel ($G^2 = 8.48$; df = 1). Hence, the proportion of items which show DIF on both occasions is lower in a unilevel model than in a multilevel model. Hence, a multilevel model provides a larger stability in DIF over time.

If we concentrate on the multilevel part of Table 4, it appears that in the course of one year four items in the logit model and three items in t!.e logistic regression model show DIF only on the first occasion. It is assumed that, due to education the different functioning is removed from these items. Take, for instance, item A in Table 5. This item proved nonuniform DIF on the first occasion only. That is, no difference for high ability students was found, but medium and low ability males outperformed medium and low ability females, and the difference between both groups decreased with ability. The item concerns the application of knowledge --one has to know the difference between eastern and western latitude and southern and northern longitude. We hypothesize that boys have a higher chance to be confronted with situations in wh:ch this kind of knowledge is relevant. For instance, in scouting or something like that. But as soon as all the students are taught the difference between latitude and longitude, the initial differences disappear.

--INSERT TABLE 5 ABOUT HERE--

The second item (B) in Table 5 only DIF showed at the second measurement. Low ability females had a higher chance of providing the right answer at the and of the third grade, whereas there was no difference between males and females at the start of the third grade --males and females performed equally poor. Perhaps, the item functions different, as only knowledge of what is meant by expressions like 'Rome' or 'Brussels' cannot solve the problem; one needs the provided contextual information as well. Since, females are better readers, it can be hypothesized that this item functions differently because of the relatively poor

[5] The analyses was done by means of a unilevel logit model with the number of common biased items as dependent variable and the total number of biased items as number of observations.

reading skills of low ability males. Hence, we hypothesize that in order to arrive at the correct answer one has to know the institutions settled in Rome and Brussels, but in addition one has to read the question well. So, the item not only appeals to certain content knowledge, but also to reading skill.

5. Discussion

It has been shown that differential item functioning, or item bias, can be detected by means of multilevel models. If the data come from a hierarchical population, as is the case in many educational studies, multilevel models have the advantage of explicitly accounting for cluster effects. Furthermore, heterogienity of variance between classes is relatively easy to model, and therefore, the model provides a better fit to the observed data.

It has been shown, by means of an example, that in a multilevel analysis less items can be proven to function differently. Moreover, DIF does seem to be more stable according to multilevel models then it seems to be according to unilevel models. Therefore, multilevel models seem better equipped for a proper assessment of DIF.

The items which show DIF in a unilevel model share a relatively large between class variance is relatively large. If the between class variance is large, the differences between both types of model are highlighted.

The between class variance of most of the DIF items is substantial. From this observation it was hypothesized that perhaps the bias of some items can be attributed to educational practice. The design of the study allows for a comparison of the DIF of the same items --taken by the same students-- at the start and at the end of the ninth grade. It was concluded that during the school year some items loose their different functioning, whereas others become to functioning differently. However, the majority of the differentially functioning items --according to either multilevel model-- were flagged DIF on both occasions. Herefrom it can be concluded that educational practice has a meritocratic effect --i.e. neutralizes DIF-- as a DIF inducing effect as well. The effects of educational practice are not as simple as we sometimes would like them to be. As education seems to have some effect on only part of the items, it can be concluded that the causes of DIF are multifactorial. Perhaps the DIF for some items are attributable to the way the subject matter was taught, whereas for other items DIF might reflect different experiences not directly related to education.

6. References

Bosker, R.J. & Snijders, T.A.J. (1990). Statistische aspecten van multinivau onderzoek (Statistical aspects of multilevel studies). Tijdschrift voor Onderwijsresearch, 15, 317-329.

Cronbach, L.J. (1976). Research in classrooms and schools: formulations of questions, designs and analysis. Occasional paper: Stanford Evaluation Consortium.

Flier, H. van der, Mellenbergh, G.J., Ader, H.J. & Wijn, M. (1984). An iterative item bias detection method. Journal of Educational Measurement, 21, 131-145.

Fienberg, S.E. (1977). The analysis of cross-classified categorical data. Cambridge: MIT-Press.

Goldstein, H. (1987). Multilevel models in educational and social research. London: Charles Griffin.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. Biometrika, 78, 45-51.

Holt, D. Scott, A. & Ewings, P. (1980). Chi-squared tests with survey data. Journal of the Royal Statistical Society, Series A, 143, 303-320.

Kok, F. (1988). Vraagpartijdigheid: Methodologische verkenningen (Item bias: Methodological research). Amsterdam: University of Amsterdam (diss.).

Kuhlemeier, H. & Bergh, H. van den, (1989). National assessment of language performance: a feasibility study in secondary education]. Arnhem: Cito.

Kuhlemeier, H., Bergh, H. van den, Notté, H., Wagenaar, H., Verstralen, H. & Cappers, R. (1994). Balans van het aardrijkskunde-onderwijs in het derde leerjaar van het voortgezet onderwijs [Balance of geography education in the third grade of secondary education]. Arnhem: Cito.

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-108.

Millsap, R.E. & Everson, H.T. (1993). Methodological review: statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17, 297-334.

Scheuneman, J.D. & Steinhaus, K.S., (1987). A theoretical framework for the study of item difficulty and discrimination. Princeton: Educational Testing Service.

Schmitt, A.P., Holland, P.W. & Dorans, N.J., (1992). Evaluating hypotheses about differential item functioning. Princeton: Educational Testing Service.

Shaggs, G. & Lissitz, R.W. (1988, April). Consistency of selected item bias indices: Implications of another failure. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Tate, R.L. & King, F.J. (1994). Factors which influence precision in school level IRT ability estimates. Journal of Educational Measurement, 31, 1-15.

Taylor, I. & Taylor, M.M., (1990). Psycholinguistics. Learning and using language. New Jersey: Prentice Hall.

Uiterwijk, H., (1994). Eindtoets Basisonderwijs [The Test at the End of Primary Education]. (diss.) Arnhem: Cito.

Van der Flier, H., Mellenbergh, G.J. Ader, H.J. & Wijn, M. (1984). An iterative item bias detection method. Journal of Educational Measurement, 21, 131-145.

Table 1 Allocation of testbooks to students

TYPE	CLASS	STUDENT	OCC1	OCC2
1	1	1	A	A
1	1	2	A	B
1	1	3	B	A
1	1	4	B	B
1	1	5	A	A
1	1	6	A	B
.	.	.	.	.
.	.	.	.	.
1	2	1	A	A
.	.	.	.	.
.	.	.	.	.
2	1	1	C	C
2	1	2	C	D
.	.	.	.	.
.	.	.	.	.
3	1	1	E	E
3	1	2	E	F
.	.	.	.	.
.	.	.	.	.

Figure 1. Three plots of the same item which showed no DIF according to a unilevel (A; Equation 1), a multilevel logit model with the restriction of homogeneity of variance over ability levels (B; Equation 3), and a multilevel model with different between class variance (C; Equation 5; (Equation 5; ▨: males; ☼: females; dashes lines 80% confidence intervals).

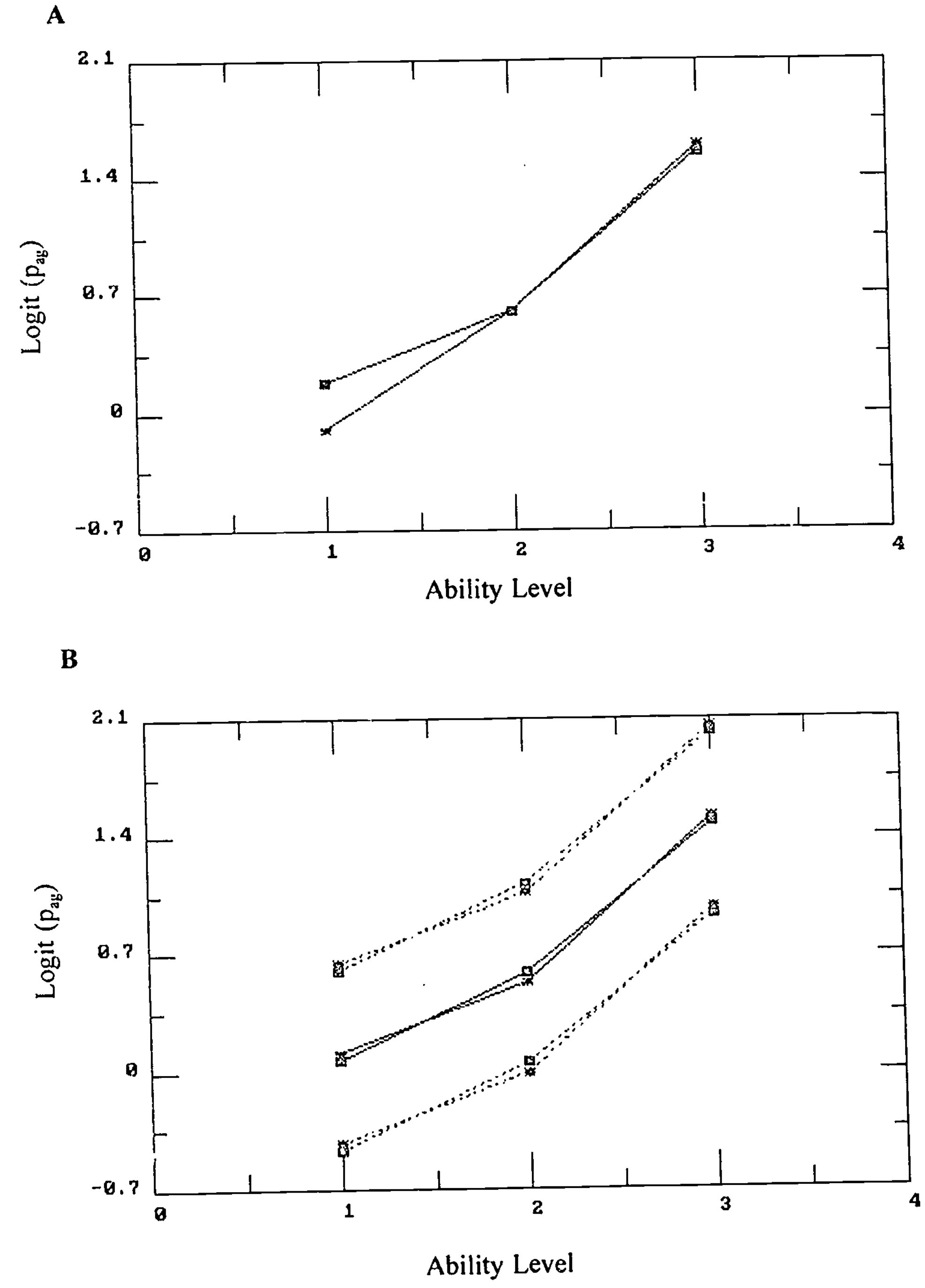
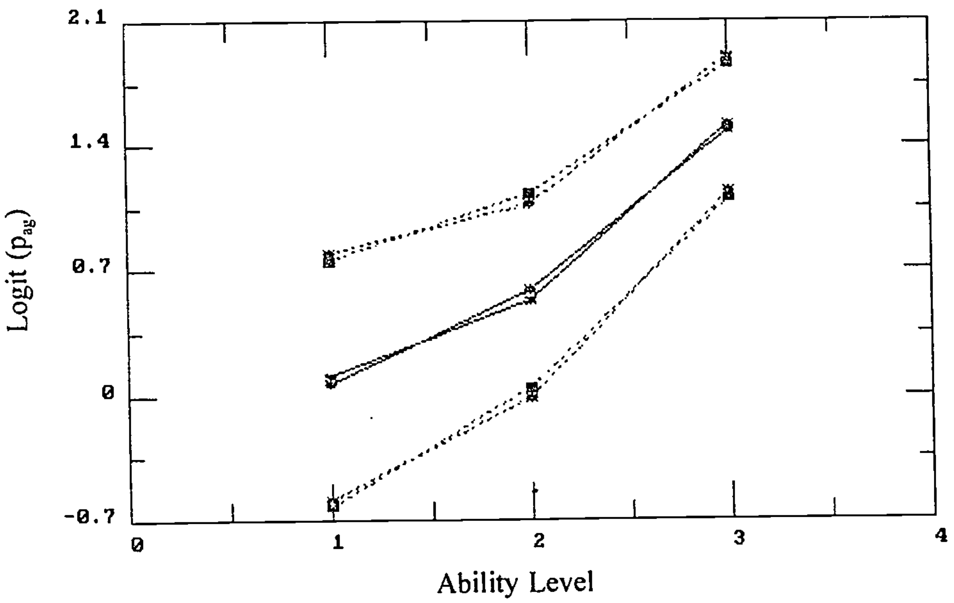


Figure 1 (continued)

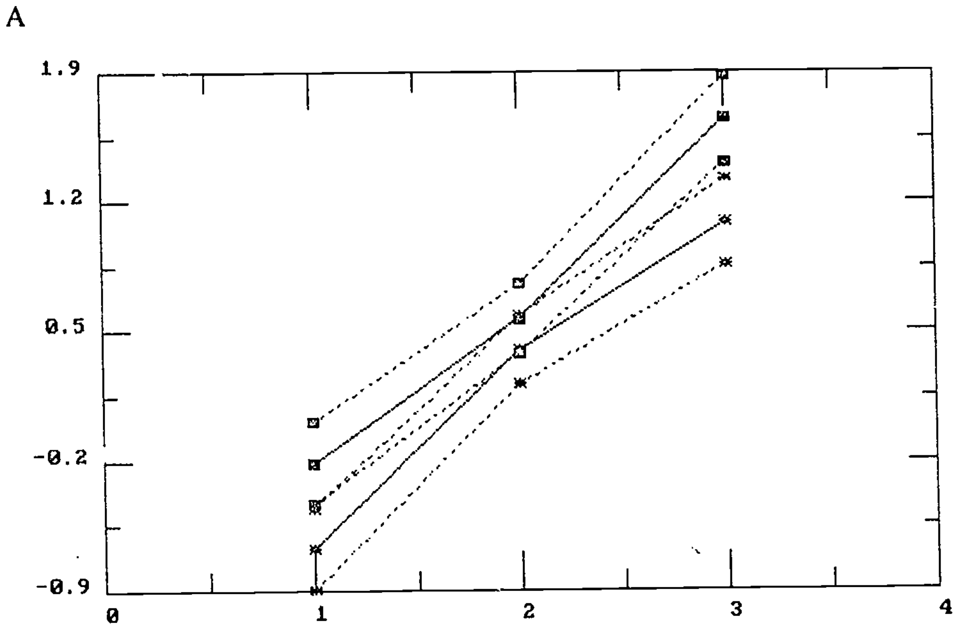
C



Janny and Pien travel from Utrecht to Rome for Holidays. They go by bus. About halfway they spent the night in a big city. Which city could that have been?

a. Berlin
b. Koln
c. Munich
d. Paris

Figure 2. A plot of a differential functioning item according to a multilevel logit model (Equation 5; ■: males; ☼: females; dashes lines 80% confidence intervals; see also Appendix 1).



Item:

German federal state	Inhabitants in millions	Area in km^2
Bayern	10.99	7.55
Niedersachsen	7.17	47.43
Baden-Württemberg	9.37	35.75
Nordrhein-Westfahlen	16.79	34.07

Which two German federal states have the highest population density

a. Bayern and Niedersachsen
b. Bayern and Nordrhein-Westfahlen
c. Niedersachsen and Baden-Württemberg
d. Nordrhein-Westfahlen and Baden-Württemberg

Figure 3 An example of an item which showed no DIF according a multilevel logistic model (dashed lines 80% confidence intervals; ▨: males; ☼: females; dashes lines 80%).

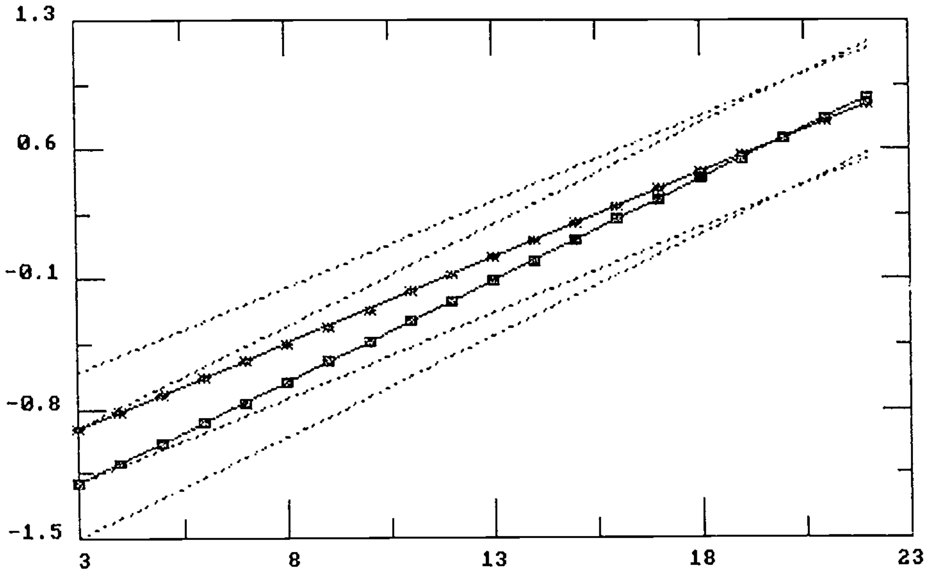


Figure 4 An example an item which showed DIF according a multilevel logistic model with (A) one random term at class level (see: Equation 4) and (B) a with two random terms at class level (dashed lines 80% confidence intervals; ■: males; ☼: females; dashes lines 80%).

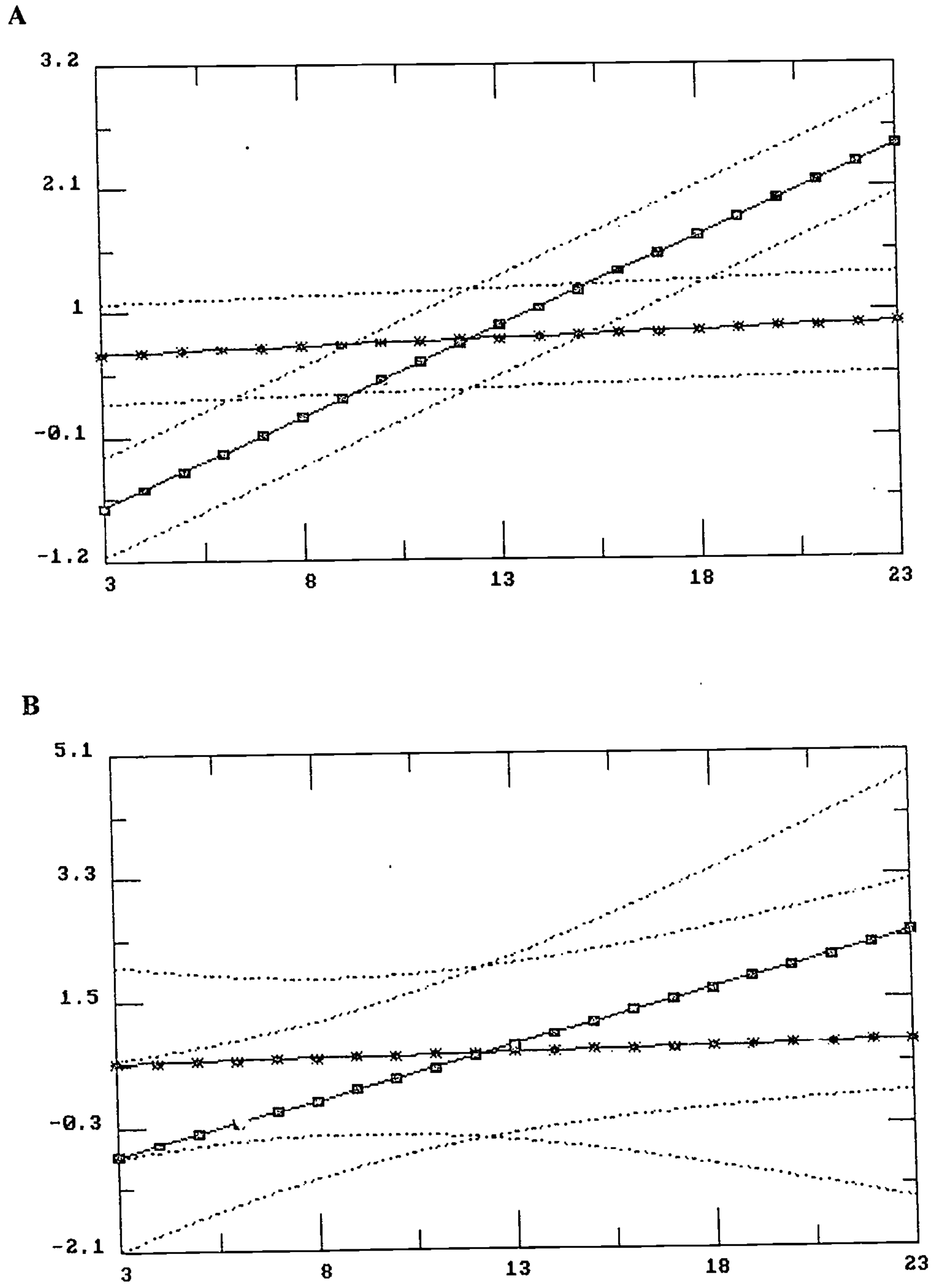


Table 2 Number (%) of items showing DIF according a logit model: unilevel versus multilevel (Totals).

LOGIT MODEL: TOTALS		UNI-LEVEL	
	DIF:	NO	YES
MULTI-LEVEL	NO	**217** (73.8)	**44** (15.0)
	YES	**9** (3.1)	**24** (8.2)

Table 3 Number (%) of items showing DIF according logistic regression: unilevel versus multilevel (Totals).

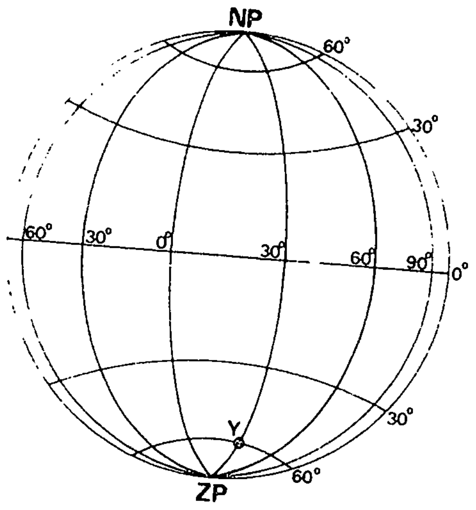
LOGISTIC REGRESS. TOTALS		UNILEVEL	
	DIF:	NO	YES
MULTI-LEVEL	NO	**244** (82.9)	**11** (3.7)
	YES	**3** (1.0)	**36** (12.2)

Table 4 Number of items showing DIF per type of model per occasion (the number of corresponding items)

		Unilevel		Multilevel	
Model	Result	T1	T2	T1	T2
Logit	NO DIF	113	113	126	125
Regress.	NO DIF	125	122	130	125
Logit	DIF	34	34 (18)	21	22 (17)
Regress	DIF	22	25 (14)	17	22 (14)

Table 5 An item which showed DIF at the first measurement occasion only (A) and an item which showed DIF at the second measurement occasion only (to both multilevel models in either case).

A What are the coordinates for position Y



A 30^0 Northern latitude; 60^0 Western longitude
B 60^0 Northern latitude; 30^0 Western longitude
C 30^0 Southern latitude; 60^0 Eastern longitude
D 60^0 Southern latitude; 30^0 Eastern longitude

B Compared to Nortern Italy Southern Italy is poor. Southern Italy has a lack of fertile farming-ground and insufficient industry to keep the people employed.

Since the second world war 'Rome', with the indispensable aid of 'Brussels', has taken action to improve the situation.

Which institutions are meant by 'Rome' and 'Brussels'?

A 'Rome': capitol of Italy; 'Brussels': capitol of Belgium
B 'Rome': Italian government; Brussels': Belgian government
C 'Rome': capitol of Italy; 'Brussels': European community
D 'Rome': Italian government; 'Brussels': capitol of Belgium

Appendix 1 Parameter estimates for the multilevel logit model with three level 2 residuals (Equation 5; standard errors between brackets).

	Fixed effects ($ß_{hg}$)
♀_Ability Lev. 1	-0.672 (0.081)
♀_Ability Lev. 2	0.405 (0.093)
♀_Ability Lev. 3	1.093 (0.118)
♂_Ability Lev. 1	-0.210 (0.101)
♂_Ability Lev. 2	0.571 (0.109)
♂_Ability Lev. 3	1.645 (0.121)

Between class covariance matrix [correlations above diagonal]

	AL1	AL2	AL3
Ability Lev. 1	0.030 (0.011)	[.62]	[.34]
Ability Lev. 2	0.016 (0.004)	0.021 (0.009)	[-.76]
Ability Lev. 3	0.011 (0.006)	-0.020 (0.009)	0.033 (0.014)

	Within class variances
♀_Ability Lev. 1	0.990 (0.057)
♀_Ability Lev. 2	0.988 (0.069)
♀_Ability Lev. 3	0.984 (0.073)
♂_Ability Lev. 1	0.986 (0.072)
♂_Ability Lev. 2	1.000 (0.078)
♂_Ability Lev. 3	0.971 (0.065)

Appendix 2 Parameter estimates (se) for a biased item according a multilevel logistic regression model (see also Figure 2)

	MODEL	
	A	B
	Fixed effects	
Intercept	0.602 (0.320)	0.605 (0.353)
Sum'	0.013 (0.022)	0.013 (0.024)
Gender	-1.802 (0.503)	-1.789 (0.506)
Sum' * Gender	0.147 (0.034)	0.146 (0.034)
	(Co)variances between classes	
$S^2_{\mu 0j}$	0.119 (0.044)	1.686 (0.535)
$S_{\mu 0j, \mu 1j}$		-0.104 (0.069)
$S^2_{\mu 1j}$		0.012 (0.005)
	Variances within classes	
S^2_{eij}	0.969 (0.028)	0.988 (0.028)