ED 308 239 TM 013 631

AUTHOR	Kelderman, Henk; Macready, George B.
TITLE	Loglinear-Latent-Class Models for Detecting Iter Bias. Project Psychometric Aspects of Item Banking No. 36. Research Report 88-10.
INSTITUTION	Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE	Nov 88
NOTE	60p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
AVAILABLE FROM	Bibliotheek, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Foreign Countries; Higher Education; *Latent Trait Theory; Mathematical Models; Multiplication; Statistical Analysis; *Statistical Bias; *Test Bias; Testing Problems; Test Items; Undergraduate Students
IDENTIFIERS	*Item Bias Detection; *Latent Class Models; Log Linear Models; Netherlands

ABSTRACT

The use of loglinear latent class models to detect item bias was studied. Purposes of the study were to: (1) develop procedures for use in assessing item bias when the grouping variable with respect to which bias occurs is not observed; (2) develop bias detection procedures that relate to a conceptually different assessed trait--a categorical attribute; and (3) exemplify the use of these developed procedures with real world data. Models are formulated so that the attribute to be measured may be continuous, as in a Rasch model, or categorical, as in latent class models. The item bias to be studied may correspond to a manifest grouping variable, a latent grouping variable, or both. Likelihood-ratio tests for assessing the presence of various types of bias are described. These methods are illustrated through analysis of a "real world" data set from a study of multiplication items administered to 286 Dutch undergraduates. Bias was related to a manifest grouping variable by giving 143 of the subjects some training in Roman numerals, in which some of the multiplication problems were written. Results indicate that it was possible to explain item bias through differences in item difficulties or error rates across levels of grouping variables. The model represented can be extended to include several observed and unobserved variables. Ten tables present information about the models and findings of the study. A 39-item list of references is included. (SLD)

Loglinear-Latent-Class Models for Detecting Item Bias

Research Report 88-10

Henk Kelderman

department of EDUCATION

Division of Educational Measurement
and Data Analysis

University of Twente

Project Psychometric Aspects of Item Banking No.36

Colofon:
Typing: Lorette Padberg
Cover design: Audiovisuele Sectie TOLAB Toegepaste Onderwijskunde
Printed by: Centrale Reproductie-afdeling

Loglinear-Latent-Class Models for Detecting Item Bias

Henk Kelderman

George B. Macready*

Paper read at the Annual Meeting of the American Educational Research Association, April 5-9, 1988.

* University of Maryland

Loglinear-latent-class models for detecting item bias / Henk Kelderman & George Macready - Enschede : University of Twente, Department of Education, November, 1988. - 52 pages

Abstract

Loglinear latent class models are used to detect item bias. These models are formulated in such a manner that the attribute to be measured or assessed may be continuous (as in a Rasch model) or categorical as in Latent Class (Mastery models). Further, the item bias to be studied may correspond to a manifest grouping variable, a latent grouping variable or both. Likelihood-ratio tests for assessing the presence of various types of bias are described and these methods are illustrated through the analysis of a "real world" data set.

Test items are biased if the item score of equally able examiners from different groups (e.g., race, sex and age) are systematically different. If several items in a test are biased in favor of a specific group, the tesc may lead to an unfair advantage for that group with regard to their assessed level of performance, when its members are compared with members of other groups. This inequity can hopefully be rectified by deleting or improving the biased items.

The basic problem in the detection of item bias is to differentiate between discrepancies in item difficulties across groups which are due to bias as opposed to differences in level on the assessed attribute. Since groups frequently differ on the asseosed attributes, bias and ability are often confounded. For this reason it is hard to tell whether observed differences in probabilities for positive item responses among groups are a result of item bias or differences in ability across the groups. Linn and Drasgow (1987) have shown that neglecting this confounding and deleting items on the basis of differences in group performance can lead to removal of valid items and may, thus, result in poor tests.

Many item bias detection methods have been proposed. Reviews are given by Osterlind (1983); Rudner, Getson and Knight (1980) and Shepard, Camilli and Averill (1981). In the earlier item bias detection methods such as the analysis-of-variance method (Cardal & Coffman, 1964; Cleary & Hilton, 1968; Hoepfner & Strickland, 1972; and Jensen, 1980) and the transformed-item-difficulty methods (Thurstone, 1925; Angoff, 198∠; Angoff & Ford, 1973) there was no rigorous control for differences in ability across groups. In chi-square methods (Scheunemann, 1979; Mellenbergh, 1982; Camilli, 1979; Nungester, 1977; Holland & Thayer, 1986) ability is controlled by comparing item performance for a given total test score. In IRT methods (Lord, 1980; Durovic, 1975) there is control for ability via the person's ability parameter in the model. Items are considered biased if the item parameters vary across group.

Kelderman (1985) proposed the use of a loglinear formulation of the Rasch (1980) model (Cressie & Holland, 1983; Duncan, 1984; Kelderman, 1984; Tjur, 1982) to study item bias. Various aspects of item bias can be modeled by adding parameters to the loglinear formulation of the Rasch model. These parameters can then be estimated using maximum likelihood procedures and hypotheses about the occurrence of item bias can be tested using likelihood-ratio tests of fit. Several grouping variables can be included in the model and various relevant models may be specified by the investigator depending on the particular problem at hand. Used in this way, loglinear models provide a very flexible modeling framework for detecting item bias. In this paper the above mentioned loglinear modeling system is extended. Our purposes are three fold: a) develop procedures for use in the assessment of item bias that may be used when the grouping variable with respects to which bias occurs is not observed, b) develop bias detection procedures that relate to a conceptually different kind of assessed trait, namely a categorical attribute, and c) exemplify the use of these developed procedures with real world data.

Haberman (1979) developed a theory of loglinear modeling that allows for the inclusion of unobserved categorical variables, or latent classes in loglinear models. This theory allows for the study of item bias with respect to unobserved or latent grouping variable. Using this kind of latent-class-loglinear model it became possible to extend the (loglinear) Rasch model to include a latent category dimension. Using this result we formulate a latent class/latent trait model where the assumption of local independence among items, which underlies the model, is conditional on the joint levels of both latent variables (i.e., the level of continuous measured trait and the level of the latent grouping variable). This extended loglinear Rasch model which incorporated a latent grouping variable may have different item difficulties for the various latent groups. If for a certain item the difficulty parameter is larger for one latent group than another, it is concluded that the item is biased with respect to the latent grouping variable.

Item bias detection procedures are also possible when the latent attribute being assessed is categorical. Under such circumstances the relation between latent and manifest variables may be specified through the use of latent class models (Lazarsfeld & Henry, 1968). In this paper, we will deal only with two state latent class models, however the procedures here described are directly applicable to other types of latent class models (e.g., Goodman, 1975 and Dayton and Macready, 1976).

The two state mastery model is particularly appropriate for assessing attributes whose acquisition is assumed to be an "all-none" process in which individuals are of one of two possible latent types : "Masters" (i.e., individuals who have the necessary and sufficient skill/ability to correctly respond to all items which are used to assess the attribute of interest) and "non-masters" (i.e., individuals who do not have the skill/ability to respond correctly to any item within the content domain of interest). However, under this model it is assumed that response "errors" may result in "masters" missing items or "non-masters" responding correctly the items.

Item bias may be investigated within a state mastery modeling framework by studying differences in omission and intrusion error rates across levels of a grouping variable with respect to which bias is suspected. If for a certain item the omission error rates and/or the intrusion error rates differ across groups, the item in question is biased with respect to the grouping variable. As in the case of a continuous measured variable, item bias may be studied with respect to either manifest or latent grouping variables, through the use of latent class loglinear models.

The use of latent grouping variables in the search for item bias, has the advantage of being applicable even when an observed grouping variable is not available. In addition, it allows for the assessment of item bias without tying that bias to any specific variables or set of variables. Thus, it may be possible following the investigation of bias to make a more definitive statements regarding its presence. Finally, the use of latent grouping variables allow an investigator to explore how various manifest grouping variables may be related to latent grouping variables with respect to which bias occurs.

In the next section of this paper, the various variables that are used in modeling are formally presented and the general loglinear model which is of interest is defined. By considering various restricted forms of this general model it is possible to make model comparisons which are useful in the assessment of item bias.

An Overall Loglinear Modeling Framework

Variables which may be Included in Models

In this paper the following types of variables may be included in the models which are considered. First the dichotomously scored responses X_j ($j=1,\ldots,k$) to each of the k test items are included within all models considered. Note that the score of any ith individual, $X_{ij}=\{0,1\}$, is 0 if the jth item is scored as incorrect, and 1 if it is scored as correct. In addition to item responses, these models include two other kinds of variables: the latent variable being measured (or assessed) and the grouping variable with respect to which bias may occur.

The measured (or assessed) variable may correspond to either a continuous or discrete categorical attribute. When this latent variable is continuous, a Rasch model (Rasch, 1980) is assumed to specify the relation between item responses and the level of the measured variable. Within the framework of loglinear modeling, this model must include as an independent variable the total score, $T=X_1+...+X_k$ (see Kelderman, 1984, for a discussion). In the case of an assessed attribute, L, which is categorical, a two state latent class model is assumed to specify the relations between item responses and the latent categories of mastery (i.e., whether an individual is a "master" or "non-master") on the assessed attribute (See Macready & Dayton, 1980, and Bergan, 1983, for general reviews of this class of models, and van der Linden, 1978, for a discussion of how they relate to IRT models).

The variables which are used to model item bias can be either observed or unobserved grouping variables. Such a variable is designated as G when its r levels are actually observed (as in the case of studying sex or race as having a possible biasing effect). Although more than one such variable may be included in these models, only one will be considered in this paper. If a grouping variable is not observed, a latent grouping variable, U, may be included in the model. In general, the number of levels of U is s and must be specified by the investigator. In this paper we will consider U to be dichotomous.

The General Model

Haberman, (1979) presents a general loglinear model which specifies the relations among a set of observable and unobservable categorical variables. Such models explain the structure of the contingency table that is formed by cross classifying the set of variables of interest. This is accomplished by specifying a linear decomposition of the natural log of expected contingency table frequencies. The components which define this dacomposition may include "main" and "interaction" type effects corresponding to various margins (or cells) of the contingency table. If all the types of variables which are mentioned above are simultaneously considered, we have a X_1 x X_2 x ... x X_k x T x G x U x L contingency table with frequencies:

$$f_{x_1 \ldots x_k tgul},$$

$x_1=0,1;\ldots;$ $x_k=0,1;$ $t=x_1+\ldots+x_k;$ $g=1,\ldots,r;$ $u=1,\ldots,s;$ $l=1,\ldots,q.$

The so-called saturated model which contains all possible main and interaction effects among the variables considered above is:

(1) $$\ln m_{x_1 \ldots x_k tgul} = \beta + \beta^{X_1}_{x_1} + \ldots + \beta^{X_k}_{x_k} + \beta^{T}_{t} + \beta^{G}_{g}$$

$$+ \beta^{U}_{u} + \beta^{L}_{l} + \beta^{X_1X_2}_{x_1x_2} + \ldots + \beta^{UL}_{ul} + \beta^{X_1X_2X_3}_{x_1x_2x_3}$$

$$+ \beta^{GUL}_{gul} + \ldots + \beta^{X_1 \ldots X_k TGUL}_{x_1 \ldots x_k tgul}$$

With the constraints

(2) $$\Sigma_{x_1} \beta^{X_1}_{x_1} = 0, \ldots x , \Sigma_{x_k} \beta^{X_k}_{x_k} = 0, \Sigma_{t} \beta^{T}_{t} = 0, \Sigma_{g} \beta^{G}_{g} = 0,$$

$$\Sigma_{u} \beta^{U}_{u} = 0, \Sigma_{l} \beta^{L}_{l} = 0, \Sigma_{x_1} \beta^{X_1X_2}_{x_1x_2} = 0, \Sigma_{x_2} \beta^{X_1X_2}_{x_1x_2} = 0$$

$$x , \ldots., \Sigma_{u} \beta^{UL}_{ul} = 0, \Sigma_{l} \beta^{UL}_{ul} = 0, \ldots., x$$

$$\Sigma_{x_1p} \beta^{X_1 \ldots X_K TGUL}_{x_1 \ldots x_k tgul} = 0, \ldots, \Sigma_{l} \beta^{X_1 \ldots X_k TGUL}_{x_1 \ldots x_k tgul} = 0,$$

where $\{m_{x_1 \ldots x_k tgul}\}$ are the expected cell frequencies obtained under the model and where $\beta^{X_1}_{x_1}$ is the parameter designating the main effect of response x_1 of item one, $\beta^{X_1X_2}_{x_1x_2}$ is the parameter designating the interaction effect of the combination of response x_1 of item one and response x_2 of item two etc.

This general model is an incomplete loglinear latent class model (see Haberman, 1979, p.554). It is termed incomplete because the contingency table contains cells with frequencies which are structurally zero. This occurs as a result of the dependence of the total score on the item responses. The cells ($x_1 \ldots x_k$tgu.l) fcr which t is not equal to $x_1+\ldots+x_k$ are by definition structurally zero. It is a loglinear model because the natural logarithm of the expected cell frequencies is specified by a linear model. Finally, it is a latent class model because the categorical variables U and L are not observed.

All models considered in this paper can be obtained from model (1) in either of two ways. First, one or more of the above types of variables may not be considered. That is, the variables in question are not used to construct the contingency table and the model does not have components related to them. For example, if G, U, and L are not considered, we have a X_1 x X_2 x ... x X_k x T contingency table, and models related to this table do not contain the components in model (1) that depend on G, U, and L.

Second, constrained forms of the saturated model defined in (1) may be specified by setting one or more of its components to zero. This will always be done in a hierarchical fashion. That is, if a component is set equal to zero, all higher order interaction components containing that component will also be set to zero. For example if $\beta^{X_1X_2}_{x_1x_2}$ is set to zero, the term $\beta^{X_1X_2X_3}_{x_1x_2x_3}$ must also be set to zero. This means that if an interaction term is present in the model, all lower order relatives must also be present. Therefore, to indicate a hierarchical model, one does not have to explicitly specify the complete model of interest. Only the highest order interaction terms found in the model need to be designated (Goodman, 1973). Thus a shorthand notation for model (1) is

(3) $\{X_1X_2...X_kTGUL\}$,

where the set of variables between braces indicates that the model contains all possible interaction effects (as well as main effects) among those variables. The notation

(4) $\{X_1\},\{X_2\}...\{X_k\},\{TGU\},\{GUL\}$

denotes a model with main effects for item 1 through k, and all possible interaction (and main) effects among T, G, and U as.well as for G, U, and L. In the remainder of this paper we will designate models of interest using this shorthand notation.

Maximum likelihood estimates of the parameters defining these models are, in general, intractable directly. However, such estimates may be obtained using a variety of iterative estimation procedures. This includes the Iterative Proportional Fitting algorithm (Goodman, 1974a,b; Haberman, 1979) as well as Fisher's (Method of) Scoring algorithm (McHugh, 1956; Haberman, 1979).

To assess the fit to data provided by a given model, the likelihood-ratio statistic, G^2, may be used. This statistic is defined as

(5) $$G^2 = \sum_{x_1 \ldots x_k t g u l} \Sigma\Sigma\Sigma\Sigma\Sigma\, f_{x_1 \ldots x_k tgul} \ln\left(\frac{f_{x_1 \ldots x_k tgul}}{m_{x_1 \ldots x_k tgul}} \right).$$

G^2 is asymptotically distributed as chi–square with degrees of freedom equal to the difference between the number of structurally nonzero cells in the contingency table and the number of independently estimated β parameters in the model of interest.

Additionally, it may be possible to assess the relative fit provided by two models, given that certain regularity conditions are met. The most important of these conditions is that the pair of models be "hierarchically" related (Alvord & Mac‹ ‹ady, 1982). This meens that one of the two models, say M, must be able to be defined in terms of the second model, say M*, by imposing one or more constraints on the parameters defining the second model (i.e., M is a special constrained form of M*). Under these circumstances it is possible to test whether M* fits the data significantly better than M. This may be statistically tested with the difference of the likelihood–ratio statistics for the two models:

(6) $$G^2_D = G^2_M - G^2_{M^*}.$$

This statistic is also asymptotically distributed as chi–quare with degrees of freedom equal to the difference in degrees of freedom for the two models in question.

In what follows we consider models that may be used to detect item bias when the measured latent variable is considered to be either continuous or categorical.

General Categories of Models to be Considered for Assessing Item Bias

Models Where the Measured Trait is Continuous

In this paper,the Rasch Model is used to specify the relation between items and the continuous latent variable being measured. When this model is specified as a loglinear model as described by Cressie and Holland (1983), Duncan (1984), Kelderman (1984), and Tjur (1982), then the model may be designated $\{X_1\},\{X_2\},\ldots,\{X_k\},\{T\}$ for a k item test (e.g., Model 1 in Table 1), where the contingency table for this model has the dimensions $X_1 \times X_2 \times \ldots \times X_k \times T$. As mentioned above, this table contains structural zeros for the cells where the sum of the item responses is not equal to the total score.

Insert Table 1 about here

The model is a quasi-independence model (see Goodman, 1968), that is, a model where there are no interactions among variables beyond those imposed by the incompleteness structure of the table (i.e., the pattern of structurally zero and non-structurally zero cells). Kelderman (1984) has shown that a quasi-independence model where there are no interactions among the item responses and the total score is equivalent to the Rasch model. By introducing one or more grouping variables in the contingency table as well as in the model, it is possible to study item bias with respect to that grouping variable.

Models Where the Grouping Variable is Manifest

When it is of interest to explore the presence of item bias relative to a specified manifest grouping variable (e.g., sex or race), we may attempt to model the frequencies in the observed $X_1 \times X_2 \times \ldots \times X_k \times T \times G$ contingency table. Using a loglinear model for this incomplete table, we can study the relation of the grouping variable G with the other variables. A general review of the procedures for assessing item bias in this case is provided by Kelderman (1985).

Models 2, 3, and 4 of Table 1 are loglinear Rasch type models which contain a manifest grouping variable. In model 2 there is only one interaction effect, {TG}. That is, the grouping variable influences the distribution of the score but not their responses to the items. This model is a Rasch model in all subgroups. Since there are no interactions between the item responses and the grouping variable, the model assumes that items have the same difficulty levels across subgroups. Therefore, if this model is able to effectively account for the contingency table data, it is reasonable to conclude that the items are not biased.

For model 3 which is described in Table 1, there are interaction effects between the item responses for each item and the grouping variable. Therefore, all items may have different difficulty levels across subgroups. Model 3 may be used to study item bias since it may be considered to be a Rasch model where the item difficulties may differ across subgroups and thus specifies the presence of item bias. The Rasch model with equal item parameters over subgroups (model 2 in Table 1) is a constrained form of the Rasch model with different item parameters over subgroups (model 3 in Table 1). Thus, the relative fit provided by these two models may be compared by using the difference likelihood-ratio statistic specified in (6). The statistic yields a test for the presence of item x subgroup interactions. If a statistically significant outcome is obtained, it may be concluded that the items have different difficulty levels for the different subgroups. (i.e., that one or more of the items is biased).

If one has concerns about bias for only some items, it would seem more appropriate to incorporate interaction terms, $\{X_jG\}$, in the model, for only those items. Following this guideline, model 4 incorporates this interaction terms for only the last three items. This model also subsumes model 2 as a constrained form. A comparison of the relative fit obtained under models 2 and 4 may be implemented to test for the presence of item bias among the last three items. If the value of the statistic is found to be significant, there is support for the contention that item difficulty levels for the last three items vary across subgroups.

Since model 4 is also a constrained form of model 3, it is possible to test for item bias in the first three items. Note that this test, however, is made conditional on the last three items being biased.

Models Where the Grouping Variable is Latent

When no grouping variables are actually observed, either because a) grouping information is not available for the variable of interest, or b) because one does not wish to tie the concept of bias to any specific manifest variable, the assessment of item bias should be based on the unobserved and incomplete X_1 x X_2 x ... x X_k x T x U contingency table. Note that what is actually observed is the incomplete X_1 x X_2 x ... x X_k x T contingency table. The categories of the latent grouping variable are then latent classes and the appropriate kind of model is an incomplete latent class model, as described by Haberman (1979, p. 554). Although several latent classes can be specified, we will limit our discussion to two such classes.

Models 5 and 6 of Table 1 are identical to models 3 and 4 respectively, except that the manifest grouping variable, G, is replaced by the latent grouping variable, U. Model 5 has interaction effects between the latent grouping variable and each item, while model 6 only has interaction effects between the latent grouping variable and the last three items.

The appropriate null model (i.e., the model corresponding to absence of item bias) to test models 5 and 6 against is model 1. Model 1 is the same as model 5 if there is only one latent class in model 5. Thus, model 1 is a restricted form of model 5. Similarly, model 1 is a restricted form of model 6. Comparing the fit of models 1 and 5 provides a test for item bias in all items. Similarly, comparing the fit of model 1 and 6 yields a test for item bias in only the last three items.

Finally, comparing the fit of models 5 and 6 yields a test for item bias in the first three items with respect to the latent grouping variable.

Models With Both a Manifest and a Latent Grouping Variable

If a grouping variable, G, is observed, but it is conjectured that the items may also be biased with respect to some unavailable or unknown (i.e., latent) grouping variable, U, we have an incomplete loglinear model for the unobserved X_1 x X_2 x ... x X_k x T x G x U contingency table. Models 7, 8, and 9, described in Table 1, are examples of this kind of model. These models explain the same observed X_1 x X_2 x ... x X_k x T x G contingency table as models 2, 3, and 4. Furthermore, models 7, 8, and 9 may be obtained from models 2. 3, and 4, respectively, by simply adding main effects for the latent grouping variable plus interaction effects which are the same as those already present, except that they also include the latent grouping variable. Note that models 7, 8, and 9 are identical to models 2, 3, and 4, respectively, if U is assumed to consist of only one latent class (i.e., U is a constant). Thus the latter models are each constrained forms of the former models. In model 7, the item responses interact with the latent grouping variable, U, but not with the manifest grouping variable, G. This indicates that, for this model, the items may be biased with respect to the latent grouping variable, but not with respect to the manifest grouping variable. To test whether the items are biased with respect to the latent grouping variable, the fit of models 2 and 7 may be compared.

In model 3, item responses may interact with both the latent and the manifest grouping variables jointly as well as separately. Thus an item may have a different difficulty for one combination of the latent and manifest grouping variable than for another combination. Model 8 may be compared with 3 to assess whether the inclusion of a latent grouping variable accounts for additional item bias that cannot be explained by the manifest grouping variable alone. On the other hand, model 8 may be tested against model 7 to explore whether the manifest grouping variable, G, accounts for item bias that cannot be explained by the latent grouping variable, U, alone.

In model 9, all item responses may interact with the unobserved grouping variable, but only the last three items may interact with the manifest grouping variable. Comparing model 9 with model 4 may be used to test the hypothesis that the latent grouping variable accounts for item bias beyond that attributable to the manifest grouping variable on the last three items. Comparing model 9 with model 7 provides a mears of assessing whether there is any manifest item bias related to the last three items in addition to the latent item bias which affects all items. Finally, comparing the fit of model 9 with that of model 8 allows for an assessment of manifest item bias for the first three items in addition to that which is related to the last three items.

Obviously the models in Table 1 are only a small selected sample of the possible models that could have been considered (see Kelderman, 1984, 1985). However, these models would appear to be some of the more useful for both the exploration and detection of item bias.

Models Where the Assessed Attribute is Discrete

Now consider models where the attribute being assessed is assumed to be categorical. We shall restrict our discussion in this paper to the case where the assessed attribute has only two levels. This class of models may be particularly appropriate when the latent variable of interest is narrow in scope (i.e., it is a highly specific skill, behavior, or attribute) and may reasonably be assumed to exist at two mutually exclusive and exhaustive levels (i.e., mastery vs non-mastery; pathologic vs non-pathologic, and domin-nt vs. recessive). The unconstrained two state latent class model which is described by Macready and Dayton (1977) may be specified as a latent loglinear model as pointed out by

Haberman (1979). This rather simple model within the loglinear modeling framework may be specified as $\{LX_1\},\ldots,\{LX_6\}$ for the unobserved X_1 x X_2 x ... x X_k x L contingency table, where L is the two state latent attribute which is to be assessed. This model may be used to explain the structure of the observed X_1 x X_2 x ... x X_k contingency table. Note that the basic underlying assumption for this model is local independence, which here means that, within each of the two latent classes, items are independent.

Within the framework of latent structure models, the parameters which may alternatively be used to define this model are a) the conditional probabilities for positive item responses given latent class membership, and b) the latent proportions of individuals within each of the latent classes. In mastery modeling, the conditional probabilities for correct item responses by individuals in the "Non-mastery" class are interpreted as "intrusion" errors (i.e., errors due to factors such as guessing and cheating). Conversely, the conditional probabilities for for incorrect item responses by individuals in the "Mastery" class are interpreted as "omission" errors (i.e., errors due to such factors as carelessness and fatigue). As was the case for a continuous measured variable, the above model and table can be extended to take into account the effects of manifest and latent grouping variables. In Table 2, some models are considered where the latent attribute being assessed is categorical. These models are formulated in an analogous fashion to those for continuous measured variables, and similar comparisons between these models may be considered. It may also be noted that models in Table 2 are assigned the same number as the model in Table 1 to which they correspond. This is because these pairs of similarly numbered models contain the same kind of bias effects (or lack thereof).

Insert Table 2 about here

The models for assessed categorical attributes differ from the models for continuous latent traits in that the relation between the item responses, X_j, and the latent trait, L, appears explicitly in the model through the interactions, $\{LX_j\}$ (see, for example, model 2 in Table 2). For the continuous latent trait models, these relations are implicitly specified by the incompleteness structure $(t=x_1+...+x_k)$ found in the models.

In addition to the nine models in Table 2 which correspond to those in Table 1, there are four additional models which are considered: models 3', 4', 4^*, and 9'. Model 3' (4') is the same as model 3 (4) in Table 2 except that an extra side condition is imposed, namely that the proportion of "Masters" is the same for both manifest groups. Similarly, model 9' is the same as model 9 in Table 2 except that an additional side condition is added. This condition specifies that the joint proportions of level of "Mastery" with the level of the latent grouping variable are the same across both levels of the manifest grouping variable. Model 4^* will be discussed in the sequel.

The relative fit provided by models 3' versus 3, 4' versus 4, and 9' versus 9, can be compared to address the hypothesis that the distribution of the latent variables are the same across both levels of the manifest grouping variable.

Suggested Strategies for Using the Proposed Modeling System

An effective, systematic investigation of the presence of item bias using the models described in Tables 1 and 2 requires some preliminary decisions regarding the general nature of the assessed trait (i.e., Is the trait more reasonably represented by a categorical or by a continuous underlying variable?) as well as the sequence of comparisons among models which should be considered. The first issue that must be addressed is whether the attribute of interest is more accurately represented by a continuous or categorical variable. Models based on a discrete underlying assessed variable may be preferred when it is reasonable to assume that a finite number of latent acquisition states underlie the attribute of interest. This may be the case, for example, when the attribute is narrow in scope. Conversely, when the assessed attribute may more reason ably be thought of as

being gradually acquired, models which incorporate a continuous measured underlying variable will be preferred.
A second factor to be addressed in choosing which models to consider is the availability of blocking variable information on variables for which the issue of bias may be of interest. If no grouping variables are avai.lable for observation, or if it is not desirable to tie the phenomenon of bias to any specific manifest variable, only models of the types 1, 5, and 6 described in Tables 1 and 2 should be considered. If the null model 1 does not fit the data, item bias with respect to a latent grouping variable may be studied by considering models 5 and 6.

If a grouping variable is observed, the remaining models 2, 3, 4, 7, 8, and 9 (in Table 1 or 2) may be considered. An investigator may choose to start by considering models with only a manifest grouping variable. If none of these yields acceptable fit, models with both manifest and latent grouping variables may be considered.

Of the models which incorporate a manifest variable, the null model 2 should be tested first. If this model does not provide adequate fit it may be compared with models 3 and 4 to see if fit is improved by taking manifest item bias into account. If neither model 3 nor model 4 provides acceptable fit, the best fitting of these three models may be compared with models 8 and 9 to investigate whether the lack of fit can be explained by item bias with respect to a latent grouping variable. In addition, it may sometimes be informative for an investigator to explore the possible presence of latent bias, even when reasonable fit is provided by models 2, 3, or 4. This may provide valuable information regarding the possible presence of bias which is independent of the manifest grouping variable being investigated.

A third consideration to be taken into account in model selection concerns prior knowledge regarding which items may be biased. If certain items are believed to be biased, first the fit of the model (e.g., model 4) with only those items biased is considered. Then the fit of this model may be compared to that of a model with all items biased. If no prior knowledge regarding possible item bias is available, an investigator may wish to first consider the model with all items biased and proceed in an exploratory fashion based on overall model fit and the observed values of parameter estimates. This may, in some cases, result in the consideration of models with one or more unbiased items.

Example Applications

Kok (1982) experimentally studied item bias in multiplication items by manipulating the test takers' skill on a possible biasing factor. Multiplication items were administered to 286 Dutch undergraduates. The items that were administered varied in format. For some items the numbers to be multiplied were written out in Dutch, while for others, Roman-numerals were used. Knowledge of Roman numerals was expected to be a biasing factor, since Dutch undergraduates show differences in their ability to decipher Roman Numerals. Bias was further related to a manifest grouping variable by giving 143 randomly selected undergraduates some training regarding Roman Numerals. It was expected that the Roman-numeral items would be more difficult for the untrained group than for the trained group.

Insert Table 3 about here

Six items were selected from the total set of items administered by Kok (1982). This set included three native-language items and three Roman-numeral items. The item contents and p-values are presented in Table 3. The six chosen items were selected based on the nature of their multiplication content. All six items had the following common properties: (a) there is a single digit multiplier which is greater than five, (b) there are three or more digits in the multiplicand, (c) there is at least one carry operation involved in correctly solving the multiplication item, and (d) the product of the highest "place" digit in the multiplicand and multiplier is a two digit number. These criteria were used to obtain a reasonably homogeneous item set. From Table 3 it can be seen that the Roman-numeral items are easier for the trained group than for the untrained group. The Roman-numeral items are, however, easier than the native-language items, even for the "untrained" students.

Since the multiplication task differed very little across items, it might reasonably be expected that there are two latent ability states, "mastery" or "non-mastery". The mastery model therefore seems most applicable in this case. The data, however, will be analyzed with both continuous and categorical models for the assessed latent attribute. Moreover, the data are analyzed both with and without a manifest grouping variable in order to better exemplify the applications of these modeling techniques.

First, consider the case of a continuous measured variable and no manifest grouping variable. In Table 1, the likelihood-ratio chi square statistics, degrees of freedom, and the corresponding right-tail probability values are presented for this case. Based on these results, it may be concluded that the Rasch model (model 1) does not adequately fit the data. Knowing that the last three items are presented as Roman-numeral items, it is hypothesized that the last three items are biased. This is tested by comparing the fit of model 1 with that of model 6.

Insert Table 4 and 5 about here

In Table 4, the differences in likelihood-ratio statistics, the difference in degrees of freedom, and the corresponding right-tail probability values are given for possible pairs of hierarchically related models. Looking at Table 4, we see that the comparison of model 1 with model 6 yields a large likelihood-ratio chi-square, relative to its degrees of freedom. This suggests that the Roman-numeral items are indeed biased. Adding item-subgroup interactions, however, for the first three items as in model 5 does not substantially improve model fit. The comparison of model 5 with model 6 in Table 4 does not show a significant difference between these models (i.e., the subsumed model 6 fits statistically no worse than model 5 which subsumes it). Since model 6 also provides satisfactory fit to the data, it may be concluded that the Rasch model in each latent subgroup with different item difficulties for the three Roman-numeral items across latent subgroups, provides an acceptable explanation for the data.

Insert Table 6 about here

In Table 6 are the Rasch item difficulty parameters that can be calculated from the β parameters of model 6. The parameters are calculated via the formula:

(7) $\delta_{iu} = \beta_0^{X_i} - \beta_1^{X_i}$ $\quad i = 1, 2, 3$

and

$\delta_{iu} = (\beta_0^{X_i} + \beta_{0u}^{X_iU}) - (\beta_1^{X_i} + \beta_{1u}^{X_iU})$ $\quad i = 4, 5, 6,$

where δ_{iu} is the item difficulty of item i for the u th latent group (Kelderman, 1985). To fix the scale, the difficulty of the first item is set equal to zero by setting the corresponding β parameter equal to zero. Looking at Table 6 we see that all Roman—numeral items are less difficult for the first latent class than for the second. This first class corresponds to what we might expect from students who have the Roman—numeral training or otherwise have acquired a skill of working with Roman Numerals, while the second class appears to contain students who do not have this skill.

Next, consideration is given to the case where the grouping variable is manifest. The null model (model 2), here, is the Rasch model with equal item difficulties across the two training groups (i.e., the trained students and the untrained students). This model does not acceptably fit the data. Comparing model 2 with model 4 (i.e., the model with the last three items biased) results in a chi—square value (see Table 4) that is large relative to its degrees of freedom (i.e., model 4 provides significantly better fit than model 2). Adding item—grouping variable interactions for the first three items as in model 3 does not give a statistically significant improvement in fit over that obtained with model 4. Moreover, model 4 provides a satisfactory fit to the data. For this reason, it is concluded that only the last three items are biased with respect to the manifest grouping variable.

Insert Table 7 about here

In Table 7, the Rasch item difficulties for each of the training groups are presented. These difficulties can also be obtained from the β parameters by use of equation (7). From Table 7 it may be seen that the Roman—numeral items are easier for the trained than for the untrained group. Furthermore, the pattern of item difficulties closely corresponds to those obtained with latent subgroups. This compatibility of outcomes should increase the confidence we have in the unobserved subgroup results.

The third case which we consider involves the Two State Mastery model which incorporates no grouping variables (see model 1 in Table 2). It may be noted that this model does not fit the data well. If item—grouping variable interactions are added for the Roman—numeral items as in model 6 (see Table 2), the model fit improves substantially. Table 5 shows that the likelihood—ratio chi square statistic related to the difference in fit provided by models 1 and 6 is large relative to its degrees of freedom. This indicates that the Roman—numeral items are biased with respect to the latent subgroups. From Table 2, however, it may also be seen that model 6 does not effectively fit the data. Adding item—subgroup interactions for items 1 through 3 to model 6, as in model 5, further improves fit. Again, it may be noted that the difference in fit provided by this pair of models is significant. Since model 5 also fits the data, it may be concluded that the latent biasing effect extends across all six items.

In Table 8, parameter estimates for latent class model 5 are presented. These estimated values correspond to the model parameters used when the model is formulated within a latent structure framework. The defining parameters within this framework are the conditional probabilities of positive item responses, given the specified latent class (i.e., "Masters" or "Non–masters") and the latent class proportions. These parameter estimates can be calculated from the β parameters via the following equations (see Haberman, 1979, p. 551):

$$\frac{\exp(\beta_{1}^{X_i} + \beta_{1u}^{X_iU})}{\exp(\beta_{1}^{X_i} + \beta_{1u}^{X_iU}) + \exp(\beta_{0}^{X_i} + \beta_{0u}^{X_iU})}$$

i=1,...,k for the conditional probabilities of having a positive response to item i given latent class u, and

$$\frac{\sum_{x_1}\cdots\sum_{x_k} \exp(\beta_{1}^{U} + \beta_{x_1 1}^{X_1U} + \ldots + \beta_{x_k 1}^{X_kU})}{\sum_{x_1}\cdots\sum_{x_k}\sum_{u} \exp(\beta_{u}^{U} + \beta_{x_1 u}^{X_1U} + \ldots + \beta_{x_k u}^{X_kU})}$$

the probability of being in latent class 1 (i.e., latent class proportion).

Insert Table 8 about here

The estimated conditional probabilities presented in Table 8 are difficult to interpret in terms of the latent 2 x 2 joint levels of mastery and grouping. A possible interpretation for each latent class is specified between parentheses (see the latent class headings in Table 8). Classes 1 and 2 have relatively high conditional probabilities for correct item responses for the Roman-numeral items, whereas classes 3 and 4 have low corresponding probabilities. It may therefore be conjectured that classes 1 and 2 correspond to latent groups of students which have some facility at working with Roman numerals (this, to a large extent, may include students in the trained group), whereas classes 3 and 4 do not have this facility. Furthermore, the native-language items tend to have higher conditional probabilities for classes 1 and 3 than for classes 2 and 4. This supports the conjecture that classes 1 and 3 correspond to "masters", and classes 2 and 4 to "non-masters". The conditional probabilities for the Roman-numeral items, however, do not conform to the mastery-nonmastery interpretation. In the "experienced/trained" group (i.e., the combined classes 1 and 2), the conditional probability for item 6 is lower in value for mastery class (1) than for non-mastery class (2). Moreover, in the "inexperienced/untrained" group, the conditional probabilities for items 4 and 5 are smaller in mastery class (3) than in non–mastery class (4). The parameters, therefore, are not fully interpretable in terms of a combination of mastery and bias classes.

We next consider the case of two–state mastery models where the grouping variable is observed. It may be seen in Table 2 that the simple two–class mastery model does not provide good fit. Adding the item–grouping variable interaction terms for bias on the Roman–numeral items, as in model 4, does not significantly improve the fit of the model (see Table 5). However, model 4 does not yield a satisfactory overall fit to the data. Adding item–grouping variable interactions for bias on the native–language items 1, 2, and 3, such as in model 3. also does not significantly improve fit. Nor does it yield a model with acceptable overall fit. It may therefore be concluded that it is not sufficient to completely attribute the lack of fit of the two–class mastery model to item bias with respect to the manifest grouping variable, "training". Note from Table 5 that differences in fit provided by models 4 and 4', and models 3 and 3' are not significant, indicating that the proportion of "masters" are equal across levels of "training". This is to be expected, given that students were randomly assigned to training groups.

Continuing to assume that the latent attribute being assessed is dichotomous, we may next consider models with both a manifest and a latent grouping variable. Adding the latent grouping variable U to all the interactions in models

2, 3, 4 and 4' from the previous analysis yields models 7, 8, 9, and 9', respectively. Note that, with the exception of model 7, all of these models provide particularly good fit to the data. To determine which of the effectively fitting models is to be preferred, tests of their relative fit were considered (see Table 5). Based on these comparisons it is concluded that model 9' is to be preferred.

Insert Table 9 about here

Presented in Table 9 are the conditional probabilities and the latent class proportions related to model 9'. Based on the magnitude of the conditional probabilities related to each latent class, it would appear that classes 3 and 4 in the "trained" group, and classes 4 and 8 in the "untrained" group might reasonably be interpreted as mastery classes. Also, the conditional probabilities for correct responses to Roman-numeral items are larger for the "trained" group than for the "untrained" group. The remaining structure, comparing class 1 with class 2, 3 with 4, 5 with 6, and 7 with 8, as the two values for an additional latent grouping variable U, is difficult to interpret. Looking at the conditional probabilities for the Roman-numeral items, the classes differentiate between individuals who answer item 5 correctly and individuals who answer items 4 and 6 correctly. Considering the native-language items, conditional probabilities increase with class 1, 2, 3, and 4 and also with class 5, 6, 7, and 8, suggesting a multi-state, discrete latent variable being measured in each group.

The model with a continuous measured trait and bias on the Roman-numeral items with respect to the manifest grouping variable (see model 4 in Table 1) did fit the data. Therefore, it may be expected that the corresponding model with a categorical assessed trait would better fit the data if the number of levels of mastery were increased. Model 4^* in Table 2 is the same as model 4', except that there are three rather than two latent levels of mastery. This new model fits the data very well.

Insert Table 10 about here

Presented in Table 10 are the conditional probabilities and the latent class proportions which correspond to model 4^*. Based on the mean values for the conditional probabilities on the native-language items for each latent class, we might interpret latent classes 1, 2, and 3, respectively, as corresponding to "non-mastery" (NM), "mixed mastery" (MM), and "mastery" (M) states. Since in this model there are no interaction effects among training, ability, and the responses to the native-language items, the same respective interpretation may be used with classes 4, 5, and 6. In considering the conditional probabilities for the

Roman-numeral items, it is seen that these items generally have higher probabilities (given the same latent class), especially for the "non-mastery" and "mixed mastery" classes. It is somewhat peculiar that item 6 has a greater probability for a correct response by a students in the "mixed mastery" class than one from the "mastery" class.

Discussion

In this paper, we have shown that it is possible to explain item bias through differences in item difficulties or error rates across levels of grouping variables. This approach is viable when the assessed attribute of interest is either continuous or categorical and the grouping variables, with respect to which bias may occur, are manifest, latent or both.

The model which is presented is quite general and can be easily extended to include several observed and unobserved grouping variables. Also this model is capable of incorporating additional interaction effects which we have not considered. One should however be cautious when considering the inclusion of additional effects within models especially when the grouping variable is latent, since many such models will not be identifiable. For example, it is easily shown that adding a term $\{X_4\ X_5\ X_6\}$ for the interaction between Roman-numeral items to model 6 which includes interaction effects $\{X_4U\}$, $\{X_5U\}$, $\{X_6U\}$ between those items and the latent grouping variable U, is not an identifiable model. This is because item interactions with U already explain the interaction among the observed responses on the Roman-numeral items.

A practical problem that occurs with this general modeling approach when latent categorical variables are present is computational infeasibility when more than just a few variables are included in a model. This is because the minimal sufficient information for parameter estimation is the contingency table frequencies the number of which increase exponentially with number of variables. Note that for k dichotomous variables the number of cells in the contingency table is 2^k. For example, if k=20 there are more than a million cells in the contingency table. For this reason it may not be feasible to analyze all items on a test simultaneously. Instead the test may need to be partitioned into carefully chosen subsets of items, where each subset is analyzed separately. The subsets may be chosen on the basis of content so that items similar in content are placed within the same subset. This increases the likelihood that unknown biasing variables might be found.

Another practical problem related to estimation is that the number of iterations required to reach a solution may be quite large or in some cases it may be difficult to reach an acceptable solution. This is especially true when the model under consideration is complex and/or the initial values used in the iterative estimation process are not themselves quite accurate. For example, 449 iterations where needed to obtain estimates for the Rasch model with the Roman-numeral items biased with respect to a latent grouping variable (see model 6 in Table 1). The starting values used in estimation for this model were arbitrary and the stopping criterion was six decimal places of precision. For the corresponding mastery model (see model 6 in Table 2), the number of iterations was 1501 to obtain a precision of five decimal places. An advantage of the Iterative Proportional Fitting algorithm, however, is that iterations may be very quickly implemented. This is because, relative to other procedures, the required operations necessary for completing an iteration are relatively simple and small in number. In the case of the mastery model 6, the required CPU time was less than 15 seconds. Additionally it may be noted, that estimation with this algorithm is far less sensitive to the values selected as initial parameter estimates than is the case with other algorithms. This dramatically reduces the likelihood of the above mention problem of not obtaining acceptable convergence.

References

Alvord, G. & Macready, G.B. (1985). Comparing fit of non-subsuming probability models. Applied Psychological Measurement, 9, 233-240.

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.) Handbook of methods for detecting test bias. Baltimore: John Hopkins University Press.

Angoff, W.H., & Ford, S.F. (1973). Item race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Bergan, J.R. (1983). Latent class models in Educational research. E.W. Gordon, (Ed.). Review of Research in Education, 10, 305-360.

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

Cardal, C. & Coffman, W.T. (1964). A Method for comparing performance of different group on the items in a test. (RM 64-61). Princeton, N.J.:Educational Testing Service.

Cleary, T.A., & Hilton, T.L.(1968). An investigation into item bias. Educational and Psychological Measurement, 8, 61-75.

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

Dayton, C.M., & Macready, G.B. (1976). A probabilistic model for a validation of behavioral hierarchies. *Psychometrika*, *41*, 189–204.

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.

Goodman, L.A. (1968). The analysis of cross–classified data: independence, quasi–independence, and interactions in contingency tables with and without missing entries. *Journal of the American Statistical Association*, *63*, 1091–1131.

Goodman, L.A. (1973). Causal analysis of data from panel studies and other kinds of surveys. *The American Journal of Sociology*, *78*.

Goodman, L.A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: a modified latent structure approach. *The American Journal of Sociology*, *79*.

Goodman, L.A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *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*.

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

Hoepfner, R., & Strickland, G.P. (1972). *Investigating test bias*. Los Angeles: Center for the Study of Evaluation, University of California.

Holland, P.W., & Thayer D.T. (1986). *Differential item functioning and the Mantel–Haenszel Procedure*. Princeton, N.J.: Educational testing Service, Research Report 86–69.

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

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

Kelderman, H. (1985). *Item bias detection using the loglinear Rasch model: Observed and Unobserved subgroups*. Paper presented at the Fiftieth Annual Meeting of the Psychometric Society, Nashville, Tennessee, june 1–4.

Kok, F. (1982). *Het partijdige item. [The biased item]* Psychological Laboratory, University of Amsterdam.

Lazarsfeld, P.F. & Henry, N.W. (1968). *Latent Structure Analysis*. Boston: Houghton Miffin.

Linn, R.L. & Drasgow, F.K. (1987). Implications of the golden rule settlement pf test construction. *Educational Measurement: Issues and Practice*, *6*.

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

Macready, G.B. & Dayton, C.M (1977). The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 2, 99-120.

Macready G.B. & Dayton, C.M. (1980). The nature and use of state mastery models. Applied Psychological Measurement, 4, 493-516.

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

McHugh, R.B. (1956). Efficient estimation and local identification in latent class analysis. Psychometrika 21, 273-274.

Nungester, R.J. (1977). An empirical examination of three models of item bias. (Doctoral dissertation Florida State University) Dissertation Abstracts International, 38, 2726 A.

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

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press.

Rudner, L.M., Getson, P.R. & Knight, D.L. (1980) Biased item detection techniques. Journal of Educational Statistics, 6, 213-233.

Scheunemann, J. (1979). A method of assessing item 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.

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. Journal of Educational Psychology, 16, 433-461.

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

van der Linden, W.J. (1978). Forgetting, guessing and mastery: The Macready and Dayton Models revisited and compared with a latent trait approach. Journal of Education Statistics, 3, 305-318.

Table 1

Models for a Continuous Measured Trait

	Model	LR	DF	p
	No Grouping Variable			
1.	{X_1},....,{X_6},{T}	86.23	52	.00
	Manifest Grouping Variable			
2.	{X_1},....,{X_6},{TG}	159.38	109	.00
3.	{GX_1},....,{GX_6},{TG}	124.08	104	.09
4.	{X_1},{X_2},{X_3},{GX_4},{GX_5},{GX_6},{TG}	128.23	106	.07
	Latent Grouping Variable			
5.	{UX_1},....,{UX_6},{TU}	51.63	40	.10
6.	{X_1},{X_2},{X_3},{UX_4},{UX_5},{UX_6},{TU}	55.55	42	.08
	Manifest and Latent Grouping Variable			
7.	{UX_1},....,{UX_6},{TGU}			
8.	{GUX_1},....,{GUX_6},{TGU}			
9.	{UX_1},{UX_2},{UX_3}, {GUX_4},{GUX_5},{GUX_6},{TGU}			

Table 2

Fit of Models for a Categorical Assessed Attribute

Model	Side Conditions	LR	DF	p
No Grouping Variable				
1. $\{LX_1\},....,\{LX_6\}$	None	91.17	48	.00
Manifest Grouping Variable				
2. $\{LX_1\},....,\{LX_6\},\{GL\}$	None	177.56	112	.00
3. $\{GLX_1\},....,\{GLX_6\}$	None	126.34	100	.04
3'. $\{GLX_1\},....,\{GLX_6\}$	Eq.L.C.Props.	128.24	102	.04
4. $\{LX_1\},\{LX_2\},\{LX_3\},$ $\{GLX_4\},\{GLX_5\},\{GLX_6\}$	None	134.92	106	.03
4'. $\{LX_1\},\{LX_2\},\{LX_3\},$ $\{GLX_4\},\{GLX_5\},\{GLX_6\}$	Eq.L.C.Props.	135.13	108	.04
4*. $\{LX_1\},\{LX_2\},\{LX_3\},$ $\{GLX_4\},\{GLX_5\},\{GLX_6\}$	Eq.L.C.Props., 3 L.C.'s Per Training Grp.	101.19	102	.50
Latent Grouping Variable				
5. $\{ULX_1\},....,\{ULX_6\}$	None	41.66	34	.17
6. $\{LX_1\},\{LX_2\},\{LX_3\},$ $\{ULX_4\},\{ULX_5\},\{ULX_6\}$	None	59.89	40	.02
Manifest and Latent Grouping Variable				
7. $\{ULX_1\},....,\{ULX_6\}$	None	143.87	100	.00
8. $\{GULX_1\},....,\{GULX_6\}$	None	76.87	72	.33
9. $\{ULX_1\},\{ULX_2\},\{ULX_3\},$ $\{GULX_4\},\{GULX_5\},\{GULX_6\}$	None	77.57	84	.68
9'. $\{ULX_1\},\{ULX_2\},\{ULX_3\},$ $\{GULX_4\},\{GULX_5\},\{GULX_6\}$	Eq.L.C.Props.	81.64	88	.67

Table 3

Homogeneous Multiplication Items Presented in Native-language and Roman-numerals Formats

Item	Multiplication	Presentation	% correct	
			Untrained	Trained
1.	6 x 4123	Native language	.37	.38
2.	7 x 974	Native language	.33	.22
3.	7 x 3423	Native language	.24	.23
4.	8 x 214	Roman Numerals	.50	.68
5.	6 x 3107	Roman Numerals	.43	.71
6.	9 x 351	Roman Numerals	.48	.66

Table 4

Comparison of Models for a Continuous Measured Variable

Subsuming Model	Subsumed Model	LR	DF	p
No Manifest Grouping Variable				
1	5	34.60	12	.00
1	6	30.68	10	.00
6	5	3.92	2	.14
Manifest Grouping Variable				
2	3	35.30	5	.00
2	4	31.15	3	.00
4	3	4.15	2	.13

Table 5

Comparison of Models for a Categorical Assessed Attribute

Subsuming Model	Subsumed Model	LR	DF	p
No Manifest Grouping Variable				
1	5	49.51	14	.00
1	6	31.28	8	.00
6	5	18.23	6	.00
Manifest Grouping Variable				
2	3	51.22	12	.00
2	4	42.64	6	.00
4	3	8.58	6	.20
2	7	33.69	12	.00
3	8	49.47	28	.01
4	9	57.35	22	.00
3'	3	1.90	2	.39
4'	4	0.21	2	.90
9'	9	4.07	4	.40

Table 6

Item Difficulty Estimates of Model {X1},{X2},{X3},{UX4}, {UX5}, {UX6}, {TU} (Model 6) from Table 1

		Item					
		Native Language			Roman Numerals		
Latent Subgr.	Subgr Prop.	1	2	3	4	5	6
1.	0.34	0.00	0.77	0.97	–1.58	–1.31	–7.39
2.	0.66	0.00	0.77	0.97	–0.93	–0.89	0.16
Difference					–0.65	–0.42	–7.55

Table 7

Item Difficulty Estimates of Model {X1}.{X2}.{X3}.{GX4}. {GX5}. {GX6}. {TG} (Model 4) from Table 1

	Item					
	Native Language			Roman Numerals		
Group	1	2	3	4	5	6
Trained	0.00	0.65	0.93	−1.80	−1.97	−1.67
Untrained	0.00	0.65	0.93	−0.59	−0.19	−0.47
Difference				−1.21	−1.78	−1.20

Table 8

Parameter Estimates for Model {ULX1}....{ULX6} (Model 5) from Table 2

		Latent Class			
Item No.	Item Format	1 (TM)	2 (TN)	3 (UM)	4 (UN)
		Conditional Probabilities			
1	Natural	0.88	0.40	0.26	0.10
2	Natural	0.77	0.21	0.40	0.00
3	Natural	0.77	0.13	0.29	0.00
4	Roman	0.85	0.81	0.28	0.35
5	Roman	0.83	0.78	0.00	0.42
6	Roman	0.71	1.00	0.42	0.18
		Latent Class Probabilities			
		0.21	0.30	0.12	0.37

Table 9

Parameter Estimates of the Model {ULX1},{ULX2},{ULX3}, {GULX4}, {GULX5}, {GULX6} with Equal Latent Class Proportions (Model 9') from Table 2

		Latent Classes							
		Trained Group				Untrained Group			
Item No.	Item Format	1	2	3	4	5	6	7	8
		Conditional Probabilities							
1.	Natural	0.10	0.19	0.50	0.88	0.10	0.19	0.50	0.88
2.	Natural	0.00	0.20	0.22	0.85	0.00	0.20	0.22	0.85
3.	Natural	0.00	0.15	0.17	0.79	0.00	0.15	0.17	0.79
4.	Roman	0.47	0.48	1.00	0.85	0.25	0.30	0.73	0.82
5.	Roman	0.54	0.61	0.80	1.00	0.39	0.00	0.78	0.65
6.	Roman	0.00	1.00	1.00	0.70	0.11	0.28	1.00	0.63
Mean		0.18	0.43	0.61	0.84	0.14	0.18	0.56	0.77
		Latent Class Probabilities							
		0.14	0.14	0.12	0.10	0.14	0.14	0.12	0.10

Table 10

Parameter Estimates of Model {LX1}.{LX_2}.{LX_3}.{GLX_4}.{GLX_5}.{GLX_6} with Three Mastery States and Equal Latent Class Proportions Across Levels of Training (Model 4*)

Table 2

		Latent Classes					
		Trained Group			Untrained Group		
Item No.	Item Format	1 (NM)	2 (MM)	3 (M)	4 (NM)	5 (MM)	6 (M)
		Conditional Probabilities					
1.	Natural	.12	.52	.87	.12	.52	.87
2.	Natural	.07	.27	.87	.07	.27	.87
3.	Natural	.04	.21	.83	.04	.21	.83
Mean		.08	.33	.86	.08	.33	.86
4.	Roman	.49	.89	.86	.27	.73	.79
5.	Roman	.59	.75	1.00	.21	.17	.61
6.	Roman	.42	1.00	.72	.17	1.00	.57
Mean		.50	.88	.86	.22	.81	.66
		Latent Class Probabilities					
		.52	.30	.18	.52	.30	.18

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

RR-87-1 R. Engelen, *Semiparametric estimation in the Rasch model*

RR-87-2 W.J. van der Linden (Ed.), *IRT-based test construction*

RR-87-3 R. Engelen, P. Thommassen, & W. Vervaat, *Ignatov's theorem: A new and short proof*

RR-87-4 E. van der Burg, & J. de Leeuw, *Use of the multinomial jackknife and bootstrap in generalized nonlinear canonical correlation analysis*

RR-87-5 H. Kelderman, *Estimating a quasi-loglinear models for the Rasch table if the number of items is large*

RR-87-6 R. Engelen, *A review of different estimation procedures in the Rasch model*

RR-87-7 D.L. Knol & J.M.F. ten Berge, *Least-squares approximation of an improper by a proper correlation matrix using a semi-infinite convex program*

RR-87-8 E. van der Burg & J. de Leeuw, *Nonlinear canonical correlation analysis with k sets of variables*

RR-87-9 W.J. van der Linden, *Applications of decision theory to test-based decision making*

RR-87-10 W.J. van der Linden & E. Boekkooi-Timminga, *A maximin model for test design with practical constraints*

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

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

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

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

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

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

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

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

RR-88-9 W.J. van der Linden & T.J.H.M. Eggen, *The Rasch model as a model for paired comparisons with an individual tie parameter*

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

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

department of
EDUCATION

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