Journal of Educational Measurement
Winter 1990, Vol. 27, No. 4, pp. 307–327

The Use of Loglinear Models for Assessing Differential Item Functioning Across Manifest and Latent Examinee Groups

Henk Kelderman
University of Twente, The Netherlands
and
George B. Macready
University of Maryland

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

Test items exhibit differential item functioning (DIF) if the item scores of equally able examinees from different groups (e.g., of different race, sex, or age) are systematically different. If several items in a test exhibit DIF in favor of a specific group, the test may lead to an unfair advantage for that group with regard to the assessed level of performance when its members are compared with members of other groups. It is expected that this inequity can be rectified by deleting or improving items exhibiting DIF.

The basic problem in the detection of DIF is to differentiate between discrepancies in item difficulties across groups that are due to DIF as opposed to differences in level on the assessed attribute. Because groups frequently differ on the assessed attributes, DIF and ability are often confounded. For this reason it is hard to tell whether observed differences in probabilities for positive item responses among groups result from DIF or from 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 DIF detection methods have been proposed. Reviews of this topic are provided by Osterlind (1983); Rudner, Getson, and Knight (1980); and Shepard, Camilli, and Averill (1981). In the earlier DIF-detection methods such as the analysis-of-variance method (Cardal & Coffman, 1964; Cleary & Hilton, 1968; Hoepfner & Strickland, 1972; Jensen, 1980) and the transformed-item-difficulty methods (Angoff, 1982; Angoff & Ford, 1973; Thurstone, 1925), there was no rigorous control for differences in true ability across groups. In chi-square methods (Camilli, 1979; Holland & Thayer, 1986; Mellenbergh, 1982; Nung-

The authors thank Frank Kok for making his data available for reanalysis.

ester, 1977; Scheunemann, 1979), 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 by means of the person's ability parameter in the model. Items exhibit DIF if the item parameters vary across groups. Thissen, Steinberg, and Gerrard (1986) discussed the separation of true ability and DIF with IRT models, and Thissen, Steinberg, and Wainer (1989) gave a review of DIF detection using IRT models.

Kelderman (1989) proposed the use of a loglinear formulation of the Rasch (1980) model (Cressie & Holland, 1983; Duncan, 1984; Kelderman, 1984; Tjur, 1982) to study DIF. Various aspects of DIF can be modeled by adding parameters to the loglinear formulation of the Rasch model. In this paper the abovementioned loglinear modeling system is extended. Our purposes are threefold: (a) Develop procedures for use in assessing DIF that may be used when the grouping variable with respect to which DIF may occur is not observed; (b) develop DIF-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 DIF with respect to unobserved or latent grouping variables. With this kind of loglinear latent class model, it is possible to extend the loglinear Rasch model to include a latent categorical dimension. Using this formulation of the latent trait/latent class model, 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 incorporates 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 for another, it is concluded that the item exhibits DIF with respect to the latent grouping variable.

DIF-detection procedures are also possible when the latent attribute is categorical. Then, 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., Dayton & Macready, 1976, 1980; Goodman, 1975).

The two-state mastery model is particularly appropriate for assessing attributes whose acquisition is assumed to be an "all or 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 that are used to assess the attribute of interest) and *nonmasters* (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 (*omission* errors) or nonmasters responding correctly to the items (*intrusion* errors).

DIF 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 DIF is suspected. If for a certain item the omission error rates or the intrusion error rates differ across groups, the item in question exhibits DIF with respect to the grouping variable. As in the case of a continuous measured variable, DIF may be studied with respect to either manifest or latent grouping variables, through the use of loglinear latent class models.

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

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

An Overall Loglinear Modeling Framework

Variables That May be Included in Models

The following types of variables may be included in the models considered in this paper. 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, the models include two other kinds of variables: the latent variable being measured (or assessed) and the grouping variable with respect to which DIF may occur.

The measured (or assessed) variable may be either a continuous or a discrete-categorical attribute. When the 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, $T = X_1 + \cdots + X_k$ (see Kelderman, 1984, for a discussion). In the case of an assessed attribute L ($l = 1, \ldots, q$), which is categorical, a 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 nonmaster) 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 that are used to model DIF can be either observed or unobserved grouping variables. Such a variable is designated as G when its levels are actually observed (as in the case of studying sex or race as having a possible DIF 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) presented a general loglinear model that 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 that define this decomposition may include *main* and *interaction* effects corresponding to various margins (or cells) of the contingency table. If all the types of variables mentioned above are simultaneously considered, we have a $X_1 \times X_2 \times \cdots \times X_k \times T \times G \times U \times L$ contingency table with frequencies

$$f_{x_1 \ldots x_k tgul}, \quad x_1 = 0, 1; \ldots; \quad x_k = 0, 1; \quad t = x_1 + \cdots + x_k;$$
$$g = 1, \ldots, r; \quad u = 1, \ldots, s; \quad l = 1, \ldots, q.$$

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

$$ln\, m_{x_1 \ldots x_k tgul} = \beta + \beta_{x_1}^{X_1} + \cdots + \beta_{x_k}^{X_k} + \beta_t^T + \beta_g^G + \beta_u^U + \beta_l^L \beta_{x_1 x_2}^{X_1 X_2}$$
$$+ \cdots + \beta_{ul}^{UL} + \beta_{x_1 x_2 x_3}^{X_1 X_2 X_3} + \cdots + \beta_{gul}^{GUL} + \cdots + \beta_{x_1 \ldots x_k tgul}^{X_1 \ldots X_k TGUL}. \quad (1)$$

With the constraints

$$\sum_{x_1} \beta_{x_1}^{X_1} = 0, \ldots, \sum_{x_k} \beta_{x_k}^{X_k} = 0, \sum_t \beta_t^T = 0, \sum_g \beta_g^G = 0, \sum_u \beta_u^U = 0,$$
$$\sum_l \beta_l^L = 0, \sum_{x_1} \beta_{x_1 x_2}^{X_1 X_2} = 0, \sum_{x_2} \beta_{x_1 x_2}^{X_1 X_2} = 0, \ldots, \sum_u \beta_{ul}^{UL} = 0, \sum_1 \beta_{ul}^{UL} = 0, \ldots,$$
$$\sum_{x_1} \beta_{x_1 \ldots x_k tgul}^{X_1 \ldots X_k TGUL} = 0, \ldots, \sum_l \beta_{x_1 \ldots x_k tgul}^{X_1 \ldots X_k TGUL} = 0, \quad (2)$$

where $\{Mx_1 \ldots x_k tgul\}$ 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 1, $\beta_{x_1 x_2}^{X_1 X_2}$ is the parameter designating the interaction effect of the combination of Response x_1 of Item 1 and Response x_2 of Item 2, 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 that are structurally zero. This occurs as a result of the dependence of the total score on the item responses. The cells $(x_1 \ldots x_k tgul)$ for which t is not equal to $x_1 + \cdots + x_k$ 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 aforementioned 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 an $X_1 \times X_2 \times \cdots \times X_k \times T$ 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

$$\{X_1X_2 \ldots X_kTGUL\}, \tag{3}$$

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

$$\{X_1\}, \{X_2\}, \ldots, \{X_k\}, \{TGU\}, \{GUL\} \tag{4}$$

denotes a model with main effects for Items 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 (IPF) algorithm (Goodman, 1974a,b; Haberman, 1979) as well as Fisher's scoring algorithm (McHugh, 1956; Haberman, 1979). In this paper we use three computer programs to obtain parameter estimates: MLLSA (Clogg, 1977), LCAG (Hagenaars, 1987), and LOGIMO (Kelderman & Steen, 1988). All three implement the IPF algorithm for some situation. Identifiability conditions for latent class models have been given by Goodman (1974b) and Clogg and Goodman (1984).

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

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

The statistic 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. This test statistic should be used with caution, however. If the expected frequencies become too small, the approximation of the statistic to the chi-square distribution is known to be bad (Lancaster, 1961). A rule of thumb is that the expected frequencies should at least be one (Cochran, 1952). Therefore, the sample size should well exceed the total number of cells of the contingency table. This means that the overall likelihood-ratio statistic is only useful if the number of items is relatively small.

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 & Macready, 1985). This means 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:

$$G^2_{\mathrm{D}} = G^2_{\mathrm{M}} - G^2_{\mathrm{M*}}. \tag{6}$$

This statistic is also asymptotically distributed as chi-square 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 DIF when the measured latent variable is considered to be either continuous or categorical.

General Categories of Models to Be Considered for Assessing DIF

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 $X_I \times X_2 \times \cdots \times X_k \times T$. 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.

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 nonstructurally 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 DIF with respect to that grouping variable.

Models Where the Grouping Variable Is Manifest

When it is of interest to explore the presence of DIF relative to a specified manifest grouping variable (e.g., Sex or Race), we may attempt to model the

Table 1
Fit of Models for a Continuous Measured Trait

	Model	G^2	df	p
	No Grouping Variable			
1.	$\{X_1\},\ldots,\{X_6\},\{T\}$	86.23	52	.00
	Manifest Grouping Variable			
2.	$\{X_1\},\ldots,\{X_6\},\{TG\}$	159.38	109	.00
3.	$\{GX_1\},\ldots,\{GX_6\},\{TG\}$	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.	$\{UX_1\},\ldots,\{UX_6\},\{TU\}$	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.	$\{UX_1\},\ldots,\{UX_6\},\{TGU\}$			
8.	$\{GUX_1\},\ldots,\{GUX_6\},\{TGU\}$			
9.	$\{UX_1\},\{UX_2\},\{UX_3\},$ $\{GUX_4\},\{GUX_5\},\{GUX_6\},\{TGU\}$			

frequencies in the observed $X_1 \times X_2 \times \cdots \times X_k \times T \times G$ 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 DIF in this case is provided by Kelderman (1989). Parameter estimates can be obtained with the computer program LOGIMO (Kelderman & Steen, 1988). LOGIMO is especially written to estimate loglinear models that include the total score T.

Models 2, 3, and 4 of Table 1 are loglinear Rasch type models that 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 the responses to the items. This model is a Rasch model in all subgroups. Because 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 can effectively account for the contingency table data, it is reasonable to conclude that the items do not exhibit DIF.

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 DIF because it may be considered to be a Rasch model where the item difficulties may differ across subgroups, and thus Model 3 specifies the presence of DIF. 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 ×

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 exhibits DIF).

If one has concerns about DIF for only some items, it would seem more appropriate to incorporate interaction terms, $\{X_jG\}$, in the model for only those items. Following this approach, Model 4 incorporates 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 DIF 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 DIF in the first three items. Note that this test, however, is made conditional on the last three items exhibiting DIF.

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 DIF to any specific manifest variable, the assessment of DIF should be based on the unobserved and incomplete $X_1 \times X_2 \times \cdots \times X_k \times T \times U$ contingency table. Note that what is actually observed is the incomplete $X_1 \times X_2 \times \cdots \times X_k \times T$ 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). The expected counts of the $X_1 \times X_2 \times \cdots \times X_k \times T \times U$ contingency table under the model may be estimated using the computer program LCAG (Hagenaars, 1987). From these expected counts, the parameter estimates may be calculated using the LOGIMO program (Kelderman & Steen, 1988).

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, whereas Model 6 has interaction effects only between the latent grouping variable and the last three items.

The appropriate null model (i.e., the model corresponding to absence of DIF) 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 DIF in all items. Similarly, comparing the fit of Models 1 and 6 yields a test for DIF in only the last three items.

Finally, comparing the fit of Models 5 and 6 yields a test for DIF in the first three items (conditional on DIF in the last 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 exhibit DIF with respect to some unavailable or unknown (i.e., latent) grouping variable U, we have an incomplete loglinear model for the unobserved $X_1 \times X_2 \times \cdots \times X_k \times T \times G \times U$ contingency table. Models 7, 8, and 9, described in Table 1, are examples of this kind of model. These models explain the same observed $X_1 \times X_2 \times \cdots \times X_k \times T \times G$ 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. It is readily seen that hierarchical relations exist between models with both manifest and latent grouping variables and models with only manifest grouping variables, so that hypotheses can be tested with respect to the influence of manifest or latent grouping variables on item difficulty.

Obviously the models in Table 1 are only a small selected sample of the possible models that could have been considered (see Kelderman, 1984, 1989). However, these models appear to be some of the more useful for both the exploration and detection of DIF.

Models Where the Assessed Attribute Is Discrete

Now consider models where the attribute being assessed is assumed to be discrete. We shall restrict our discussion 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. nonmastery; pathological vs. nonpathological; and dominant vs. recessive). The unconstrained two-state latent class model described by Macready and Dayton (1977) may be specified as a latent loglinear model, as pointed out by Haberman (1979). The parameter estimates of models with discrete latent variables can be obtained with the computer program MLLSA (Clogg, 1977). This rather simple model may be specified as $\{LX_1\}, \ldots, \{LX_6\}$ for the unobserved $X_1 \times X_2 \times \cdots \times X_k \times L$ contingency table, where L is the two-state latent attribute that is to be assessed. This model may be used to explain the structure of the observed $X_1 \times X_2 \times \cdots \times X_k$ 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 proportions of individuals within each of the latent classes. In mastery modeling, the conditional probabilities for correct item responses by individuals in the nonmastery class are interpreted as intrusion errors (i.e., errors due to factors such as guessing and cheating). Conversely, the conditional probabilities 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 model and table above 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 DIF effects (or lack thereof).

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 assessed attribute 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 $(\mathrm{t} = x_1 + \cdot \cdot \cdot + x_k)$ found in the models.

Suggested Strategies for Using the Proposed Modeling System

An effective, systematic investigation of the presence of DIF using the models described in Tables 1 and 2 requires some preliminary decisions. The first issue is whether the attribute of interest is more accurately represented by a continuous or a 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

Table 2
Fit of Models for a Categorical Assessed Attribute

	Model	G^2	df	p
	No Grouping Variable			
1.	$\{LX_1\},\ldots,\{LX_6\}$	91.17	50	.00
	Manifest Grouping Variable			
2.	$\{LX_1\},\ldots,\{LX_6\},\{GL\}$	177.56	112	.00
3.	$\{GLX_1\},\ldots,\{GLX_6\}$	126.34	100	.04
4.	$\{LX_1\},\{LX_2\},\{LX_3\},\{GLX_4\},\{GLX_5\},\{GLX_6\}$	134.92	106	.03
4*.	$\{LX_1\},\{LX_2\},\{LX_3\},\{GLX_4\},\{GLX_5\},\{GLX_6\}$	96.10	95	.45
	Latent Grouping Variable			
5.	$\{ULX_1\},\ldots,\{ULX_6\}$	41.66	36	.24
6.	$\{LX_1\},\{LX_2\},\{LX_3\}\{ULX_4\},\{ULX_5\},\{ULX_6\}$	59.89	42	.04
	Manifest and Latent Grouping Variable			
7.	$\{ULX_1\},\ldots,\{ULX_6\},\{G\}$			
8.	$\{GULX_1\},\ldots,\{GULX_6\}$			
9.	$\{ULX_1\},\{ULX_2\},\{ULX_3\},\{GULX_4\},\{GULX_5\},\{GULX_6\}$			

* For this model, there are three latent levels of mastery rather than two as was the case for all other latent class models considered.

attribute may more reasonably be thought of as being gradually acquired, models that incorporate a continuous measured underlying variable will be preferred.

A second issue in choosing models is the availability of blocking variable information on variables for which the issue of DIF may be of interest. If no grouping variables are available for observation, or if it is not desirable to tie the phenomenon of DIF to any specific manifest variable, only Models 1, 5, and 6 described in Tables 1 and 2 should be considered. If the null Model 1 does not fit the data, DIF 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, and 4, and 7, 8, and 9 (in Tables 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 that incorporate a manifest variable, the null Model 2 should be tested for fit. In addition, this null model may be compared with Models 3 and 4 to see if fit is improved by taking manifest DIF 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 DIF with respect to a latent grouping variable. Alternatively, it may sometimes be informative for an investigator to explore the possible presence of latent DIF, even when reasonable fit is provided by Models 2, 3, or 4. This may provide valuable information regarding the possible presence of DIF that is independent of the manifest grouping variable being investigated.

A third consideration in model selection concerns prior knowledge regarding which items may suffer from DIF. If certain items are believed to be subject to DIF, first the fit of the model (e.g., Model 4) with only those DIF items is considered. Then the fit of this model may be compared to that of a model in which all items are hypothesized to exhibit DIF. If no prior knowledge regarding possible DIF is available, an investigator may wish to first consider the model in which all items are hypothesized to exhibit DIF and proceed in an exploratory fashion on the basis of 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 DIF items.

Example Applications

Kok (1982) experimentally studied DIF in multiplication items by manipulating the test takers' skill on a possible DIF 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 in Dutch, whereas for others, Roman numerals were used. Knowledge of Roman numerals was expected to be a DIF factor, because Dutch undergraduates show differences in their ability to decipher Roman numerals. DIF was further related to a manifest grouping variable by giving 143 randomly selected undergraduates some training regarding Roman numerals. It was of course expected that the

Roman numeral items would be more difficult for the untrained group than for the trained group.

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 content and proportions of correct answers are presented in Table 3. The six chosen items were selected on the basis of the nature of their multiplication content. All six items had the following common properties: (a) There is a single-digit multiplier that 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 were easier for the trained group than for the untrained group. The Roman numeral items were, however, easier than the native-language items, even for the untrained students.

Because the multiplication task differed very little across items, it might reasonably be expected that there are two latent ability states, mastery and nonmastery. 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 to better exemplify the applications of these modeling techniques.

Because there is apparently only one DIF factor in this data (Roman numerals decoding), models with a combination of manifest and latent subgroups are not appropriate. Additionally, these models (both for the continuous and discrete cases) were not identified. For both reasons, these models will not be further addressed in this example.

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. On the basis of these results, it may be concluded that the Rasch model (Model 1) does not adequately fit the data.

Considering a latent grouping variable, we see that Models 5 and 6 marginally fit the data. Furthermore, they do not differ significantly ($G_D^2 = 3.92$, $df = 2$,

Table 3
Homogeneous Multiplication Items Presented in Native-language and Roman-numerals Formats

Item	Multiplication	Presentation	Proportion 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

$p = .14$), so we choose the more parsimonious Model 6 as the preferred model for this pair. Recall that this preferred Model 6 allows only the Roman numeral items to exhibit DIF.

Table 4 presents the Rasch item difficulty parameters that can be calculated from the β parameters of Model 6. The parameters are calculated by means of the following formula:

$$\begin{aligned} &\delta_{iu} = \beta_0^{X_i} - \beta_1^{X_i} && i = 1, 2, 3; \\ &\text{and} && \qquad (7) \\ &\delta_{iu} = (\beta_0^{X_i} + \beta_{0u}^{X_iU}) - (\beta_1^{X_i} + \beta_{1u}^{X_iU}) && i = 4, 5, 6, \end{aligned}$$

where δ_{iu} is the item difficulty of item i for the uth latent group (Kelderman, 1989). To fix the scale, the difficulty of the first item is set equal to zero by setting the corresponding β parameters equal to zero. Looking at Table 4 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 in working with Roman numerals, whereas the second class appears to contain students who do not have this skill. Note that the difference in difficulty between both latent classes is considerably larger for the last Roman numeral item than for the other Roman numeral items. It therefore seems that the latent class variable is highly correlated with the last Roman numeral item.

Next, consider the case where the grouping variable is manifest. Models 3 and 4 (Table 1) both marginally fit the data, and their difference is not significant ($G_D^2 = 4.15$, $df = 2$, $p = .13$). So again (as in the case of the latent grouping variable models) we choose the more parsimonious Model 4.

In Table 5, the Rasch item difficulties for Model 4 are presented. From Table 5 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 corresponds to those obtained with latent subgroups. However, a marked difference between the latent subgroups solution and the manifest subgroup solution is found in the last Roman numeral item: The difference in item difficulty between the latent subgroups is much larger than for the manifest subgroups, whereas for the remaining Roman numeral items, the difference in item difficulty between the

Table 4
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 5
Item Difficulty Estimates of Model {X1},{X2},{X3},{GX4},
{GX5},{GX6},{TG} (Model 4) from Table 1

	Item					
	Native Language			Roman Numerals		
Observed Subgroup	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

latent subgroups is smaller than for the manifest subgroups. This suggests that the strong relationship between the latent class variable and the last Roman numeral item cannot be explained entirely by the effect of Roman numerals training alone. The latent class variable seems to pick up an effect that is peculiar to Item 6. The marginality of the fit of Models 4 and 6 may very well have resulted from inadequate explanation of this effect.

Consider now the case where the assessed attribute has two states: mastery or nonmastery. Table 2 shows that the Two-State Mastery model does not fit the data (see Model 1), nor do any of the models with a manifest grouping variable (i.e., Models 2, 3, and 4). Of the models with latent grouping variables, only Model 5 has an acceptable fit.

In Table 6, 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 nonmasters) and the latent class proportions. These parameter estimates can be calculated from the β parameters by means of the following equations (see Haberman, 1979, p. 551):

$$\frac{\exp\,(\beta_1^{X_i} + \beta_{1u}^{X_iU})}{\exp\,(\beta_1^{X_i} + \beta_{1u}^{X_iU}) + \exp\,(\beta_0^{X_i} + \beta_{0u}^{X_iU})}$$

$i = 1, \ldots, k$ for the conditional probabilities of having a positive response to item i given Latent Class u, and

$$\frac{\sum\limits_{x_1} \cdots \sum\limits_{x_k} \exp\,(\beta_1^U + \beta_{x_1 1}^{X_1U} + \cdots + \beta_{x_k 1}^{X_kU})}{\sum\limits_{x_1} \cdots \sum\limits_{x_k} \sum\limits_{u} \exp\,(\beta_u^U + \beta_{x_1 u}^{X_1U} + \cdots + \beta_{x_k u}^{X_kU})}$$

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

The estimated conditional probabilities presented in Table 6 are difficult to interpret in terms of the latent 2 × 2 joint levels of mastery and grouping. A

Table 6
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	Native	0.88	0.40	0.26	0.10
2	Native	0.77	0.21	0.40	0.00
3	Native	0.77	0.13	0.29	0.00
	Mean	0.81	0.25	0.31	0.03
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
	Mean	0.80	0.86	0.23	0.32
		Latent Class Proportions			
		0.21	0.30	0.12	0.37

possible interpretation for each latent class is specified between parentheses (see the latent class headings in Table 6). 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 who have some facility at working with Roman numerals (this, to a large extent, may include students in the trained group), whereas students in 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 nonmasters. 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 the mastery class (1) than for the nonmastery class (2). Moreover, in the inexperienced/untrained group, the conditional probabilities for Items 4 and 5 are smaller in the mastery class (3) than in the nonmastery class (4). The parameters, therefore, are not fully interpretable in terms of a combination of mastery and DIF classes.

The model with a continuous measured trait and DIF in 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.

Presented in Table 7 are the conditional probabilities and the latent class proportions that correspond to Model 4*. On the basis of 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 nonmastery (NM); mixed mastery (MM; i.e., a latent class containing individuals who have mastered or partially mastered some of the items while not mastering the remaining items); and mastery (M) states. Because 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 may be seen that in the nonmasters class and the masters class, the trained subjects have higher conditional probabilities than the untrained subjects. However, the mixed-masters conditional probabilities do not seem to be affected by Roman numerals training. The conditional probabilities of both trained and untrained subjects are about the same. Furthermore, it is noteworthy that the conditional probability of the last Roman numerals item is equal to one, whereas other conditional probabilities of this item are considerably lower. It seems, therefore, that this item is strongly related to the mixed-masters class. As was also suggested by the analysis with a continuous latent trait, the last Roman numerals item seems to measure an effect that is not adequately explained by training or the latent trait.

Another possible interpretation of the conditional probabilities of the Roman numeral items given the mixed-mastery latent class is in terms of Bergan and

Table 7
Parameter Estimates of Model {LX1},{LX2},{LX3},{GLX4},{GLX5},{GLX6} with Three Mastery States and (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	Native	.11	.45	.85	.11	.45	.85
2	Native	.07	.17	.88	.07	.17	.88
3	Native	.03	.21	.71	.03	.21	.71
Mean		.07	.27	.81	.07	.28	.81
4	Roman	.46	.82	.86	.27	.78	.75
5	Roman	.56	.73	1.00	.21	.77	.61
6	Roman	.29	1.00	.71	.20	1.00	.66
Mean		.44	.85	.86	.23	.85	.67
		Latent Class Proportions					
		.41	.42	.16	.54	.19	.27

Stone's (1985) hierarchical ordering of items. The conditional probabilities of the Roman numerals items in the mixed-mastery class are about as large as those in the mastery class, whereas in the native-language items they are smaller. Therefore, Roman numeral items appear to be mastered before native language items.

For the case of a latent grouping variable, a model with three mastery states would not be identifiable when only six items are considered. Thus, we do not consider a Model 6 with three levels of assessed mastery for these data.

Discussion

In this paper, we have shown that it is possible to explain DIF 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 DIF may occur, are manifest, latent, or both.

The modeling framework that we have 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 that 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, because many such models will not be identifiable. For example, it is easily shown that adding a term $\{X_4X_5X_6\}$ 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 problem occurs because the minimum sufficient information for parameter estimation are the contingency table frequencies. The number of these frequencies increases exponentially with the 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 procedure increases the likelihood that unknown DIF factors 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 or the initial values used in the iterative estimation process are not themselves reasonably accurate. For example, 449 iterations were needed to obtain estimates for the Rasch model with the Roman numeral items showing DIF 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 1,501 to obtain a precision of five decimal places. An advantage of the IPF algorithm, however, is that iterations may be very quickly implemented 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 on a VAX8650 computer 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-mentioned problem of not obtaining acceptable convergence.

In the case of a continuous latent trait, a Rasch model is assumed, which implies the assumption that the nonbiased items all have the same discrimination parameter. This may be felt as a limitation of this item-bias detection method. For the case of a manifest grouping variable, Kelderman (1989) performed a sensitivity analysis focusing on this feature. His simulation results showed that the test for one biased item is rather robust for deviations of the equal-discrimination assumption in the remaining items. For the case of a latent grouping variable, the inclusion of discrimination parameters gives rise to near-identification problems leading to estimation problems, unless parameters are fixed in advance to a certain value.

Models where more complicated IRT models are combined with latent subgroups and where model parameters are fixed have been described by several authors. Mislevy and Verhelst (1987) chose a linear-logistic test model and Yamamoto (1987, 1988) a two-parameter-logistic model. These models have the same basic philosophy: an IRT model combined with latent classes.

References

Alvord, G., & Macready, G. B. (1985). Comparing fit of nonsubsuming 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, MD: Johns 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. *Review of Research in Education, 10,* 305–360.

Bergan, J. R., & Stone, C. A. (1985). Latent class models for knowledge domains. *Psychological Bulletin, 98,* 166–184.

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

Cardal, C., & Coffman, W. T. (1964). *A method for comparing performance of different groups on the items in a test* (RM No. 64-61). Princeton, NJ: Educational Testing Service.

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

Clogg, C. C. (1977). *Unrestricted and restricted maximum likelihood latent structure analysis: A manual for users* (Working paper No. 1977-09). University Park: The Pennsylvania State University, Population Issues Research Office.

Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association, 79,* 762–771.

Cochran, W. G. (1952). The chi-squared test of goodness of fit. *Annals of Mathematical Statistics, 23,* 315–345.

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 behavioural hierarchies. *Psychometrika, 41,* 189–204.

Dayton, C. M., & Macready, G. B. (1980). A scaling model with response errors and intrinsically unscalable respondents. *Psychometrika, 45,* 343–356.

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—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. *American Journal of Sociology, 78,* 135–191.

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. *American Journal of Sociology, 79,* 1179–1259.

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,* 755–768.

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

Hagenaars, J. A. (1988). LCAG—Loglinear modeling with latent variables: A modified LISREL approach. In W. Saris and I. Gallhofer (Eds.), *Sociometric research.* (Vol. 2, pp. 111–130). London: MacMillan.

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

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel Procedure (Research Report No. 86-69). Princeton, NJ: Educational Testing Service.

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

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

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

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

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

Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association, 56,* 223–234.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Linn, R. L., & Drasgow, F. K. (1987). Implications of the golden rule settlement of test construction. *Educational Measurement: Issues and Practice, 6,* 13–17.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: 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.

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

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

Mislevy, R. J., & Verhelst, N. (1987). Modeling item responses when different subjects employ different solution strategies. *Research Bulletin,* No. RR-87-47-ONR. Princeton, NJ: Educational Testing Service.

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

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: 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.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128.

Thissen, D., Steinberg, L., & Wainer, H. (1989). *Detection of differential item functioning using the parameters of item response models.* Lawrence: University of Kansas.

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 Educational Statistics, 3,* 305–318.

Yamamoto, K. (1987). *A hybrid model for item responses.* Unpublished doctoral dissertation, University of Illinois.

Yamamoto, K. (1988). *Hybrid model of IRT and latent class models.* Princeton, NJ: Educational Testing Service.

Authors

HENK KELDERMAN is Assistant Professor, Department of Educational Measurement and Data-Analysis, University of Twente, Postbus 217, 7500 AE, Enschede, The Netherlands. *Degrees:* MA, University of Amsterdam; PhD, University of Twente. *Specializations:* psychometrics and loglinear IRT.

GEORGE B. MACREADY is Professor, Department of Measurement, Statistics, and Evaluation, College of Education, University of Maryland, College Park, MD 20742. *Degrees:* BA, Willamette University; MA, University of Oregon; PhD, University of Minnesota. *Specialization:* latent structure modeling.