Explanatory Secondary Dimension Modeling of Latent Differential Item Functioning

Paul De Boeck, Sun-Joo Cho and Mark Wilson

Applied Psychological Measurement 2011 35: 583
DOI: 10.1177/0146621611428446

Article

Applied Psychological Measurement
35(8) 583–603
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621611428446
http://apm.sagepub.com

Explanatory Secondary Dimension Modeling of Latent Differential Item Functioning

Paul De Boeck[1,2], Sun-Joo Cho[3], and Mark Wilson[4]

Abstract

The models used in this article are secondary dimension mixture models with the potential to explain differential item functioning (DIF) between latent classes, called latent DIF. The focus is on models with a secondary dimension that is at the same time specific to the DIF latent class and linked to an item property. A description of the models is provided along with a means of estimating model parameters using easily available software and a description of how the models behave in two applications. One application concerns a test that is sensitive to speededness and the other is based on an arithmetic operations test where the division items show latent DIF.

Keywords

differential item functioning, dimensionality, item response model, latent class model

For *manifest groups*, differential item functioning (DIF) means that, conditioning on the person trait (e.g., ability), one or more items function differently in a focal group in comparison with a reference group (Scheuneman, 1979). In other words, equal levels of the person trait lead to different response probabilities depending on the group one belongs to. The difference relates most commonly to the item location and more rarely to the slope or gradient of the item, which, in an item response context, are commonly called the difficulty and degree of discrimination of an item, respectively. Reviews and discussions of the topic can be found in Millsap and Everson (1993) and Holland and Wainer (1993) and more recently in Teresi and Fleishman (2007).

It is common to view DIF as the consequence of one or more dimensions not accounted for and, consequently, as the failure to account for *secondary dimensions*, also called ''nuisance dimensions'' that can result in DIF (Ackerman, 1992; Bolt & Stout, 1996; Camilli, 1992;

[1] University of Amsterdam, Netherlands
[2] K. U. Leuven, Belgium
[3] Vanderbilt University, Nashville, TN, USA
[4] University of California, Berkeley, USA

Corresponding author:
Sun-Joo Cho, Department of Psychology and Human Development, Vanderbilt University, Peabody #H213A, 230 Appleton Place, Nashville, TN 37203, USA
Email: sj.cho@vanderbilt.edu

Douglas, Roussos, & Stout, 1996; Finch, 2005; Kok, 1988; MacIntosh & Hashim, 2003; Oort, 1998; Oshima & Miller, 1992; Roussos & Stout, 1996; Shealy & Stout, 1993a, 1993b). Ackerman (1992) described nuisance dimensions as skills, which an examinee uses to solve particular items but which are not part of the intended construct, that is, the primary dimension, measured by the test. Reading ability, for example, can be considered a nuisance dimension in a test designed to measure algebraic symbol manipulation as the primary dimension.

In this article, the authors take a somewhat different approach on these two previous points: manifest groups and secondary dimensions. First, DIF is understood as also encompassing *latent DIF* (DIF between latent classes), as in Cho and Cohen (2010), and Cohen and Bolt (2005). Instead of working with manifest reference and focal groups, this study works with latent classes: a non-DIF latent class and a DIF latent class. In terms of Meredith (1993), the selection variable that defines the subpopulations for measurement invariance is a discrete latent variable: membership in one of the two latent classes. Second, this study takes a *mixed dimensionality* approach, where the secondary dimension plays a role in one group (the DIF latent class) but not in the other (the non-DIF latent class). One class is one-dimensional, and the other is two-dimensional. In contrast with most studies, the focus is on the modeling and explanation of DIF, rather than on detection of items that display DIF.

The idea of mixed dimensionality can also be applied to manifest groups, and the mixture approach can be used without mixed dimensionality. Here, the study addresses the more general case of latent DIF *and* mixed dimensionality. Using just one element, either latent DIF or mixed dimensionality, is straightforward. The following sections will explain why each of these elements, and therefore also the combination, is of interest.

Latent DIF

The notion of DIF applies when items function differently, comparing latent classes and conditioning on the person trait (Cho & Cohen, 2010; Cohen & Bolt, 2005). There can be four a priori reasons to consider latent DIF. The first two involve situations where no observation can be made about group membership, either because the observer has no idea what the crucial groups would be (i.e., the ‘‘no idea’’ reason) or because the group of interest is not observable (i.e., ‘‘unobservable’’ reason). An example of the latter is a cognitive strategy. The last two reasons refer to cases where observations of group membership are made or could have been made, but those observations may not be perfectly reliable (i.e., the ‘‘reliability’’ reason) or they may not provide a perfectly valid indication of the true group membership (i.e., the ‘‘validity’’ reason). In all such cases, the notion of latent DIF can be useful. It is possible to combine the latent and manifest, that is, the manifest (or observed) group memberships as predictors of latent class probabilities. However, for reasons of simplicity, and because the two illustrations provided are applications of the ‘‘unobservable’’ type, in this study latent DIF is used, and thus a mixture approach, without a role for manifest groups.

Secondary Dimensions

Shealy and Stout (1993a, 1993b) described a *multidimensional model for DIF* (MMD). Based on MMD, Roussos and Stout (1996) developed methods to detect DIF, with two assumptions: (a) in addition to the intended θ_1 construct, the test also measures secondary constructs, θ_2, θ_3, and so on and (b) the focal and reference groups differ with respect to the level of the secondary constructs. The secondary dimension is assumed to play a role in both groups and is a factor that plays a role in the creation of DIF because the two groups differ with respect to this secondary dimension. This approach combines elements from a substantive analysis of the items, and statistical methods such as DIMTEST (Stout, 1987) to identify multidimensionality, and SIBTEST (Stout & Roussos, 1995) for a DIF analysis. The examples Roussos and Stout gave stem from Douglas et al. (1996). The approach is used as a basis for DIF detection in the first place, not to estimate a model for the data that can account for DIF. Although this approach has been described for manifest DIF, it can easily be generalized to latent DIF.

This article shares the idea of a substantive analysis, not as a basis for detecting DIF but rather to model and explain DIF on the basis of item properties. Explanation of DIF is one of the important trends in the study of DIF (Cho & Cohen, 2010; Cohen & Bolt, 2005; Cohen, Gregg, & Deng, 2005; DeAyala, Kim, Stapleton, & Dayton, 2002; Muthén, Kao, & Burstein, 1991; Muthén & Lehman, 1985; Penfield, 2010; Samuelsen, 2005; Zumbo & Gelin, 2005). The explanatory approach taken in this study hinges on the availability of a known item property that can explain latent DIF and that is at the same time the source of a random effect called a secondary dimension.

Another approach that could be seen as a dimensionality approach is the *multiple-indicator multiple-cause* (*MIMIC*) *latent variable model*. This model uses multiple indicators to measure a latent trait and also one or more external causes with an effect on the latent trait and on the indicators. If external variables, such as gender, have a direct effect on an indicator, which cannot be explained through the latent trait, then one may conclude that there is DIF. The MIMIC model has become a popular approach to DIF (Finch, 2005; Fleishman, Spector, & Altman, 2002; MacIntosh & Hashim, 2003; Muthén et al., 1991; Muthén & Lehman, 1985; Woods, 2009). However, although the MIMIC approach relies on a dimensional view, only the intended latent trait has been used (the primary dimension), without a secondary dimension. A remarkable exception can be found in Glöckner-Rist and Hoijtink (2003). These authors set up a two-dimensional model and regress both latent traits on external variables. However, they do not go as far as invoking the secondary dimension as the basis for DIF in the same test.

Mixed Dimensionality as an Alternative

The approach proposed here differs from the MMD and MIMIC approaches. First, the MMD approach assumes a secondary dimension in the reference group and the focal group. Translated for latent DIF, this implies a secondary dimension in the non-DIF and DIF latent classes. A mixed dimensionality approach is characterized by a secondary dimension that is limited to the DIF latent class. The factor at the root of the DIF does not play a role in the non-DIF class, but it does in the DIF class, while there are no observations apart from the responses as an indication for the group in which it does or does not play a role.

Second, the MIMIC approach does not imply a secondary dimension but a direct effect of the group variable on the indicators beyond a possible effect on the latent trait (the primary dimension). Translated for latent DIF, this means that neither of the classes has a secondary dimension but that the classes differ in their item parameters. In contrast with the MIMIC approach, mixed dimensionality does imply a secondary dimension but only in the DIF latent class.

The authors believe that mixed dimensionality is a valuable alternative, with a secondary dimension in one class, instead of a secondary dimension in either both (MMD), or in neither (MIMIC), class. However, unlike the MMD and MIMIC approaches, the mixed dimensionality approach is primarily meant for DIF explanation, after one has established DIF or for DIF hypothesis testing. For an exploratory detection approach, one would need a plausible set of item properties that can be tried out in an exploratory way.

It is true that mixed dimensionality, with a secondary dimension in just one class, is perhaps unusual. First, there may seem to be a conflict, because, on one hand, a secondary dimension is invoked to explain a difference between two classes, and, on the other hand, the dimension exists only in one of these classes. The reason that this is not really a conflict will become clear from a more detailed explanation of the models, but as a brief explanation, note that the absence of a dimension means zero variance, while the presence of a dimension means a nonzero variance and possibly a mean that is different from zero.

A second possible issue is whether the introduction of a nuisance dimension implies local dependency. The authors believe it does, as explained by Bradlow, Wainer, and Wang (1999); Tuerlinckx and De Boeck (2004); and Ip (2010), among others. If a set of DIF items shares an item property that can explain the DIF, and this item property is also the source of a nuisance dimension, then DIF and local dependency are formally equivalent. This makes the mixed dimensionality approach an interesting illustration of how phenomena that are traditionally labeled differently, such as DIF and local dependency, can sometimes in fact be seen as unitary.

In Terms of Measurement Invariance Notions

DIF implies lack of measurement invariance. Compared with factor models, item response theory (IRT) models have a somewhat different parameterization, so that the notions developed for factor models require some explanation when transposed to IRT models. Lack of ''scalar invariance'' (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000) relates to differences in item difficulties (intercepts) and implies, therefore, uniform DIF. Lack of ''metric invariance'' (Horn & McArdle, 1992; Millsap & Kwok, 2004; Vandenberg & Lance, 2000) relates to differences in item discriminations and therefore implies nonuniform DIF. However, because for binary items, discriminations and specific variance are two sides of the same coin, it is not possible to differentiate between ''metric invariance'' and ''uniqueness invariance.'' In terms of Meredith (1993), metric and scalar invariance are implied by strong factorial invariance.

The idea of mixed dimensionality can be related to the notion of measurement invariance as follows. Making use of mixed dimensionality, it is possible to regain scalar invariance (equivalence of difficulties) because the secondary dimension is added instead of inequivalence of difficulties in the unidimensional model. The price that is paid in both cases is the loss of configural invariance (similar discrimination pattern; Horn & McArdle, 1992; Vandenberg & Lance, 2000) of the global model because the class-specific secondary dimension distorts the similarity of the discrimination patterns of the two classes. However, because the absence of the secondary dimension in one class is formally equivalent with a zero variance of this dimension, it is strictly speaking not a distortion of metric invariance and, hence, also not a distortion of configural invariance. The discriminations of the primary dimension are the same in the two latent classes and those of the second are as well but with zero variance in the non-DIF latent class. The greater complexity of the model must not be seen as a disadvantage because the model provides an explanation for the DIF and thus for the lack of measurement invariance.

Aim of the Study

The aim of this study is to show that a mixed dimensionality model can explain latent DIF in the statistical and substantive sense. In the statistical sense, a model with a secondary dimension in just the DIF latent class is expected to fit the data reasonably well if latent DIF has been established. In the substantive sense, the basis for the nuisance dimension is an item property that is linked with the DIF and that has the potential to generate a nuisance dimension. As this approach relies totally on item properties, it is less appropriate for latent DIF detection but is rather appropriate for DIF explanation. Two applications are described where latent DIF has been established before and where there seems to be an item property that may be at the basis of the DIF, while it also can be expected to generate a secondary dimension. Whether the model with a secondary dimension in just the DIF latent class fits the data as well or better, and also whether including a secondary dimension in the non-DIF latent class improves the goodness of fit, is checked.

In the next section, the models to be considered are described and the aim of the studies to be reported is explained in terms of these models. This is followed by a section on model estimation and evaluation, and by two sections delineating the applications. In the first, speededness is at the basis of the DIF, and in the second, DIF can be explained by the kind of arithmetic operation required by the item.

Models

The models to be discussed in the following are all mixture item response models, either fully one-dimensional (with no secondary dimension, for example, Cohen & Bolt, 2005), mixed one-dimensional and two-dimensional (with a secondary dimension limited to the DIF latent class, as in the proposed models in this study), or fully two-dimensional (with a secondary dimension for the non-DIF latent class and the DIF latent class, for example, Cho, Cohen, & Templin, 2008). The one-dimensional models are used as reference models: as instances where DIF is detected before a secondary dimension is introduced. The mixed dimensionality models are the authors' alternative to a fully two-dimensional model, developed for reasons explained earlier. The extension to fully two-dimensional models are checked to see whether the data require the use of a secondary dimension for both latent groups. The models are presented for the case of one item feature suspected of leading to DIF.

All models are confirmatory models because they are all based on an explanatory item property. For the sake of simplicity, only one such property is considered, but in principle, more properties can be focused on. However, to have explanatory power, the properties must be shared by more than one item.

Apart from the mixture models, and for reasons of comparison, two two-dimensional models without classes will also be estimated: a confirmatory model and an exploratory model, with the constraints that the variances are one and the correlation is zero. The confirmatory model shares with the previous models that, when the items do not have the property, the loadings on the second dimension are constrained to be zero. In the exploratory model, all loadings are free parameters.

One-Dimensional Mixture Models

Model 0: 1d-non-DIF. The first model is the mixture 2-parameter logistic (2PL) model for two latent groups, which is a non-DIF one-dimensional mixture model (*1d-non-DIF*) and possesses measurement invariance. It functions as the reference model instead of the simple 2PL model because it is, in common with the models to be applied, a mixture model. The logit of the probability of success of person j on item i is defined as follows in Model 0:

(a) if person j is a member of the non-DIF latent class ($\gamma_j = 0$), indicated with subscript r:

$$\eta_{ijr} = \alpha_i \theta_{jr} + \beta_i, \tag{1}$$

and (b) if person j is a member of the DIF latent class ($\gamma_j = 1$), indicated with subscript f:

$$\eta_{ijf} = \alpha_i \theta_{jf} + \beta_i, \tag{2}$$

with η_{ijr} and η_{ijf} as the two logits of the probability of success;

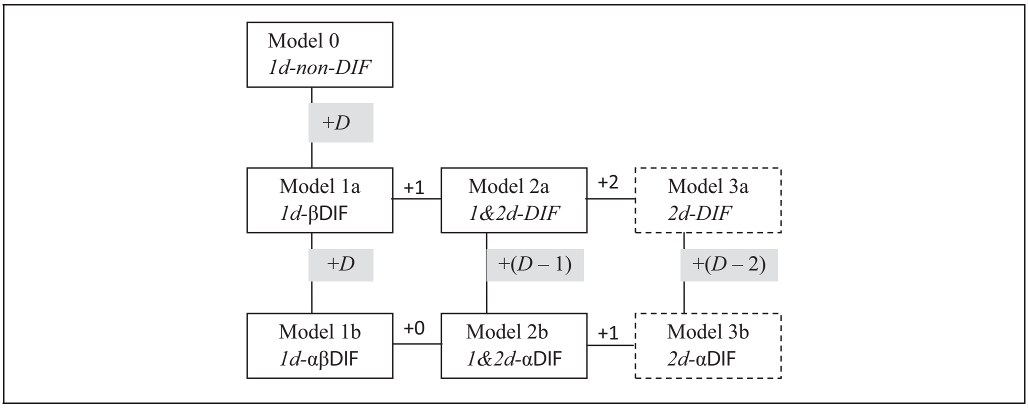

Figure 1.
Note: Relationships between dimensional latent DIF models. The numbers of free parameters for the seven models are obtained starting by the number for Model 0 and adding the numbers on the connecting lines between the models. Hierarchical relationships are indicated by lines flagged with the difference in number of free parameters (D is the number of DIF items). Models 3a and 3b are investigated only as extensions of the Models 2a and 2b, and they are therefore put in boxes with dashed lines.

with α_i as the degree of discrimination of item i, the same in both groups;
with β_i as the difficulty of item i, the same in both groups; and
with θ_{jr} and θ_{jf} as the person trait in the non-DIF latent class and the DIF latent class, respectively, $\theta_{jr} \sim N(\mu_r, \sigma_r^2)$ and $\theta_{jf} \sim N(\mu_f, \sigma_f^2)$.

For reasons of identification, $\sigma_r^2 = 1$ and $\mu_r = 0$. Model 0 has $2I + 3$ free parameters: I difficulties and I degrees of discrimination, a mean and variance of the DIF latent class, and a mixing probability $\pi_r = 1 - \pi_f$.

Model 1a: 1d-βDIF. Next, consider a one-dimensional difficulty DIF mixture model (*1d-βDIF*). It differs from Model 0 in that Equation 2 for the DIF latent class now becomes

$$\eta_{ijf} = \alpha_i \theta_{jf} + \beta_{if}, \tag{3}$$

with β_{if} as the difficulty of item i if $\gamma_j = 1$ and $\beta_{if} = \beta_i$ if item i does not have the item property that is expected to induce DIF.

Model 1a has $2I + D + 3$ free parameters, with D as the number of items with the property in question and therefore also the number of items suspected of showing latent DIF. Whether they actually show DIF is an empirical issue. Model 1a has D more free parameters than Model 0, as shown in Figure 1. Because there is a hypothesis about which items are DIF items, based on the item property under consideration, their number is known. Note, also, that (partly) continuous item properties can be used for this purpose but that for the models described here and for the applications, binary item properties are used. Model 1a does not fulfill scalar invariance, but it is metric invariant.

Model 1b: 1d-αβDIF. This is also a one-dimensional mixture model but now with difficulty DIF and discrimination DIF (*1d-αβDIF*). It differs from Model 1a in that also the discrimination of the DIF items is different in the DIF latent class:

$$\eta_{ijf} = \alpha_{if} \theta_{jf} + \beta_{if}, \tag{4}$$

with α_{if} as the discrimination of item i, if $\gamma_j = 1$. The number of free parameters is now $2(I + D) + 3$. The difference in the number of free parameters with Model 1a is D, as shown in Figure 1. Of course, this implies that there can be discrimination DIF whenever there is difficulty DIF and vice versa. As in the previous case, DIF is linked to a binary item property. However, whether the two kinds of DIF do co-occur is an empirical issue. Model 1b does not fulfill scalar invariance; neither does it fulfill metric invariance.

Mixture of One-Dimensional and Two-Dimensional Model

Model 2a: 1&2d-DIF. As explained earlier, it is possible that DIF can be explained by a secondary (nuisance) dimension present only in the DIF latent class (*1&2d-DIF*), while the same primary dimension applies to both latent groups. In Model 2a, the item difficulties, as well as the discriminations on the primary dimension, are the same in the two latent classes. For the non-DIF latent class, if $\gamma_j = 0$, Equation 1 applies. However, for the DIF latent class, if $\gamma_j = 1$, the logit is formulated as follows:

$$\eta_{ijf} = \alpha_{i1}\theta_{jf1} + \alpha_{i2}\theta_{jf2} + \beta_i, \tag{5}$$

with α_{i1} as the discrimination of item i for the primary dimension, $\alpha_{i1} = \alpha_i$, and α_{i2} as the discrimination of item i for the secondary dimension, but $\alpha_{i2} = 0$, if item i is not a DIF item;
with $(\theta_{pf1}, \theta_{pf2})' \sim BVN\left(\mu_{f1}, \mu_{f2}, \sigma^2_{f1}, \sigma^2_{f2}, \sigma_{f12}\right)$; and
with μ_{f1}, μ_{f2} as the DIF latent class means of the primary and secondary dimensions, respectively;
with $\sigma^2_{f1}, \sigma^2_{f2}, \sigma_{f12}$ as the DIF latent class variance and covariance of the primary and secondary dimensions, respectively, assuming, for identification reasons, that $\sigma^2_{f2} = 1$.

In addition, one discrimination parameter, for a DIF item $i*$ and Dimension 2, α_{i*2}, is fixed to 1.00. Also, for identification reasons, the previous restrictions regarding the non-DIF latent class still apply. Model 2a has $2I + D + 4$ free parameters, as can be derived from Figure 1. In comparison with Model 1a, it has one more mean (of Dimension 2 in the DIF latent class), and it has a covariance parameter for the latent DIF class, but one parameter is lost by fixing the discrimination of a DIF item on the second dimension.

Note that in Model 2a, there is no longer a difference in difficulty between the two latent classes. Interestingly, Models 1a and 2a have the feature of metric invariance (for Model 2a this applies only to the primary dimension), but Model 1a lacks scalar invariance. The scalar invariance is regained in Model 2a, at the cost of introducing a secondary dimension. The implications of Model 2a are twofold:

(a) The respondents from the DIF latent class differ with respect to how much DIF they show (depending on θ_{pf2}). The idea of individual differences in DIF has been used earlier (Camilli & Penfield, 1997; Longford, Holland, & Thayer, 1993; Van den Noortgate & De Boeck, 2005), but it also follows, of course, from a nuisance dimension view on DIF.
(b) All DIF is assumed to rely on one secondary dimension. This assumption will certainly not hold in all cases. It is in line with an explanatory approach to DIF that the source of DIF is specified, but there is not necessarily just a single source.

Model 2b: 1&2d-αDIF. As an extension of the previous model, the discrimination of the DIF items may be different for the primary dimension depending on the latent class (*1&2d-αDIF*). The role of the secondary dimension may have an impact on how well the DIF items measure the primary dimension. This leads to the following for $\gamma_j = 1$:

$$\eta_{ijf} = \alpha_{if1}\theta_{jf1} + \alpha_{i2}\theta_{jf2} + \beta_i, \tag{6}$$

with α_{if1} as the discrimination of item i for dimension 1, but $\alpha_{if1} = \alpha_i$ for all non-DIF items.

The covariance σ_{f12} is fixed to zero for identification reasons. Therefore, the difference in the number of parameters between Models 2a and 2b is $D-1$, as shown in Figure 1, with a total of $2(I+D)+3$ parameters. Model 1b does not imply metric invariance; neither does it imply scalar invariance. Interestingly, also Model 2b regains the feature of scalar invariance at the cost of introducing a secondary dimension.

Two-Dimensional Mixture Models

Models 3a and 3b: 2d-DIF and 2d-αDIF. Models 2a and 2b can be extended such that the secondary dimension also applies to the non-DIF latent class, so that Models 3a and 3b are obtained (*2d-DIF and 2d-*α*DIF*). As a consequence, Model 3a has two more free parameters than Model 2a: a secondary dimension variance σ^2_{r2} and a covariance σ_{r12}, while the mean of the secondary dimension in the non-DIF latent class is fixed $\mu_{r2} = 0$. Similarly, Model 3b has a variance σ^2_{r2}, but $\sigma_{r12} = 0$, in line with $\sigma_{f12} = 0$ in Model 2b and again $\mu_{r2} = 0$. Note that Model 3a is itself not formally a DIF model because the within-latent class item parameters are identical for the two latent classes. It is still denoted here as *2d-DIF* because it can explain DIF that would appear when a one-dimensional model is used. The latent classes of Model 3a differ only in the means and variance of the latent traits.

Models 3a and 3b share the feature of scalar invariance. Model 3a is also metric invariant; it is a fully measurement-invariant model. Because of differences in discriminations, Model 3b is not metric invariant, but it shows configural invariance instead.

Aim of the Applications

In the following, two applications are described. They share that one source of latent DIF may be expected and that the item property at the basis of this latent DIF is known, while it is unclear which subset of the person sample actually shows DIF. That the subset is unclear means that membership of the subset is not given. This is precisely what latent class models can take into account, so that a latent DIF approach seems appropriate. The first application is on speededness. For the data to be analyzed, which items suffer from speededness is known (Bolt, Cohen, & Wollack, 2002) but not which respondents show speededness. The second application is on arithmetic operations (addition, subtraction, multiplication, division). From an earlier analysis of the same data (Van Nijlen & Janssen, 2010), it is clear that items requiring division are more difficult for some of the students, while it is not known for which students that is.

An important preliminary question is whether any DIF occurs at the level of latent classes. For that reason Model 0 (*1d-non-DIF*) will be compared with Models 1a and 1b (*1d-*β*DIF* and *1d-*αβ*DIF*). Given that there are indeed indications for DIF, three main questions are formulated, referring to the horizontal and vertical comparisons in Figure 1.

1. Do the goodness of fit results confirm the theoretical analysis that a secondary dimension in the DIF latent class can explain the DIF (β*DIF*) that is found in a one-dimensional model? This implies that the *first* (*and leftmost*) *horizontal comparison* in Figure 1 should be investigated, between Models 1a and 2a on one hand (*1d-*β*DIF* and *1&2d-DIF*) and Models 1b and 2b on the other hand (*1d-*αβ*DIF* and *1&2d-*α*DIF*).

2. Is the DIF of a kind that the discrimination on the primary dimension is affected? This implies that *the vertical comparison* in Figure 1 should be investigated, of Model 1a with 1b (*1d*-β*DIF* vs. *1d*-αβ*DIF*) and Model 2a with 2b (*1&2d-DIF* vs. *1&2d*-α*DIF*).
3. Is the secondary dimension limited to the DIF latent class or does it also apply to the non-DIF latent class? This issue implies the *second* (*and rightmost*) *horizontal comparison* in Figure 1, comparing Model 2a with Model 3a (*1&2d-DIF* and *2d-DIF*) or Model 2b with Model 3b (*1&2d*-α*DIF* and *2d*-α*DIF*). For this comparison, the best fitting of Models 2a and 2b is taken and a nonzero variance of the secondary dimension in the non-DIF latent class is allowed for.

For an interpretation of the DIF latent class in Models 2a and 2b, the proposed mixed dimensionality models, the following results are important. First, the DIF latent class, with the secondary dimension, is expected to be the minority class (although this is not necessary). Given that the authors work with latent classes, there is no external basis for deciding what the reference group is, other than the majority. Second, it is expected that the discriminations of items for the secondary dimension are larger than zero when the items have the crucial item property. This means that, for these items, an additional factor affects the DIF class, compared with the non-DIF class. This other factor may be at the cost of the primary dimension, which would then result in decreased discriminations for the primary dimension. Third, the means of the primary dimension (μ_{f1}) and of the secondary dimension (μ_{f2}) inform us about advantages and disadvantages for the DIF latent class. The mean of the DIF latent class on the primary dimension informs us about the level of the primary latent trait in the DIF latent class, and the mean on the secondary dimension tells us what the advantage or disadvantage is of relying on the secondary dimension. If both means are smaller than zero, then the DIF class has a double disadvantage. For each of the two applications, these three points are discussed. One should realize, however, that alternative identification constraints may lead to a somewhat different interpretation.

Model Estimation and Evaluation

Estimation

The computer program LatentGOLD 4.5 Syntax module (Vermunt & Magidson, 2007) was used to estimate the model parameters. LatentGOLD 4.5 Syntax module uses the marginal maximum likelihood estimation. Details on using LatentGOLD for item response models are provided in Vermunt and Magidson (2005).

There are well-known estimation problems in mixture modeling (Congdon, 2003; Frühwirth-Schnatter, 2006; McLachlan & Peel, 2000; Vermunt & Magidson, 2005). Issues that affect estimation of the models are briefly described and a method to deal with them is indicated. The first problem is a permutation of the classes, called label switching. Label switching can occur across different runs starting from different initial values. As there are class-specific constraints for model identification, label switching can be easily detected and then the labels of the latent classes can be reordered across different runs and models. A second problem is that the estimation of mixture models, in general, is prone to yielding multiple local maxima (Frühwirth-Schnatter, 2006; McLachlan & Peel, 2000; Muthén et al., 2002). The usual method of checking the local maxima is to run the model with multiple different starting values (McLachlan & Peel, 2000). In all, 10 different sets of initial values were used to investigate this problem. All results reported in this article are from convergent solutions.

Model Specification in LatentGOLD

Models 0 to 3 were implemented using the Cluster module of LatentGOLD. A way to transform parameters from the Cluster module of LatentGOLD to obtain parameters from Model 2b is described as an example, from which a way of deriving the parameters from the other models can also be obtained. Note that subscript c is used for the latent classes, with $c = r, f$. The formulation of Model 2b in LatentGOLD is as follows:

$$\eta_{ijc} = \alpha_{ic1}\theta_{jc1} + \alpha_{ic2}\theta_{jc2} + (\beta_{i01} - \beta_{i00}) + \beta_{ic1}, \tag{7}$$

with β_{i01} as the intercept for response 1 for item i, and β_{i00} as the intercept for response 0 for item i;
with β_{ic1} as the slope for item i in class c;
with θ_{jc1} and θ_{jc2} as the class-specific abilities for the primary and the secondary dimensions, respectively;
and with α_{ic1} and α_{ic2} as the item discriminations for the primary and the secondary dimensions, respectively.

The class-specific item difficulty can be calculated as $\beta_{ic} = -(\beta_{i10} - \beta_{i00}) - \beta_{ic1}$. Parameter α_{ic2} is estimated only for DIF items in the DIF latent class. Note that only Models 1a and 1b have class-specific item difficulties, but Models 2a and 2b do not. Thus, for Model 2b, $\beta_{if} = \beta_{ir}$.

Model Evaluation

When models are nested and the difference does not concern a parameter constraint on the boundary of the parameter space, then the models can be compared using a likelihood ratio (LR) test. Figure 1 shows the comparisons where a LR test can be used as follows. The test does not apply for horizontal comparisons in Figure 1 because models toward the left have one free variance parameter less than the next model to their right. For all models, Akaike's (1974) information criterion (AIC) and Schwarz's (1978) Bayesian information criterion (BIC) are computed, so that models can be compared on that basis as well. As each of the evaluation statistics has its own strengths, the authors prefer to take a combined approach. For the second horizontal comparison, the best fitting of Models 2a and 2b are extended with a variance parameter for the secondary dimension in the non-DIF latent class, to check the improvement in goodness of fit.

Speededness Application

Data and Earlier Modeling

A sample of 3,872 examinees was randomly drawn from a total group of 24,753 freshmen entering a midwestern university system who took Form A and Form B of a college-level mathematics placement test in the spring of 1997. The forms were presented consecutively in one timed session. All items are multiple choice. Forms A and B consisted of 32 and 36 items, respectively. Results from this test were used to place entering college students into remedial mathematics or precalculus courses. For the present application, the authors use data from an item subset from the beginning of the test and another item subset from the end of the test. Being located toward the end of the test can be considered an item property in the sense explained in the section on secondary dimensions. Following Bolt et al. (2002), the authors assumed absence of speededness for the first 23 items from Form A (first subset, omitting 2 items, 10 and 20, as these authors did), and they assumed that the last 8 items from Form B (second subset) did show speededness for a

Table 1. Model Evaluation for Speededness Data

	Log likelihood	AIC	BIC	No. of free parameters	Proportion DIF class
Model 0: non-DIF	−64,004.4	128,130.8	128,512.8	61	.236
Model 1a: *1d-βDIF*	−63,694.5	127,526.9	127,958.9	69	.130
Model 1b: *1d-αβDIF*	−63,678.5	127,511.1	127,993.2	77	.224
Model 2a: *1&2d-DIF*	−63,688.1	127,516.2	127,954.5	70	.323
Model 2b: *1&2d-αDIF*[a]	−63,657.9	127,469.9	127,952.0	77	.311

Note: AIC = Akaike's information criterion; BIC = Bayesian information criterion; DIF = differential item functioning.
[a]The log likelihood after extension of the secondary dimension to the non-DIF latent class, while fixing the other estimates from Model 2b, is −63,656.8.

subgroup of the respondents. The last 9 items on Form A and the first 28 items on Form B were not analyzed, as these items were assumed to be not sufficiently speeded to affect the model in a substantial way, as Bolt et al. did for the same data.

In the mixture Rasch model used by Bolt et al. (2002), a higher relative difficulty for end-of-test items was found for the speeded latent class (a minority class) compared with the non-speeded latent class. This finding is formally equivalent with latent DIF, while the speeded latent class is the DIF latent class. Wollack, Cohen, and Wells (2003) used the Bolt et al. method to study the effect of removing speeded examinees from the calibration sample on the stability of a score scale maintained over an 11-year period. Using only the nonspeeded examinees for estimating item parameters and equating, Wollack et al. obtained a more stable one-dimensional scale, with fewer drifting items and less scale variability compared with the scale, based on responses for the total group, which included the speeded and nonspeeded examinees. Parameter estimates for end-of-test items using only responses from the nonspeeded group were found to be very similar to estimates for those same items when they were administered in non-speeded locations on a different form of the test.

Alternative ways of modeling the effect of speededness can be found in the literature. For example, Goegebeur, De Boeck, Wollack, and Cohen (2008) introduced a model where the speededness effect was a random variable and therefore a source of individual differences. Considering speededness as a random effect does not contradict the present approach. In fact, the nuisance dimension can be seen as exactly the random speededness effect. Therefore, that approach is also in line with Semmes, Davison, and Close (2011). In fact, Suh, Cho, and Wollack (2011) simulated data on the basis of a random effects model as used by Goegebeur et al., and they found that the true parameters from the nonspeeded state could be recovered quite well on the basis of a mixed dimensionality approach. Another alternative view is local dependence (Douglas, Rim, Habing, & Gao, 1998). This view does not conflict with the present approach either because the secondary dimension explains DIF and can explain local dependence. The DIF items share an item property (i.e., being located at the end of the test) that can explain the DIF and the local dependency. It has been shown that random effects of a shared item property can explain local dependence (Bradlow et al., 1999; Tuerlinckx & De Boeck, 2004).

Results

Table 1 gives the log likelihood, AIC and BIC values of the estimated models. An important preliminary result is that Models 1a and 1b fit much better than Model 0. The LR tests comparing Model 1a and Model 1b to Model 0 are highly significant: $\chi^2(8) = 619.8$ ($p \leq .001$), and $\chi^2(16) = 651.8$ ($p \leq .001$), respectively. These results are confirmed by the AIC and BIC.

1. Regarding Research Issue 1, whether a secondary dimension helps (first horizontal comparison in Figure 1), it was found that Model 2a fits better than Model 1a and that Model 2b fits better than Model 1b, as can be derived from the AIC and BIC values. This implies empirical evidence in favor of mixed dimensionality as an explanation of latent DIF.
2. Regarding Research Issue 2, whether there is discrimination DIF related to the primary dimension (vertical comparisons in Figure 1), the results are as follows. Model 1b has a better fit than Model 1a, $\chi^2(8) = 32, p \leq .001$, although its BIC is higher, and Model 2b has a better fit than Model 2a, $\chi^2(7) = 60.4, p \leq .001$. This latter result is largely confirmed by the AIC, and also by the BIC, but with a smaller difference. Therefore, the authors rely on an extension of Model 2b for the next research issue.
3. Regarding Research Issue 3, whether the secondary dimension extends to the non-DIF latent class (second horizontal comparison in Figure 1), it turns out that the log likelihood does not really improve when a secondary variance parameter is estimated for the non-DIF latent class (as in Model 3b). This means that Model 2b does not need to be expanded to include a secondary dimension in the non-DIF latent class.

The item parameter estimates of Model 2b are shown in Table 2. The two latent classes share their difficulties for all items and also the discriminations of the non-DIF items for the primary dimension. In addition to the difference in dimensionality, the two latent classes differ with respect to the estimated discriminations for the items with DIF, on the primary dimension (see Table 2), and with respect to their estimated means and the variances of the two dimensions. In the non-DIF class, there is only one dimension, with a fixed mean (0.00) and a fixed variance (1.00). In the DIF class, there are two dimensions, with estimated means of -0.634 and -0.987, respectively. The corresponding variance estimate for the primary dimension is 0.787, while the variance for the secondary dimension is fixed to 1.00. The estimation of the DIF class proportion is .311.

The two latent classes differ in three respects. First of all, the DIF latent class is a minority class. Only a minority of the respondents seems sensitive to speededness. Second, the DIF latent class has nonzero discriminations on the secondary dimension for the items toward the end of the test, in agreement with the interpretation of the class as being speeded. Also, the discriminations of these items on the primary dimension differ. This may be the result of the secondary dimension playing a role toward the end of the test, at the cost of the primary dimension, which makes sense if the respondents switch to another strategy. It may seem surprising that the discriminations are so high, but it should be noted that for reasons of identification, several constraints have been introduced with possible consequences for the size of the discriminations. A positive correlation between the two dimensions would yield decreased discrimination estimates, but the choice of a specific positive correlation would of course be arbitrary. Third, the DIF latent class seems to have a somewhat lower ability, and its mean on the secondary dimension is negative. This means that the speeded class has a double disadvantage. On average, its ability seems to be lower, and being speeded toward the end of the test seems not to help. The effect size of the secondary dimension for the average person from the DIF latent class is 1/2.68 (odds ratio).

Omitted responses are sometimes considered useful indicators of speededness (Mroch & Bolt, 2006). Therefore, a validation check, looking at the number of nonresponses among the last eight items in the DIF and the non-DIF latent classes, was conducted. The last eight items on the test were examined as a group to determine whether particular response patterns might emerge

Table 2. Mixed Dimensionality Model 2b (*1&2d-αDIF*): Item Estimates for Speededness Data

	Discrimination primary dimension			
	Non-DIF class	DIF class	Discrimination secondary dimension	Difficulty
Item 1	0.782			0.782
Item 2	1.005			1.509
Item 3	1.063			1.584
Item 4	1.272			0.537
Item 5	1.238			1.645
Item 6	0.531			1.644
Item 7	0.695			0.879
Item 8	1.183			−0.666
Item 9	0.668			0.987
Item 11	0.877			0.317
Item 12	0.996			1.324
Item 13	0.926			0.613
Item 14	0.990			−0.329
Item 15	0.907			−0.414
Item 16	1.116			0.102
Item 17	1.335			0.472
Item 18	0.842			−0.234
Item 19	0.631			1.368
Item 21	0.987			0.640
Item 22	0.705			−1.132
Item 23	0.805			−0.213
Item 38	1.047	0.510[a]	0.881[a]	0.535
Item 39	0.780	0.372[a]	0.661[a]	−0.476
Item 41	0.570	1.151[a]	1.361[a]	−1.125
Item 42	0.278	1.663[a]	1.652[a]	−1.511
Item 43	0.673	0.426[a]	1.000[a]	0.250
Item 44	1.125	1.976[a]	2.305[a]	−0.445
Item 45	0.587	0.501[a]	1.290[a]	−0.312
Item 46	0.510	2.300[a]	2.535[a]	−1.638

Note: DIF = differential item functioning.

[a] The estimates apply solely to the DIF latent class.

conditional on latent group membership. The numbers of omitted items were cross-tabulated with latent group membership from Model 2b, based on the posterior probability. It was found that when the total number of nonresponses was considered, the proportions from the non-speeded latent group decreased as the frequency of nonresponse increased, whereas the proportions of the latent speeded group increased when the frequencies of nonresponse were higher.

As planned, the confirmatory and exploratory two-dimensional models without latent classes were also estimated. Both models had a higher log likelihood than Model 0 but a lower one than all other models (−63,755.8 and −63,733.6, respectively). Also, their AIC and BIC values were higher than any of the other models, including the secondary dimension models.

Arithmetic Operations Application

Data and Earlier Modeling

From a national assessment in 2002 in Flanders, Belgium, data are available from 1,004 second-grade students in a prevocational track of secondary education (ages 13-14) who took a test with

Table 3. Model Evaluation for Arithmetic Operations Data

	Log likelihood	AIC	BIC	No. of free parameters	Proportion of DIF class
Model 0: non-DIF	−15,304.8	30,727.6	31,017.4	59	.080
Model 1a: *1d*-β*DIF*	−15,172.1	30,476.2	30,800.4	66	.230
Model 1b: *1d*-αβ*DIF*	−15,165.3	30,476.6	30,835.1	73	.240
Model 2a: *1&2d-DIF*[a]	−15,166.8	30,467.6	30,796.7	67	.292
Model 2b: *1&2d*-α*DIF*[b]	−15,155.1	30,456.2	30,814.8	73	.440

Note: AIC = Akaike's information criterion; BIC = Bayesian information criterion; DIF = differential item functioning.

[a]The log likelihood after extension of the secondary dimension to the non-DIF latent class, while fixing the other estimates from Model 2a is −15,166.0.

[b]The log likelihood after extension of the secondary dimension to the non-DIF latent class, while fixing the other estimates from Model 2b is −15,152.0.

28 items, assessing their skills in arithmetic operations (7 items each of addition, subtraction, multiplication, and division). The use of calculators was not allowed.

The data from this test were analyzed by Van Nijlen and Janssen (2010) together with data from a parallel test with contextualized items, in which the use of a calculator was allowed. Using a mixture Rasch model, a model with three latent classes for the two data sets was chosen as the best fitting, on the basis of the BIC. The three classes can be interpreted as a proficient class, a class with differences between contextualized and noncontextualized performance, and a class with an especially poor performance on the division items. The authors will concentrate here on the first data set, which stems from the noncontextualized form. From the earlier results, they expected two latent classes: a non-DIF latent class and a DIF latent class, with an average lower level of performance and with a relatively poorer performance on the division items than in the non-DIF latent class.

Results

Table 3 gives the log likelihood, the AIC and the BIC values of the estimated models. An important preliminary result is that Models 1a and 1b fit much better than did Model 0. The LR tests comparing Models 1a and 1b to Model 0 are highly significant: $\chi^2(7) = 265.4$ ($p \leq .001$) and $\chi^2(14) = 279$ (p £ .001), respectively. These results are confirmed by the AIC and BIC.

1. Regarding the question of whether a secondary dimension helps (first horizontal comparison in Figure 1), it was found that Model 2a fits better than does Model 1a and that Model 2b fits better than does Model 1b, as can be derived from the AIC and BIC values. This is, again, a result in favor of the secondary dimension hypothesis.
2. Regarding the question of whether there is discrimination DIF related to the primary dimension (vertical comparisons in Figure 1), the results are as follows. Based on the LR test, Model 1b fits better than Model 1a, $\chi^2(7) = 13.6$, $p \leq .05$, but its AIC and BIC values are higher. The evidence is also mixed for Models 2a and 2b. Following the LR test, the model with discrimination DIF on the primary dimension (Model 2b) has a better fit than does the equivalence model (Model 2a), $\chi^2(6) = 23.4$, $p \leq .01$, but the BIC value (although not the AIC value) of the latter is better. The evidence in favor of primary dimension discrimination DIF is not very strong—certainly not as strong as in the

Table 4. Mixed Dimensionality Model 2a (*1&2d-DIF*): Estimation for the Arithmetic Operations Data

	Discrimination primary dimension	Discrimination secondary dimension	Difficulty
Item 2	0.847		1.762
Item 3	0.882		1.262
Item 4	0.782		0.533
Item 5	0.587		2.750
Item 7	1.100		0.754
Item 8	0.816		0.183
Item 10	1.006		1.605
Item 11	0.939		0.741
Item 13	0.778		1.890
Item 14	1.348		1.695
Item 15	0.849		0.698
Item 16	0.690		2.008
Item 18	1.338		0.942
Item 19	1.122		1.651
Item 20	1.501		2.449
Item 23	1.515		2.232
Item 24	0.821		0.217
Item 25	1.106		0.816
Item 26	1.421		1.814
Item 27	1.013		1.493
Item 28	0.679		1.077
Item 1	0.154	0.513[a]	−0.191
Item 6	0.536	0.956[a]	1.161
Item 9	0.695	1.000[a]	1.399
Item 12	0.767	0.575[a]	0.758
Item 17	0.868	0.836[a]	2.138
Item 21	0.642	1.197[a]	0.289
Item 22	0.808	0.621[a]	0.794

[a] The estimates apply solely to the DIF latent class.

speededness study. Given these results, Models 2a and 2b have been extended with a secondary dimension variance in the non-DIF latent class to deal with Research Issue 3.

3. Regarding the question of whether the secondary dimension extends to the non-DIF latent class (second horizontal comparison in Figure 1), the log likelihood did not improve much for either Model 2a or 2b when allowing for a secondary dimension variance in the non-DIF latent class (as in Models 3a and 3b).

The item parameter estimates of Model 2a are shown in Table 4 to illustrate the difference from the speededness study, although Model 2b would be a legitimate choice as well. Because of the nature of the model, the item parameter estimates are common to the two latent classes, except for the presence of a secondary dimension in the DIF latent class. The difference between the two latent classes can be seen in the estimated means and variances of the two dimensions. In the non-DIF class, there is only one dimension, with a fixed mean (0.00) and variance (1.00). In the DIF class, there are two dimensions, with means of −1.150, and −2.945, respectively. The corresponding variance on the primary dimension is 1.034, and the variance on the secondary dimension is fixed to 1.00. The correlation between the two dimensions in the DIF latent class is .353. The estimation of the DIF class proportion is 0.292.

The two latent classes differ in three respects. First, the DIF latent class is a minority class. Only a minority of the respondents seem to rely on a division-specific dimension. Second, the DIF latent class has nonzero discriminations on the secondary dimension for the division items, in agreement with the interpretation of the class as using a division-specific dimension. Third, the DIF latent class seems to have a somewhat lower arithmetic ability level, and its mean on the secondary dimension is also negative. This means that the class with a division-specific dimension has a double disadvantage. On average, arithmetic ability seems to be lower, and using a second dimension for division operations indicates the poor quality of the division strategy in the DIF latent class. The effect size of the secondary dimension for the average person from the DIF latent class is 1/19.01 (odds ratio). As explained earlier, the precise value of the mean on the secondary dimension depends on the fixed value of the variance and on the one fixed discrimination. The rather extreme value of the mean must be reduced for items with a secondary dimension discrimination smaller than one.

As announced, the confirmatory and exploratory two-dimensional models without latent classes were also estimated. The confirmatory model has a higher log likelihood (−15,200.2) than Model 0 but a lower one than all other models. Also, the AIC and BIC values of the confirmatory model are higher. In contrast, the exploratory two-dimensional model does have a higher log likelihood (−15,131.9) and a better AIC value (30,431.9) than all other models, but clearly, it also has more free parameters (84). This explains why its BIC (30,844.4) is higher than that of most other models, including the secondary dimension models. As this two-dimensional model is purely exploratory, it has a higher risk of overfitting, so it seems reasonable to use a higher penalty for the number of free parameters, as established in the BIC.

Discussion and Conclusion

From the results obtained in both applications, it can be concluded that a mixture approach makes sense and that a secondary dimension leads to a better goodness of fit for the data sets under consideration. The mixture approach is further supported by the generally better goodness of fit compared with the two-dimensional model without latent classes. In the second application, the exploratory (but not confirmatory) two-dimensional model without latent classes yielded a better log likelihood, but when the much higher number of free parameters was penalized as in the BIC, the mixture models were again superior.

It was sufficient to include the secondary dimension in only one of the two classes. This means that the theoretical analysis in terms of mixed dimensionality received empirical support. Furthermore, the DIF latent class with a secondary dimension turned out to be a minority class. This class had a lower mean on the primary dimension and a negative mean on the secondary dimension. It can therefore be interpreted as a class with less able respondents who made use of an alternative strategy or alternative processes, which turned out not to be really helpful, given the negative mean of the secondary dimension.

The mixed dimensionality approach has several merits. First, it illustrates how the theoretical analysis of DIF in terms of multidimensionality (Ackerman, 1992; Shealy & Stout, 1993a, 1993b) can be used, not only for DIF detection, as illustrated elsewhere (Roussos & Stout, 1996), but also for modeling and explaining DIF. However, the type of model this study focused on was somewhat different, with a secondary dimension in only one of two classes.

Second, the models in this article are for latent DIF. Both applications exemplify the possibility that items function differently between latent groups of respondents, which are difficult or impossible to observe independently of the item responses. Only the responses can indicate whether a respondent has been speeded or has problems with division operations. One may perhaps prefer not to use a DIF framework for applications with latent classes. However, formally speaking, the models are the same as for cases where one could or does have observed group memberships.

Third, the proposed approach is helpful in explaining DIF. Other authors have initiated a multidimensional explanatory methodology (among others, Adams, Wilson, & Wang, 1997; Douglas et al., 1996; Muthén & Lehman, 1985; Penfield, 2010; Roussos & Stout, 1996; Shealy & Stout, 1993a, 1993b). In addition, the possibility of individual differences in DIF has been given attention (Camilli & Penfield, 1997; Longford et al., 1993). The mixed dimensionality approach adds to the toolkit and combines ideas from several of these approaches.

Fourth, the mixed dimensionality approach and the applications illustrate how DIF, dimensionality (random effects), and local dependence can all be linked without the need to choose between these psychometric phenomena. It does not follow that this is true in all situations; for example, it is sometimes simpler or more interpretable to go one way or the other for how to model the data. For the particular situations investigated here, it was indeed possible to develop a unitary approach.

Fifth, the results of the mixed dimensionality approach illustrate that a test that may be considered unidimensional can show a two-dimensional or, more generally, a multidimensional structure, depending on the subgroup of respondents, their learning history and development, and the test-taking situation—or a combination of these. From this perspective, the dimensionality of a test is not a feature of the test as such but depends on how the respondents approach the items in the test. However, this is formally equivalent to saying that the test has the potential to be multidimensional. There must be a basis in the test for the multidimensionality to show.

The mixed dimensionality approach also has limitations. It cannot easily be used for all situations. The case considered here is one in which one item property was available as a suspect for DIF or for a lack of measurement invariance in general. More positively formulated, the usefulness of the multidimensionality approach is more evident when one has a good and rather simple explanatory hypothesis. In the two applications, the hypothesis concerning DIF (speededness and lack of division mastery) made sense as such and was also suggested by previous analyses by other authors. In many cases, a researcher has no clue about how to define an item property at the basis of a secondary dimension, or the clue is not sufficiently specific.

If a limited set of property suspects is not available to a researcher, further study would be required before the secondary dimension approach can be implemented. For example, a Q matrix can be constructed indicating which skills are involved in the items, and the skills from this matrix can be used as properties in the secondary dimension analysis. Alternatively, a more exploratory approach can be followed, some examples of which have been suggested by Roussos and Stout (1996) to find one or more potential sources of DIF. Given that this would lead to a clue as to the source of the suspected DIF, the corresponding item properties can be used in the next step as the basis for one or more secondary dimensions. An important criterion is that the detected DIF pattern makes sense from an interpretational point of view. With a preliminary and explorative step, the described approach could also be used without a clear prior hypothesis and thus also as a second step in the context of DIF detection.

Although the mixed dimensionality approach has limitations, it also has potential for interesting extensions and broader use. To begin with, the approach can be applied to manifest groups in a straightforward way. A further possibility is to map the estimated group membership from the data into manifest group membership in cases where it is available, to check the degree of proximity of latent classes to manifest groups. In the same line, the manifest grouping, if available, can be included in the model to explain the latent class proportions, as in the multinomial logistic regression model (e.g., Cho & Cohen, 2010).

As an example of the broader use, consider the Saltus model (Mislevy & Wilson, 1996; Wilson, 1989) but extended with a random Saltus effect. This extension is formally equivalent to the mixed dimensionality as used here. Also for the larger study of cognitive development, one may be interested in the concept of increasing dimensionality. Considering Siegler's (1976) balance items, the subsequent rules that are mastered through one's cognitive development can be linked to item properties. It can therefore be investigated whether these subsequent rules each generate a new dimension reflecting the degree of mastery of the new rule, perhaps at the cost of a shrunken variance of dimensions that reflect developmentally earlier rules. Jansen and van der Maas (2002); van der Maas, Quinlan, and Jansen (2007); and Mislevy and Wilson (1996) have defended the use of mixture models for the study of balance task data and cognitive development, although not of the mixed dimensionality type as proposed here.

Finally, more in general, the mixed dimensionality approach can be considered for cases where respondents differ in the use of multiple strategies. A single-strategy method for solving problems can perhaps be captured with just one dimension, whereas a multiple-strategy method, used for example because the primary strategy is not adequate, implies more than one dimension if the respondents differ in how able they are with respect to the different strategies. It was in fact shown by Goegebeur et al. (2008) that when the speeded strategy was in use by the respondents, it was only used gradually, while the regular strategy remained in place to some extent.

In sum, the mixed dimensionality approach is useful beyond the DIF context, and within the DIF context, it can be used for the common concept of DIF (manifest DIF) and also for latent DIF. However, it should primarily be used for confirmatory and explanatory purposes, and not for exploratory or detection purposes.

Acknowledgment

The authors are grateful to Allan Cohen and to Daniel Van Nijlen and Rianne Janssen for making the data available. The authors thank two anonymous reviewers and the editor for their insightful comments and suggestions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The first author received the following financial support for the research, authorship, and/or publication of this article: Part of the research is supported by the GOA/2005/04 grant from the K. U. Leuven.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331-348.

Bolt, D., & Stout, W. F. (1996). Differential item functioning: Its multidimensional model and resulting Sibtest detection procedure. *Behaviormetrika, 23*, 67-95.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*, 129-147.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on the Mantel-Haenszel log-odds ratio. *Journal of Educational Measurement, 34*, 123-139.

Cho, S.-J., & Cohen, A. S. (2010). Multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics, 35*, 336-370.

Cho, S.-J., Cohen, A. S., & Templin, J. (2008, April). *A multidimensional mixture IRT model for DIF analysis*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133-148.

Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice, 20*, 225-233.

Congdon, P. (2003). *Applied Bayesian modelling*. New York, NY: Wiley.

DeAyala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2*, 243-276.

Douglas, J., Rim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics, 23*, 129-151.

Douglas, J., Roussos, L. A., & Stout, W. F. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33*, 465-485.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278-295.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology, 57B*, S275-S284.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York, NY: Springer.

Glöckner-Rist, A., & Hoijtink, H. (2003). The best of two worlds: Factor analysis of dichotomous data using item response theory and structural equation. *Structural Equation Modelling, 10*, 544-565.

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika, 73*, 65-87.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Horn, J. L., & McArdle, J. J. (1992). A practical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology, 63*, 395-416.

Jansen, B. R. J., & van der Maas, H. L. J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology, 81*, 383-416.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine, & J. Rost (Eds.), *Latent trait and latent class models* (pp. 263-274). New York, NY: Plenum.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171-196). Hillsdale, NJ: Erlbaum.

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement, 27*, 372-379.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543.

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

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*, 93-115.

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61*, 41-71.

Mroch, A. A., & Bolt, D. M. (2006, April). *An IRT-based response likelihood approach for addressing test speededness*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Muthén, B., Brown, C. H., Booil, J. K. M., Khoo, S.-T., Yang, C. C., Wang, C.-P., & Kellam, S. G. (2002). General growth mixture modelling for randomized preventive interventions. *Biostatistics, 3*, 459-475.

Muthén, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: An application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*, 1-22.

Muthén, B. O., & Lehman, J. (1985). Multiple group IRT modelling: Applications to item bias analysis. *Journal of Educational Statistics, 10*, 133-142.

Oort, F. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modelling, 5*, 107-124.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16*, 237-248.

Penfield, R. D. (2010). Modelling DIF effects using distractor-level invariance effects: Implications for understanding the causes of DIF. *Applied Psychological Measurement, 34*, 151-165.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

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

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Semmes, R., Davison, M., & Close, C. (2011). Modeling individual differences in numerical reasoning speed as a random effect of response time limits. *Applied Psychological Measurement, 35*, 433-446.

Shealy, R., & Stout, W. F. (1993a). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239.). Hillsdale, NJ: Erlbaum.

Shealy, R., & Stout, W. F. (1993b). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8*, 481-520.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78-90.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W. F., & Roussos, L. A. (1995). *SIBTEST users manual* (2nd ed.) [Computer program manual]. Urbana-Champaign: University of Illinois.

Suh, Y., Cho, S.-J., & Wollack, J. A. (2011, April). *A comparison of item calibration procedures in the presence of test speededness*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research, 16*, 33-42.

Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 289-316). New York, NY: Springer.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics, 30*, 443-464.

van der Maas, H. L. J., Quinlan, P. T., & Jansen, B. R. J. (2007). Towards better computational models of the balance scale task: A reply to Schultz and Takana. *Cognition, 103*, 473-479.

Van Nijlen, D., & Janssen, R. (2010). Contextualized and non-contextualized mathematics items in large-scale assessment (Unpublished manuscript). K. U. Leuven, Belgium.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced*. Belmont, MA: Statistical Innovations.

Vermunt, J. K., & Magidson, J. (2007). Latent GOLD 4.5 syntax module [Computer program]. Belmont, MA: Statistical Innovations.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276-289.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40*, 307-330.

Woods, C. M. (2009). Evaluation of the MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*, 1-27.

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research & Policy Studies, 5*, 1-23.