TECHNICAL

REPORT

0692

THE DIMCAT METHODOLOGY : EXTENSIONS AND DIFFERENTIATION CAPACITY

HIDEGKUTI, I. and P. DE BOECK

IAP STATISTICS

NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Running head: Extensions and differentiation of DIMCAT

The DIMCAT methodology:

Extensions and differentiation capacity

István Hidegkuti and Paul De Boeck

K.U.Leuven

Belgium

Abstract

The dimension/category framework (DIMCAT) is used to differentiate the categorical versus dimensional nature of psychological phenomena. In this paper an extension of the DIMCAT methodology is described for rating-scale data and latent categories, as well as a simulation study that was conducted with 80 to 200 data sets for each of four pairs of differentiations in DIMCAT. The four pairs of differentiations were studied for different data and category types as well as for different sizes of category effect. In case of manifest categories the AIC and the likelihood-ratio test (LR) made a very good differentiation for all of the eight differentiations, for most combinations of data type, size of category difference and type of category. For six out of the eight differentiations the BIC also made the differentiation in a correct way, but for the other two – both related to degrees of discrimination – the BIC results were rather poor. In case of latent categories the results were very similar except that for one differentiation the performance of the LR and the AIC was also rather poor.

Introduction

In many fields of psychology there is a long-lasting debate about whether a certain psychological phenomenon is categorical or dimensional. Often what is considered as a category, such as a psychiatric diagnosis, is not, and often what is not considered categorical, such as cognitive development, actually is. When talking about categories an important distinction has to be made; a distinction between the manifest and latent nature of the phenomena in question. Often when categories are used, an observed continuum is divided into parts, which are labeled as categories. These observed or manifest categories are nothing more than a convenient description tool. In other cases, the categories are an artefact of the measurement approach, such as when a rating scale is used. As a consequence, it is often assumed without any test that there is a categorical entity behind such a ”category” (Meehl, 1995). Therefore, it is of importance to investigate the latent nature of these phenomena to find out the underlying ”truth”. The Dimension/Category (DIMCAT) framework addresses the very question whether a given (psychological) phenomenon is more category-like or more dimension-like, when looked at from a latent perspective, as in statistical modeling with latent variables.

The most widely used approach to differentiate between categorical and dimensional is *taxometrics* (see, e.g., Waller & Meehl, 1998), a family of methods of which maximum covariance (MAXCOV) is the most well-known. Recently, an alternative approach, DIMCAT, was formulated by De Boeck, Wilson and Acton (2005), based on item response theory. On the one hand, DIMCAT is more specific in that it is based on an explicit measurement model, and on the other hand, it is more general in that more aspects of what it means to be categorical are investigated.

The importance of the distinction between the categorical versus dimensional nature is at least twofold. On the one hand it is an interesting question from a purely theoretical point of view, to better understand the phenomena one encounters. On the other hand, the knowledge about the latent nature of phenomena may have important practical consequences, for instance the therapeutical approach for psychiatric syndromes may depend on how a certain disorder is conceptualized, and the approaches just mentioned can help to find out how phenomena are best conceptualized.

In the following we will first describe the DIMCAT framework and its extension to rating-scale data and unobserved (latent) categories. Next, a simulation study is described to investigate the differentiation capacity of the approach.

The Dimension/Category (DIMCAT) framework

The DIMCAT framework for manifest categories and binary data was formulated by De Boeck, Wilson and Acton (2005) for manifest categories and binary data, but it can easily be extended to latent categories and rating-scale data. The present description of the framework is heavily based on De Boeck et al. (2005). DIMCAT is based on item response modeling (IRT), more specifically on the two parameter logistic model (2PL) and its restricted variants. DIMCAT models are meant to describe the latent structure of categories that are defined on the basis of observed indicators. Examples of manifest categories are gender, diagnosed personality disorders and categories based on cut-offs, for example developmental stages based on a manifest index. Examples of latent categories are non-diagnosed personality types or personality disorders, clinical syndromes, discrete underlying developmental stages. The data relate to a set of indicators, the value of which is observed for units of observation (mostly persons). For example, the units of observations can be patients, the indicators a set of symptoms, the manifest categories can be different diagnosed personality disorder categories, and the latent categories can be the unobserved alternative for manifest personality disorder categories.

The DIMCAT framework basically rests on two distinctions. The first distinction is between *quantitative differences* and *qualitative differences.* Quantitative means that the categories are different by a degree of something; e.g., category B having more of the same category A has. Qualitative means that the differences cannot be reduced to a degree of something identical but that something of a different kind is involved. There are two types of qualitative differences: lack of parallelism of indicator profiles in the various categories, and differences in within-category indicator relevance (how much the indicators differentiate within categories). Qualitative differences are more category-like than quantitative differences.

In a somewhat different context, the two types of qualitative differences correspond to two types of differential item functioning (DIF; Holland & Wainer, 1993; Mellenbergh, 1982; Millsap & Everson, 1993). A clear example of lack of parallelism is that the ordering of the indicators according to the value they have in the categories differs depending on the category. Differences in indicator relevance means that depending on the category other indicators are more relevant in determining a degree within the category, as when the nature of a factor from factor analysis depends on the category. De Boeck et al. (2005) distinguish in their applications between genuinely qualitative differences, which are related to the nature of the categories, and more tangential differences that stem from circumstances. The latter can hardly be used as evidence for the categorical nature. For example, the life events of persons with an antisocial personality disorder lead to circumstances and effects of these circumstances that are on average different from those of persons without the disorder. These effects may be considered tangential differences and too narrow to conclude that a phenomenon (e.g.,antisocial personality disorder) is categorical.

The second distinction is between *heterogeneity* and *homogeneity*. Heterogeneity means that there are systematic differences in degree within the categories, whereas homogeneity means there is not. Within-category homogeneity is more category-like than within-category heterogeneity.

These two distinctions define four combinations. The only combination that may be considered as purely categorical is the one with homogeneity within categories and qualitative differences between categories. The opposite case, heterogeneity with quantitative differences is considered dimension-like because the differences within and between categories are a matter of degree of the same thing, whereas the remaining two cases are hybrid combinations in the sense that they combine category-like and dimension-like features. It is important that the differences between these four combinations are not absolute, but gradual, because heterogeneity and the size of qualitative differences can be small or larger, and that also within each combination graduality applies. The question of dimension-likeness or category-likeness is not a simple one and in most of the cases, only a relative answer can be given, as dimension- or category-likeness is a matter of degree (De Boeck et al., 2005; Meehl, 1979; Waller & Meehl, 1998).

To find out which combination is the most appropriate, a series of models is fitted, and then the decision about what kind of category-likeness the phenomenon in question shows, is based on the comparison of the goodness-of-fit statistics of the models.
The basic DIMCAT model for manifest as well as for latent categories is:

$$P(Y_{pi} = 1 | \theta_{pc}, c_p) = \frac{\exp(\alpha_{ic}(\theta_{pc} + \beta_i + \delta_{ic} + \gamma_c))}{1 + \exp(\alpha_{ic}(\theta_{pc} + \beta_i + \delta_{ic} + \gamma_c))} \tag{1}$$

where: Y_{pi} is the observation of person p $(p = 1, \ldots, P)$ on indicator i $(i = 1, \ldots, I)$, α_{ic} is the discrimination of indicator i within category c $(c = 1, \ldots, C)$, θ_{pc} is the latent variable for observation p from category c $(\theta_{pc} \sim N(0, \sigma_c^2))$, β_i is the location of indicator i, δ_{ic} is the Saltus parameter (Wilson, 1989) and γ_c is the category (group) difference parameter for category c. The Saltus (Latin for leap) parameter (δ_{ic}) accounts for shifts in the item locations when going from one category to another. The Saltus parameter was introduced by Wilson (1989) to capture a homogeneous shift of a subset of the indicators, but it is used here in a generalized sense. The Saltus parameters can be interpreted as interactions between indicators and categories. In order for the model in Equation 1 to be identified, not just the common restrictions are needed for location (through θ or β) and for discrimination (through α or the variance of θ), but also restrictions on δ_{ic} and γ_c are needed. For a reference category formulation, all δ's in one category equal to zero, and also one δ in each of the other categories, and one γ are equal to zero. The parameterization that will be used is one with γ_c to be interpreted as a category effect; see the Method section.

Extension of DIMCAT

Although in its original form DIMCAT was described for binary data and manifest categories, the DIMCAT framework includes also models for rating-scale data and models with latent categories. Because these extensions are not described in an explicit way by De Boeck et al. (2005) they will be described below.

Rating-scale data

There are several models for rating-scale indicators that can be implemented in the DIMCAT framework, but it seems reasonable to use models with a constrained indicator response structure in order to reduce the number of parameters. Of these models we opted for the Modified Graded Response Model (MGRM; Muraki, 1990), which – in contrast with the Partial Credit Model (PCM; Masters, 1982) and the Graded Response Model (GRM; Samejima, 1996) – assumes that the steps between the response categories are invariant throughout the items, which can be expected when working with the same response format (e.g., disagree - neutral - agree; Embretson & Reise, 2000; Muraki, 1990). A possible alternative is Andrich's (1978) Rating Scale Model (RSM) which is a restriction on the PCM. The choice of the MGRM is threefold. First, the MGRM, in contrast with the RSM, includes discrimination parameters which is an important feature in the DIMCAT framework. Second, the thresholds for the points on the rating scales are easy to interpret as a set of subsequent cut-offs in a common distribution. Finally, all models needed for the study can be estimated with the M*plus* software (Muthén & Muthén, 2005) including the restricted latent class versions and M*plus* requires the use of GRM, and can easily be restricted to obtain MGRM.

Given M response alternatives indexed by m ($m = 1, \ldots, M$) in the MGRM the cumulative (conditional) probability of responding in response category m ($m = 1, \ldots, M$) or higher given a category c_p is defined as:

$$P(Y_{pi} \geqslant m|\theta_{pc}, c_p) = \frac{\exp(\alpha_{ic}(\theta_{pc} + \beta_i + \delta_{ic} + \omega_m + \gamma_c))}{1 + \exp(\alpha_{ic}(\theta_{pc} + \beta_i + \delta_{ic} + \omega_m + \gamma_c))} \tag{2}$$

and the probability of responding in response category m is:

$$P(Y_{pi} = m|\theta_{pc}, c_p) = P(Y_{pi} \geqslant m|\theta_{pc}, c_p) - P(Y_{pi} \geqslant m + 1|\theta_{pc}, c_p), \tag{3}$$

except that for the lowest rating-scale category $P(Y_{pi} \geqslant m|\theta_{pc}, c_p) = 1$ and for the highest rating-scale category $P(Y_{pi} \geqslant m + 1|\theta_{pc}, c_p) = 0$. The ω_m parameter denotes the rating-scale category threshold parameter, separating category $m-1$ and category m. (ω_1 is not used because $P(Y_{pi} \geqslant m|\theta_{pc}, c_p) = 1$.)

Latent classes

The unconditional version of the model for an individual p and item i can be expressed as follows:

$$P(Y_{pi} = m|\boldsymbol{\theta}_p) = \sum_{c=1}^{C} \pi_c P(Y_{pi} = m|\theta_{pc}, c_p). \tag{4}$$

That is, the probability that person p responds in category m on item i can be written as a weighted sum of conditional probabilities (conditional on the latent class) over the classes. The conditional probability $P(Y_{pi} \geqslant m|\theta_{pc}, c_p)$ is defined as in (2). The marginal version for a response pattern $\mathbf{y}_p$ reads as:

$$P(\mathbf{y}_p) = \sum_{c=1}^{C} \pi_c \int \prod_{i=1}^{I} P(Y_{pi} = m|C_p = c) N(\theta_{pc}|0, \sigma^2) d\theta_{pi}, \tag{5}$$

where π_c is the overall probability of belonging to latent class c ($c = 1, \ldots, C$). For the manifest case, one $\pi_c = 1$ and the other values are 0, because the category is known. Similar models have been developed for example by Wilson (1989), Rost (1990), and Mislevy and Wilson (1996) for dichotomous data and by Rost (1991) for polytomous data. To estimate this model, an Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) can be used.

Several computer programs have been developed by now that use the EM algorithm to estimate mixture models. Examples are ℓEM (Vermunt, 1997), M*plus* (Muthén & Muthén, 2005) or WINMIRA (von Davier, 2001).

Simulation study

As it was mentioned above, in DIMCAT the following distinctions are made: (a) between homogeneity and heterogeneity, and (b) between quantitative and qualitative differences, more specifically for the latter, between (b.1) location equivalence and lack of location equivalence, and between (b.2) discrimination equivalence and lack of discrimination equivalence. Location equivalence refers to parallelism in the profiles, and discrimination equivalence refers to equivalence of indicator relevance within categories. These three distinctions (a, b.1, b.2) and the distinction between the model with equal discriminations (called the one parameter logistic model (1PL) in case of binary data) and the model with different discriminations (called the two parameter logistic model (2PL) in case of binary data) are at the basis of the simulation study.

An important issue in modeling is how well one can differentiate between related models one of which is the true model – in other words, whether the true model can be recognized from the data. This is important for several reasons. One reason is that when interested in the estimates, one can trust these estimates better when they stem from a model one may consider the true model. In some cases it might be possible to obtain reasonably good estimates from the wrong model when the estimates are not sensitive to misspecifications of other aspects, but, of course, a true model is to be preferred. Another reason is that one may be interested in which model is better – for example, because of the theoretical relevance of the models. The aim of this simulation study relates to the second reason, because we want to investigate how well the true model can be identified among a set of related models. However, each of the model features is actually a continuum; the differences between models are gradual – that is, they can be very small or very large. It is evident that the power to differentiate between models will depend on the size of the difference. Consequently, the differentiation issue cannot be given an absolute answer. One will always be able to differentiate if one chooses a sufficiently large difference or a sufficiently large sample, and one can also come up with data sets that do not allow for a differentiation because the differences are too small. It may also be impossible to find general differentiation rules, because too many factors may be involved.

The simulation study is based on the following principles.

1. The study should be representative of the broader category of models in which one is interested. The broader category is the category of item response models. Because the distinction between models without and models with discrimination parameter (1PL vs. 2PL) is a basic feature, it will be incorporated into the simulation study.

2. The data sets should be representative of data sets in psychological research. Often the number of respondents is much smaller in a psychological study than in an educational measurement study. For example, 200 respondents is perhaps not a sufficiently large sample in an educational measurement context, but for psychological studies it is often

a reasonable number. Similarly, it is quite common that the number of indicators (e.g., items) is not very large. For example, many personality inventories, especially for research, have scales with 10 or fewer items. When not measurement of individual persons is the purpose, but when rather the relations between variables or the differences between groups or experimental conditions is the focus of interest, it is not crucial to have highly reliable measures.

3. In most cases a large number of factors may be considered in a simulation study – number of respondents, number of items, true parameter values, etc. – and of these factors many levels may be included. On the other hand, the data that are generated according to a given design, can be analyzed with a variety of models, which leads to an explosion of the size of the study.

For the models to analyze the data (the analysis models) a single-difference strategy will be followed. Not all models are used for the analysis, but all models are used that deviate from the true model in only one single respect. As will be seen below, not all analysis models that are defined in that way are meaningful, but most are. If the difference in question does not interact with other factors in the design in the respect that is under consideration, then the inference can be based on an aggregation of data sets from more than one cell in the design, as will be shown. Some deviations from the absence of interaction will be reported.

Method

Simulated Data Sets

To study the differentiation capacity of the DIMCAT approach both binary and rating-scale data were generated. The data sets were generated based on (2) and (4), with $M = 2$ for binary data and $M = 3$ for rating-scale data.

Common features of the simulated data sets

All data sets consist of ten indicators ($I = 10$), and two categories ($C = 2$). The fixed item locations were generated on the basis of a standard normal distribution ($\beta_i \sim N(0, 1)$). The category effect is either 0.5 or 1, with $\gamma_1 = 0$, and either $\gamma_2 = 0.5$ or $\gamma_2 = 1$, so that one may consider γ_2 the category difference parameter. In the case of rating-scale data the number of response categories is three ($M = 3$), the rating scale category threshold parameters are 0 and -1 for ω_2 and ω_3, respectively. Note that unidimensionality applies within categories, in line with the applications of De Boeck et al. (2005). This is not a necessary assumption, but it worked quite well in the applications in question.

The factors of the design are as follows:

1. Within-group homogeneity vs. within-group heterogeneity,

   $\theta_{pc} \sim N(0, 0)$ vs. $\theta_{pc} \sim N(0, 1)$.

2. Equal discriminations vs. different discriminations, generated as follows (but treated as fixed),

   $\alpha_{ic} \sim LogN(1, 0)$ vs. $\alpha_{ic} \sim LogN(1, 0.25)$. (That is, 1PL model vs. 2PL model.) LogN denotes the lognormal distribution. This second design factor has to be seen in combination with the fourth. In case of discrimination equivalence, only one set of discriminations was drawn for both groups, whereas in case of lack of discrimination equivalence, two separate draws were made, one for each group.

3. Location equivalence vs. group-specific locations,

   $\delta_{ic} = 0$ for all i and c vs. $\delta_{ic} = 0$ for all i and c , except $\delta_{12} = -1$, and $\delta_{22} = 1$.

4. Discrimination equivalence vs. group-specific discriminations,

   $\alpha_{i1} = \alpha_{i2}$ vs. $\alpha_{i1} \neq \alpha_{i2}$, the latter being implemented by redrawing from the lognormal distribution $LogN(1, 0.25)$.

5. 100, 300, or 500 as number of persons in each category.

In order to enhance the generalizability of the study, each data set was generated from a different set of difficulties and a different set of discriminations (if applicable), as explained earlier. The levels of the different factors were chosen to be representative of the applications in De Boeck et al. (2005). The chosen parameter values (variance of θ and α, the values of δ and γ) were certainly not extreme but rather moderate or even small. Using a probability of 0.5 as a reference level, an effect of 1 on the logit scale (the scale of our parameters) implies a difference of 0.23 on the probability scale, and an effect of 0.5 on the logit scale represents 0.12 on the probability scale. The effects are maximal when applied to the reference level of 0.5. They are smaller for other probabilities.

As one may notice, the degree of location nonequivalence and the degree of discrimination nonequivalence is not the same. The locations are only different for two items out of ten, whereas for the discriminations two independent draws were made. The reason for that is that the estimation of the discrimination parameters is less reliable than the estimation of the location parameters, hence, a more substantial nonequivalence is needed for discriminations in order to detect the nonequivalence.

The combination of the four main factors (apart from the number of respondents in each category) would result in sixteen cells in the data generation design, but, as one shall see, not all of these combinations result in feasible models. The data generation design appears in Table 1. The nature of the categories, manifest or latent, is not really a factor of the simulation design as will be explained.

In the first part of the design, the data were generated from the model with equal discriminations (i.e., the 1PL model in case of binary data), and the following two factors

Table 1: The design for data generation

	Equal discriminations (1PL)		Different discriminations (2PL)			
	Discrimination equivalence		Location equivalence		Location nonequivalence	
	Location equivalence	Location nonequivalence	Discrimination equivalence	Discrimination nonequivalence	Discrimination equivalence	Discrimination nonequivalence
Within-group homogeneity	cell a	cell c	cell e	cell g*[a]	cell i*	cell k*
Within-group heterogeneity	cell b	cell d	cell f	cell h	cell j	cell l

[a] The cells indicated with * are not used in the study for reasons explained in the text.

were varied, so that four cells were obtained:

(a) homogeneity and location equivalence,

(b) heterogeneity and location equivalence,

(c) homogeneity and no location equivalence,

(d) heterogeneity and no location equivalence.

Discrimination nonequivalence does not play here, as in case of the 1PL model the discrimination parameters equal for all items.

In the second part of the design, the data were generated from the model with different discriminations (i.e., the 2PL model in case of binary data), and the following three factors were varied, yielding the following eight cells:

(e) homogeneity with location and discrimination equivalence,

(f) heterogeneity with location and discrimination equivalence,

(g) homogeneity with location equivalence and no discrimination equivalence,

(h) heterogeneity with location equivalence and no discrimination equivalence.

(i) homogeneity with discrimination equivalence and no location equivalence,

(j) heterogeneity with discrimination equivalence and no location equivalence,

(k) homogeneity with no location equivalence and no discrimination equivalence,

(l) heterogeneity with no location equivalence and no discrimination equivalence.

The combination of the model with different discriminations with homogeneity might seem strange, because discriminations are in fact weights for variation in the latent variable. However, given that the categories differ, the discriminations can refer to the weights of this difference, and, thus, to an indicator-dependent size of the difference between categories. This is the true model for cell e and i. When this is combined with the absence of discrimination equivalence, however, an absurd (unidentified) combination is obtained. The item-dependent category difference cannot be different depending on the manifest category. This means that cells g and k make no sense. In addition, when the model with different discriminations and homogeneity is combined with no location equivalence, it will not be possible to identify where the differences in item locations come from (the model is not identified), consequently cell i (where the model combines these features) is excluded from the study. In this way, only nine cells (4+5) remain in the data generation design.

These nine cells were repeated for each combination of the other design factors. These other factors are data type (binary or rating-scale), category main difference (0.5 or 1) and number of persons in each category (100, 300, 500). The data for the case of manifest and latent categories are generated in the same way, and therefore this is, strictly speaking, not a design factor, but we will treat the distinction between manifest and latent as if it were an overarching factor. The effects of the other factors will be investigated within the levels of this factor: separately for manifest and latent categories.

Analyses

In principle, each data set was analyzed with four or five models: the true model and three alternative models (first part of the design) or four alternative models (second part of the design) that each differed from the true model in one single respect. For example, when the model for the data generation is a 2PL model with heterogeneity in the categories and with no location equivalence but discrimination equivalence of the indicators (cell j in Table 1), then five analysis models will be used: (a) the true model; (b) the 1PL with heterogeneity and no location equivalence but discrimination equivalence; (c) the 2PL with homogeneity, no location equivalence but discrimination equivalence; (d) the 2PL with heterogeneity, location equivalence and discrimination equivalence; (e) and the 2PL with heterogeneity, no location equivalence and no discrimination equivalence.

For the first part of the design (equal discriminations), the respects which were varied in the analysis models were: (a) equal discriminations vs. different discriminations, (b) homogeneity vs. heterogeneity, and (c) location equivalence vs. no equivalence. As noted earlier, discrimination equivalence vs. no equivalence is not considered here, as in this part of the design the discriminations are equal. For each of the three respects, one of the three alternative analysis models differs from the true model in the respect in question, yielding four analysis models. For the second part of the design (different discriminations), the respects which were varied in the analysis models were: (a) different discriminations vs. equal discriminations, (b) homogeneity vs. heterogeneity, (c) location equivalence vs. no equivalence and (d) discrimination equivalence vs. no equivalence. For each of the four respects, one of the four alternative analysis models differs from the true model in the respect in question, yielding five analysis models. This design for the analysis, however, led to the earlier-mentioned absurd combination of the 2PL model with homogeneity and absence of discrimination equivalence (one of the theoretical analysis models in cells e, h and l).

All the analyses were accomplished using the M*plus* software (Muthén & Muthén, 2005) with the restrictions that the mean of θ is zero in one of the two (latent) categories (reference category), the mean of θ in the other category showing the size of the category difference, that the variance of θ is one if the model with different discriminations is used, and that all the δ's are zero in a reference (latent) category. When performing a latent class analysis (for latent categories), a possible risk is that the optimization process converges to a local minimum. To avoid this possibility for each analysis one hundred starting value sets were used. After ten iterations per set of starting values the ten best sets of starting values (based on the likelihood value) were selected to run the optimization process till convergence was reached for each of these ten. The final solution is the best among these ten.

When carrying out latent class analysis, another important question may be the number of latent classes underlying the data. To find out what the optimal number is, one has to analyze the data at hand with different numbers of classes in order to compare the fit of those models. However, in the major application field of DIMCAT, investigating the dimensional or categorical nature of different psychological phenomena, it seems reasonable to assume that the number of classes is known because of the focus of the study (e.g., a class of depressed people, and a class of 'normal' people), and the aim of the researcher is to analyze the data given the assumed number of classes. For this reason and for reasons of simplicity the question concerning the number of latent classes is not dealt with in the present study.

Combination of the Two Designs

As a result of the combination of the designs for data generation and for data analysis, several cells are relevant for the same differentiation and each cell is relevant for several differentiations (see Table 1). All the cells generated with the same feature are relevant for the differentiation of that feature (in principle only, because some analysis models do not make sense for reasons noted above). Let us take heterogeneity as an example. There are six cells in the design in which the data are generated with heterogeneity: cells b, d, f, h, j, l; and all the data sets in these cells are analyzed with a model with heterogeneity (the true model), and also with an alternative model with homogeneity. In cells h and l, however, the alternative model with homogeneity does not make sense because of the combination of homogeneity with no discrimination equivalence (see above). That is, in the example case of heterogeneity, four cells are relevant for the differentiation from homogeneity. The following eight differentiations were investigated (with the corresponding number of cells):

(a) true 1PL model to be differentiated from the 2PL model: four cells,

(b) true 2PL model to be differentiated from the 1PL model: five cells,

(c) true homogeneity to be differentiated from heterogeneity: three cells,

(d) true heterogeneity to be differentiated from homogeneity: four cells,

(e) true location equivalence to be differentiated from absence of location equivalence: five cells,
(f) true absence of location equivalence to be differentiated from location equivalence: four cells,
(g) true discrimination equivalence to be differentiated from absence of discrimination equivalence: two cells,
(h) true absence of discrimination equivalence to be differentiated from discrimination equivalence: two cells.

For the case of 2 x 100 persons, 20 data sets were generated, and for the two other cases (2 x 300 and 2 x 500), 10 data sets were generated. As a result, the eight differentiations, from (a) to (h), were based on the following numbers of data sets: 160, 200, 120, 200, 200, 160, 80, and 80 (in principle only, because convergence was not always reached, mainly in the latent class analyses.)

Criteria

Three differentiation criteria were used:

(a) the likelihood ratio test (LR) for each comparison of the true model with an alternative model, either a more restricted or a more general model.
(b) the Akaike information criterion (AIC; Akaike, 1993): the deviance plus two times the number of parameters,
(c) the Bayesian information criterion (BIC; Schwarz, 1978): the deviance plus the number of parameters times the natural logarithm of the number of respondents.

For the LR, a successful differentiation would mean that the LR was not statistically significant if the true model was the more restricted model, and that the LR was statistically significant indeed if the true model was the more relaxed model. For the AIC and BIC, a successful differentiation would mean that the value of the true model was lower. The null hypothesis will each time be the simpler model. When looking at the differentiations in terms of Type I and Type II errors, for some differentiations, the Type I error rate is relevant (when the more relaxed model is the true one), whereas for other differentiations, the Type II error rate is relevant (when the more restricted model is the true one).

Results

The results for the different combinations of category type (manifest or latent), data type (binary or rating-scale) and category difference (0.5 or 1) appear in Tables 2 to 9. The results of the eight types of differentiations are discussed here in four steps. (1) In general, the differentiations can be made very well for half of the cases: 1PL (vs. 2PL), homogeneity (vs. heterogeneity), heterogeneity (vs. homogeneity), and location equivalence (vs. no). (2) For binary data, in two cases, when the true 2PL model is to be differentiated from the 1PL model, and when the true lack of discrimination equivalence is to be differentiated from discrimination equivalence, the performance of the BIC is rather poor, whereas the other two criteria perform quite well. This poorer performance of the BIC is consistent with the tendency of the BIC to prefer simpler models, as was stated already in the original publication on the BIC by Schwarz (1978, p. 463) "...our procedure leans more than Akaike's towards lower-dimensional models (when there are 8 or more observations)." (3) The complementary finding of the BIC's poor performance when the true model is more complex is that it performs better than the LR and the AIC when the true model is simpler. This is very clear for the case of discrimination equivalence (vs. no) where the results are sometimes (in the case of latent classes) poor for the LR and AIC but not for the BIC. (4) Finally, the detection of no location equivalence is problematic, especially for latent categories. This may be due to the deviance from location equivalence being very limited (two items only).

When the decision is based on the LR or the AIC the results are very similar. In general, both the Type I and the Type II error rates are low, except for the following cases. For latent classes, lack of location equivalence is not easily detected, meaning that the Type II error rate is high. However, also for latent classes discrimination equivalence is rejected too easily, meaning that the Type I error rate is high. This phenomenon is also present to a lesser extent for manifest categories, especially when the category effect is 1. The difference between location and discrimination is perhaps a consequence of the fact that the deviation from location equivalence is small.

The decisions based on the BIC show a somewhat different picture as described above. The poorer results of the BIC all relate to a high Type II error rate, and hence, imply that the BIC tends to be conservative.

Although in most of the cases there are no apparent interactions, there are some instances where the combination of the design factors matters. Two such combinations were found, which are described in the following.

(A) As it was pointed out above, in case of binary data the performance of the BIC in recognizing the 2PL model is rather poor, but in many cases it is also more challenging for the LR and for the AIC than the other differentiations. This finding may stem from the earlier mentioned fact, that in cell e there is a strange combination of the 2PL model and homogeneity. When checking the results per cell it becomes clear that this strange combination hinders the recognition of the 2PL model. For example, for manifest categories, binary data and a group main difference of 0.5 (Table 2), the 2PL model is correctly identified in 72.1% of the data sets (based on the LR), but if one takes a look at the results for the five cells which are relevant for this differentiation, it can be seen that the poor performance in cell e has a major effect on the overall result. The percentage of correct differentiations based on the LR are 7.5, 85, 93.9, 90 and 89.2 for cells e, f, h, j and l, respectively. This is a clear case where the variation of one factor has an effect that differs depending on the value of the other factors. However, the uncommon true model of cell e does not always have a detrimental effect on the recognition of the 2PL model. When rating-scale data are generated and analyzed the results for cell e are as good as in the other relevant cells (for the LR): 97.5, 100, 100, 100 and 100 for cells e, f, h, j and l, respectively (for the case of manifest categories and a main effect of 0.5).

(B) True location equivalence can be identified quite well, based on the aggregated results but it turns out that the model used for data generation affects the identification. When the 2PL model (with heterogeneity) is used to generate the data the recognition of location equivalence becomes less successful. In the case of manifest categories, binary data and a category effect of 1 (Table 3), the percentage of successful differentiations (based on the LR) are 97.5, 97.5, 100, 50 and 67.5 for cells a, b, e, f and h, of Table 1, respectively. It is not unexpected that interactions with respect to locations are more difficult to be established in a model with varying discriminations, especially when the discrimination for the items in question is small. Note that the discriminations are drawn from a lognormal distribution and can be small for the items in question.

A general result is that, in case of manifest categories (Tables 2 to 5), the differentiations are stronger than in case of latent categories (Tables 6 to 9). When the categories are latent, especially the recognition of the lack of location equivalence and the recognition of discrimination equivalence becomes problematic. The former problem is most likely to stem from the fact that in case of locations a milder form of nonequivalence was used (the locations of two items differ in the categories), whereas in case of lack of discrimination equivalence, a more substantial form of nonequivalence (two independent draws) was used. This was a deliberate choice for a reason mentioned earlier. To investigate this issue we generated new data sets in the relevant cells (fourty data sets in each cell) for the case of latent categories, binary data and a group main effect of 0.5 (compare with Table 6), but now with independent draws from a standard normal distribution for the locations depending on the category. Not surprisingly, this more substantial lack of location equivalence largely improves the recognition of the true model. In the original study the percentage of the correct differentiation for no location equivalence (vs. yes) is 34.8, 49.3 and 14.5 for the LR, AIC and BIC, respectively, whereas the corresponding result for the newly generated data sets were 100, 93.3 and 63.7 for the LR, AIC and BIC, respectively. As for the lack of discrimination equivalence, instead of two independent draws, new data sets were generated (forty in each relevant cell) with the nonequivalence being restricted for the first two items so that $\alpha_{12} = \alpha_{11} + 1$ and $\alpha_{22} = \alpha_{21} + 1$. The original results with two independent draws (compare with Table 6) for no discrimination equivalence (vs. yes) were 90, 85 and 1.7 for the LR, AIC and BIC, respectively, whereas the results for the newly generated milder form of nonequivalence were 0, 33.3 and 2.4 for the LR, AIC and BIC, respectively. It is clear that the differentiation success depends on the size of the true difference.

As for the second problem, the recognition of discrimination equivalence, it is interesting to note that when the categories are latent, an asymmetry is found in the results (Tables 6 to 9). In contrast with the cases involving manifest categories, true discrimination equivalence is detected well when using the BIC, but not with LR or BIC, whereas true discrimination nonequivalence is identified well with the LR and AIC, but not with BIC (as mentioned above). This phenomenon is most likely to stem from the fact that BIC tends to favor simpler models (as explained above), whereas LR and AIC tend to favor more complex models. With the increased uncertainty of model estimation introduced by latent categories, these tendencies play a more important role.

Discussion

The results of the simulation study show that the differentiations relevant to DIMCAT can be made very well. As could be expected, the differentiation capacity depends on

Table 2: Results for eight types of differentiation: manifest categories, binary data, category effect of 0.5 (percentages)

Differentiation	LR	AIC	BIC
1PL (vs. 2PL)	96.9	94.4	100
2PL (vs. 1PL)	72.1	68.4	18.9
Homogeneity (vs. Heterogeneity)	95.8	87.4	99.2
Heterogeneity (vs. Homogeneity)	100	100	100
Location equivalence (vs. no)	91.2	88.1	99
No location equivalence (vs. yes)	91.2	93.2	77.7
Discrimination equivalence (vs. no)	82.4	86.5	100
No discrimination equivalence (vs. yes)	87.1	82.9	31.4

Table 3: Results for eight types of differentiation: manifest categories, binary data, category effect of 1 (percentages)

Differentiation	LR	AIC	BIC
1PL (vs. 2PL)	95.6	94.4	100
2PL (vs. 1PL)	64.8	61.1	11.9
Homogeneity (vs. Heterogeneity)	100	94.1	99.2
Heterogeneity (vs. Homogeneity)	100	100	100
Location equivalence (vs. no)	84.2	76.5	96.4
No location equivalence (vs. yes)	91.2	93.2	75.7
Discrimination equivalence (vs. no)	36.2	44.9	100
No discrimination equivalence (vs. yes)	89	84.9	26

Table 4: Results for eight types of differentiation: manifest categories, rating-scale data, category effect of 0.5 (percentages)

Differentiation	LR	AIC	BIC
1PL (vs. 2PL)	98.1	98.8	100
2PL (vs. 1PL)	99.5	99.5	83.9
Homogeneity (vs. Heterogeneity)	99.1	72.3	76.8
Heterogeneity (vs. Homogeneity)	100	100	100
Location equivalence (vs. no)	83	74	94.5
No location equivalence (vs. yes)	96.8	99.4	93.5
Discrimination equivalence (vs. no)	79.7	84.8	100
No discrimination equivalence (vs. yes)	100	100	75.9

Table 5: Results for eight types of differentiation: manifest categories, rating-scale data, category effect of 1 (percentages)

Differentiation	LR	AIC	BIC
1PL (vs. 2PL)	96.9	97.5	100
2PL (vs. 1PL)	99.5	99.5	78.2
Homogeneity (vs. Heterogeneity)	99	84.5	84.5
Heterogeneity (vs. Homogeneity)	100	100	100
Location equivalence (vs. no)	68.7	54.5	84.3
No location equivalence (vs. yes)	97.4	99.4	91
Discrimination equivalence (vs. no)	57.5	68.5	100
No discrimination equivalence (vs. yes)	100	100	72.7

Table 6: Results for eight types of differentiation: latent categories, binary data, category effect of 0.5 (percentages)

Differentiation	LR	AIC	BIC
1PL (vs. 2PL)	81.6	85.7	95.2
2PL (vs. 1PL)	80.1	73.1	14.6
Homogeneity (vs. Heterogeneity)	88.2	82.4	92.9
Heterogeneity (vs. Homogeneity)	91.9	93.2	89.2
Location equivalence (vs. no)	85.1	86.6	100
No location equivalence (vs. yes)	34.8	49.3	14.5
Discrimination equivalence (vs. no)	30.5	44.1	94.9
No discrimination equivalence (vs. yes)	90	85	1.7

Table 7: Results for eight types of differentiation: latent categories, binary data, category effect of 1 (percentages)

Differentiation	LR	AIC	BIC
1PL (vs. 2PL)	92	90.7	100
2PL (vs. 1PL)	73.1	63.7	13.7
Homogeneity (vs. Heterogeneity)	93.8	85.6	95.9
Heterogeneity (vs. Homogeneity)	96.7	98.7	96
Location equivalence (vs. no)	88.8	79.8	97.2
No location equivalence (vs. yes)	48.2	58.3	28.8
Discrimination equivalence (vs. no)	14.3	21.4	100
No discrimination equivalence (vs. yes)	90.8	83.1	6.2

Table 8: Results for eight types of differentiation: latent categories, rating-scale data, category effect of 0.5 (percentages)

Differentiation	LR	AIC	BIC
1PL (vs. 2PL)	94.2	94.2	96.1
2PL (vs. 1PL)	98.8	97.1	63.7
Homogeneity (vs. Heterogeneity)	99.1	90.7	92.5
Heterogeneity (vs. Homogeneity)	98.7	99.4	98.7
Location equivalence (vs. no)	82.3	68.6	96.6
No location equivalence (vs. yes)	41.5	56.3	20
Discrimination equivalence (vs. no)	29.2	43.8	100
No discrimination equivalence (vs. yes)	96	92	8

Table 9: Results for eight types of differentiation: latent categories, rating-scale data, category effect of 1 (percentages)

Differentiation	LR	AIC	BIC
1PL (vs. 2PL)	89.8	86.8	99.4
2PL (vs. 1PL)	99.4	95.9	56.1
Homogeneity (vs. Heterogeneity)	99	67.7	74
Heterogeneity (vs. Homogeneity)	100	100	100
Location equivalence (vs. no)	74.9	67.4	94.9
No location equivalence (vs. yes)	60.3	71	38.2
Discrimination equivalence (vs. no)	21.4	42.9	100
No discrimination equivalence (vs. yes)	95.1	88.5	9.8

how much the true model differs from the competing models. The performance of the LR and AIC were similarly good, but the BIC tended to prefer simpler models in two cases in which the true model was more complex. This result reflects the tendency of the BIC to prefer the more parsimonious models. As it was emphasized above, the differentiation capacity is largely determined by the differences one is interested in. In the simulation study, the data generation values were chosen to be in line with the empirical values of the applications from De Boeck et al. (2005).

The more detailed investigation of the differentiation results suggest that the correctness of most decisions is not influenced by interactions in a substantial way, and that therefore the aggregation principle is justified indeed. However, the differentiation of location equivalence is problematic when the true model is the 2PL. The other problem, related to the strange combination of 2PL and homogeneity, does not posit a problem in real-life applications because this model is not of much interest.

On the other hand, the simulation study has several limitations. First, in case of the latent class analyses, only the two class (category) solution was taken into account, and the models were compared only for that case. Second, when the class membership is known beforehand (manifest categories), classification is not an important issue (as far as the analyses are concerned), but in case of latent class analyses it also can be an important issue how well the observations are classified, that is, how well the class sizes can be recovered, and how well the subjects can be assigned to their true classes. (These questions are interrelated, of course.) The question of classification is not dealt with here, because the focus of the present simulation study is the differentiation capacity of DIMCAT for the various models.

As for the kind of differentiation test, it can be concluded that when comparing the relative fit of DIMCAT models one should trust more the LR and the AIC than the BIC. That being said, it has to be emphasized that in case of latent categories, the BIC is to be preferred when differentiating discrimination equivalence versus the lack of discrimination equivalence.

When concentrating on the DIMCAT dimensions of homogeneity versus heterogeneity and equivalence versus nonequivalence, the conclusions are as follows. Homogeneity can be very well differentiated from heterogeneity and vice versa. For equivalence and nonequivalence, the situation is more complicated. For manifest categories, the only problem seems to be the use of the BIC for the detection of discrimination nonequivalence for binary data, but the LR test and the AIC are performing quite well. For latent categories, both location and discrimination nonequivalence need a stronger extent then just two indicators in order to be detected, and by preference the LR test and the AIC are used instead of the BIC, in order to detect the nonequivalence.

These conclusions are reassuring and informative in that they point out that the differentiation capacity is rather high in general, but lower in some cases that could be specified.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood theory. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–381). Budapest: Akadémiai Kiadó.

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika, 43*, 567–573.

De Boeck, P., Wilson, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review, 112*, 129–158.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM-algorithm (with discussion). *Journal of the Royal Statistical Society, B, 39*, 1–38.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* London: Erlbaum.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

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

Meehl, P. E. (1979). A funny thing happened on the way to the latent entities. *Journal of Personality Assessment, 43*, 563–581.

Meehl, P. E. (1995). Bootsraps taxometrics. Solving the classification problem in psychopathology. *American Psychologist, 50*, 266–275.

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

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 55*, 195–215.

Muraki, E. (1990). Fitting a polytomous item response model for Likert-type data. *Applied Psychological Measurement, 14*, 59–71.

Muthén, L. K., & Muthén, B. O. (2006). M*plus* 4.0 [Computer software]. Los Angeles,CA: Muthén & Muthén.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271–282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology, 44*, 75–92.

Samejima, F. (1996). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer Verlag.

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

Vermunt, J. K. (1997). *ℓEM: A general program for the analysis of categorical data.* Tilburg University, The Netherlands.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua.* London: Sage.

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