

Detecting Heterogeneity in Logistic Regression Models

Katalin Balázs, István Hidegkuti, and Paul De Boeck,
K. U. Leuven, Belgium

In the context of item response theory, it is not uncommon that person-by-item data are correlated beyond the correlation that is captured by the model—in other words, there is extra binomial variation. Heterogeneity of the parameters can explain this variation. There is a need for proper statistical methods to indicate possible extra heterogeneity and its location because investigating all different combinations of random parameters is not practical and sometimes even unfeasible. The ignored random person effects are the focus of this study. Considering the random

weights linear logistic test model, random effects can occur as a general latent trait and as weights of covariates. A simulation study was conducted with different sources and degrees of heterogeneity to investigate and compare various methods: individual analyses (one per person), marginal modeling, principal component analysis of the raw data, DIMTEST, and DETECT. The methods are illustrated with an application on deductive reasoning. *Index terms: heterogeneity, binary data, covariates, PCA, marginal modeling, DIMTEST, DETECT*

Test data are often of a binary type and may be considered as repeated measures because different items are presented to the same persons. The focus of this article is on binary repeated measures with a design. A test has a design when the items are carefully controlled, selected, and counterbalanced based on a number of design factors (Embretson, 1985). Test design became a popular approach in cognitive psychology, for example, for the study of intelligence (Sternberg, 1977a), deductive reasoning (Rijmen & De Boeck, 2002; Sternberg, 1979), analogies (Whitely, 1978), and spatial representations (Egan, 1979; Embretson, 1985, chap. 3).

The design can be represented as a set of item features, or design variables, functioning as predictors and providing a potential basis to explain the data. Tests do not always have such a design, and the individual items enter the psychometric model instead of the design factors. In contrast, when the test is based on a design, the design variables can be used as predictors in a mixed logistic regression model for the data. An item response model with explanatory item covariates for the binary data is a logistic regression model, as in equation (1):

$$\log \left(\frac{P(Y_{pi} = 1 | \theta_p, \boldsymbol{\beta}_p)}{1 - P(Y_{pi} = 1 | \theta_p, \boldsymbol{\beta}_p)} \right) = \theta_p + \beta_{1p}x_{1i} + \dots + \beta_{kp}x_{ki} + \dots + \beta_{Kp}x_{Ki}. \quad (1)$$

The model assumes binary response variables, which are nonlinearly related to the covariates. Y_{pi} is the response of person p ($p = 1, \dots, P$) to item i ($i = 1, \dots, I$) and follows a Bernoulli distribution (binomial distribution with $n_{pi} = 1$). $P(Y_{pi} = 1 | \theta_p, \boldsymbol{\beta}_p)$ is the success probability for person p and

item i , modeled as a function of the covariates. x_{ki} is the k th covariate ($k = 1, \dots, K$), changing its value over items, and the β_{kp} is the associated random weight. θ_p is the random intercept that is the so-called ability of the person in the context of achievement tests. When the intercept is random, while the weights of the covariates are fixed, the resulting model is the linear logistic test model (LLTM; Fischer, 1973). The LLTM has been commonly used without assuming heterogeneity in the weights of the covariates but with heterogeneity being restricted to the intercept. When the effects of the covariates are random over persons, which is indicated with subscript p , then the random weights LLTM is obtained (RWLLTM; Rijmen & De Boeck, 2002), or in other words, the resulting model is a logistic mixed model. The term *mixed* refers to the combination of fixed and random effects.

The term *heterogeneity* refers to any source of the binomial variance beyond the fixed effects, and the more specific term *extra heterogeneity* denotes the heterogeneity that is not yet included in the model, when the model does not specify all sources of variance in the data. Ignored variance causes overdispersion (Collett, 1991). Overdispersion may occur for many reasons, such as an additional explanatory variable is ignored, there are hidden clusters, interviewer effects are present, there is an insufficient number of interaction terms, or the choice of link function is inappropriate (Collett, 1991; Fitzmaurice, Heath, & Cox, 1997; Williams, 1982). In principle, heterogeneity may stem from the persons or from the items. Underdispersion can also occur, but that is a rare phenomenon (Collett, 1991; Fitzmaurice et al., 1997). Normally, several factors play a role in the measurement and can cause overdispersion (Collett, 1991; Fitzmaurice et al., 1997; Hattie, 1985; McCullagh & Nelder, 1989). In this study, the focus is on person-based heterogeneity, which can occur as a random effect in the intercept and/or in the weights of the covariates (also called random coefficients) in an RWLLTM.

Neglected random effects yield a kind of local item dependency, but the notion of local item dependency is more general. Item response dependencies can be studied through the correlations of the residuals of the applied item response theory (IRT) models, providing indices for item dependencies: Q_2 (Van den Wollenberg, 1982; Yen, 1984) and Q_3 (Yen, 1984). A specialized computer software was developed (IRT LD) for the detection of local dependencies (Chen & Thisen, 1997), and graphical techniques were also proposed for detecting residual dependencies (e.g., Landwehr, Pregibon, & Shoemaker, 1984). These are valuable methods for dependencies in general, but they do not aim directly at item covariates as a source of heterogeneity.

The aim of the present study is to investigate methods for detecting heterogeneity in data with item covariates. The motivation for this interest is that in psychology, it is not uncommon that item covariates have a person-based effect. Often, one is precisely interested in the individual differences in these effects. In personality psychology, the study of individual differences in the effect that situational features have is called interactionism (e.g., Blumer, 1969; Pervin, 1977). In the domain of intelligence, the study of cognitive processes, as initiated by Sternberg (1977b) and Embretson (1985), is based on item covariates indicating how much of a certain process is required to succeed in the item. The weights of these covariates are assumed to show individual differences in the ability for dealing with the difficulty represented by the covariate. A similar idea is behind the development of a cognitive diagnostic approach, as initiated by Tatsuoka and Tatsuoka (1982), which represents the item covariates in the so-called Q-matrix (Tatsuoka, 1990). Although in further developments (DiBello, Stout, & Roussos, 1995), a formalization is chosen that is different from the one in equation (1), individual differences with respect to the item covariates, as defined in the Q-matrix, are an important ingredient of the approach.

Because a well-established theory that specifies the sources of heterogeneity is often not available, one may consider to include random effects for all possible covariates. However, this leads to models with high dimensionality, which require high-dimensional integrals to be solved for

a successful estimation. An interesting alternative to deal with high dimensionality is a Bayesian approach (Béguin & Glas, 2001; Segall, 2001). However, high-dimensional models may require larger sample sizes than in a typical study in psychology, where a few hundred or even less than 100 is a common practice. For these reasons, a diagnostic approach of heterogeneity without estimating all possible random effects seems useful. As a first step in the diagnostic approach, it can be investigated whether there are individual differences and where they are, so that in a next step, a specific model can be estimated.

There is a wide range of literature on the diagnosis of heterogeneity in biometrics, with several procedures for dealing with heterogeneity. Unfortunately, most of these procedures cannot be implemented in the field of psychometrics because they are developed for data following a binomial distribution with $n > 1$ (Collett, 1991). In psychometrics, one often has only one observation for each combination of a person and an item.

On the other hand, several methods were developed in psychometrics for indicating multidimensionality in an item set, independently of a possibly available test design. An early overview of unidimensionality assessment is provided by Hattie (1985). At present, the most prominent methods are DETECT (Zhang & Stout, 1999), a method to reveal the dimensionality structure of the data, and DIMTEST (Stout, Douglas, Junker, & Roussos, 1993), a method for testing the unidimensionality of a test. Both methods are nonparametric. Because they are developed to investigate the dimensionality of the data, and because the dimensions refer to individual differences, such as random weights of the item covariates do, these methods are possible candidates for a diagnostic approach to heterogeneity. DIMTEST and DETECT do not use item covariates; hence, part of the information is not used in the data analysis. Although principal components analysis (PCA) is not really appropriate for binary data, it can also detect dimensional variance. Neither of these methods makes use of these item covariates. Nevertheless, all three undirected methods—DIMTEST, DETECT, and PCA—are investigated on their performance for data with a design.

Also, two directed methods are investigated. They are directed to the item covariates but without an actual estimation of the possible random effects of the item covariates. First, logistic regressions are used for each individual separately, and the variance of the weights is investigated. Second, marginal modeling is performed with a separate modeling of the association structure (Hardin & Hilbe, 2003). None of the methods may be appropriate for the estimation of item response models, but they may be useful for the detection of heterogeneity nevertheless.

The methods that are applied differ in several respects. The differences are summarized here but will be further explained when the methods are described more in detail. The first respect in which these methods can differ is whether they indicate the *localization of the heterogeneity* in terms of the item covariates (individual differences in intercept and slopes). DIMTEST, DETECT, and the (size of) eigenvalues of a PCA indicate (extra)heterogeneity. The DETECT partition and the PCA loadings can also give an indication of where the heterogeneity is located. Finally, the individual analyses and marginal modeling explicitly locate the (extra) heterogeneity in terms of the covariates.

The second respect in which methods can differ is whether they provide an *absolute or relative decision* about the presence of extra heterogeneity. In some cases, an associated significance test is available, such as for DIMTEST and marginal modeling, whereas in other cases, a rule of thumb has been proposed in the literature, such as for DETECT. For other methods, no evident decision rule exists. The loadings of the PCA may be interpretable in terms of the item covariates, and corresponding eigenvalues may give an indication of the relative size of heterogeneity in the weights of the corresponding item covariates. In a similar but more direct way, also the variance of the individual estimates (from individual logistic regressions) give such an indication. However, the

critical values are unknown. PCA and individual logistic regressions can still be used for a relative decision because they indicate for which covariates the weights are more likely to be heterogeneous than for others. Based on this indication, random parameters can be included in the model in the order suggested by the diagnostic analyses, until the fit statistic of the random effects model does not improve any more. Finally, marginal modeling yields an absolute statistical test to decide whether the localized heterogeneity is significant. From the size of the parameters, one can also make a rank order and, hence, relative decisions. Among the investigated methods, the two directed methods, individual analyses and marginal modeling, will be described and studied first. In the last section, an application to real data will be presented.

Overview of the Methods

Individual Analyses

As explained, the heterogeneity that will be studied implies individual differences in the intercept and/or in the slope(s). A very simple approach would be to perform a logistic regression analysis for each single person and then to inspect the variance of the regression weights and the intercept. Apart from the fact that the separate analyses do not take advantage of information from other individuals, this method has the drawback that complete or quasi-complete separation (see e.g., Webb, Wilson, & Chong, 2004) may occur rather easily. For binary data, complete separation is realized when the 0 and 1 responses can be perfectly separated by the weighted sum of the covariates in the regression equation. When the overlap is limited to the weighted sum of zero, then the separation is quasi-complete. Complete and quasi-complete separation do not give unique, finite maximum likelihood estimates. Therefore, data that result in complete or quasi-complete separation in the logistic regression analysis have to be omitted, but this omission is not without consequences for the variance of the estimates.

On one hand, the method is limited to cases when there is information on the item covariates a priori because it uses the information of the item covariates. On the other hand, based on the variance of the regression weights and the variance of the intercept, the method provides a direct indication of where the heterogeneity is located. Furthermore, based on the order of the variance estimates, it can be used for a relative decision about which random effects should be included first in the model.

Marginal Models

In general, one can follow a marginal modeling approach as an alternative to an IRT model when one is not interested in the measurement of the latent traits reflected in the random effects. Because the detection of heterogeneity does not require such measurement, this approach can be applied in this study. The primary aim of marginal models is to find the relationship between the expected value of the response variable and the covariates (i.e., to find an appropriate model for the mean). Using generalized estimating equations (GEE) and, more particularly, the GEE2 variant (Hardin & Hilbe, 2003), besides an estimation for the mean, an estimation for the association structure also is obtained. For binary data, it is appropriate to use odds ratios instead of correlations as the measure of associations:

$$OR(Y_{pi}Y_{pi'}) = \frac{P(Y_{pi} = 1, Y_{pi'} = 1)P(Y_{pi} = 0, Y_{pi'} = 0)}{P(Y_{pi} = 1, Y_{pi'} = 0)P(Y_{pi} = 0, Y_{pi'} = 1)}, \quad (2)$$

where p refers to a cluster (i.e., a person in this case), i is the first item of the item pair, and i' is the second item of the item pair.

In alternating logistic regression (ALR) (Carey, Zeger, & Diggle, 1993), a logistic regression model is fitted to obtain an estimation of the effects that covariates have on odds ratios (ORs):

$$\log(OR(Y_{pi}Y_{pi'})) = \sum \alpha_k x_{ki} x_{ki'}, \quad (3)$$

where x_{ki} and $x_{ki'}$ are the values of items i and i' on the k th item covariate (one covariate is an overall 1 vector), and α_k is the association parameter belonging to the k th item covariate. In other words, α_k is a weight that indicates how much item covariate k contributes to the log odds ratio. Heterogeneity based on covariate k is shown in α_k being larger than zero. Negative values of α_k are possible but not really meaningful because it implies that sharing an item covariate value yields a negative association. In this study, a $-1/+1$ coding was used for the covariates in the ALR (in equation (3)). The $+1/-1$ coding implies a positive association for same-sign values of x_{ki} and $x_{ki'}$ and a negative association for opposite-sign values. The GENMOD procedure of the SAS software (SAS Institute, 1999) was applied for ALR analyses. In general, the interpretation of α_k depends on how the covariates are coded.

The method of marginal modeling provides the location of heterogeneity in terms of the covariates through the α parameters (equation (3)) and their statistical significance. The asset of the marginal modeling approach is that it can localize the heterogeneity based on advance knowledge of the item covariates (which is also a limitation of the method) and that a statistical test for the parameter estimates is available through the standard error estimate.

PCA for the Raw Data

Although PCA is not an orthodox method for the analysis of binary data, it may be a useful and quite easy technique for detecting heterogeneity in practice. In case of heterogeneity that stems from multidimensionality, the data are correlated, and the underlying dimensions correspond to the sources of heterogeneity. Earlier, several attempts were made to find an index that would reflect unidimensionality, based on the idea that the larger the variance explained by the first principal component, the better the assumption of unidimensionality (Hattie, 1985). However, it is well known that PCA for binary data may lead to artifacts, especially when the proportions of response values are extreme, but this study nevertheless explores how PCA behaves for detecting heterogeneity as in a logistic model.

PCA is an undirected approach that can be used as a detection method in several ways. First, the eigenvalues give an indication of the size of the heterogeneity but without a statistical test or a clear absolute decision criterion. Second, from the loadings the items have on the components, one can derive an indication of where the heterogeneity occurs. When item covariates are known and are sources of heterogeneity, the loadings should show specific patterns, as explained in the Results section. The order of the eigenvalues of corresponding random effects could be a criterion for a relative decision on the heterogeneity. PCA does not use advance knowledge of the item covariates. However, the method should be used with caution because of the possibility of artifacts.

DIMTEST

DIMTEST (Stout et al., 1993) is a nonparametric statistical approach for assessing unidimensionality of dichotomously scored test items. DIMTEST provides a tool for assessing extra heterogeneity beyond a general underlying latent trait. The original DIMTEST (DIMTEST 1) requires three item groups: AT1, a unidimensional subset of items, and two subsets from the remaining items: AT2, which contains the same number of items as AT1 and has a similar difficulty structure, and PT, the rest of the items. The unidimensional subset (AT1) can be selected based on expert opinion or factor analysis. When the data are unidimensional, the sum of the standardized differences between the

observed and expected variance of the AT1 scores must be close to zero for each PT score group. However, for short tests, this statistic is positively biased (Nandakumar & Stout, 1993). The analogous statistic is calculated for the AT2 scores for each PT group. This statistic for AT2 is sensitive to the bias but not to the multidimensionality because AT2 has the same multidimensionality structure as PT. The DIMTEST T statistic is based on the discrepancy between the standardized difference of the variances for AT1 and AT2 and hence is an unbiased estimator of multidimensionality. The DIMTEST T statistic is asymptotically normally distributed when the data are unidimensional so that a significance test can be obtained.

In an early simulation study (Stout, 1987), the DIMTEST procedure was shown to have good power in detecting multidimensionality when the sample size was very large (750, 2,000, 20,000). DIMTEST performs not as well for smaller sample sizes, for example, 200 (van Abswoude, van der Ark, & Sijtsma, 2004). It is important to note that in psychological studies, 200 is already a large sample size.

Recently, Stout, Froelich, and Gao (2001) developed a resampling method, in which unidimensional data are generated based on the estimated item characteristic curves of the investigated data estimated by a combination of kernel smoothing methods. The generated data are used instead of AT2 for bias correction. Furthermore, Froelich and Habing (2003) suggested implementing DETECT (Stout & Zhang, 1999) and HCA/CCPROX (Roussos, Stout, & Marden, 1998) for the selection of a unidimensional item cluster. This method has a higher power than linear factor analysis. A new version of the DIMTEST software (DIMTEST 2), which seems to have an improved efficiency in detecting multidimensionality, was released in 2005. Both versions of DIMTEST are investigated.

Both DIMTEST procedures provide a criterion for an absolute decision on extra heterogeneity beyond a general dimension. DIMTEST provides a statistical test but does not give an indication of where the extra heterogeneity is located. For using DIMTEST, no advance knowledge of item covariates is required.

DETECT

The DETECT procedure is a nonparametric method developed for detecting test dimensionality or, more precisely, for disclosing the dimensionally homogeneous item clusters of a test. The DETECT procedure was established originally by Kim (1994), and later the theory of DETECT was developed more extensively (see, e.g., Stout et al., 1996; Zhang & Stout, 1999). In the case of sufficiently separated, strongly homogeneous item clusters (as in a simple structure), the procedure finds the exact number of latent dimensions and the true latent structure of the test. When the clusters are clearly separable (as in an approximate simple structure) but the item vectors differ considerably in their angles in the test space, DETECT finds the crucial clusters but not necessarily the number of dimensions. The $R(P)$ index provides information about the degree to which simple structure is realized. Approximate simple structure is assumed in practice when the estimated $R(P) \geq 0.8$, and simple structure is assumed when $R(P)$ equals 1.

The theoretical DETECT index for a given partition (P) is based on the sum of the conditional covariances (conditional on the test composite) of item pairs belonging to the same cluster minus the conditional covariances of item pairs belonging to different clusters. A DETECT value between 0 and 0.1 may indicate unidimensionality; higher values, between 0.1 and 0.5, 0.5 and 1, 1 and 1.5, and 1.5 or higher, correspond to weak, moderate, strong, and very strong multidimensionality, respectively (Douglas, Kim, Roussos, Stout, & Zhang, 1999).

An iterative procedure is used to obtain the partition with the highest DETECT index for a given maximal number (chosen by the user) of nonoverlapping clusters (Zhang & Stout, 1999). The current version of DETECT starts with hierarchical cluster analyses and then uses a generic algorithm to obtain the global maximum DETECT value. Cross-validation is an option

in the procedure. In the cross-validation, two subsets are used with approximately equal size. First, the DETECT value is calculated for the first subset, which is called the *maximum DETECT value*. In addition, a partitioning is obtained that is applied on the second subset. The resulting DETECT value is called the *reference DETECT value*. Zhang and Stout (1999) suggest that when the discrepancy between the reference DETECT value and the maximum DETECT value is large, one should suspect unidimensionality. They define a *discrepancy measure* as the difference between the maximum DETECT value and the reference DETECT value divided by the reference DETECT value (Zhang & Stout, 1999). In their study, a *combined criterion* was used. When the discrepancy exceeded 0.5 or the reference DETECT value was smaller than or equal to 0.1, the data set was judged unidimensional. This decision rule worked perfectly in their case. Zhang and Stout warn that DETECT might not perform well for a small sample size or when the number of items is small.

DETECT provides an absolute decision on extra heterogeneity, but the criterion is a rule of thumb and not a statistical test. An important asset of the DETECT method is that it yields a cluster structure and therefore may give an indication of not just whether extra heterogeneity occurs but also where it is located. The method does not require advance knowledge of item covariates, though. When item covariates are available and the clusters can be linked to there, the method can be informative also for the relative decisions on heterogeneity.

The Simulation Study

To test the methods, a simulation study was carried out. A quite modest problem size was chosen, with 32 items, 3 covariates, and 200 persons. The size of the data set is rather typical in psychology when a test or inventory is used, and it is rather large in comparison to most experiments. The covariates were binary and were crossed in an orthogonal way so that there were eight types of items, with 4 items of each type. From an experimental point of view, this is a $2 \times 2 \times 2$ within-subject repeated-measures design. In contrast with experiments, tests often do not have a design, but having a design is a desirable feature for a test (Embretson, 1985) and also for purposes of cognitive diagnosis, as noted in the introduction (Tatsuoka, 1990; Tatsuoka & Tatsuoka, 1982). In psychological experiments with repeated measures, a design is often used, and the approach described here is relevant for repeated-measures experiments as well.

For the data generation, the coding of the covariates was +1 and -1. When the effects were fixed, the coding did not matter, as any change in the coding could be adapted through the intercept. However, when the effect was random over persons, opposite signs of the covariate values led to a negative association, whereas same signs led to a positive correlation. In combination with a random intercept (with a coding of +1 for all items), opposite signs and random weights for the corresponding covariate yielded a simple structure because a general factor and a bipolar factor were formally equivalent with two unipolar factors. This particular structure of item covariates makes sense for both the ability and the personality domains. Perhaps bipolar item covariates as such are not evident in the ability domain, but it is common in an unrotated factor solution to find a general dimension (random intercept) and a bipolar dimension, so that a simple structure also is obtained. For the personality domain, contrasts do make sense as item covariates, and of course, the simple structure is not uncommon for personality (e.g., Goldberg, 1993).

One of the slopes (β_{p1}) and/or the intercept (θ_p) were defined to be random over persons; the other slopes had a fixed value for all individuals. The model for the data generation was the following:

$$\text{logit}(P(Y_{pi} = 1 | \theta_p, \beta_{p1})) = \theta_p + \beta_{p1}x_{i1} + \beta_2x_{i2} + \beta_3x_{i3}. \quad (4)$$

When only one random effect was used, the variance was varied between 0 and 1.2 with steps of 0.2 (0, 0.2, 0.4, 0.6, 0.8, 1, 1.2). The mean of the intercept was always zero, and the means of the slopes were 1. The theoretical mean of the raw data for each data set as a whole was .5. In total, there are 14 cells of the design for the 7 values of the intercept variance and for the 7 values of the slope variance. In 2 cells of the design, there is no random effect present, namely, when the variance of the manipulated parameter (intercept or slope) is zero. This part of the simulation study will further be referred to as the *single-effect design*.

When both the intercept and one slope were random, three variance values were used –0, 0.2, and 1.2 – so that nine variance combinations were obtained from crossing the three levels. These values represent three kinds of effects of the covariates: a fixed effect (variance is zero), a minor source of heterogeneity (variance is 0.2), and a major source of heterogeneity (variance is 1.2). With two random effects, the distribution was bivariate normal with zero correlation. In one cell of the design, there is no random effect, namely, when all variances are zero. This part of the simulation study will be called the *combined-effect design*.

The above-described variance values are the theoretical values with which the data were generated. The actual variance of the random effects may be different due to the sampling that is inherent to the generation procedure. The corresponding two variances are denoted as *theoretical variance and real variance*, respectively.

In general, a relatively small number of data sets (10) was generated per cell because the results seemed rather stable over these 10 data sets. Only for the investigation of DIMTEST was a larger number of data sets used (100).

Results

Individual Analyses

For all individual analyses, the same covariate coding was used as for the data generation of the data. For the combined-effect design, the results are given in Table 1. The results are very similar for the single-effect design. Complete or quasi-complete separation occurred in 6.9% of the individual logistic regression analyses for the single-effect design, and this ratio was 9.5% for the combined-effect design. The corresponding estimates were not considered in the calculation of the variances.

First, it is clear that the larger the theoretical variance, the larger the variance of the individual estimates. For theoretical values of 0 (including the fixed effects), mean variances of 0.2 (e.g., for $\hat{\sigma}_{\beta_3}^2$ when $\sigma_{\theta}^2 = 0.2$ and $\sigma_{\beta_1}^2 = 0$) to 0.25 (for $\hat{\sigma}_{\beta_3}^2$ when $\sigma_{\theta}^2 = 1.2$ and $\sigma_{\beta_1}^2 = 1.2$) were found. For theoretical values of 0.2, mean variances of 0.43 (for $\hat{\sigma}_{\theta}^2$ when $\sigma_{\theta}^2 = 0.2$ and $\sigma_{\beta_1}^2 = 1.2$) to 0.50 (for $\hat{\sigma}_{\theta}^2$ when $\sigma_{\theta}^2 = 0.2$ and $\sigma_{\beta_1}^2 = 0.2$) were found, and finally, for theoretical values of 1.2, mean variances of 1.09 (for $\hat{\sigma}_{\beta_1}^2$ when $\sigma_{\theta}^2 = 1.2$ and $\sigma_{\beta_1}^2 = 1.2$) to 1.22 (for $\hat{\sigma}_{\beta_1}^2$ when $\sigma_{\theta}^2 = 0$ and $\sigma_{\beta_1}^2 = 1.2$) were found. When the theoretical (and real) variance was zero, the mean of the variance estimates from the individual analyses was still 0.20 or somewhat higher. One may not generalize this value for the general case of homogeneity. A general and absolute criterion for heterogeneity is not available for the estimated variance values.

Considering the 140 data sets from the single-effect design and taking into account only the intercept and the first slope, the highest estimated variance for a parameter with a zero theoretical variance was 0.25, whereas the smallest estimated variance for a parameter with nonzero theoretical variance was 0.37. Considering the 90 data sets from the combined-effect design, the highest estimated variance value for a parameter with zero theoretical variance was 0.28, and the smallest estimated variance value for a parameter with nonzero theoretical variance was 0.36. According to these results, any value between 0.28 and 0.36 as a decision rule of thumb would result in a perfect

Table 1
 Mean Variance of the Individual Estimates for the Combined Effect Design

Intercept Variance (σ_θ^2)	Slope Variance ($\sigma_{\beta_1}^2$)		
	0	0.2	1.2
0			
Intercept (θ_p)	0.21	0.22	0.23
Slope (β_{1p})	0.22	0.45	1.22
Slope (β_2)	0.22	0.22	0.22
Slope (β_3)	0.22	0.22	0.22
0.2			
Intercept (θ_p)	0.48	0.50	0.43
Slope (β_{1p})	0.22	0.44	1.19
Slope (β_2)	0.21	0.21	0.22
Slope (β_3)	0.20	0.23	0.23
1.2			
Intercept (θ_p)	1.21	1.17	1.10
Slope (β_{1p})	0.23	0.47	1.09
Slope (β_2)	0.23	0.23	0.20
Slope (β_3)	0.24	0.23	0.25

decision for these data sets. When the variances of the second and third slopes were considered, the smallest critical value with perfect predictions was 0.33.

Second, from a further analysis, it seems that the relation between real variance and the estimated variance is very strong and linear when there is only one random effect ($R^2 = .87$ for the intercept, and $R^2 = .88$ for the slope). When two random effects were present, the real variance was again linearly related to the estimated variance ($R^2 = .98$ for the slope, $R^2 = .97$ for the intercept). Although these linear relations are of interest, the weights of the prediction function may not be generalized to other kinds of data sets.

Third, from a more detailed inspection of the results, it was concluded that a wrong decision was never made when the variance of the individual estimates was used to decide which theoretical variance is larger (the intercept or the slope). Therefore, the method of individual analyses seems to be successful in deciding on the order of random effects to be included in the model.

In sum, although an absolute general criterion for the individual logistic regression analyses method is unknown, the method can be used for relative decisions on heterogeneity and, hence, for determining the order of random effects to be included in the model.

Marginal Modeling

In Figures 1 and 2, box plots for the estimated association parameters (α) belonging to the random intercept and random slope are shown for the single-effect design.

It is clear that the values of the association parameters are increasing with the theoretical variance of the random effect (indicated on the right-hand side of the figures). There is overlap among estimated association parameters of different theoretical variances, but the real variances do also overlap.

In Figure 3, the estimated association parameters are displayed for the combined-effect design. In each panel of Figure 3, all four association estimates of the corresponding 10 data sets are

Figure 1
Box Plots for the Estimated Association Parameter Values Referring to the
Random Intercept for Different Theoretical Values of the Variance

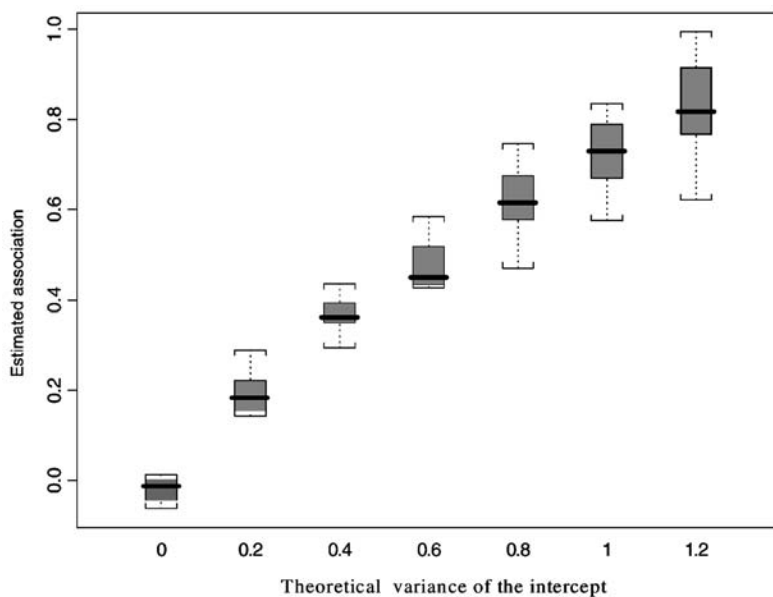


Figure 2
Box Plots for the Estimated Association Parameter Values Referring to the
Random Slope for Different Theoretical Values of the Variance

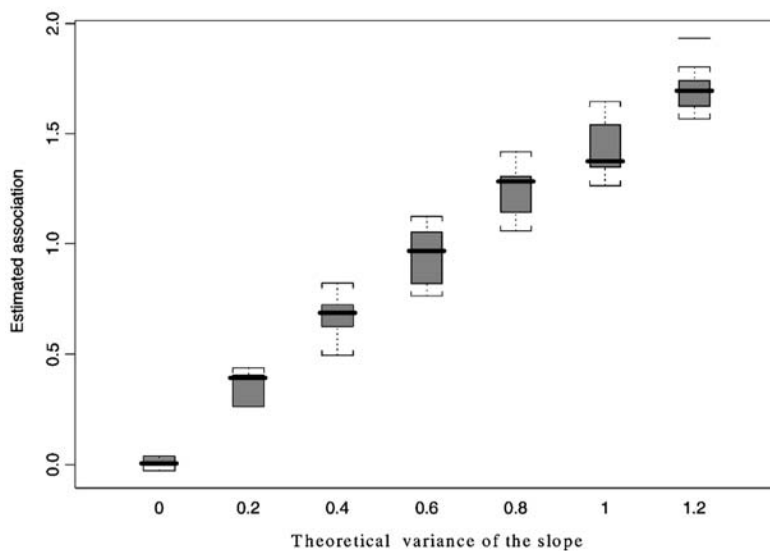
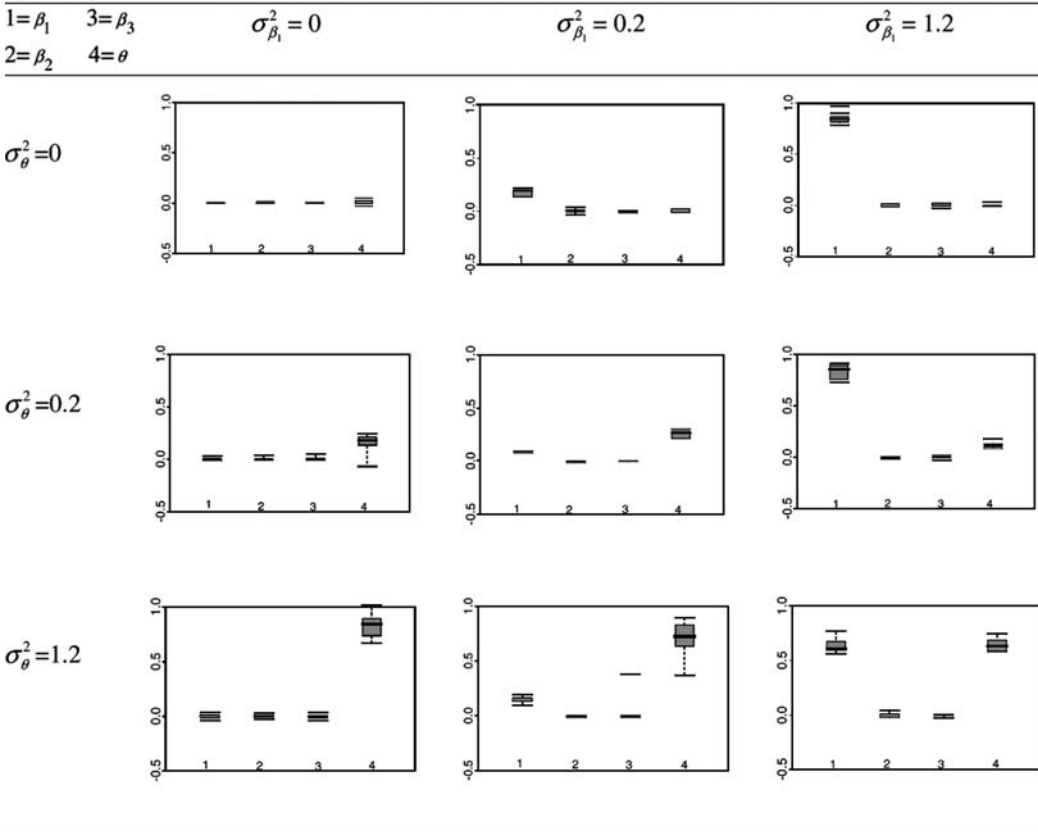


Figure 3
 Box Plots for the Estimated Association Parameters in the Combined-Effect Design



plotted (related to the intercept and to the three slopes; see equation (3)). The first refers to the association parameter of the first (sometimes random) slope, the second and the third refer to the next (and always fixed) slopes, and the fourth refers to the (sometimes random) intercept. As can be seen, only the association parameters belonging to random effects differ substantially from zero. The obtained values are also closely related to the amount of heterogeneity. This is a clear and unambiguous result.

First, when $p \leq .05$ was used as the critical value for an absolute decision criterion, the marginal modeling approach resulted in 23 (5.23%) false alarms (Type I error) and zero misses (Type II error) in the single-effect design and 12 (5%) false alarms and zero misses in the combined effect design. The number of false alarms corresponds to the p value. Note that by counting all variables with zero variance, false alarms are possible for 440 variables in the single-effect design and for 240 variables in the combined-effect design.

Second, it is clear that the estimated association was a function of the real variance. For the single-effect design, the R^2 statistic for the association estimates and the real variance was .95 for the intercept and .97 for the slope. The relation between the association parameters and the real variance was also linear for the combined-effect design. The corresponding R^2 statistic was .93 for the intercept and .97 for the slope.

In sum, marginal modeling with ALR seems to be a very good method to detect heterogeneity. The statistical test provides an absolute decision criterion, and based on the high values of R^2 , a relative decision based on the ordering of the variances also seems to work well. The procedure requires the item covariates to be known, and the heterogeneity can be located by the marginal modeling.

Principal Components Analysis

PCA Eigenvalues

When only one effect is random, only one salient principal component is expected because there is one source of heterogeneity. In a similar way, two salient components are expected when both the intercept and the slope are random. The results confirmed these expectations.

First, the best cutoff value for the eigenvalues was selected in terms of percentage of correct decisions, and 1.9 seemed to be the best cutoff value for these data. When 1.9 was used to decide whether the eigenvalue represents a true source of heterogeneity, 5% false alarms (1 out of 20 data sets) and 0% missers (out of 120 data sets) were obtained for the single-effect design. For the combined-effect design, 1.67% false alarms (1 out of 60) and 8.33% missers (10 out of 120) were obtained. In addition, the elbow criterion (based on a judgment by the first author) also indicated the correct number of dimensions (for 100% of the data sets in the single-effect design but only for 72% in the combined-effect design).

Second, when only one effect was random, a linear relation was obtained between the eigenvalues and the real variances. However, there are overlaps between the eigenvalues of data sets with high but different theoretical variances (above 0.6) due to overlapping real variance values. The real variance was linearly related to the first eigenvalue, $R^2 = .97$ for the intercept and also $R^2 = .97$ for the slope. When both the slope and the intercept were random, the corresponding eigenvalues did not have such a nice interpretation. The higher the variance of one random effect, the larger (but still moderate) the decrease in the eigenvalue of the other random effect.

Although the PCA approach also suffers from the absence of an absolute criterion, because a general reference eigenvalue is not available for all types of data sets, the procedure seems rather effective for relative decisions on heterogeneity. The PCA eigenvalues do not locate the heterogeneity, but an inspection of the PCA loadings may help to locate the heterogeneity, as explained next.

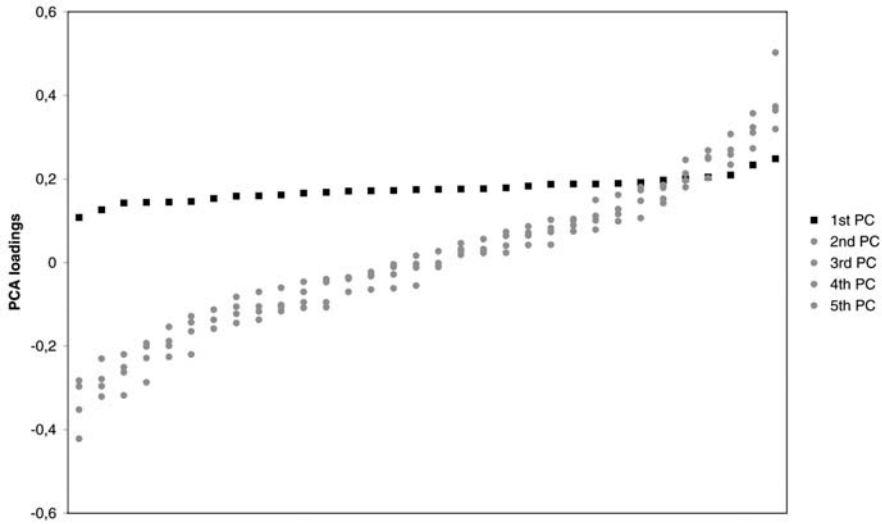
PCA Loadings

The PCA loadings were found to show the hypothesized pattern. Figures 4 and 5 show two representative cases for the single-effect design, one for the random intercept (Figure 4) and another for the random slope (Figure 5). The PCA loadings belonging to the 32 items are ordered in the figures. Each line represents a series of PCA loadings belonging to one principal component. For simplicity's sake, only the PCA loadings belonging to the first five principal components are plotted. For the random intercept, an almost horizontal line with only positive values can easily be noticed (in black). For the random slope, the line in black shows the hypothesized pattern, with a jump from negative to positive values (because of the opposite-signs coding).

In the case of the combined-effect design, the same effects were observed as earlier. Figure 6 shows a representative case as an illustration. One line is horizontal (for the intercept component), and the other one shows a jump (for the slope component). When the theoretical variances of the intercept and slope were equal, which of the two random effects showed in the first component depended in fact on the real variances.

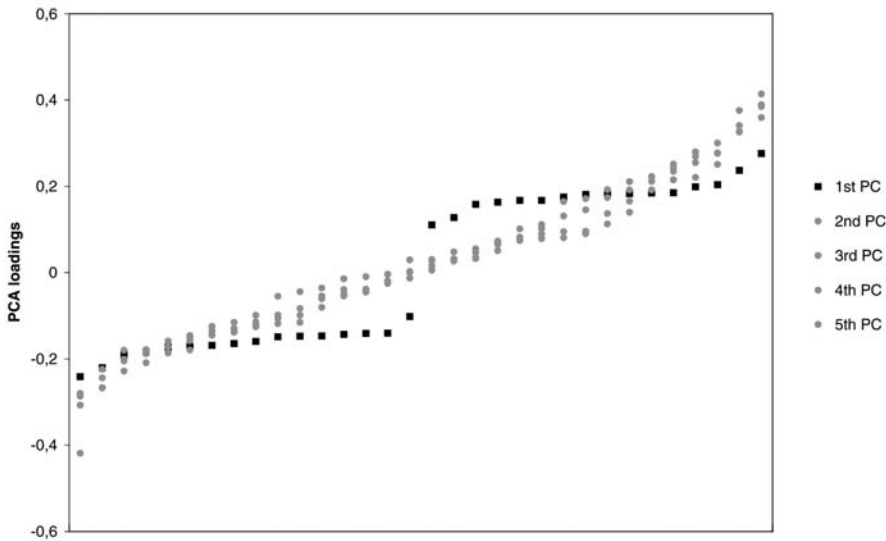
In an additional parallel simulation study with unipolar coding instead of bipolar coding for the item covariate with random effects, the results were similar. The only difference was that the PCA

Figure 4
Principal Components Analysis (PCA) Loadings for the First Five Principal Components (PCs),
Ordered as a Function of the Size of the Loadings



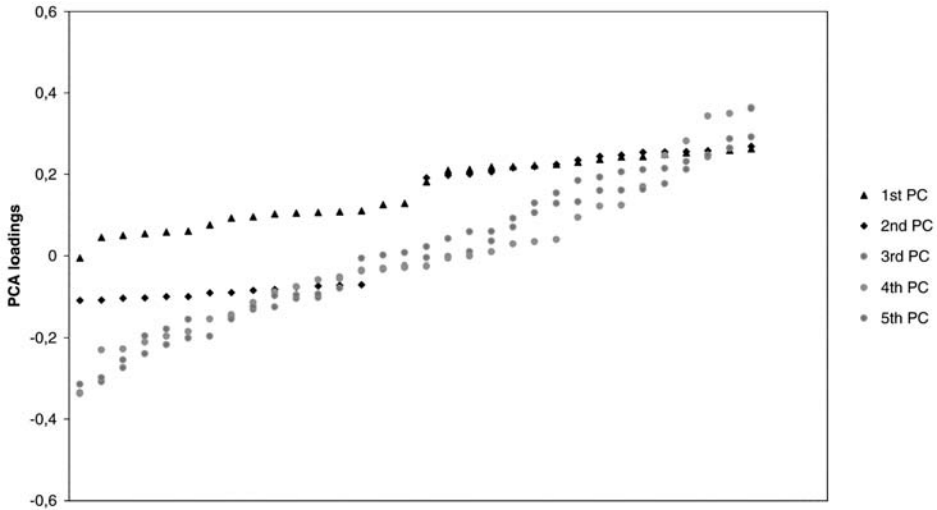
Note. The intercept variance is 0.6; all slope variances are zero.

Figure 5
Principal Components Analysis (PCA) Loadings for the First Five Principal Components (PCs),
Ordered as a Function of the Size of the Loadings



Note. The slope variance is 0.6; the other variances are zero.

Figure 6
 Principal Components Analysis (PCA) Loadings of the First Five Principal Components (PCs)



Note. The intercept variance and the slope variance are both 1.2; the other variances are zero.

loadings referring to the random slope were mostly positive, and the jump of the ordered loadings was more moderate than in the Figures 5 and 6.

These results show that one may derive the source of the variance from the pattern of the loadings. When the variance of the slope is concerned, one needs of course advance knowledge of the item covariates to interpret the pattern of the loadings in terms of slope variance. Although the results suggest that PCA is an easy and good method to detect and locate heterogeneity for the considered category of problems, the success of this approach is limited because the PCA of binary data is subject to artifacts when extreme means of items occur.

DIMTEST

Because DIMTEST concentrates on extra heterogeneity beyond a general underlying trait, and because this kind of extra heterogeneity occurs in this study only in the combined-effect design, that part of the design was used for investigating the two versions of the DIMTEST procedure. For this part of the simulation study, 100 data sets were generated in each cell because the results were not as clear-cut as for the previous methods. The sorting of the items into two of the three subsets required for DIMTEST was always made by the automatic item selection option of the DIMTEST software. The desired significance level of the DIMTEST statistic was set to $\alpha = .05$.

Table 2 contains the number of data sets indicated to be multidimensional (out of 100) when DIMTEST 1 was applied (Stout et al., 1993). Because the cells in the first column and first row of Table 2 are unidimensional, ideally, one would expect 5 out of 100 data sets to be indicated as multidimensional, with much higher frequencies than 5 in the other four cells. In fact, the frequencies are slightly higher in the first column and much higher for the rest of the unidimensional cells than expected. When one or both of theoretical variances are 0.2 and the variance of the other random parameter is 1.2, the frequency of the data sets indicated as multidimensional is very low, meaning that the detection of multidimensionality is rather poor. Finally, when both theoretical variances

Table 2
 Number of Data Sets Indicated as Multidimensional by DIMTEST 1

Intercept Variance (σ_0^2)	Slope Variance ($\sigma_{\beta_1}^2$)		
	0	0.2	1.2
0	6	12	21
0.2	10	18	21
1.2	8	20	82

Table 3
 Number of Data Sets Indicated as Multidimensional by DIMTEST 2

Intercept Variance (σ_0^2)	Slope Variance ($\sigma_{\beta_1}^2$)		
	0	0.2	1.2
0	2	33	62
0.2	6	58	100
1.2	2	39	99

are 1.2, still only 82 out of the 100 data sets were identified as multidimensional, meaning that 18% of cases of strong heterogeneity went undetected. DIMTEST 1 resulted in 64.75% missers (259 out of 400 data sets) and 11.4% false alarms (57 out of 500 data sets).

Note that from the point of view of DIMTEST, a bipolar coding with random weights of an item covariate also leads to multidimensionality. Taking this into account, one may expect multidimensionality in the first row, except for the first cell. However, as can be seen in Table 2, the detection of multidimensionality based on the bipolar covariate is quite poor, and the global results improve only slightly. Considering the DIMTEST perspective on bipolarity, 51.3% missers (308 out of 600) and 8% false alarms (24 out of 300) were obtained.

When DIMTEST 2 (William Stout Institute for Measurement, 2005) was applied, the detection rate of multidimensionality improved a lot, as can be seen in Table 3, but the power of the procedure is still not ideal for data with a small sample size and with a structure as in this study, except for a slope variance of 1.2 combined with an intercept variance of 0.2 or higher. This is not a bad result because these are the common cases that researchers would be interested in to detect a general underlying dimension (random intercept) combined with a substantial other source of heterogeneity.

These results are not unexpected. As mentioned earlier, DIMTEST 1 underperforms for small samples (van Abswoude et al., 2004). Furthermore, equal numbers of items loading on the different dimensions lead to less stable DIMTEST 1 results than unequal numbers of items (van Abswoude et al., 2004). However, DIMTEST 2 performs much better, unless the additional source of heterogeneity is not really substantial. In sum, one should be aware that, when the sample size is small or when the dimensional variance is small, DIMTEST may overlook small variances.

DETECT

As for DIMTEST, the combined-effect design is also the focus for DETECT because it is a method to detect extra heterogeneity. For DETECT, only 10 data sets per cell were used because

Table 4
 Mean DETECT and R Values With Cross-Validation for an Analysis With Two Dimensions

Intercept Variance (σ_θ^2)	Slope Variance ($\sigma_{\beta_1}^2$)					
	0		0.2		1.2	
	Maximum Value	Reference Value	Maximum Value	Reference Value	Maximum Value	Reference Value
0						
Detect	0.583	0.036	0.857	0.470	4.117	3.821
R	0.357	0.000	0.476	0.250	0.986	0.981
0.2						
Detect	0.628	0.008	0.858	0.353	3.514	3.864
R	0.389	0.006	0.465	0.196	0.906	0.902
1.2						
Detect	0.595	-0.002	0.875	0.278	2.907	3.430
R	0.381	0.014	0.458	0.144	0.906	0.919

the variance of the test statistics was small. The DETECT procedure can be applied in different ways, which will be reflected in this study.

First, the DETECT analyses were limited to *two latent dimensions*, but later a larger number of dimensions were allowed, although the true dimensionality was never larger than two. DETECT was applied *with cross-validation* first because using the cross-validation option is strongly recommended (Zhang & Stout, 1999). The examinees of each data set were randomly assigned to two subsets (with equal size). The results are shown in Table 4. Both the average of the maximum DETECT values, the reference DETECT values, and the associated *R* values for each subset are given.

According to the results, the DETECT statistics are not as sensitive to the intercept variance as they are to the slope variance. The slope variance is clearly linearly related to the maximum DETECT value and also to the reference DETECT value ($R^2 = .94$ and $R^2 = .95$, respectively), but the intercept variance is not ($R^2 = .02$ and $R^2 = .004$, respectively). This is because the intercept is not a source of item clusters, whereas the slope certainly is because of the bipolar coding of the corresponding item covariate. Based on the bipolar coding, multidimensionality should be detected when the slope variance is larger than zero. When the slope variance is 0 or 0.2, the reference DETECT values are much smaller than the maximum DETECT values. This finding shows the effect of cross-validation.

According to the DETECT manual, when the DETECT procedure indicates multidimensionality, multidimensionality can be concluded only when the data have at least an approximate simple structure ($R(P) \geq 0.8$). Consequently, strictly speaking, only 28 data sets may be considered as multidimensional, all with a theoretical slope variance of 1.2. Because the DETECT values turn out to be a linear function of the slope variance, which is the source of multidimensionality in these data sets, it makes sense to consider all DETECT values, as well as those with an $R(P)$ lower than the critical value.

In the following, different decision rules are compared concerning the inferences to be made regarding unidimensionality and extra heterogeneity. According to the DETECT manual, the critical values for indicating unidimensionality and strong multidimensionality are < 0.1 and ≥ 1 , respectively. Applying these critical values, 6.67% false alarms (2 out of 30 data sets) and 50% missers (30 out of 60 data sets) were obtained. Using a critical value of 0.5 for multidimensionality, still 33.33% missers (20 out of 60) were found). When a cutoff value of 0.1 was used, only 10% false alarms (3 out of 30) and 8.33% missers (5 out of 60) were found. Based on these results,

Table 5
 Number of Clusters Found by DETECT with Cross-Validation When 12 Dimensions Were Allowed

Intercept Variance (σ_θ^2)	Slope Variance ($\sigma_{\beta_1}^2$)														
	0					.2					1.2				
	Number of Clusters														
	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
0				5	4	1			7	3					10
0.2				7	3		2	3	4	1					10
1.2		1	2	6	1		1	4	5						10

a cutoff value of 0.1 seems to be successful for deciding on extra heterogeneity. Note that these results are obtained while considering a bipolar unidimensional structure as multidimensional (in the sense of DETECT). When such a structure is considered as unidimensional, the number of false alarms is of course higher.

When the discrepancy measure was applied, it was calculated as the difference of the two DETECT values divided by the absolute value of the reference DETECT value because the reference DETECT value was often negative in this study. The application of the combined criteria resulted in zero false alarms and 35% missers (21 out of 60 multidimensional data sets).

When DETECT was used *without cross-validation* and when *two dimensions*, for slope variances of 0 and 0.2, were allowed, the maximal DETECT values were much higher than the reference DETECT values obtained in the cross-validation procedure. For a slope variance of 1.2, the DETECT values are similar to the ones obtained earlier. For the analyses without cross-validation, the optimal critical DETECT value turned out to be 0.5, yielding 3.33% missers (2 out of 60) and zero false alarms. With the critical value of 0.1, as recommended in the manual, a remarkable amount of unidimensional cases was overlooked. In line with the results of this study, van Abswoude et al. (2004) suggested that the upper bound of unidimensionality might be too low. With a higher critical value, the procedure without cross-validation seems to work well for the kind of data used in this study.

Because in practice, one may not have an idea about the number of latent dimensions, this study investigated how the method works when more than two dimensions are assumed. For such an analysis, the highest possible number of dimensions in the DETECT program, *12 dimensions*, was allowed. First, *cross-validation* was applied. The results concerning the number of clusters are reported in Table 5. The *R* values indicated simple structure only when the theoretical slope variance was 1.2. For that case, the correct partition with two clusters based on the bipolar covariate was always found. When the theoretical slope variance was 0.2, the highest maximum DETECT value for the data was obtained for two to five clusters. When the theoretical slope variance was 0, the number of clusters was between two and five, and for one data set (with zero intercept variance), even six clusters were found.

However, it is mentioned in the DETECT manual that a considerably higher maximum DETECT value than the reference DETECT value indicates unidimensionality. The combined decision rule (discrepancy larger than 0.5 or reference DETECT smaller than or equal to 0.1) resulted in zero false alarms and 33.33% missers (20 out of 60).

A new decision algorithm was developed as follows:

1. Choose the highest maximum number of dimensions (12) in the DETECT program and run the DETECT procedure.
2. When the dimensionality indicated by DETECT is 2, $k = 2$, and the reference DETECT value for $k = 2$ is higher than 0.1, the true dimensionality is 2; if it is smaller or equal, the test is unidimensional. When the indicated dimension is higher than 2, $k > 2$, go to Step 3.
3. Calculate the discrepancy measure for dimensionality k . If it is smaller than or equal to its critical value, the true dimensionality is k . If the discrepancy is higher than the critical value, choose $k = k - 1$ as maximal dimensionality and return to Step 2.

When 0.5 was chosen as a critical value in Step 3, 8.33% missers (5 out of 60) were found, and the dimensionality was overestimated for 40% of the data sets (12 out of 30). With 0.3 as the critical value, again 8.33% missers (5 out of 60) were found, and the dimensionality was still overestimated for 13.33% data sets (4 out of 30).

When the DETECT procedure was used *without cross-validation* and 12 dimensions were allowed, the true number of clusters was always found when the slope variance was 1.2. For a slope variance of 0.2, two to six clusters were found, and for a slope variance of 0, three to seven clusters were indicated by DETECT. When a cutoff value of 1.1 was used for deciding on unidimensionality and multidimensionality, perfect decisions were obtained. This cutoff value is much higher than the optimal critical value for DETECT without cross-validation and allowing for only two dimensions because many dimensions were obtained when DETECT had maximal freedom in choosing the number of dimensions. It seems that when higher critical values are used than those provided by the DETECT manual, the procedure also works well without cross-validation and perhaps even better. The problem is that the proper critical values are not known a priori.

There are some remaining problems. The estimated DETECT value is based on conditional covariances calculated for each item pair and for each total score group and the rest score group. The minimum number of examinees for each cell is defined by the user in the input of the DETECT procedure. Only those total score groups are considered that contain at least as many examinees as the reference value defined in the input. The recommended value is 20, which should be lowered if fewer than 85% of the examinees are used for the covariance calculations. In the present study, the minimum number of examinees per cell had to be decreased for each data set. This may be the consequence of the small ratio of examinees versus items (200 to 32), although it is a common ratio in psychological research. To reach the recommended percentage of observations used in the covariance calculation, one should have at least $20 \times I$ observations, where I is the number of items. In psychological research, this condition is often not fulfilled.

In this study, the R values were a function of the variance and often did not reach the criterion value for multidimensionality. In general, when all DETECT values are interpreted, it is not easy to find cutoff values. Apart from these problems, DETECT turned out to be a reasonably good method to detect extra heterogeneity and also to explore its location.

In sum, DETECT performs quite well as an indicator of extra heterogeneity when a higher critical value is chosen than is suggested in the manual, but it does not link the heterogeneity in a direct way to item covariates.

Application of the Methods for Real Data

The deductive reasoning data from Rijmen and De Boeck (2002) are used here for an application. It was found in their study that some of the item covariates were sources of heterogeneity. This kind of structure is described in Rijmen and De Boeck as an RWLLTM.

The data concern 30 items on deductive reasoning, and 214 high school students participated in the study. For solving each item, two inferences were needed. Both inferences are based on two premises; thus, 2×2 premises were presented in each item. Six of the items are filler items and were not considered further in the analyses. The design for the remaining 24 items was as follows.

1. For the first inference, one premise has an “if x then y ” structure. The other premise defines the type of inference either as a *modus ponens* (MP) or a *modus tollens* (MT). For an MP, the second premise is “ x ,” so that the combination of “if x then y ” and “ x ” yields the inference “ y .” For an MT, the second premise is “ y ”; hence, the correct inference is “not x .”
2. Also, the second inference is based on two premises. The inference is one of the following kinds:
 - A *disjunctive syllogism* with “ p or q ” and “not p .” The resulting inference is “ q .”
 - A *disjunctive modus ponens* with “if p or q then r ” and “ q .” The resulting inference is “ r .”
 - A *modus ponens* with “if p then q ” and “ p .” The resulting inference is “ q .”

When the disjunctive syllogism or the disjunctive modus ponens is the second premise, the first inference is a premise for this second inference. For example, the premise of a modus ponens type of first inference combined with the premise of a disjunctive syllogism reads as “if x then y ”; “ x ”; “ y or q .” The correct first inference is “ y ,” and therefore “not q ” follows. The modus ponens as a second inference is combined with either another modus ponens or with a modus tollens as a first premise, so that the overall inference is a conjunction—for example, “if x then y ”; “ x ”; “if p then q ”; “not q ”; therefore, “ y and not p .”

3. The correct answer for the item is either the acceptance or the rejection of the presented inference.
4. The content of the item is related to either people or objects.
5. Within each content, the premises are grouped for half of the items (first inference followed by the second inference) and mixed for the second half of the items (the first and second premises of the first inference were separated from each other with a premise of the second inference). The design was not fully crossed, but Factors 4 and 5 were confounded deliberately to reduce the number of items.

Here is an item of the test (the language of the test was Dutch): “Alex is not in Ecuador” (“not y ”), “Jonas is not in Switzerland or Hedwig is in Algeria” (“not p or q ,” where $q = x$), and “If Hedwig is in Algeria, Alex is in Ecuador” (“if x then y ”). The inference presented to the respondents is “Jonas is in Switzerland” (“ p ”). Because “Alex is not in Ecuador” and “if Hedwig is in Algeria, Alex is in Ecuador,” Hedwig is not in Algeria. Furthermore, “Jonas is not in Switzerland or Hedwig is in Algeria”; hence, Jonas is not in Switzerland. Consequently, the offered inference should be rejected (the correct inference is “not p ”).

This item contains a modus tollens inference regarding Alex (“if x then y ”; “not y ”; therefore “not x ”). The second inference (about Jonas and Hedwig) is a disjunctive syllogism (“not p or x ” and “not x ” as the first inference and therefore “not p ”). The correct answer is the rejection of the inference (“ p ”). The content concerns people, and the premises are ungrouped because the first and the third premises (“Alex is not in Ecuador” and “If Hedwig is in Algeria, Alex is in Ecuador”) define a modus tollens, but they are separated by a premise of the second inference.

There were four response alternatives: true, not true, undecidable, and don’t know. There were no undecidable items among the 24 items. The correct answer was coded as 1, and the other answers were coded as 0.

For the analyses, six covariates were used based on five factors. All factors except for the second factor were coded as a binary covariate. For the second factor, two binary covariates were needed because this factor has three levels (disjunctive syllogism, disjunctive modus ponens, modus ponens with conjunction). Rijmen and De Boeck (2002) found that all covariates, except for the combination of the confounded item content and grouping (Factors 4 and 5), were significantly related to the probability of success on the items. Therefore, and because of the confounding, the covariates regarding the content and the grouping of the items were not used in the analyses. Several models were fitted to the data with zero to four random slopes (Rijmen & De Boeck, 2002). Each covariate was either fixed or random, defining 2^4 different combinations, resulting in 16 models. Based on the Bayesian information criterion (BIC) statistics, a model with a random intercept and a random weight for the disjunctive syllogism covariate was selected. However, based on the Akaike information criterion (AIC) statistic, a model with a random intercept and random weights for all covariates except for acceptance versus rejection seems to be the best model. The AIC results imply four sources of heterogeneity in the data: the overall level (intercept), the first type of inference (Covariate 1), and the second type of inference (Covariates 2 and 3) but not acceptance versus rejection (Covariate 4). The results of this full modeling approach will now be compared to the results obtained from the studied methods.

When individual logistic regression models were fitted to the data, the analyses resulted in complete separations for 6 examinees and quasi-complete separations for 114 examinees. Only the analyses for 94 individuals (43.9% of the data) could be considered, which makes the results doubtful. One strategy would be to include random effects in the order of the variance of the individual estimates—namely, first for Covariate 2 ($\sigma^2 = 1.987$), then for the intercept ($\sigma^2 = 1.502$), then for Covariate 1 ($\sigma^2 = 1.478$) and Covariate 4 ($\sigma^2 = 1.131$), and finally for Covariate 3 ($\sigma^2 = 1.068$). This would be a different result than obtained by Rijmen and De Boeck (2002).

The marginal modeling approach was applied on the data with $+1/-1$ coding for the covariates. The association parameters for the intercept and the first two covariates were significantly different from zero, and the third covariate came close to being significant ($p < .001$, $p = .004$, $p < .001$, and $p = .053$, respectively). This result is in line with the results obtained by Rijmen and De Boeck (2002).

The PCA also indicated four random factors. The graph of the loadings showed a general factor, and based on the elbow criterion applied on the graph of the eigenvalues, there are three more specific factors. As far as the number of sources of heterogeneity is concerned, this result is in agreement with the conclusions of the marginal modeling approach. However, the PCA solution is not easy to interpret in terms of the covariates.

Using both versions of DIMTEST (DIMTEST 1 and DIMTEST 2), the analyses indicated that the data are multidimensional ($p = .020$ and $p = 0.049$, respectively). In the DETECT procedure *without cross-validation*, the R index value was only 0.604. If the simple structure index is ignored, the procedure indicates three clusters, using DETECT $> .5$ as a criterion (DETECT value

is 0.627), as suggested by the simulation study. Unfortunately, the partitioning of the items is not clearly interpretable in terms of the covariates.

The DETECT results based on cross-validation could not be interpreted either when the R index is taken into account ($R = .496$). When the R index was not considered, three criteria could be applied that were investigated in the simulation study (the reference DETECT, the discrepancy measure, and the decision algorithm proposed on the basis of the simulation study). All these decision rules confirmed three clusters: the reference DETECT value was .33, which is higher than .1; the discrepancy measure was .66, which is higher than .5; and the same conclusion can be drawn from the decision algorithm. However, the interpretation of the resulting partition in terms of covariates was again not easy.

In sum, the marginal modeling approach yielded results that are very close to those obtained by Rijmen and De Boeck (2002). DIMTEST indicated multidimensionality, but DIMTEST does not provide the number of heterogeneity sources. The PCA indicated the same number of heterogeneity sources as the AIC in the study by Rijmen and De Boeck but did not locate them clearly. The other methods each show some differences compared to the results from Rijmen and De Boeck, with respect to either the order of the importance of the effects (such as the individual regression analyses) or the number of heterogeneity sources (DETECT). In sum, although both versions of the DIMTEST procedure and the PCA provided information on the data structure in line with the full modeling approach, the marginal modeling approach seems to be the best approach. Marginal modeling not only locates the heterogeneity but also provides a significance test.

Discussion and Conclusion

Various methods were investigated for detecting heterogeneity in moderately small data sets with binary repeated measures and with item covariates. Both a simulation study and an application were presented, where data with a design were analyzed. This is perhaps not a very common problem in educational measurement because in that context, the data sets tend to be much larger than $n = 200$, and item covariates are not so common. But it is a rather common structure for a within-subjects psychological experiment or for a psychological test with a design (e.g., a test with subscales). Furthermore, in psychological measurement, the assumption of design factors with effects that differ depending on the person makes sense as a structure with person-by-item interaction.

As mentioned earlier, there are important differences between the investigated methods from a practical point of view. Among the methods that require the availability of item covariates, marginal modeling gave excellent results, both in the simulation study and in the application. Marginal modeling provides a statistical test for the association parameters and also locates heterogeneity. Also, individual analyses seem to be quite sensitive to the size of the heterogeneity. However, this method can be used only for a relative decision, for deciding on the order of the random effects that should be included in the model. The addition of a bootstrap procedure could help to find an appropriate cutoff value for different data structures.

It is difficult to differentiate among the methods that do not require item covariates. PCA seemed to be an effective method in the simulation study, but PCA has the drawback that it is vulnerable to artifacts. DIMTEST 2 seems to be more effective than DIMTEST 1, but it still seems less sensitive than PCA and DETECT because it tends to overlook small variances. DETECT would be a preferable method, in principle, because it does not require advance knowledge of the item covariates and still can locate heterogeneity in an indirect way. However, the decision criterion for DETECT is not evident, and its performance in the application was not in line with a full modeling approach.

References

- Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541-562.
- Blumer, H. (1969). *Symbolic interactionism: Perspective and method*. Englewood Cliffs, NJ: Prentice Hall.
- Carey, V. J., Zeger, S. L., & Diggle, P. J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, *80*, 517-526.
- Chen, W.-H. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265-289.
- Collett, D. (1991). *Modelling binary data*. London: Chapman & Hall.
- DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-389). Mahwah, NJ: Lawrence Erlbaum.
- Douglas, J., Kim, H.-R., Roussos, L., Stout, W. F., & Zhang, J. (1999). *LSAT dimensionality analysis for December 1991, June 1992, and October 1992 administrations* (Law School Admission Council Statistical Report 95-05). Newton, PA: LSAT.
- Egan, D. E. (1979). Testing based on understanding: Implications from studies of spatial ability. *Intelligence*, *3*, 1-15.
- Embretson, S. E. (1985). *Test design: Developments in psychology and psychometrics*. London: Academic Press.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374.
- Fitzmaurice, G. M., Heath, A. F., & Cox, D. R. (1997). Detecting overdispersion in large scale surveys: Application to a study of education and social class in Britain. *Applied Statistics*, *46*, 415-432.
- Froelich, A. G., & Habing B. (2003). Conditional covariance based subtest selection for DIMTEST. In William Stout Institute for Measurement, *DIMTEST Version 2.0* [Software manual]. Urbana-Champaign, IL: William Stout Institute for Measurement.
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*, 26-34.
- Hardin, J. W., & Hilbe, J. M. (2003). *Generalized estimating equations*. London: Chapman & Hall.
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139-164.
- Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data. (Doctoral dissertation, University Illinois at Urbana-Champaign). *Dissertation Abstracts International*, *55-12B*, 5598.
- Landwehr, J. M., Pregibon, D., & Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, *79*, 61-71.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman & Hall.
- Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*, 41-68.
- Pervin, L. A. (1977). The representative design of person-situation research. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, NJ: Lawrence Erlbaum.
- Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, *26*, 271-285.
- Roussos, L., Stout, W. F., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, *35*, 1-30.
- SAS Institute. (1999). *SAS online doc* (Version 8) [Software manual on CD-ROM]. Cary, NC: Author.
- Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, *66*, 79-97.
- Sternberg, R. J. (1977a). Component processes in analogical reasoning. *Psychological Review*, *34*, 356-378.
- Sternberg, R. J. (1977b). *Intelligence, information processing, and analogical reasoning*. Hillsdale, NJ: Lawrence Erlbaum.
- Sternberg, R. J. (1979). The nature of mental abilities. *American Psychologist*, *34*, 214-230.
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617.
- Stout, W. F., Douglas, J., Junker, B., & Roussos, L. (1993). *DIMTEST user's manual*. Urbana-Champaign: University of Illinois at Urbana-Champaign, Department of Statistics.

- Stout, W. F., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory*. New York: Springer-Verlag.
- Stout, W. F., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331-354.
- Stout, W. F., & Zhang, J. (1999). The theoretical index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213-249.
- Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Mahwah, NJ: Lawrence Erlbaum.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7*, 215-231.
- van Abswoude, A. A. H., van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*, 3-24.
- Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.
- Webb, M. C., Wilson, J. R., & Chong, J. (2004). An analysis of quasi-complete binary data with logistic models: Applications to alcohol abuse data. *Journal of Data Science, 2*, 273-285.
- Whitely, S. E. (1978). Information-processing on intelligence test items: Some response components. *Applied Psychological Measurement, 1*, 465-476.
- Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics, 31*, 144-148.
- William Stout Institute for Measurement. (2005). *DIMTEST Version 2.0* [Software manual]. Urbana-Champaign, IL: Author.
- Yen, W. M. (1984). Effect of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.
- Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213-249.

Acknowledgments

The authors acknowledge the financial support from K. U. Leuven, the OFOE grant to Katalin Balázs, and the IAP/5 network grant to Paul De Boeck. They thank Frank Rijmen for his permission to analyze the deductive reasoning data and Kristof Meers for his help in DOS programming.

Author's Address

Address correspondence to Katalin Balázs, Department of Psychology, K. U. Leuven, Tiensestraat 102, B-3000 Leuven, Belgium; e-mail: Katalin.Balazs@psy.kuleuven.be.