Running head: DETECTING HETEROGENEITY IN LOGISTIC REGRESSION MODELS

# Detecting Heterogeneity in Logistic Regression Models

Katalin Balázs, István Hidegkuti and Paul De Boeck

K.U. Leuven, Belgium

Abstract

It is not uncommon that person-by-item data in the context of Item Response Theory are correlated beyond the correlation that is captured by the model, or in other words stated, that 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, since investigating all different combinations of random parameters is not very practical or 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 in order to investigate and to compare various methods: individual analyses (one per person), marginal modeling, principal component analysis of the raw data, DIMTEST and DETECT.

Test data are often of a binary type, and may be considered as repeated measures, since different items are presented to the same persons. The focus of this paper is on binary repeated measures with a design. The availability of a design is not very common, but it is interesting, because a design is a potential basis to explain the data. Many tests do not have 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 items are characterized by corresponding item features, and these can be used as covariates in a mixed logistic regression model for the data. The resulting item response model, with explanatory item covariates for the binary data, is a logistic regression model, as in Equation 1:

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

The model assumes binary response variables, which are nonlinearly related to the covariates. $Y_{pi}$ is the response of person $p$ ($p$ = 1,…, $P$) to item $i$ ($i$ = 1,…, $I$), and follows a binomial distribution with $n_{pi}$ = 1, and parameter $P(Y_{pi}=1|\theta_p,\boldsymbol{\beta}_p)$ which 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,…, $K$) changing its value over items, and the $\beta_{kp}$ is the associated random weight. $\theta_p$ is the random intercept that is the so-called ability of the person in the context of achievement tests. When the intercept is the only random effect, meaning that the weights of the covariates are fixed, the resulting model is the Linear Logistic Test Model (LLTM; Fischer, 1973). 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 resulted model is a logistic mixed model. The term

*mixed* refers to the combination of fixed and random effects. Note, that also the intercept can bee seen as a weight of a covariate that is an overall 1-vector. The LLTM has been commonly used without assumming heterogeneity in the weights of the covariates, but with heterogeneity being restricted to the intercept.

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, which means that the model does not specify all sources of variance in the data. Ignored variance causes overdispersion. In the context of logistic regression models, the notion for a "too large variance" is overdispersion (Collett, 1991). Underdispersion can also occur, but that is a rare phenomenon. In principle, heterogeneity may stem from the persons or from the items. In this study, the focus is on person-based heterogeneity, which can occur either in the intercept or in the weights of the covariates, more precisely when the intercept or the weights are random effects (also called random coefficients).

In general, extra heterogeneity implies local item dependency. In this study dependencies are investigated that stem from random effects in a logistic regression model, but in practice, other sources of dependency may also occur. Item response dependencies can be studied through the correlations of the residuals of the applied 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 also developed (IRT LD) for the detection of local dependencies (Chen & Thissen, 1997), and graphical techniques were proposed for detecting residual dependencies (e.g., Landwehr, Pregibon & Shoemaker, 1984). These are valuable alternatives to the approach for detecting heterogeneity that is followed here. These approaches are relevant to dependencies in general and are therefore less specific than the aim of the present study.

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 from a psychological point of view, it is not uncommon that factors such as 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 such interactions is called interactionism (see, Blumer, 1969; Pervin, 1977). In the domain of intelligence, the study of cognitive processes, as initiated by Sternberg (1977) and Embretson (1985), is based on item covariates indicating how much of a certain process is required to succeed in the item. The random 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 different formalisation is chosen than 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 (Beguin & 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 one hundred 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, one can investigate whether there are

random effects and where they are, so that in a next step one can estimate a more directed model.

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 the heterogeneity of the item covariate weight does, these methods are possible candidates for a diagnostic approach to heterogeneity. However, they are less directed than it is possible when applied to data with a design, because no use is made of the item covariates. Although PCA is not really appropriate for binary data, it can also detect dimensional variance, but neither this method makes use of item covariates. Nevertheless, all three undirected methods, DIMTEST, DETECT and PCA will be investigated on their performance for data with a design.

As directed methods, two will be 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 will be used for each individual separately, and the variance of the weights will be checked. Second, a marginal modeling will be 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 both may be useful for the detection of heterogeneity.

The methods that will be 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 or not they indicate the localisation of the heterogeneity in terms of the item covariates. Some methods detect (extra)heterogeneity, such as DIMTEST, the DETECT statistic, and the (size of) eigenvalues of a PCA. Other methods can also give an indication of where the heterogeneity is located, such as DETECT clusters and the PCA loadings. Finally, the individual analyses and marginal modeling are an explicit way to the location of (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. Some methods provide a test statistic to make an absolute decision about the occurrence or absence of heterogeneity or extra heterogeneity. In some cases, an associated significance test is available, such as for DIMTEST and marginal modeling, and 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, so that the corresponding eigenvalues may give an indication of the 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 be still used for a relative decision, because they indicate for which covariates the weights are more likely to be heterogeneous than for others. Random parameters can be included in the model in the order

that is suggested by the diagnostic analyses, until the fit statistic of the random effects model does not improve any more.

## Overview of the Methods

### *Individual Analyses*

As explained, the heterogeneity that will be studied implies individual differences in the intercept or / and in the slope(s). A very simple logistic regression approach would be to do 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 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. 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 from persons for whom the logistic regression analysis results in complete or quasi-complete separation, have to be omitted, but this omission is not without consequences for the variance of the estimates.

In general, because the method uses the information of the item covariates, it is limited to cases when there is information on the item covariates a priori, but this is not a problem for this study. Based on the variance of the regression weights this method provides a direct indication of where the heterogeneity is located, and based on the ordering of these variables, it can used for a relative decision about which random effects should be included 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 latent traits. Since 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 in more particular the GEE2 variant (Hardin and Hilbe, 2003), beside estimation for the mean, also an estimation for the association structure is obtained. For binary data, it is appropriate to use odds ratios instead of correlations as the measure of associations:

$$OR(Y_{pi}Y_{pi'}) = \frac{P(Y_{pi}=1,Y_{pi'}=1)P(Y_{pi}=0,Y_{pi'}=0)}{P(Y_{pi}=1,Y_{pi'}=0)P(Y_{pi}=0,Y_{pi'}=1)}\ , \qquad (3)$$

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 covariates have on odds ratios (OR):

$$\log(\ OR\ (Y_{pi}Y_{pi'}))\ =\ \Sigma\,\alpha_k\, x_{ki}\, x_{ki'}\,, \qquad (4)$$

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 $\alpha_k$ is the association parameter belonging to the *k*-th item covariate. In

other words, $\alpha_k$ is a weight that indicates how much item covariate *k* contributes to the log odds ratio. Heterogeneity based on covariate k is shown, in $\alpha_k$ being larger than zero. Negative values of $\alpha_k$ are possible, but often not really meaningful, because it implies that positive products $x_{ki}\, x_{ki'}$ yield negative associations. In this study a -1/+1 coding was used for the covariates in the ALR (in Equation 4). 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 Inc., 1999) was applied for ALR analyses. When the estimation did not converge (in about 65 % of the cases), dummy coding was used (with success in all cases) and the resulting estimates were transformed to obtain in an indirect way the corresponding weights for a +1/-1 coding. In general the interpretation of $\alpha_k$ depends on how the covariates are coded, and the coding is also important for the kind of correlation that can be modelled. For example, a +1/-1 coding is appropriate for a covariate that induced bipolarity (positive *and* negative correlations), but a dummy coding is not appropriate for a direct modelling of bipolarity, and as a consequence, a transformation is required.

The method of marginal modeling provides the localisation of the heterogeneity in terms of the covariates through the α-parameters (Equation 4). It will be derived from the α -estimates and their statistical significance whether there is heterogeneity and where it is. In principle, it would be possible to elaborate a system of pairwise likelihood ratio tests to find out whether there is extra heterogeneity in comparison to a number of reference models, but the approach that will be followed here is simpler and is only based on the α-estimates for all item covariates. The asset of the marginal modeling approach is that it can localize the heterogeneity, because there is advance knowledge of the item covariates, but the requirement to have this advanced knowledge is a limitation of the method for general use.

### *PCA for the Raw Data*

Although Principal Component Analysis 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, 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 (Hattie, 1985), based on the idea that the larger is the variance which is explained by the first principal component, the better the assumption of unidimensionality. It is well-known that PCA for binary data may lead to artifacts especially when the proportions of response values are extreme, but we will nevertheless explore how it 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 of all, 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 used, and they are sources of heterogeneity, the loadings should show specific patterns, as it will be explained in the result section. The order of the eigenvalues could be a criterion for a relative decision on the heterogeneity. PCA does not require advance knowledge of the item covariates. However, the method should be used with caution, because of the possibility of artifacts.

### *Dimensionality Test (DIMTEST)*

DIMTEST (Stout, Douglas, Junker & Roussos, 1993) is a nonparametric statistical approach for assessing unidimensionality of dichotomously scored test items. This technique

provides a tool for assessing extra heterogeneity beyond the general underlying latent trait. The method is based on the principle that for two parallel subsets of items the variance of the sum scores should be about equal in homogeneous subgroups of persons, as defined on the basis of a third subset of items. A T statistic, which can be tested on its significance, is used to decide upon the presence of extra heterogeneity.

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 good 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.

DIMTEST provides a criterion for an absolute decision on extra heterogeneity beyond one underlying dimension, based on a statistical test. It does not give an indication of where the extra heterogeneity is located, and no advance knowledge of item covariates is required.

### *Dimensionality Evaluation To Enumerate Contributing Traits (DETECT)*

The DETECT procedure is a nonparametric IRT based method developed for detecting the latent dimensionality of a test, or more precisely for disclosing the dimensionally homogeneous item clusters of a test. The DETECT procedure was developed originally by Kim (1994), and its theory was further adapted by Zhang and Stout (see, e.g., Zhang & Stout, 1999, or Stout, Habing, Douglas, Kim Roussos & Zhang, 1996). In case of sufficiently separated, strongly homogeneous item clusters (as in a “simple structure”), the procedure is able to find the exact number of latent dimensions and the true latent structure of the test. Even if the item vectors in the test space are considerably differing in their angles, but when the clusters are still clearly separable (an approximate simple structure type), DETECT

still finds the crucial clusters. For a simple structure, the number of clusters indicates the dimensionality; for an approximate simple structure, the number of clusters found by DETECT may be smaller than the number of the latent dimensions, because DETECT finds only the substantively distinct dimensions. The R(P) index provides information about the degree to which simple structure is realized. As a guideline, the authors of DETECT recommend to assume approximate simple structure in practice when the estimated R(P) ≥ 0.8. In that case the DETECT statistic can be used. In case of simple structure the R value equals to one.

The statistic is based on the covariance within the item pairs, conditional upon the test composite, $\theta_\alpha$ (the standardized linear combination of the underlying latent traits or dimensions). The items are clustered in an iterative procedure to obtain the partition with the highest value of the DETECT statistic for a given maximal number (chosen by the user) of non-overlapping clusters (Zhang and Stout, 1999). The theoretical DETECT index for a given partition (P) is based on the sum of the conditional covariances (conditional upon $\theta_\alpha$) 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 indicates 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. The authors emphasize that these categories may depend on the particular application, and may deviate from the above described ones (Douglas, Kim, Roussos, Stout & Zhang, 1999).

The current version of DETECT starts with a hierarchical cluster analysis (HCA) and then uses a generic algorithm to obtain the global maximum DETECT value. A cross-validation is also build into 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*. Afterwards, a partitioning of the items is obtained based

on the second subset, and this partitioning is applied for the first subset. The obtained 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, disregarding the partitioning provided by DETECT. Zhang and Stout (1999) define a discrepancy measure to decide on unidimensionality in their simulation study as the difference between the maximum DETECT value and the reference DETECT value divided by the reference DETECT value. In their study, when the discrepancy exceeded the critical value of 0.5 or the reference DETECT value was smaller or equal than 0.1, the data sets was judged unidimensional. This decision rule worked perfectly in their case.

The authors (Zhang & Stout, 1999) warn that DETECT might not perform well in case of a small sample size, or a small number of items. Items close to the test composite, and items with small discrimination parameters may be incorrectly classified. It is also important to note that in case of an approximate simple structure, the partition which maximizes the DETECT index, does not necessarily indicate the number of dimensions of the data, and that the indicated number may be smaller than the actual number of dimensions. On the other hand, those items that have a relatively small discrimination parameter and are close to other clusters may form a new cluster in the DETECT analyses. These clusters are not sizable and should not be considered, but they explain why DETECT may suggest more dimensions than there are in the data (Zhang & Stout, 1999).

DETECT provides an absolute decision on extra heterogeneity beyond one underlying dimension (i.e., heterogeneity), but the criterion is a rule of thumb and not a statistical test. The method does not require advance knowledge of item covariates. 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, however, without an

explicit link to the item covariates. When item clusters can be linked to the item covariates indeed, the method can be informative also for the relative decisions on heterogeneity.

## The Simulation Study

In order 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 with most experiments. The covariates were binary and were crossed in an orthogonal way, so that there were eight types of items, and four items of each type. From an experimental point of view this is a 2x2x2 within-subject repeated measures design. In contrast with experiments, tests often do not have a design, but it is a desirable feature to have for a test (Embretson, 1985), also for purposes of cognitive diagnosis, as noted in the introduction (Tatsuoka & Tatsuoka, 1982; Tatsuoka, 1990), and in psychological experiments with repeated measures a design is often used, indeed.

For the generation of the data, 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, if the effect was random over persons, opposite signs of the covariate values lead to a negative association, whereas same signs lead 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 yield a simple structure (+1, +1 and +1, -1). This particular structure of item covariates makes sense for both the ability and the personality domain. 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 is obtained.

For the personality domain, contrasts do make sense as item covariates, and of course also the simple structure is not uncommon for personality.

One of the slopes and/or the intercept was defined to be random over persons. The general model for the data generation was the following:

$$\text{logit}\left(P(Y_{pi}=1\middle|\theta_p,\beta_{p1})\right)=\theta_p+\beta_{p1}x_{i1}+\beta_2x_{i2}+\beta_3x_{i3} \qquad (5)$$

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, the means of the slopes were 1. The theoretical mean for each data set as a whole was .5. This part of the simulation study will be referred further on as the *single effect design*, however, in two cells of the design (with zero variance value for the slope and for the intercept) there is no random effect present.

When both the intercept and one slope were random, three variance values were used: 0, 0.2, and 1.2, so that nine combinations were obtained from crossing the three levels. These values represent three kinds of effects of the covariates: fixed effects, a minor source of heterogeneity, and a major source of heterogeneity, respectively. With two random effects, the distribution was bivariate normal with zero correlation. This second kind of design will be called the *combined effect design*.

The above described variance values are the theoretical values the data were generated with. The actual variance of the random effects may be different due to the sampling that is inherent to the generation procedure. These 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 ten 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 coding of the covariates was used as for the generation of the data .For the combined effect design, the results are given in Table 1. The results are 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.

---

Insert Table 1 about here.

---

First, it is clear that the larger the theoretical variance is, the larger the variance of the individual estimates is. For theoretical values of 0, mean variances of 0.2 to 0.25 were found, for theoretical values of 0.2, mean variances of 0.43 to 0.50 were found, and finally, for theoretical values of 1.2, mean variances of 1.09 to 1.22 were found. When the theoretical (and real) variance was zero, the mean established variance based on the estimates from the individual analyses was still .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, while the smallest estimated variance for a parameter with non-zero 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 non-zero theoretical variance was 0.36. According to these results, any value between 0.28 and 0.36 as a rule of thumb would result in a perfect decision for these data sets. When 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 both the slope and the intercept variance were random, 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 by definition 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 the larger (of the intercept or slope). Therefore the method of individual analyses can be used to decide on the order in which random effects are included in the model until the model fit would be sufficient. It is clear from the results that theoretical values of variance as small as 0.2 lead to variances of the estimates that are larger than when the theoretical values of variance is zero.

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, or in other words, for determining the order in which to include effects as random effects in

the model. Given that one has a priori knowledge of the item covariates, an advantage of the method is that it locates the heterogeneity.

### *Marginal Modeling*

In Figures 1 and 2 the estimated association parameters $\alpha$ belonging to the random intercept and random slope are plotted, for the single effect design. The ten data sets per theoretical variance are shown on the approximately horizontal curves in the figures. The data sets are ordered on the x-axis based on the association estimates.

---

Insert Figure 1 about here.

---

---

Insert Figure 2 about here.

---

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 figure). The series belonging to adjacent theoretical values show some overlap, but also the real variances do overlap.

---

Insert Figure 3 about here.

---

In Figure 3, the estimated association parameters are displayed for the combined effect design. In each of the panels of Figure 3, all four association estimates of the corresponding ten data sets are plotted (related to the intercept and to the three slopes, see Equation 4). Each line consist of ten parameter estimates belonging to different data sets. The triangles denote the association parameters of the intercepts, the circles refer to the association parameters of the manipulated slopes and the squares to the ones of the fixed slopes. The estimates are ordered on the x-axes according to their value on the y-axes, in a different panel for each pair of a theoretical value of the intercept and slope variance. As it can be seen, only the association parameters belonging to random effects differ from zero. The obtained values are also closely related to the amount of heterogeneity. This is a clear and unambiguous result.

First, as it was mentioned before, the +1/-1 coding did not converge for 65% of the data sets, therefore the 0/1 coding was used instead, and from these results the corresponding α-estimates of the +1/-1 coding were calculated. As a consequence, the corresponding significance test could not been used for these data sets. For 40 data sets in both parts of the study with maximum 0.2 variance in the random parameters, the +1/-1 coding did converge indeed. In each data set there were four parameters, so that all in all, there were 160 observations available for investigating the behavior of the significance test of the α-values for both parts of the study. With $p \leq .05$ as the critical value for an absolute decision criterion, only one false alarm (out of 120 cases with a true α of zero) was found, and no misser was found (out of 40 cases with a true α larger than zero) for the combined effect design.

The significance test for the α-values could not be used for all data, but an ad hoc critical value for α can be applied. The α-values of the random parameters with zero and 0.2 theoretical variance did not overlap, for the single effect design, so that a critical value for the α-values between 0.018 and 0.14 would be perfectly suitable. For the combined effect design, any critical value between 0.05 and 0.12 would lead to perfect predictions. Therefore in this

study any critical value for the for the α-values between 0.05 and 0.12 would result in perfect predictions. Comparing the α-values for the second and the third covariates to the ad hoc critical value of 0.06, perfect predictions could obtained.

Second, it is clear that the estimated association was a function of the real variance. For the single effect design $R^2$ 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 $R^2$ was .93 for the intercept, and .97 for the slope.

In sum, marginal modeling with ALR seems to be a quite effective method to detect heterogeneity in an absolute sense, when convergence is obtained (and a statistical test is possible). Given the high values of $R^2$, also a relative decision based on the ordering of the variances, seems to be a good procedure. Given that one has item covariates available, an advantage is that the heterogeneity can be located .

## *Principal Component Analysis*

### *PCA Eigenvalues*

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

First, when 1.9 was used as the best criterion for an eigenvalue to decide whether it represents a true source of heterogeneity, 5% false alarms (one out of 20 data sets) and 0% missers (out of 120 data sets) was obtained for the single effect design. For the combined

effect design, 1.67 % (one out of 60) false alarms and 8.33% (ten out of 120) missers were obtained. Taking into account both parts of the simulation study in most of the cases, the elbow criterion (based on a judgment by the first author) also indicated the correct number of dimensions (100% of the data sets for the single effect design, but only 72% for 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 a high, but different theoretical variance (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 was, the larger (but still moderate) the decrease was in the eigenvalue of the other random effect.

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

*PCA Loadings*

The PCA loadings were found to show the hypothesized pattern. Figure 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, the 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).

---

Insert Figure 4 about here.

---

---

Insert Figure 5 about here.

---

In case of the combined effect design, the same effects were observed as earlier. Figure 6 is provided for illustrative purposes. One line is horizontal (for the intercept component) and the other one shows a jump (for the slope component). When the intercept and slope variances were equal, it depended on the data set which of the two random effects showed in the first component, because the order of the real variances is a matter of chance given that the generation values are equal.

---

Insert Figure 6 about here.

---

In an additional parallel simulation study with an unipolar coding instead of a bipolar coding for the item covariate with random effects, the results were similar. The only difference was that the PCA loadings referring to the random slope were mostly positive and the jump of the ordered loadings was more moderate than in 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 advanced knowledge of the item covariates to interpret the pattern of the loadings in terms of slope variance.

While the results suggest that PCA is a quite easy and good method to detect and to 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 when both manipulated parameters are random, the combined effect design was used for investigating DIMTEST. 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 made by the automatic item selection option of the DIMTEST software. As a first step, a factor analysis was used, and the items for the first subset were selected based on their second factor loadings. The desired significance level of the DIMTEST statistic was set to $\alpha = .05$.

---

Insert Table 2 about here.

---

Table 2 contains the number of the data sets indicated to be multidimensional (out of 100). Since the cells in the first column and the first row of Table 2 are unidimensional, ideally, one would expect 5 out of 100 data sets to be indicated as multidimensional in those cells and much higher frequencies than five 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 it is expected. For three of the four remaining cells, the frequency is rather low, meaning that the detection of multidimensionality is rather poor. Finally, when both theoretical variances are 1.2, still only 81 out of the 100 data sets were identified as multidimensional, meaning that 19% cases of strong heterogeneity went undetected. DIMTEST resulted in 64.75% (259 out of 400 data sets) missers and 11.4% (57 out of 500 data sets) false alarms.

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 quite poor, and the global results improve only slightly. Considering the DIMTEST perspective on bipolarity 51.3% (308 out of 600) missers and 8% (24 out of 300) false alarms were obtained.

These results do not come unexpected. As it was mentioned earlier, DIMTEST underperforms for small samples (van Abswoude, van der Ark & Sijtsma, 2004). Furthermore, equal numbers of items loading on the different dimensions leads to less stable DIMTEST results than unequal numbers of items (van Abswoude, van der Ark & Sijtsma, 2004). These may be the reasons for the moderate detection rate of heterogeneity in this study. One should be aware that, when the sample size is small, DIMTEST may overlook small variances.

## *DETECT*

As for DIMTEST, also for DETECT the focus is on the combined effect design, because it is a method to detect extra heterogeneity beyond one underlying dimension. For DETECT only 10 data sets per cell were used, because the variance of the test statistics was small. There are different ways to apply the DETECT procedure, and this will be reflected in this study.

In a first step, the DETECT analyses were limited to *two latent dimensions*, but in order to gain a better understanding of the results, later the analyses were repeated for a larger number of dimensions (whereas the true dimensionality was never larger than two). DETECT was first applied with cross-validation, because using the *cross-validation* option in the DETECT procedure is strongly recommended (Zhang & Stout, 1999). The examinees of each data set were randomly assigned to two subsets (with equal size). The results of the first step are shown in Table 3. Both the maximum DETECT value and the reference DETECT value, and the associated R-values for each subset are given. In this first step, different decision rules will be compared, concerning the inferences to be made regarding unidimensionality and extra heterogeneity.

---

Insert Table 3 about here.

---

According to the results, the DETECT statistic is not so sensitive to the intercept variance than to the slope variance. The maximum DETECT value and the reference DETECT value increase with the slope variance, but not with the intercept variance. 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. 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). When the slope variance is 0 or 0.2, the reference DETECT values are much smaller than those of the first subset. This shows the effect of cross-validation.

According to the DETECT manual, when the DETECT procedure indicates multidimensionality, the DETECT value can be interpreted only in case of simple structure or approximate simple structure. For an approximate simple structure, the R-value should be higher than 0.8. For unidimensional data sets this condition is not required for interpreting the DETECT value. Because simple structure is a condition for the interpretation of the DETECT value, strictly speaking only 28 data sets could be considered for multidimensional (out of the 60 multidimensional data sets), all with theoretical slope variance of 1.2. All 28 data sets should be detected as showing extra heterogeneity when DETECT is used, because of a bipolar covariate with random weight.

Since the DETECT values are linear functions of the slope variance that is the source of multidimensionality in these data sets, it makes sense to consider all DETECT values. The theoretical critical values for indicating unidimensionality and strong multidimensionality are $<0.1$ and $\geq 1$, respectively. Applying these values for the 28 data sets with approximate simple structure, not one misser was found (and false alarms were not possible). Applying these critical values to all 90 data sets, two false alarms on a total of 30 data sets and 30 missers on a total of 60 data sets were obtained. Using the critical value of 0.5 for moderate multidimensionality, still 20 missers were found. When only one critical value was used (0.1) for deciding upon unidimensionality, only three false alarms and five missers were found for the 90 data sets. Based on these results, the value of 0.1 seems to be a successful criterion to

find out whether there is extra heterogeneity beyond an underlying dimension. Note that these results are obtained while considering a bipolar unidimensional structure as multidimensional (in the sense of DETECT). When such structure is considered as unidimensional, the number of false alarms is of course higher.

When also the discrepancy measure was considered in the decisions about extra heterogeneity, in our study the discrepancy 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. The application of the combined criteria resulted in zero false alarm and 21 missers (out of 60 multidimensional data sets).

When DETECT was used *without cross-validation,* and allowing for *two dimensions*, the DETECT values and the R-values for slope variances of 0 and 0.2 were much higher than the reference DETECT value in the cross-validation procedure. For a slope variance of 1.2, values similar to the reference values were obtained. For these analyses without cross-validation, the optimal critical DETECT value turned out to be 0.5, yielding two missers and zero false alarm. With the critical value of 0.1, as recommended in the manual, a remarkable amount of unidimensional cases were overlooked. The same was found in the comparative study of van Abswoude et al. (2004), who noted that the suggested upper bound of unidimensionality might be too low. However, with a higher critical value, the procedure without cross-validation also seems to work well for our problem.

Because in practice one may not have an idea about the number of latent dimensions, it is interesting to see how the method works when more then two dimensions are assumed. For an analysis with more than two dimensions, *12 dimensions,* were allowed, the highest possible number of dimensions in the DETECT program, in order to give maximal freedom to DETECT in finding clusters. Also in this case, first the *cross-validation* procedure was followed. The results concerning the number of clusters are reported in Table 4. 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.

---

Insert Table 4 about here

---

However, it is mentioned in the DETECT manual, that a considerably higher maximum DETECT value than the reference DETECT value indicates that the clusters may stem from capitalization upon chance, and the data set may be unidimensional. The earlier described combined decision rule (discrepancy of DETECT values larger than 0.5 or the reference DETECT value is smaller or equal than 0.1) resulted in zero false alarms and 20 missers (out of 60 multidimensional data sets).

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 k=2, and the reference DETECT value for k is higher than 0.1, the true dimensionality is two, if it is smaller or equal, the test is unidimensional.

When k>2, go to (3).

(3) Calculate the discrepancy measure for dimensionality k. If it is smaller or equal than 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 (2).

When for the discrepancy measure 0.5 was chosen as a critical value five missers were found and the dimensionality was overestimated for 12 data sets. With 0.3 as critical value, again five missers were found, and the dimensionality was still overestimated for four data sets.

When the DETECT procedure was used *without cross-validation*, and *12 dimensions* were allowed, the true dimensionality 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 critical DETECT value of 1.1 was used, perfect decisions were obtained, identifying cases with zero slope variance as one-dimensional if the value was lower than 1.1, and identifying cases with slope variance .2 or 1.2 if the DETECT value was equal to or larger than 1.1. 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 know a priori.

There are some remaining problems. The DETECT value is based on conditional covariances calculated for each item pair and for each total score group based on the remaining items. The minimum number of examinees for each total score group is defined by the user in the input of the DETECT procedure. The recommended value is 20. Only those total score groups are considered for the covariance calculation which contain at leas as many examinees as the reference value defined in the input. This value should be lowered if the minimum percentage of examinees used for the covariance calculations is lower than 85%. 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. In order to reach the recommended percentage of observations used in the covariance calculation, one should have 20(I-2) observations, where I is the number of items. In psychological research this condition is often not fulfilled.

The R-value seems to be a function of the variance, and in this study the criterion value was often not reached for multidimensional data sets. In general, when we interpret all DETECT values, it seems that cut-off values are not easy to find. Apart from these problems, DETECT turned out to be a reasonably good method to detect extra heterogeneity and also to explore where it is located.

## Discussion and Conclusion

Various methods were investigated for detecting heterogeneity in small data sets with binary repeated measures and with item covariates. 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, for example a test with subscales. Furthermore, in psychological measurement the assumption of design factors with effects that differ depending on the person make sense, as a structure with person-by-item interaction.

As it was 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. Marginal modeling provides a statistical test for the association parameters and also locates heterogeneity. Also, the results of individual analyses seem to be quite sensitive to the size of the heterogeneity, but 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. Although additional bootstrap could help to find appropriate cut off value for procedures with different cut-off values for different data structures.

Among the methods that do not require item covariates, it is difficult to differentiate. PCA seemed to be an effective method in this study, but PCA has the drawback that it is vulnerable to artifacts. DIMTEST 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 a priori information about the item covariates and still can locate heterogeneity. But although DETECT seemed to be a quite effective method in this study, the decision criteria for DETECT are not always evident.

## Author Note

Katalin Balázs, István Hidegkuti and Paul De Boeck, Department of Psychology, K. U. Leuven, Belgium. We acknoledge the financial support from the IAP/5 network grant to Paul De Boeck.

Correspondence concerning this article should be addressed to Katalin Balázs, Department of Psychology, K. U. Leuven, Tiensestraat 102, B-3001, Leuven, Belgium, Katalin.Balazs@psy.kuleuven.ac.be.

# 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, Chapman and 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: Erlbaum.

Douglas, J., Kim, H.-R., Roussous, 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]

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.

Hardin, J. W., Hilbe, J. M. (2003) Generalized estimating equations, 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 Campaign). *Dissertation Abstracts International, 55-12B*, 5598

Landwehr, J. M., Pregibon, D. & Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models, *Journal of American Statistical Association, 79*, 61-71.

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*.

Rijmen. F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement. 26,* 271-285.

SAS Institute, Inc. (1999). *SAS online doc* (Version 8). [Software manual on CD-ROM]. Cary, NC: SAS Institute, Inc.

Segall, D. O. (2001). General ability measurement: an application of multidimensional item response theory. *Psychometrika*, 66, 79-97.

Sternberg, R.J. (1977). Intelligence, Information Processing, and Analogical Reasoning. Hillsdale, NJ: Erlbaum.

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.

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.

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: 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.

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.

Table 1

*The mean of variance of the individual estimates for the combined effect design*

| Intercept variance | First slope variance | | |
|---|---|---|---|
| | 0 | 0.2 | 1.2 |
| 0 | | | |
| Intercept | 0.21 | 0.22 | 0.23 |
| Slope 1 | 0.22 | 0.45 | 1.22 |
| Slope 2 | 0.22 | 0.22 | 0.22 |
| Slope 3 | 0.22 | 0.22 | 0.22 |
| 0.2 | | | |
| Intercept | 0.48 | 0.50 | 0.43 |
| Slope 1 | 0.22 | 0.44 | 1.19 |
| Slope 2 | 0.21 | 0.21 | 0.22 |
| Slope 3 | 0.20 | 0.23 | 0.23 |
| 1.2 | | | |
| Intercept | 1.21 | 1.17 | 1.1 |
| Slope 1 | 0.23 | 0.47 | 1.09 |
| Slope 2 | 0.23 | 0.23 | 0.20 |
| Slope 3 | 0.24 | 0.23 | 0.25 |

Table 2

*The number of data sets indicated as multidimensional by DIMTEST*

| | First slope variance | | |
|---|---|---|---|
| Intercept variance | 0 | 0.2 | 1.2 |
| 0 | 6 | 12 | 21 |
| 0.2 | 10 | 18 | 21 |
| 1.2 | 8 | 20 | 82 |

Table 3

*The mean DETECT and R-values with cross-validation for an analysis with 2dimensions*

| Intercept variance | First slope variance | | | | | |
|---|---|---|---|---|---|---|
| | 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 |

Table 4

*The number of clusters found by DETECT with cross-validation when 12 dimensions were allowed*

| | First slope variance | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | | | | | .2 | | | | | 1.2 | | | | |
| | Number of clusters | | | | | | | | | | | | | | |
| Intercept variance | 2 | 3 | 4 | 5 | 6 | 2 | 3 | 4 | 5 | 6 | 2 | 3 | 4 | 5 | 6 |
| 0 | | | 5 | 4 | 1 | | | 7 | 3 | | 10 | | | | |
| .2 | | | 7 | 3 | | 2 | 3 | 4 | 1 | | 10 | | | | |
| 1.2 | 1 | 2 | 6 | 1 | | 1 | 4 | 5 | | | 10 | | | | |

# Figure Captions

*Figure 1.* The estimated association parameter values referring to the random intercept, for different values of the variance, ordered according to the size of the estimated association

*Figure 2.* The estimated association parameter values referring to the random slope, for different values of the variance, ordered according to the size of the estimated association

*Figure 3*. The association parameter estimated in the combined effect design

*Figure 4.* The PCA loadings for the first five principal components, ordered as a function of the size of the loadings (the intercept variance is 0.6, the other variances are zero)

*Figure 5.* The PCA loadings for the first five principal components, ordered as a function of the size of the loadings (the slope variance is 0.6, the other variances are zero)

*Figure 6.* The PCA loadings of the first five principal components (the intercept variance and the slope variance are both 1.2, the other variances are zero)

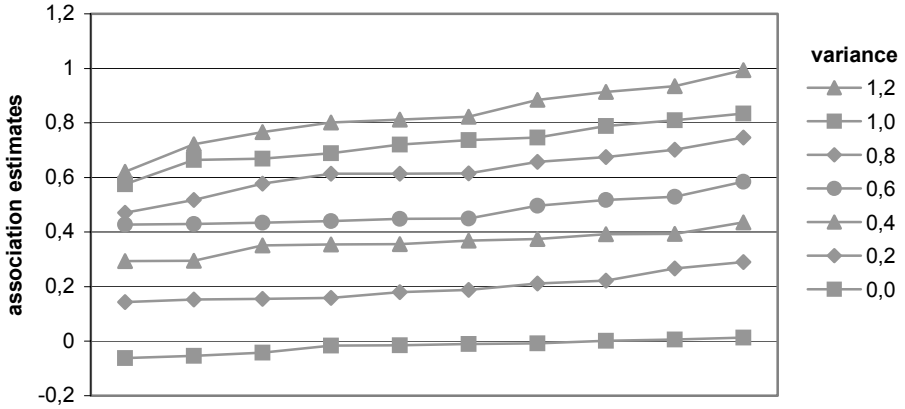



Figure 1

The estimated association parameter values referring to the random intercept, for different values of the variance, ordered according to the size of the estimated association

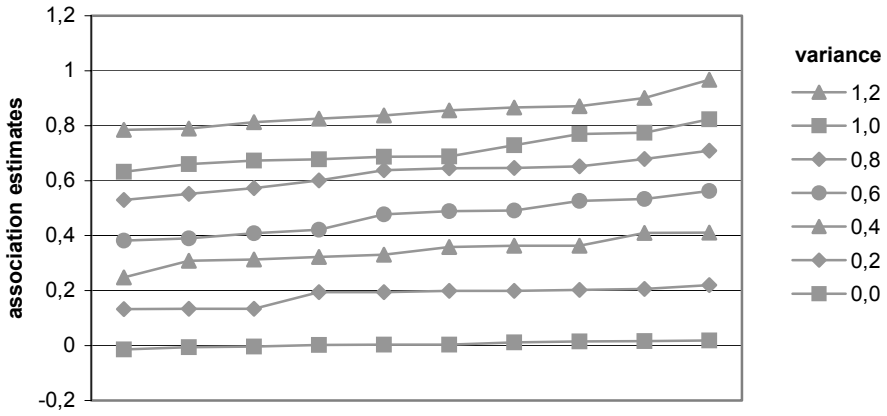
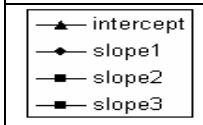



Figure 2

The estimated association parameter values referring to the random slope, for different values of the variance, ordered according to the size of the estimated association

Figure 3: The association parameter estimated in the combined effect design




First slope variance=0

First slope variance=0.2

First slope variance=1.2

Int. variance=0

Int. variance=0.2

Int. variance=1.2

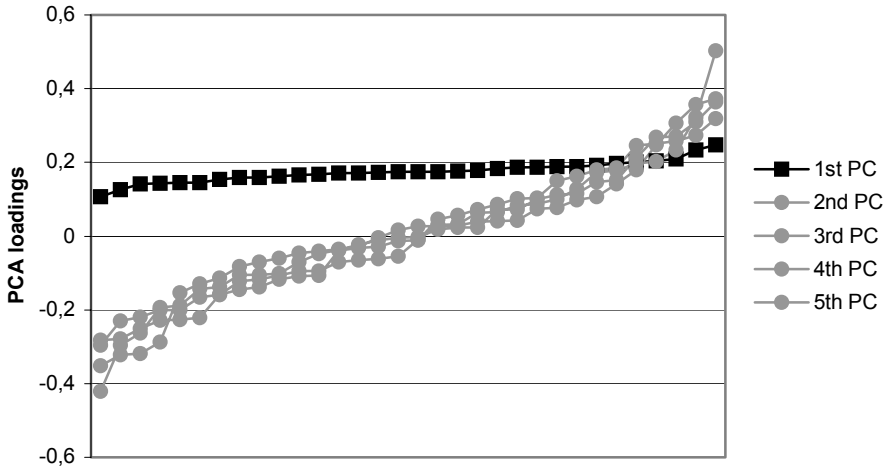



Figure 4

The PCA loadings for the first five principal components, ordered as a function of the size of the loadings (the intercept variance is 0.6, the other variances are zero)

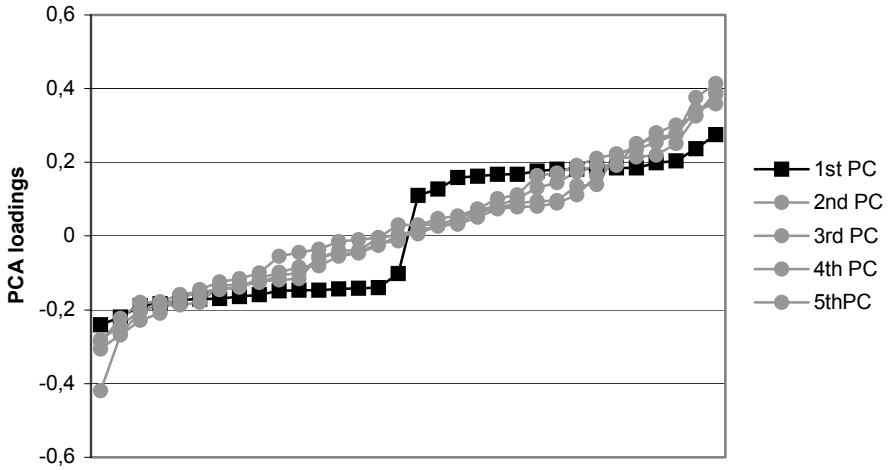



Figure 5

The PCA loadings for the first five principal components, ordered as a function of the size of the loadings (the first slope variance is 0.6, the other variances are zero)

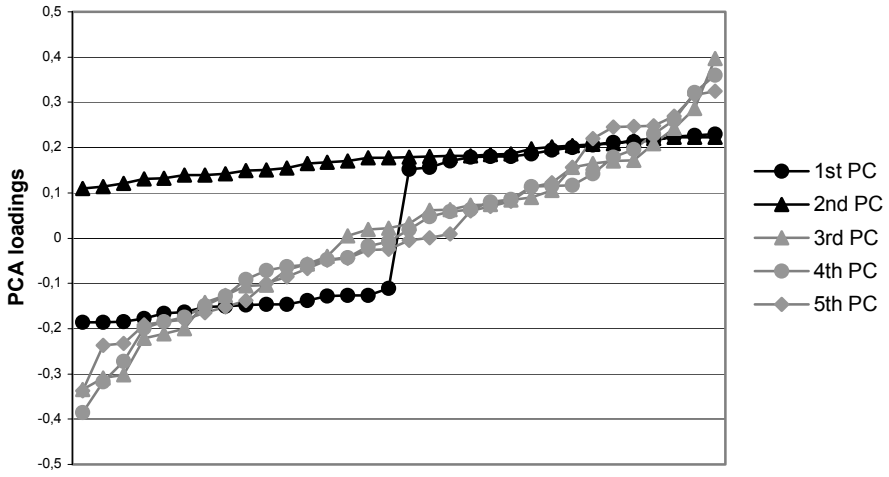



Figure 6

The PCA loadings of the first five principal components (the intercept variance and the first slope variance are both 1.2, the other variances are zero)