A misspecification test for finite-mixture logistic models for clustered binary and ordered responses

Francesco, Bartolucci and Silvia, Bacci and Claudia, Pigini

University of Perugia, Università Politecnica delle Marche

2015

A misspecification test for finite-mixture logistic models for clustered binary and ordered responses

Francesco Bartolucci*† Silvia Bacci*‡§ Claudia Pigini*¶

May 8, 2015

Abstract

An alternative to using normally distributed random effects in modeling clustered binary and ordered responses is based on using a finite-mixture. This approach gives rise to a flexible class of generalized linear mixed models for item responses, multilevel data, and longitudinal data. A test of misspecification for these finite-mixture models is proposed which is based on the comparison between the Marginal and the Conditional Maximum Likelihood estimates of the fixed effects as in the Hausman's test. The asymptotic distribution of the test statistic is derived; it is of chi-squared type with a number of degrees of freedom equal to the number of covariates that vary within the cluster. It turns out that the test is simple to perform and may also be used to select the number of components of the finite-mixture, when this number is unknown. The approach is illustrated by a series of simulations and three empirical examples covering the main fields of application.

Keywords: Generalized Linear Mixed Models, Hausman Test, Item Response Theory, Latent Class model, Longitudinal data, Multilevel data

*Department of Economics, University of Perugia, Via A. Pascoli, 20, 06123 Perugia.

†*email*: bart@stat.unipg.it

‡Department of Economics, University of Perugia, Via A. Pascoli, 20, 06123 Perugia.

§*email*: silvia.bacci@unipg.it

¶Department of Economics and Social Sciences, Unversità Politecnica delle Marche, P.le Martelli 8, 60121 Ancona

‖*email*: c.pigini@univpm.it

1 Introduction

Generalized Linear Mixed Models (GLMMs, Skrondal and Rabe-Hesketh, 2004; McCulloch et al., 2008; Stroup, 2012) represent a very useful instrument for the analysis of clustered data, as they use random effects to account for the dependence between observations within the same cluster. This structure of the data arise in Item Response Theory (IRT) applications (Hambleton and Swaminathan, 1985; De Boeck and Wilson, 2004), in the multilevel context where individuals are collected in groups (Goldstein, 2003), and in longitudinal/panel studies in which repeated responses on the same individuals are available (Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2009). In this article, we focus on logistic regression models for binary and ordered responses; for one of the first applications see Stiratelli et al. (1984) and Anderson and Aitkin (1985).

The random effects in a GLMM are typically assumed to have a normal distribution and the consequences of non-normality have been receiving considerable attention in the literature, especially for nonlinear models. In fact, in linear models, the wrong specification of the random effect distribution tends to have minor consequences, as the maximum likelihood estimators are consistent and asymptotically normally distributed under mild conditions. In particular, recent studies conclude that the consequences of violations of normality on the quality of the estimates and on random effects predictions are rather severe (Heagerty, 1999; Heagerty and Kurland, 2001; Rabe-Hesketh et al., 2003; Agresti et al., 2004; Litière et al., 2008). The negative effects of distributional misspecification motivate the development of alternative approaches to formulate and test hypotheses about this latent (also called mixing) distribution.

A well known approach, that formulates in a flexible way the random effect distribution, is based on assuming a discrete distribution that leads to a finite-mixture model. This approach is seen as semiparametric because a discrete distribution may approximate arbitrary well any continuous distribution. Nevertheless, the idea of approximating the true mixing distribution by a discrete one goes back to studies preceding the development of the class of GLMMs and, then, in the context of simpler models involving incidental parameters. In particular, we refer to the nonparametric maximum likelihood approach (Kiefer and Wolfowitz, 1956; Laird, 1978; Lindsay, 1983).

The first applications of random effects with discrete distribution in the GLMM context are Lindsay et al. (1991) in the IRT context, Aitkin (1999) in the general context of clustered data, and Vermunt (2003) with multilevel data. Heckman and Singer (1984) used the finite-mixture approach to formulate a flexible model for survival data, and Aitkin (1996) used this approach to create overdispersion in a generalized linear model.

In addition to a greater flexibility, the finite-mixture approach has some advantages over the normal approach. Mainly, it avoids integrating out the random effects, which may be complex when random effects are multidimensional, and a rather simple Expectation Maximization (EM) algorithm (Dempster et al., 1977) may be used instead. Moreover, the approach leads to a natural clustering of sample units that may be of main interest in

certain relevant applications (e.g., Deb, 2001). In fact, a GLMM based on finite-mixture formulation may be seen as a latent class model (Lazarsfeld and Henry, 1968; Goodman, 1974; Hagenaars and McCutcheon, 2002) extended with the inclusion of covariates. The finite-mixture approach has also some limitations with respect to the normal approach for the distribution of the random effects, such as the difficult interpretation in certain contexts, in which these effects represent missing covariates that are naturally seen as continuous. Moreover, there is the need to choose the number of mixture components (also called latent classes or support points), and some instability problems in estimation often arise due to multimodality of the likelihood function. For a comparison between the normal and the finite-mixture approaches we refer the reader to Skrondal and Rabe-Hesketh (2004) and Bartolucci et al. (2014a). Nevertheless, the finite-mixture approach is the main alternative to the normal approach to formulate the distribution of the random effects for GLMMs, and in particular for logistic regression models. This is testified by recent applications such as Jain et al. (1994) and Kim et al. (1995) in the context of brand preferences, Pudney et al. (1998) for the analysis of data about farm tenure contracts, and Deb (2001) for the study of the demand for preventive care. Several further applications are described in Skrondal and Rabe-Hesketh (2004); see also Azzimonti et al. (2013) and Heinzl and Tutz (2013).

Testing hypotheses about the mixing distribution, and in particular the normality, has attracted a considerable attention in the recent statistical literature. A standard method to check for normality of the mixing distribution is based on empirical Bayes estimates of the individual effects (Lange and Ryan, 1989). However, this method has been criticized because of its lack of power (Verbeke and Lesaffre, 1996; Verbeke and Molenberghs, 2013). Among other methods, it is worth mentioning the method based on residuals (Ritz, 2004; Pan and Lin, 2005), the method based on simulating the random effects from their posterior distribution given the observed data (Waagepetersen, 2006), the method based on comparing Marginal Maximum Likelihood (MML) and Conditional Maximum Likelihood (CML) estimates (Tchetgen and Coull, 2006), methods based on the covariance matrix of the parameter estimates and the information matrix (Alonso et al., 2008, 2010), and that based on the gradient function (Verbeke and Molenberghs, 2013).

In the present article, we propose a general test for misspecification of the discrete mixing distribution in logistic models with binary and ordered responses. We extend the approach developed by Tchetgen and Coull (2006) which, as mentioned above, is based on the comparison of CML and MML estimates for the fixed effects, as in the Hausman's test (Hausman, 1978); the difference between the two estimates is normalized on the basis of an estimate of the variance-covariance matrix of this difference. The test relies on the consistency of the CML estimator that is attained under mild distributional assumptions; essentially the only requirement is that the random effects are constant within each cluster. At least to our knowledge, this approach has not been developed in the context of finite-mixture models. Moreover, with respect to the approach of Tchetgen and Coull (2006),

which is only referred to the case of normally distributed random effects, our approach presents some novelties and peculiarities deriving from the finite-mixture nature of the models of interest, as we argument below.

First of all, since none of the two estimators compared is ensured to be fully efficient, we use a generalized estimate of the variance-covariance matrix of the difference through a method adopted, in a related context, by Bartolucci et al. (2014c). This also ensures stable results in small samples, while retaining the simplicity of the approach and its low computational complexity. Second, the proposed test may also be used to select the number of support points of the discrete distribution, which is alternative to commonly used selection criteria, such as the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978). This is a crucial issue in the use of the models of our interest that, obviously, does not arise when random effects are normally distributed. Third, an issue that is typically ignored in the statistical field is that one of the possible sources of misspecification is the dependence between the random effects and the observable covariates, that is, a problem of endogeneity. In the finite-mixture approach, a greater variety of methods to model this dependence is available with respect to the normal approach, and the proposed test has an important role in this regard, as will be clear in the following.

The paper is organized as follows. In the next section we describe the class of GLMMs with a special focus on the case of binary and ordinal response variables. In Section 3, the two estimation methods applied for the test are described, that is, the MML method under the discreteness assumption of random effects and the CML method. In Section 4 we recall the traditional Hausman test and, then, we illustrate the proposed test in the finite-mixture context. Application on real data are provided in Section 6 and some final remarks conclude the work in Section 7. The finite-sample properties of the proposed test are investigated through an extensive Monte Carlo study, whose design and main results are reported in the Appendix, whereas a summary of this study is provided in Section 5. Upon request, we also make available the `R` codes we used to implement the proposed approach.

2 The class of GLMMs of interest

The class of GLMMs is highly flexible, because it allows us to accommodate several types of response variables (e.g., continuous, binary, count) and to account for different hierarchical data structures (i.e., multilevel data, longitudinal data, and item response data). These models are based on a link function (McCullagh and Nelder, 1989) applied to the conditional expected value of each response variable given the available covariates and a set of random effects having a suitable distribution, which is typically normal. As mentioned in Section 1, we focus in particular on versions of these models for binary and ordinal response variables, which are based on a logit link function.

Let n denote the number of clusters and, for each cluster i, let J_i denote the number of units in the cluster and let $\boldsymbol{x}_i$ be a column vector of covariates. Moreover, for each unit j in cluster i, let y_{ij} denote the response variable of interest and let $\boldsymbol{z}_{ij}$ be the corresponding column vector of specific covariates. In the binary case we have $y_{ij} = 0, 1$, whereas in the more general ordinal case we have $y_{ij} = 0, \ldots, L-1$, where L is the number of categories. In any case the response variables are collected in the cluster-specific vectors $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ_i})'$, $i = 1, \ldots, n$. Similarly, the unit-specific covariates are collected in the matrices $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iJ_i})$, $i = 1, \ldots, n$.

The notation defined above is completely general as it is suitable for different settings of interest. In the multilevel setting, units j refer to individuals, each of them being nested in a given group i (e.g., pupils within schools, patients within hospitals). In such a case, $\boldsymbol{x}_i$ is a vector of group-specific characteristics (e.g., number of pupils per school, number of hospital beds), whereas $\boldsymbol{z}_{ij}$ is a vector of individual-specific covariates (e.g., age, gender). In the case of longitudinal data, index j refers to time occasions and i identifies different individuals. In such a context, $\boldsymbol{x}_i$ is a vector of time-constant individual covariates (e.g., gender) and $\boldsymbol{z}_{ij}$ is a vector of time-varying individual covariates (e.g., income). Finally, in the similar context of item responses data, j denotes the item to which individual i answers, but $\boldsymbol{z}_{ij}$ is simply a vector of dummies and $\boldsymbol{x}_i$ is usually null.

In the case of binary responses, the basic model we consider is the random intercept logit model based on the assumption

$$\log \frac{p(y_{ij} = 1|\alpha_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})}{p(y_{ij} = 0|\alpha_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})} = \alpha_i + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{\gamma}, \quad i = 1, \ldots, n,\ j = 1, \ldots, J_i, \tag{1}$$

where α_i is the random effect for cluster i, $\boldsymbol{\beta}$ is the vector of regression parameters for the cluster-specific covariates, and $\boldsymbol{\gamma}$ is that for the unit-specific covariates. The random effects α_i are typically assumed to have distribution $N(0, \sigma^2)$, so that the common intercept is absorbed in $\boldsymbol{\beta}$. The alternative approach, that is of main interest in the present paper, assumes that the distribution of each of these random parameters is discrete with k support points $\xi_1, \ldots, \xi_k$ and corresponding probabilities $\pi_h = p(\alpha_i = \xi_h)$, $h = 1, \ldots, k$, so that the result is a finite-mixture model (McLachlan and Peel, 2000). In each case, *local independence* is assumed, that is, the response variables in each $\boldsymbol{y}_i$ are conditionally independent given α_i, $\boldsymbol{x}_i$, and $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iJ_i})$. It is also well-known that an alternative to random-effects approaches is a fixed-effects approach in which the α_i parameters are estimated together with $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ or are eliminated by conditioning on suitable sufficient statistics, as in the CML method. For an up to date review see Bartolucci et al. (2014b).

With ordinal responses, the above model is extended as follows:

$$\log \frac{p(y_{ij} \geq l|\alpha_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})}{p(y_{ij} < l|\alpha_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})} = \alpha_i + \delta_y + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{\gamma}, \quad l = 1, \ldots, L-1, \tag{2}$$

on the basis of cumulative logits in increasing order, also known as global logits (McCullagh, 1980; Agresti, 2002). In the above expression, the cut-points are in suitable

order, that is, $\delta_1 < \ldots < \delta_{L-1}$. A more general formulation is based on substituting the cut-points δ_l with cluster-specific cut-points α_{il}, as follows:

$$\log \frac{p(y_{ij} \geq l|\boldsymbol{\alpha}_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})}{p(y_{ij} < l|\boldsymbol{\alpha}_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})} = \alpha_{il} + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{\gamma}, \quad l = 1, \ldots, L-1, \tag{3}$$

with $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{i,L-1})$ having multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ or a discrete distribution with support points $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k$ and corresponding probabilities $\pi_h = p(\boldsymbol{\alpha}_i = \boldsymbol{\xi}_h)$, $h = 1, \ldots, k$.

It is worth noting that, in the IRT setting, models (1) and (2) correspond to the Rasch model (Rasch, 1960) in the binary case and to the graded response model with fixed discriminating parameters (Samejima, 1969) in the ordinal case, being $\boldsymbol{x}_i$ the null vector and the elements of $\boldsymbol{\gamma}$ corresponding to item difficulty parameters. For details on the possible parameterizations for polytomous IRT models, see Bacci et al. (2014). Moreover, a nice interpretation of these models is provided by introducing an underlying continuous response y_{ij}^* defined as

$$y_{ij}^* = \alpha_i + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{\gamma} + \varepsilon_{ij}.$$

This variable is related to the observed response y_{ij} through a suitable function, that is, $y_{ij} = G(y_{ij}^*)$, defining an observation rule. In particular, $G(\cdot)$ is a parametric function which depends in a suitable way on specific parameters according to the different nature of y_{ij}. With binary responses, we have

$$G(y_{ij}^*) = I\{y_{ij}^* > 0\}, \tag{4}$$

where $I\{\cdot\}$ is an indicator function assuming value 1 when its argument is true and value 0 otherwise, so that the model defined in (1) results, provided that ε_{ij} has a standard logistic distribution. More generally, when y_{ij} is an ordinal variable with l categories, the model in (2) derives when

$$G(y_{ij}^*) = \begin{cases} 0, & y_{ij}^* \leq -\delta_1, \\ 1, & -\delta_1 < y_{ij}^* \leq -\delta_2, \\ \vdots & \vdots \\ L-1, & y_{ij}^* > -\delta_{L-1}, \end{cases} \tag{5}$$

with ε_{ij} still having standard logistic distribution.

All the above models may be extended to deal with the dependence of the random effects on one or more cluster-specific covariates $\boldsymbol{w}_i$, which may be seen as a form of endogeneity. Two approaches are here considered. First, an interaction term may be included so that, in the binary case, we have

$$\log \frac{p(y_{ij} = 1|\alpha_i, \boldsymbol{w}_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})}{p(y_{ij} = 0|\alpha_i, \boldsymbol{w}_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})} = \boldsymbol{w}_i'\boldsymbol{\alpha}_i + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{\gamma}, \quad i = 1, \ldots, n, \; j = 1, \ldots, J_i, \tag{6}$$

where the cluster-specific covariates in $\boldsymbol{w}_i$ may be a subset of those in $\boldsymbol{x}_i$. Another possible extension consists in assuming that mass probabilities may depend on the covariates in $\boldsymbol{w}_i$ by a multinomial logit parametrization:

$$\log \frac{p(\alpha_i = \xi_{h+1}|\boldsymbol{w}_i)}{p(\alpha_i = \xi_1|\boldsymbol{w}_i)} = \phi_h + \boldsymbol{w}_i'\boldsymbol{\psi}_h, \quad h = 1, \ldots, k-1, \tag{7}$$

where ϕ_h are intercepts and $\boldsymbol{\psi}_h$ are vectors of regression parameters (see also Huang and Bandeen-Roche, 2004). When the support points ξ_h are suitably ordered, then alternative parametrizations based on cumulative logits may be adopted, which are more parsimonious and easier to interpret than the multinomial logits.

3 Estimation methods

In this section we describe two estimation methods for the GLMM parameters that will be used for the proposed Hausman test. First, we illustrate the MML method under the assumption of the discreteness of α_i. Then, a description of the CML method is provided, which is based on a fixed-effects approach.

3.1 Discrete Marginal Maximum Likelihood

The assumption of local independence implies that

$$p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{x}_i, \boldsymbol{Z}_i) = \prod_j p(y_{ij}|\alpha_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij}), \quad i = 1, \ldots, n, \tag{8}$$

where $p(y_{ij}|\alpha_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})$ depends on the model specification; see, for instance, equation (1) for the random-intercept model for binary responses. Then, the *manifest distribution* of $\boldsymbol{y}_i$ given the covariates is obtained by marginalizing $p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{x}_i, \boldsymbol{Z}_i)$ with respect to α_i:

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{Z}_i) = \sum_h \left[\prod_j p(y_{ij}|\xi_h, \boldsymbol{x}_i, \boldsymbol{z}_{ij}) \right] \pi_h, \quad i = 1, \ldots, n,$$

that provides the following marginal log-likelihood function

$$\ell_M(\boldsymbol{\theta}) = \sum_i \log p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{Z}_i) = \sum_i \log \sum_h \left[\prod_j p(y_{ij}|\xi_h, \boldsymbol{x}_i, \boldsymbol{z}_{ij}) \right] \pi_h,$$

with $\boldsymbol{\theta}$ denoting the overall vector of free parameters including $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, the support points ξ_h, $h = 1, \ldots, k$, and $k-1$ logits for the probabilities π_h.

The maximization of function $\ell_M(\boldsymbol{\theta})$ may be efficiently performed through an Expectation Maximization (EM) algorithm (Dempster et al., 1977), based on the *complete data* log-likelihood function, that is the log-likelihood that could be computed knowing the latent class from which every unit comes (i.e., knowing the discrete random effects α_i):

$$\ell_M^*(\boldsymbol{\theta}) = \sum_i a_{hi} \left[\log \pi_h + \sum_j \log p(y_{ij}|\xi_h, \boldsymbol{x}_i, \boldsymbol{z}_{ij}) \right], \tag{9}$$

with a_{hi} being an indicator variable equal to 1 if $\alpha_i = \xi_h$ and to 0 otherwise.

The EM algorithm is implemented along the usual lines, alternating two steps. The E-step consists in computing the posterior expected value of each a_{hi}, which is equal to the posterior probability of belonging to a certain latent class given the response configuration he/she provided, that is,

$$\hat{a}_{hi} = p(\alpha_i = \xi_h | \boldsymbol{x}_i, \boldsymbol{Z}_i, \boldsymbol{y}_i) = \frac{p(\boldsymbol{y}_i | \xi_h, \boldsymbol{x}_i, \boldsymbol{Z}_i)\pi_h}{p(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{Z}_i)}.$$

The resulting values $\hat{a}_{hi}$ are then substituted in (9) so as to obtain $\hat{\ell}^*_M(\boldsymbol{\theta})$. The following M-step consists in maximizing function $\hat{\ell}^*_M(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and the result is used to update the estimates at the E-step. This iterative process continues until convergence so as to obtain the MML estimate $\hat{\boldsymbol{\theta}}_M$. Besides, this scheme may be easily adapted to estimate extended models based on assumptions (6) and (7).

For deriving the Hausman test, it is important to recall that the asymptotic variance-covariance matrix for $\hat{\boldsymbol{\theta}}_M$ may be estimated by the sandwich formula (White, 1982), as follows:

$$\widehat{\boldsymbol{V}}_M(\hat{\boldsymbol{\theta}}_M) = \boldsymbol{H}_M(\hat{\boldsymbol{\theta}}_M)^{-1} \boldsymbol{S}_M(\hat{\boldsymbol{\theta}}_M) \boldsymbol{H}_M(\hat{\boldsymbol{\theta}}_M)^{-1}, \tag{10}$$

with

$$\begin{aligned}
\boldsymbol{H}_M(\boldsymbol{\theta}) &= \sum_i \frac{\partial^2 \log p(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{Z}_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}, \\
\boldsymbol{S}_M(\boldsymbol{\theta}) &= \sum_i \boldsymbol{u}_{M,i}(\boldsymbol{\theta}) [\boldsymbol{u}_{M,i}(\boldsymbol{\theta})]', \\
\boldsymbol{u}_{M,i}(\boldsymbol{\theta}) &= \frac{\partial \log p(\boldsymbol{y}_i | \alpha_i, \boldsymbol{x}_i, \boldsymbol{Z}_i)}{\partial \boldsymbol{\theta}}.
\end{aligned}$$

In particular, $\boldsymbol{H}_M(\boldsymbol{\theta})$ is the Hessian of the log-likelihood function, whereas $\boldsymbol{S}_M(\boldsymbol{\theta})$ is equal to n times the empirical variance-covariance matrix of the score vector. From the matrix $\widehat{\boldsymbol{V}}_M(\hat{\boldsymbol{\theta}}_M)$ we can extract in the usual way the standard errors for the parameter estimates.

3.2 Conditional Maximum Likelihood (CML)

An alternative to the semi-parametric MML approach described above is given by the CML method (Andersen, 1970, 1972; Chamberlain, 1980), which is based on considering intercepts α_i as fixed parameters rather than random effects. This method gives a consistent estimator of the $\boldsymbol{\gamma}$ parameters for the covariates in $\boldsymbol{Z}_i$ under mild regularity conditions and independently of the true distribution from which values α_i come, as it relies on conditioning on a sufficient statistic for α_i.

In presence of binary data, the CML approach consists of maximizing the conditional log-likelihood function

$$\ell_C(\boldsymbol{\gamma}) = \sum_i \log p(\boldsymbol{y}_i | y_{i+}, \boldsymbol{Z}_i), \quad y_{i+} = \sum_{j=1}^{J_i} y_{ij},$$

where

$$p(\boldsymbol{y}_i|y_{i+},\boldsymbol{Z}_i)=\frac{\exp\left(\sum_j y_{ij}\,\boldsymbol{z}'_{ij}\boldsymbol{\gamma}\right)}{\sum_{\boldsymbol{s}\in\mathcal{S}_{J_i}(y_{i+})}\exp\left(\sum_j s_j\,\boldsymbol{z}'_{ij}\boldsymbol{\gamma}\right)} \tag{11}$$

and the sum $\sum_{\boldsymbol{s}\in\mathcal{S}_{J_i}(y_{i+})}$ is extended to all binary vectors $\boldsymbol{s}=(s_1,\ldots,s_{J_i})'$ with sum equal to y_{i+}. We observe that $p(\boldsymbol{y}_i|y_{i+},\boldsymbol{Z}_i)$ does not depend anymore on α_i and $\boldsymbol{x}_i$ (and, possibly, on $\boldsymbol{w}_i$ under extended model based on assumption (7)), but only on the regression parameters $\boldsymbol{\gamma}$ for the unit-specific covariates in $\boldsymbol{Z}_i$.

The conditional log-likelihood $\ell_C(\boldsymbol{\gamma})$ may be simply maximized by a Newton-Raphson algorithm, based on the score vector

$$\begin{aligned}\boldsymbol{u}_C(\boldsymbol{\gamma}) &= \sum_i \boldsymbol{u}_{C,i}(\boldsymbol{\gamma}),\\ \boldsymbol{u}_{C,i}(\boldsymbol{\gamma}) &= \frac{\partial \log p(\boldsymbol{y}_i|y_{i+},\boldsymbol{Z}_i)}{\partial\boldsymbol{\gamma}},\end{aligned}$$

and Hessian matrix

$$\boldsymbol{H}_C(\boldsymbol{\gamma})=\sum_i \frac{\partial^2 \log p(\boldsymbol{y}_i|y_{i+},\boldsymbol{Z}_i)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'},$$

so as to obtain the CML estimate $\hat{\boldsymbol{\theta}}_C$. Finally, the asymptotic variance-covariance matrix for $\hat{\boldsymbol{\gamma}}_C$ is estimated as

$$\widehat{\boldsymbol{V}}_C(\hat{\boldsymbol{\gamma}}_C)=\boldsymbol{H}_C(\hat{\boldsymbol{\gamma}}_C)^{-1}\boldsymbol{S}_C(\hat{\boldsymbol{\gamma}}_C)\boldsymbol{H}_C(\hat{\boldsymbol{\gamma}}_C)^{-1}, \tag{12}$$

with

$$\boldsymbol{S}_C(\boldsymbol{\gamma})=\sum_i \boldsymbol{u}_{C,i}(\boldsymbol{\gamma})[\boldsymbol{u}_{C,i}(\boldsymbol{\gamma})]'.$$

There exist several ways to implement the CML method in the presence of ordinal response variables; see Baetschmann et al. (2011) for a review. Here we rely on the idea of reducing the model of interest to a model for binary data by suitably dichotomizing the response variables and considering the contributions to the conditional log-likelihood as those resulting from all the possible dichotomizations of these variables (Chamberlain, 1980); see also Bartolucci et al. (2014c).

More in detail, we consider the $L-1$ possible dichotomizations and for each of them we transform the response variables y_{ij} in the binary variables

$$y_{ij}^{(l)}=I\{y_{ij}\geq l\},\quad i=1,\ldots,n,\ j=1,\ldots,J_i.$$

The sum of the conditional log-likelihood functions corresponding to each dichotomization provides the *pseudo conditional log-likelihood* function

$$\tilde{\ell}_C(\boldsymbol{\gamma})=\sum_i\sum_l \log p(\boldsymbol{y}_i^{(l)}|y_{i+}^{(l)},\boldsymbol{Z}_i),\quad y_{i+}^{(l)}=\sum_{j=1}^{J_i} y_{ij}^{(l)}, \tag{13}$$

where $\boldsymbol{y}_i^{(l)} = (y_{i1}^{(l)}, \ldots, y_{iJ_i}^{(l)})$ and $p(\boldsymbol{y}_i^{(l)}|y_{i+}^{(l)}, \boldsymbol{Z}_i)$ is defined as in (11) substituting each y_{ij} with $y_{ij}^{(l)}$.

The pseudo conditional log-likelihood $\tilde{\ell}_C(\boldsymbol{\gamma})$ may be maximized by a simple extension of the Newton-Raphson algorithm implemented for the binary case, using the *pseudo-score* vector

$$\begin{aligned} \tilde{\boldsymbol{u}}_C(\boldsymbol{\gamma}) &= \sum_i \tilde{\boldsymbol{u}}_{C,i}(\boldsymbol{\gamma}), \\ \tilde{\boldsymbol{u}}_{C,i}(\boldsymbol{\gamma}) &= \sum_l \frac{\partial \log p(\boldsymbol{y}_i^{(l)}|y_{i+}^{(l)}, \boldsymbol{Z}_i)}{\partial \boldsymbol{\gamma}} \end{aligned}$$

and the *pseudo-observed information* matrix

$$\tilde{\boldsymbol{H}}_C(\boldsymbol{\gamma}) = \sum_i \sum_l \frac{\partial^2 \log p(\boldsymbol{y}_i^{(l)}|y_{i+}^{(l)}, \boldsymbol{Z}_i)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'}.$$

Finally, the sandwich estimator of the variance-covariance matrix of the resulting pseudo CML estimator, still denoted by $\hat{\boldsymbol{\theta}}_C$ to simplify the notation, has the same expression as in (12) with the appropriate adjustments.

To conclude, we remind that the CML method does not allow to estimate the effect of cluster-specific covariates in vector $\boldsymbol{x}_i$, differently from the MML method. Moreover, it is known to provide less efficient estimators than MML if the distribution of the random effects is correctly specified. Nonetheless, the robustness of CML estimator to misspecification of the distribution of α_i makes it appropriate for the comparison with MML estimator in finite-mixture models, as will be illustrated in the following.

4 The proposed Hausman-type test for misspecification

In this section we describe the proposed Hausman-type test for misspecification of the distribution of the random effects for the finite-mixture GLMMs illustrated in Section 2. We also discuss its use for selecting the number of mixture components.

4.1 Test formulation

The traditional Hausman test (Hausman, 1978) is typically used to test the assumption of normality of the random effects in linear mixed models, which are a special case of GLMMs for normal responses. The test is based on the comparison of two estimators that under the null hypothesis of correct model specification (H_0) are both consistent, but if the model is misspecified (H_1) only one of them remains consistent. Consequently, we have evidence of misspecification from the distance between the two estimators as they

converge to two different points in the parameter space under H_1. Moreover, it is required that one of the two estimators is asymptotically efficient under H_0, so as to simplify the estimation of the variance-covariance matrix of the difference between them.

In the present context, H_0 corresponds to a model of type (1) for binary data or (2) for ordinal data, or its extended versions defined in Section 2, in which the distribution of the random effects α_i is discrete with k support points. Moreover, under the basic formulations there is independence of these random effects from the observable covariates, so as to rule out endogeneity. In this context, the Hausman test is based on the statistic

$$\begin{aligned} T_1 &= n(\hat{\boldsymbol{\gamma}}_M - \hat{\boldsymbol{\gamma}}_C)' \widehat{\boldsymbol{W}}_1^{-1} (\hat{\boldsymbol{\gamma}}_M - \hat{\boldsymbol{\gamma}}_C), \qquad (14) \\ \widehat{\boldsymbol{W}}_1 &= \widehat{\boldsymbol{V}}_C(\hat{\boldsymbol{\gamma}}_C) - \widehat{\boldsymbol{V}}_M(\hat{\boldsymbol{\gamma}}_M), \qquad (15) \end{aligned}$$

which has asymptotical χ^2_c distribution under H_0, where c is the dimension of parameter vector $\boldsymbol{\gamma}$ or, equivalently, the number of unit-specific covariates in $\boldsymbol{z}_{ij}$. Note that the test is based on a comparison between the CML and MML estimators for the unit-specific covariates, as only these parameters are estimable under the CML approach. In this regard, we have to clarify that $\boldsymbol{V}_M(\hat{\boldsymbol{\gamma}}_M)$ is a suitable block of matrix $\boldsymbol{V}_M(\hat{\boldsymbol{\theta}}_M)$ defined in (10). Moreover, the variance-covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\gamma}}_M - \hat{\boldsymbol{\gamma}}_C)$, denoted by $\boldsymbol{W}$, is estimated as the difference between $\widehat{\boldsymbol{V}}_C(\hat{\boldsymbol{\gamma}}_C)$ and $\widehat{\boldsymbol{V}}_M(\hat{\boldsymbol{\gamma}}_M)$ due to the efficiency of $\hat{\boldsymbol{\gamma}}_M$ under H_0 which, in turn, implies that the covariance matrix between $\hat{\boldsymbol{\gamma}}_M$ and $\hat{\boldsymbol{\gamma}}_C$ is $\boldsymbol{C}(\hat{\boldsymbol{\gamma}}_M, \hat{\boldsymbol{\gamma}}_C) = \boldsymbol{V}_M(\hat{\boldsymbol{\gamma}}_M)$.

It is worth noting that, in the present context, the formula to estimate $\boldsymbol{W}$ may rise some instability problems for small samples in which the difference between $\boldsymbol{V}_C(\hat{\boldsymbol{\gamma}}_C)$ and $\boldsymbol{V}_M(\hat{\boldsymbol{\theta}}_M)$ is not ensured to be positive definite; see also Vijverberg (2011) for related problems. Therefore, we rely on a generalized version of the test based on a different way of estimating $\boldsymbol{W}$ that has been used by Bartolucci et al. (2014c) in a related context. In particular, we propose to use the following estimator:

$$\begin{aligned} \widehat{\boldsymbol{W}}_2 &= n\, \boldsymbol{D}\, \widehat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\gamma}}_C)\, \boldsymbol{D}', \\ \boldsymbol{D} &= (\boldsymbol{E}, -\boldsymbol{I}), \end{aligned}$$

with $\boldsymbol{I}$ being the identity matrix of dimension q and $\boldsymbol{E}$ a matrix such that $\hat{\boldsymbol{\gamma}}_M = \boldsymbol{E}\, \hat{\boldsymbol{\theta}}_M$. Moreover, the joint variance-covariance matrix of $\hat{\boldsymbol{\gamma}}_C$ and $\hat{\boldsymbol{\theta}}_M$ is obtained by the generalized sandwich formula

$$\widehat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\gamma}}_C) = \begin{pmatrix} \boldsymbol{H}_M(\hat{\boldsymbol{\theta}}_M) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_C(\hat{\boldsymbol{\gamma}}_C) \end{pmatrix}^{-1} \boldsymbol{S}\left(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\gamma}}_C\right) \begin{pmatrix} \boldsymbol{H}_M(\hat{\boldsymbol{\theta}}_M) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_C(\hat{\boldsymbol{\gamma}}_C) \end{pmatrix}^{-1},$$

with

$$\boldsymbol{S}\left(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\gamma}}_C\right) = \sum_i \begin{pmatrix} \boldsymbol{u}_{M,i}(\hat{\boldsymbol{\theta}}_M) \\ \boldsymbol{u}_{C,i}(\hat{\boldsymbol{\gamma}}_C) \end{pmatrix} \left(\boldsymbol{u}_{M,i}(\hat{\boldsymbol{\theta}}_M)' \quad \boldsymbol{u}_{C,i}(\hat{\boldsymbol{\gamma}}_C)'\right),$$

and $\boldsymbol{H}_M(\hat{\boldsymbol{\theta}}_M)$, $\boldsymbol{H}_C(\hat{\boldsymbol{\gamma}}_C)$, $\boldsymbol{u}_{M,i}(\hat{\boldsymbol{\theta}}_M)$, and $\boldsymbol{u}_{C,i}(\hat{\boldsymbol{\gamma}}_C)$ defined as in Sections 3.1 and 3.2. $\boldsymbol{H}_C(\hat{\boldsymbol{\gamma}}_C)$ and $\boldsymbol{u}_{C,i}(\hat{\boldsymbol{\gamma}}_C)$ are substituted by $\tilde{\boldsymbol{H}}_C(\hat{\boldsymbol{\gamma}}_C)$ and $\tilde{\boldsymbol{u}}_{C,i}(\hat{\boldsymbol{\gamma}}_C)$ in case of ordinal responses, as illustrated in Section 3.2.

Overall, the test statistic defined as

$$T_2 = n(\hat{\boldsymbol{\gamma}}_M - \hat{\boldsymbol{\gamma}}_C)' \widehat{\boldsymbol{W}}_2^{-1} (\hat{\boldsymbol{\gamma}}_M - \hat{\boldsymbol{\gamma}}_C) \tag{16}$$

has still an asymptotically distribution of type χ^2_c under H_0, but it gives more stable results, while being easy to compute. This is the approach that we adopt in the following.

4.2 Use of the proposed test for finite-mixture GLMMs

The traditional Hausman test is typically employed to investigate about the possible sources of misspecification of the distribution of the random effects, being the absence of normality and the possible dependence between the random effects and the covariates (i.e., endogeneity) the most relevant ones. Similarly, the proposed test based on statistic T_2 allows us to assess the model specification in the general setting of GLLMs, with some peculiarities deriving from the discrete nature of the distribution of the random effects, as we argument in the following.

A crucial aspect related to the models with discrete random effects is the choice of the number of latent classes (or mixture components), denoted by k. In general, the prevailing approaches which have been adopted in the literature balance model fit and parsimony and are based on information criteria, obtained through penalization of the maximum log-likelihood. Among these criteria, the most common are the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978), which are based on the minimization of the following indices:

$$\begin{aligned} AIC &= -2\,\hat{\ell} + 2\,\#\text{par}, \\ BIC &= -2\,\hat{\ell} + \log(n)\,\#\text{par}, \end{aligned}$$

where $\hat{\ell} = \ell(\hat{\boldsymbol{\theta}}_M)$ is the maximum of the marginal log-likelihood of the model of interest and #par stands for the number of free parameters. Several alternatives have been developed in the literature, which are based on different penalization terms; see the Appendix for a detailed description. Unfortunately, there is not any result in the literature that indicates one of these criteria as clearly outperforming the others, although there is a certain evidence in favor of BIC. Among the most recent comparative studies, see Dias (2006), Nylund et al. (2007), and Yang and Yang (2007).

The proposed Hausman test based on T_2 represents an interesting alternative to the information criteria mentioned above to select the number of mixture components, when this number is unknown. In this regard, we suggest to adopt a sequential strategy consisting in increasing k until the test does not stop to reject H_0. We expect that the selection criterion for k based on T_2 is more parsimonious with respect to the available criteria mentioned above, provided that the assumption of independence between the random effects and the covariates hold. In particular, this is expected to happen when the distribution of the random effects is continuous rather than discrete. In this situation, the estimator $\hat{\boldsymbol{\gamma}}_M$

may attain values very close to the estimator $\hat{\boldsymbol{\gamma}}_C$, which is consistent, even for small values of k; see also Lindsay et al. (1991). On the other hand, note that the above information criteria are typically considered unsatisfactory by applied researches because they tend to select large values of k and then non parsimonious models, especially with large samples.

It is also worth noting that, while the other criteria to select k only perform relative comparisons among differently specified models, the proposed test allows us to formulate an absolute judgment about the appropriateness of the model based on a certain number of mixture components. In fact, a sufficiently high p-value for a certain k leads to conclude for the correct specification of such a model in the complex.

Finally note that, with longitudinal data, the proposed test can be used in connection with that proposed by Bartolucci et al. (2014c) to test the assumption that the random-effects are time-constant rather than time-varying. More in detail, we may adopt a two-step procedure consisting in testing first the assumption of time-constant random effects and, only if this hypothesis is not rejected, the modified Hausman test here proposed is applied to select the correct number of mixture components.

5 Simulation study

In order to analyze the proposed approach, we performed a Monte Carlo simulation study. A detailed description of the design and results of this study is reported in the Appendix, whereas we provide a brief summary in the following.

The simulation study is based on the random intercept model specified in Section 2 by assumptions (1) and (2). We consider two scenarios: one refers to the longitudinal setting and the other to the IRT setting. In our benchmark design, the distribution of the random effects α_i has $k_0 = 3$ support points $\left[-\sqrt{3/2}, 0, \sqrt{3/2}\right]$ with probabilities 0.25, 0.50, and 0.25 respectively. For the longitudinal setting, we consider one cluster-specific covariate x_i following a standard normal distribution together with one unit-specific covariate z_{ij}, with $j = 1, \ldots, J$ denoting the time occasions, generated from an AR(1) process with correlation $\rho = 0.5$. The parameters of the mean specification are both scalars and equal to 1. The Hausman test statistic will therefore be asymptotically distributed as a χ^2_1. For the IRT setting, model (1) based on a logit link function simplifies to a Rasch model. Therefore, the Hausman test statistic T_2 will have null asymptotic distribution of type χ^2_{J-1}, where J is the number of items.

The experiment on the two models is repeated with different discrete distributions for α_i, including a shift in the original distribution, $\alpha_i \in \left[1 - \sqrt{3/2}, 1, 1 + \sqrt{3/2}\right]$ with probabilities 0.25, 0.50, and 0.25, and formulating a strongly asymmetric distribution, $\alpha_i \in [-5, 0, 25]$ with probabilities 0.33, 0.50, and 0.17 respectively.

The second part of our simulation study deals with possible misspecifications of the random effect distribution. First, a case where the true distribution of α_i is continuous is considered: the data are generated as above with the exception of the random effects

which are now $\alpha_i \sim N(0, 3)$. Second, the analysis considers a case where the random effects are correlated with the regression covariates.

In terms of results, the proposed test presents good size properties under the null hypothesis of correct specification of the number of mixture components of the distribution of the random effects. If the number of classes is underspecified, the Hausman test's rejection rate considerably increases when the distribution of the random effects is skewed. Instead, if the random effects follow a continuous distribution, a situation that is likely to occur with real data, the proposed Hausman test typically chooses a more parsimonious model in comparison to standard model selection criteria. This is particularly true for large values of J, which usually leads to a clearer interpretation of the results, especially when the aim is data classification or when the interest in on the regression parameters. In the presence of correlation between the random effects with the regression covariates, rejection rates are remarkably high even in very small samples. In addition, the power of the test increases in the intensity of the correlation, while an increasing number of occasions J seems to only slightly affect the rejection rates.

6 Applications

We illustrate three applications of the Hausman test in different settings. We first describe the problem of choosing the number of mixture components in a Rasch model and in a random intercept logit model for clustered data. Then, we deal with the proper specification of a model for ordered longitudinal data.

Example 1: Rasch model for the assessment of ability in mathematics

We illustrate the proposed Hausman test by using a dataset concerning the responses of a sample of 1510 examines to 12 binary items on Mathematics, which has been extrapolated from a larger dataset collected in 1996 by the Educational Testing Service within the National Assessment of Educational Progress (NAEP) project. The same set of data was also analyzed by Bartolucci and Forcina (2005) and Bartolucci (2007). In particular, Bartolucci and Forcina (2005) fitted some types of LC models under different constraints.

The Hausman test and the information criteria described in the Appendix are applied to a sequence of Rasch models with an increasing number of latent classes. As shown in Table 1, the Hausman test selects $k = 3$ latent classes, as well as BIC, CAIC, and the corresponding modified versions BIC* and CAIC*, whereas the other criteria detect four or more classes.

Intuitively, the correct specification of the Rasch model is confirmed by the results in Table 3, which show the item difficulty estimates obtained with the CML approach and with the MML approach. In fact, we observe that with $k = 3$ mixture components

Table 1: *Naep data, Rasch model: selection of the number k of mixture components.*

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Hausman T_2	414.850	90.071	6.721	2.895	1.639
Hausman p-value	0.000	0.000	**0.821**	0.992	0.999
AIC	22042.3	20511.4	20364.6	**20361.8**	20365.0
BIC	22106.2	20585.9	**20449.7**	20457.6	20471.4
AIC_3	22054.3	20525.4	20380.6	**20379.8**	20385.0
CAIC	22118.2	20599.9	**20465.7**	20475.6	20491.4
HTAIC	22042.6	20511.7	20365.0	**20362.3**	20365.6
AIC_c	22018.5	20483.6	20332.9	20326.2	20325.5
BIC^*	22068.1	20541.4	**20398.9**	20400.4	20407.8
$CAIC^*$	22080.1	20555.4	**20414.9**	20418.4	20427.8

Table 2: *Naep data, Rasch model: item difficulty estimates under CML ($\hat{\boldsymbol{\gamma}}_C$) and under MML with $k = 1, \ldots, 5$ ($\hat{\boldsymbol{\gamma}}_M$).*

	CML	MML				
		$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Item 1	**0.000**	0.000	0.000	**0.000**	0.000	0.000
Item 2	**-0.047**	-0.038	-0.045	**-0.047**	-0.047	-0.047
Item 3	**0.691**	0.549	0.670	**0.689**	0.691	0.691
Item 4	**-1.040**	-0.855	-0.984	**-1.032**	-1.037	-1.040
Item 5	**1.521**	1.207	1.478	**1.518**	1.521	1.521
Item 6	**0.013**	0.010	0.012	**0.013**	0.013	0.013
Item 7	**0.662**	0.527	0.642	**0.661**	0.662	0.662
Item 8	**1.191**	0.945	1.158	**1.189**	1.191	1.191
Item 9	**0.334**	0.267	0.323	**0.333**	0.334	0.334
Item 10	**0.525**	0.418	0.508	**0.524**	0.525	0.525
Item 11	**2.427**	1.945	2.339	**2.418**	2.427	2.427
Item 12	**2.474**	1.984	2.383	**2.464**	2.474	2.474

the item estimates by MML are already very close to those obtained with CML; see also Lindsay (1983).

We also perform the Hausman test for the Rasch model based on the assumption of normality of the distribution of the random effects (Tchetgen and Coull, 2006). A value of the test statistic T_2 equal to 10.230 with a p-value equal to 0.510 lead to accept the null hypothesis of correct model specification. However, the normality assumption does not allow us to cluster subjects in homogeneous classes in an easy way, differently from the discreteness assumption. Indeed, according to the Rasch model with $k = 3$, we observe (Table 3) that the 37.9% of subjects is allocated to class 3, which identifies the best performers, whereas the 16.4% of subjects belongs to the worst performers' class, that is class 1.

Table 3: *Naep data, Rasch model with $k = 3$: estimated support points and weights (standard errors in brackets).*

	$h = 1$	$h = 2$	$h = 3$
$\hat{\xi}_h$	-0.647 (0.138)	0.967 (0.131)	2.430 (0.120)
$\hat{\pi}_h$	0.164 (–)	0.457 (0.154)	0.379 (0.251)

Example 2: a random intercept logit model for the use of contraceptives in Bangladesh

The data come from a study about fertility in Bangladesh carried out by the Bangladesh National Institute of Population Research and Training. It collects information on the knowledge and use of family planning methods of a sample of ever-married women. For a detailed description of data see Huq and Cleland (1990); see also Mazharul Islam and Mahmud (1995).

Here we consider a subset of 1934 women nested in 60 administrative districts (clusters).[1] The response of interest is a binary variable denoting whether the interviewed woman is currently using contraception. The unit-specific covariates correspond to the following women's characteristics: geographical residence area (0= rural, 1=urban), age, number of children (no children, a single child, two children, three or more children; no children is the reference category). No variable describing the district characteristics is available.

The Hausman test and the information criteria are applied to a sequence of random intercept logit models with an increasing number of latent classes. As shown in Table 4, all information criteria agree in selecting two latent classes, whereas the Hausman test is more parsimonious and gives evidence for just one latent class at 5% level. In other words, according to the proposed Hausman-type test, the detection of a latent structure seems to be superfluous with data at issue and this also simplifies the interpretation of the results.

Parameter estimates for $k = 1$, reported in Table 5, show that contraceptive use is higher in urban than in rural areas (odds ratio = 2.219) and it declines a little with age (2.358% lower odds per year of age). Besides, contraceptive use is higher among women with a child and much higher among women with two or more children, than among those with no children, with odds ratios of almost three and more. Note that, if we adopt $k = 2$ as suggested by the information criteria, we obtain regression parameter estimates very similar to those shown in Table 5 (output here omitted).

[1] Data freely downloadable from `http://www.stata-press.com/data/r11/bangladesh.dta`

Table 4: *Bangladesh contraceptive data, random intercept logit model: selection of the number k of mixture components.*

	$k=1$	$k=2$	$k=3$	$k=4$
Hausman T_2	10.160	9.778	5.164	5.163
Hausman p-value	**0.071**	0.082	0.400	0.396
AIC	2469.1	**2427.2**	2430.0	2434.0
BIC	2481.7	**2444.1**	2451.1	2459.4
AIC_3	2475.1	**2435.2**	2440.0	2446.0
CAIC	2487.7	**2452.1**	2461.1	2471.4
HTAIC	2471.2	**2430.8**	2435.4	2441.8
AIC_c	2458.2	**2413.4**	2413.6	2415.5
BIC^*	2462.8	**2418.9**	2419.7	2421.6
$CAIC^*$	2468.8	**2426.9**	2429.7	2433.6

Table 5: *Bangladesh contraceptive data, random intercept logit model with $k = 1$: estimates of regression coefficients ($\hat{\boldsymbol{\gamma}}$), standard errors, odds ratios ($\exp(\hat{\boldsymbol{\gamma}})$).*

	$\hat{\boldsymbol{\gamma}}$	st.err.($\hat{\boldsymbol{\gamma}}$)	$\exp(\hat{\boldsymbol{\gamma}})$
urban area	0.800	0.189	2.218
age	-0.024	0.007	0.976
one child	1.067	0.183	2.906
two children	1.276	0.170	3.582
three or more children	1.214	0.201	3.368

Example 3: random intercept global logit models for the assessment of self-reported health status

The third example we propose is based on a longitudinal dataset about Self-Reported Health Status (SRHS), which derives from a subset of version I of the Health and Retirement Study (HRS)[2] (Juster and Suzman, 1995), conducted by the University of Michigan and supported by the US National Institute on Aging and the Social Security Administration. Our data comprise 1308 individuals who were asked to express opinions on their health status at 4 equally spaced time occasions, from 2000 to 2006. The response variable (SRHS) is measured on a Likert type scale based on 5 ordered categories (poor, fair, good, very good, and excellent). A longer version of the the same set of data was analyzed by Bartolucci et al. (2014c), who performed a test for the null hypothesis of time-constant random effects, versus the hypothesis of time-varying random effects, rejecting the null hypothesis (for more details about the data characteristics, see also Heiss, 2008; Bartolucci et al., 2014a). For our illustrative example, we reduced the panel length so as to minimize the impact of possible time-varying random effects.

We consider three time-constant covariates, describing gender, race, and educational

[2]See `http://www.nia.nih.gov/health/publication/growing-older-america-health-and-retirement-study`

Table 6: *HRS data, random intercept global logit model with free cut-points and with endogeneity of type* (7)*: selection of the number* k *of latent classes.*

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Hausman T_2	75.483	59.454	19.484	22.274	13.767	9.003	5.994	3.374
Hausman p-value	0.000	0.000	0.000	0.000	0.001	0.011	**0.050**	0.185
AIC	14879.9	13355.1	12852.8	12636.9	12497.6	12486.4	12457.8	12449.3
BIC	14948.6	13499.2	13072.5	12932.1	**12868.3**	12932.6	12979.4	13046.4
AIC_3	14889.9	13376.1	12884.8	12679.9	12551.6	12551.4	**12533.8**	12536.3
CAIC	14958.6	13520.2	13104.5	12975.1	**12922.3**	12997.6	13055.4	13133.4
HTAIC	14880.0	13355.2	12853.2	12637.5	12498.5	12487.7	12459.5	12451.5
AIC_c	14859.9	13313.2	12789.1	12551.4	12390.4	12357.6	12307.4	12277.4
BIC^*	14916.8	13432.5	12970.8	12795.4	**12696.7**	12726.0	12737.9	12770.0
CAI^*	14926.8	13453.5	13002.8	12838.4	**12750.7**	12791.0	12813.9	12857.0

level of individuals, and two time-varying covariates, corresponding to age and squared age. We first formulate the random intercept global logit model (2), having constant shift in the cut-points, and the global logit model (3) with free cut-points. In both cases, the proposed Hausman test repeatedly rejects the null hypothesis of correct model specification, despite most information criteria tend to choose 5 latent classes (outputs here omitted). Note that also the traditional Hausman test for the assumption of normally distributed random effects (Tchetgen and Coull, 2006) strongly rejects the model with $T_2 = 32.158$ and a p-value smaller than 0.001.

A possible problem with the data at issue may be due to the presence of endogeneity, that is dependence between the random effects and the time-varying covariates. For this reason, we extend models (2) and (3) to account for a possible effect of age and squared age on the mixture components weights, as in equation (7). In particular, the model based on assumptions (3) and (7) is not rejected with $k = 7$, as the corresponding p-value is around 5% (see Table 6). On the other hand, BIC and several other information criteria tend again to choose $k = 5$ components.

We conclude highlighting that, on one side, the traditional Hausman test recognizes the misspecification of the model, but does not detect a valid alternative, and, on the other side, the information criteria lead to select a misspecified model since they rely on a relative comparison between models.

7 Conclusions

We propose a misspecification test for Generalized Linear Mixed Models (GLMMs) for clustered binary and ordinal responses, which modifies the traditional Hausman test to account for the assumption of discrete, instead of normal, random effects. The proposed approach is easy to implement and may also be used to select the number of latent classes (or mixture components or support points), characterizing the models at issue.

The proposed Hausman-type test represents an element of novelty in the context of

model selection for finite-mixture models which is mainly based on information criteria, such as the Bayesian Information Criterion (BIC). With respect to these selection criteria, our proposal is expected to lead to more parsimonious models when the true distribution of the random effects is continuous and the dependence between these effects and the covariates is correctly specified. This is particularly useful in applications, where information criteria tend to choose a large number of components, especially with large samples. Moreover, the proposed test may reject all models having a different number of mixture components, so detecting misspecification problems (e.g., the presence of endogeneity), that are completely ignored by the information criteria. Finally, while these criteria are only based on relative comparisons among differently specified models, our proposal allows us to formulate an absolute assessment about the appropriateness of a given model, relying on the value of the test statistic and the corresponding p-value.

The performance of the proposed approach is evaluated through a Monte Carlo simulation study that provides satisfactorily results under different scenarios. In particular, we observe good size properties under the null hypothesis of correct specification of the number of support points of the random effect distribution. The results of the power analysis suggest that: (*i*) when the number of classes is underspecified, rejection rates are particularly high especially when the random effect distribution is skewed and has a large variance; (*ii*) when the random effect distribution is continuous, the Hausman test tends to select a more parsimonious specification of the number of support points, with respect to standard selection criteria, especially with many units per cluster; (*iii*) in the presence of correlation of the random effects with the regression covariates, rejection rates are remarkably high even in very small samples and increase for higher correlation values.

The approach is also illustrated by three applications covering different settings, that is, multilevel data, longitudinal data, item responses. Interestingly, each application presents a different potentiality of the proposed approach. In fact, in the first application we obtain the same results of selection criteria such as BIC in terms of number of mixture components. In the second application, contrary to the BIC, the proposed test leads to the conclusion that a latent structure is not necessary, and then to a very parsimonious and easily interpretable model. In the third application, the proposed approach leads to reject all models in which the random effects are assumed to be independent of the covariates, considering therefore a form of endogeneity.

Regarding the comparison with the available statistical literature, the proposed approach can be seen as a development of Tchetgen and Coull (2006), whose proposal is based on the comparison of MML and CML estimates of models with normally distributed random effects. We acknowledge that the approach of Tchetgen and Coull (2006) has been criticized by some authors. We refer, in particular, to Alonso et al. (2008) that, to motivate the need of alternative approaches, stated that the approach of Tchetgen and Coull (2006) can only be applied when there is at least one unit-specific covariate and that cannot be used for the Rasch model and other IRT models. Moreover, they state that the test cannot be applied when auto-regressive random effects are present. Regarding the

first aspect, we do not agree with Alonso et al. (2008) for two reasons: first, in models for item responses a covariate indeed exists and this is the indicator variable for the item, making our test easily usable, as we show by an empirical example; second, even if unit-specific covariates (which vary within the cluster) do not exist, they can be "artificially" created (e.g., in a longitudinal dataset, interactions of time-constant covariates with time dummies). Finally, our test is intended to be used when the assumption that the random-effects are time-constant is realistic. However, if this assumption is questionable, the proposed test can be used in connection with that proposed by Bartolucci et al. (2014c), the latter being specifically devoted to test the assumption that the random effects are time-constant rather than time-varying by comparing differently formulated conditional maximum likelihood estimators.

We conclude outlining that the applicability of the modified Hausman test is limited to certain finite-mixture GLLM based on a canonical link function. We also evaluated the performance of the test through simulation studies in case of linear and Poisson models, but we did not obtained interesting results. However, other cases to try are represented by survival data and by zero inflated Poisson models.

References

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Hoboken.

Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis*, 47:639–653.

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and computing*, 6:251–262.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalised linear models. *Biometrics*, 55:218–234.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademiai Kiado.

Alonso, A., Litière, S., and Molenberghs, G. (2008). A family of tests to detect misspecifications in the random-effects structure of generalized linear mixed models. *Computational statistics & data analysis*, 52:4474–4486.

Alonso, A. A., Litière, S., and Molenberghs, G. (2010). Testing for misspecification in generalized linear mixed models. *Biostatistics*, 11:771–786.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of Royal Statistical Society, Series B*, 32:283–301.

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of Royal Statistical Society, Series B*, 34:42–54.

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society, Series B*, pages 203–210.

Azzimonti, L., Ieva, F., and Paganoni, A. M. (2013). Nonlinear nonparametric mixed-effects models for unsupervised classification. *Computational Statistics*, 28:1549–1570.

Bacci, S., Bartolucci, F., and Gnaldi, M. (2014). A class of multidimensional latent class irt models for ordinal polytomous item responses. *Communication in Statistics - Theory and Methods*, 43:787–800.

Baetschmann, G., Staub, K. E., and Winkelmann, R. (2011). Consistent estimation of the fixed effects ordered logit model. Technical Report 5443, IZA.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*, 72:141–157.

Bartolucci, F., Bacci, S., and Pennoni, F. (2014a). Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 63:267–288.

Bartolucci, F., Bellio, R., Sartori, N., and Salvan, A. (2014b). Modified profile likelihood for fixed-effects panel data models. *Econometric Reviews*, in press.

Bartolucci, F., Belotti, F., and Peracchi, F. (2014c). Testing for time-invariant unobserved heterogeneity in generalized linear models for panel data. *Journal of Econometrics*, in press.

Bartolucci, F. and Forcina, A. (2005). Likelihood inference on the underlying structure of IRT models. *Psychometrika*, 70:31–43.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52:345–370.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the Inverse-Fisher information matrix. In Opitz, O., Lausen, B., and Klar, R., editors, *Information and Classification, Concepts, Methods and Applications*, pages 40–54. Springer, Berlin.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, 47:225–238.

De Boeck, P. and Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer-Verlag, New York.

Deb, P. (2001). A discrete random effects probit model with application to the demand for preventive care. *Health Economics*, 10:371–383.

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

Dias, J. (2006). Model selection for the binary latent class model: A Monte Carlo simulation. In Batagelj, V., Bock, H.-H., Ferligoj, A., and Žiberna, A., editors, *Data Science and Classification*, pages 91–99. Springer Berlin Heidelberg.

Goldstein, H. (2003). *Multilevel Statistical Models*. Arnold, London.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215–231.

Hagenaars, J. and McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge University Press, Cambridge, MA.

Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer Nijhoff, Boston.

Hausman, J. (1978). Specification tests in econometrics. *Econometrica*, 46:1251–1271.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, 55:688–698.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88:973–985.

Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric model for duration data. *Econometrica*, 52:271–320.

Heinzl, F. and Tutz, G. (2013). Clustering in linear mixed models with approximate Dirichlet process mixtures using EM algorithm. *Statistical Modelling*, 13:41–67.

Heiss, F. (2008). Sequential numerical integration in nonlinear state space models for microeconometric panel data. *Journal of Applied Econometrics*, 23:373–389.

Huang, G.-H. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69:5–32.

Huq, M. N. and Cleland, J. (1990). Bangladesh fertility survey, 1989. Technical report, Main Report.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307.

Hurvich, C. M. and Tsai, C.-L. (1993). A corrected Akaike information criterion for vector autoregressive model selection. *Journal of Time Series Analysis*, 14:271–279.

Jain, D. C., Vilcassim, N. J., and Chintagunta, P. K. (1994). A random-coefficients logit brand-choice model applied to panel data. *Journal of Business & Economic Statistics*, 12:317–328.

Juster, F. T. and Suzman, R. (1995). An overview of the health and retirement study. *The Journal of Human Resources*, 30:S7–S56.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906.

Kim, B.-D., Blattberg, R. C., and Rossi, P. E. (1995). Modeling the distribution of price sensitivity and implications for optimal retail pricing. *Journal of Business & Economic Statistics*, 13:291–303.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixture distribution. *Journal of the American Statistical Association*, 73:805–811.

Lange, N. and Ryan, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, pages 624–642.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

Lindsay, B., Clogg, C. C., and Grego, J. (1991). Semiparametric estimation in the rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86:96–107.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Annals of Statistics*, 11:86–94.

Litière, S., Alonso, A., and Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine*, 27:3125–3144.

Mazharul Islam, M. and Mahmud, M. (1995). Contraceptions among adolescents in Bangladesh. *Asia Pacific Population Journal*, 10:21–38.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42:109–142.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, 2nd Edition*. Chapman and Hall, CRC, London.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. Wiley.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer.

Nylund, K., Asparouhov, T., and Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14:535–569.

Pan, Z. and Lin, D. (2005). Goodness-of-fit methods for generalized linear mixed models. *Biometrics*, 61:1000–1009.

Pudney, S., Galassi, F. L., and Mealli, F. (1998). An econometric model of farm tenures in fifteenth-century Florence. *Economica*, 65:535–556.

Rabe-Hesketh, S., Pickles, A., and Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, 3:215–232.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Intitute for Educational Reserch, Copenhagen.

Ritz, C. (2004). Goodness-of-fit tests for mixed models. *Scandinavian journal of statistics*, 31:443–458.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, 17.

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

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Pychometrika*, 52:333–343.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling. Multilevel, Longitudinal and Structural Equation Models*. Chapman and Hall/CRC, London.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, pages 961–971.

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. CRC Press.

Tchetgen, E. J. and Coull, B. A. (2006). A diagnostic test for the mixing distribution in a generalised linear mixed model. *Biometrika*, 93:1003–1010.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91:217–221.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer.

Verbeke, G. and Molenberghs, G. (2013). The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models. *Biostatistics*, 14:477–490.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, 33:213–239.

Vijverberg, W. P. (2011). Testing for IIA with the Hausman-McFadden Test. IZA Discussion Papers 5826, Institute for the Study of Labor (IZA).

Waagepetersen, R. (2006). A simulation-based goodness-of-fit test for random effects in generalized linear mixed models. *Scandinavian journal of statistics*, 33:721–731.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25.

Yang, C.-C. and Yang, C.-C. (2007). Separating latent classes by information criteria. *Journal of Classification*, 24:183–203.

Appendix

The simulation study

In this section, we describe the setting and present the main results of a Monte Carlo study limited to models for binary responses. We first investigate the finite-sample properties of the proposed test in finite-mixture models for correct specification of the number of mixture components. Then, we study the power properties of the proposed test when the distribution of the random effects is misspecified, namely when the true distribution is continuous and in presence of correlation between the random effects and the regression covariates.

The simulation design

The simulation study is based on the random intercept model specified in Section 2 with a logit link function. We consider two scenarios: one refers to the longitudinal setting and the other one refers to the IRT setting. In the longitudinal design, the model is specified as follows

$$y^*_{ij} = \theta_i + x'_i\beta + z'_{it}\gamma + \varepsilon_{ij} \tag{17}$$

$$y_{ij} = \mathbf{I}(y^*_{ij} > 0) \quad \text{for } i = 1, \ldots, n \;\; j = 1, \ldots, J \tag{18}$$

where the distribution of the random effects α_i has $k_0 = 3$ support points $\left[-\sqrt{3/2}, 0, \sqrt{3/2}\right]$ with probabilities 0.25, 0.50, and 0.25 respectively. We also consider one observation-specific covariate z_{ij}, with $j = 1, \ldots, J_i$ denoting the time occasions and $J_i = J$ for $i = 1, \ldots, n$, generated as

$$\begin{aligned} z_{i0} &\sim N\left(0, \pi^2/3\right), \\ z_{ij} &= z_{i,j-1}\rho + u_{ij}, \\ u_{ij} &\sim N\left(0, \left(1-\rho^2\right)\pi^2/3\right), \end{aligned}$$

with $\rho = 0.5$. We also consider a cluster-specific covariate x_i following a standard normal distribution. Besides, the error terms ε_{ij} are i.i.d. with zero-mean logistic distribution with variance $\pi^2/3$, whereas γ and β are both scalars and equal 1. As outlined in Section 4, the Hausman test will compare only the estimators of γ, since the CML approach does not allow for the identification of cluster-specific effects. The Hausman test statistic will therefore be asymptotically distributed as a χ^2_1.

In the IRT scenario, model (17) with link a logit function simplifies in a Rasch model, as follows

$$\log \frac{p(y_{ij} = 1|\alpha_i, \boldsymbol{z}_{ij})}{p(y_{ij} = 0|\alpha_i, \boldsymbol{z}_{ij})} = \alpha_i - \boldsymbol{z}'_{ij}\boldsymbol{\gamma}.$$

where $j = 2, \ldots, J_i$ with $J_i = J$ for $i = 1, \ldots, n$ and $\boldsymbol{\gamma}$ is a $J-1$-dimensional vector of item difficulty parameters; these parameters are taken as equidistant points in the interval

$[-2, 2]$. Therefore, the Hausman test statistic will be asymptotically distributed as a χ^2_{J-1}. We repeat the experiment on the two models with different discrete distributions for α_i: we consider a shift in the original distribution, $\alpha_i \in \left[1 - \sqrt{3/2}, 1, 1 + \sqrt{3/2}\right]$ with probabilities 0.25, 0.50, and 0.25, and a strongly asymmetric distribution, $\alpha_i \in [-5, 0, 25]$ with probabilities 0.33, 0.50, and 0.17 respectively.

In each scenario, we compare the performance of the proposed test with that of standard selection criteria by estimating finite-mixture models under the assumption of k number of support points for α, with $k = 1, \ldots, 6$. More in detail, apart from AIC and BIC (see Section 4.2), we consider the following information criteria: Consistent AIC (CAIC; Bozdogan, 1987), AIC_3 (Bozdogan, 1993), HT-AIC (Hurvich and Tsai, 1989), AIC_c (Hurvich and Tsai, 1993), the adjusted CAIC (CAIC^*; Yang and Yang, 2007), and adjusted BIC (BIC^*; Sclove, 1987). Overall, they are based on the following indices:

$$
\begin{aligned}
\text{AIC} &= -2\,\hat{\ell}_M + 2\#\text{par}, \\
\text{BIC} &= -2\,\hat{\ell}_M + \#\text{par}\log(n), \\
\text{AIC}_3 &= -2\,\hat{\ell}_M + 3\#\text{par}, \\
\text{CAIC} &= -2\,\hat{\ell}_M + \#\text{par}(\log(n) + 1), \\
\text{HT-AIC} &= -2\,\hat{\ell}_M + 2\#\text{par} + \frac{2(\#\text{par} + 1)(\#\text{par} + 2)}{n - \#\text{par} - 2}, \\
\text{AIC}_c &= -2\,\hat{\ell}_M + 2\frac{\#\text{par}(\#\text{par} - 1)}{n - \#\text{par} - 1}, \\
\text{BIC}^* &= -2\,\hat{\ell}_M + \#\text{par}\log\frac{n+2}{24}, \\
\text{CAIC}^* &= -2\,\hat{\ell}_M + \#\text{par}\left(\log\frac{n+2}{24} + 1\right)
\end{aligned}
$$

with $\hat{\ell}_M$ denoting the maximum of log-likelihood and #par is the number of free parameters. As all these criteria consist in penalized versions of the maximum log-likelihood, the optimal number of latent classes is that corresponding to the minimum value of the corresponding index. In practice, we fit a given discrete GLMM for increasing values of k until the index does not start to increase. Then, we select the previous k as the optimal number of latent states, which guarantees the best compromise between goodness-of-fit and model parsimony.

The second part of our simulation study deals with possible misspecification of the random effect distribution. First, we analyze a case where the true distribution of α_i is continuous: the data are generated as above with the exception of the random effects which are now $\alpha_i \sim N(0, 3)$. Secondly, the analysis considers a case where the random effects are correlated with the regression covariates. In this scenario, α_i is generated starting from a Gaussian copula: we generate continuous random effects as $\alpha_i^* = \tau\bar{z}_i + w_i\sqrt{1 - \tau^2}$, where $\bar{z}_i = (1/J_i)\sum_{j=1}^{J_i} z_{ij}$ and $w_i \sim N(0, 1)$. We then obtain the discrete random effects α_i from α_i^* so that α_i has $k_0 = 3$ support points $\left[-\sqrt{3/2}, 0, \sqrt{3/2}\right]$ with probabilities

0.25, 0.50, and 0.25 respectively. The parameter τ controls the correlation between α_i and z_{ij} and we analyze the situations where $\tau = 0, 0.5, 0.8$.

The main results

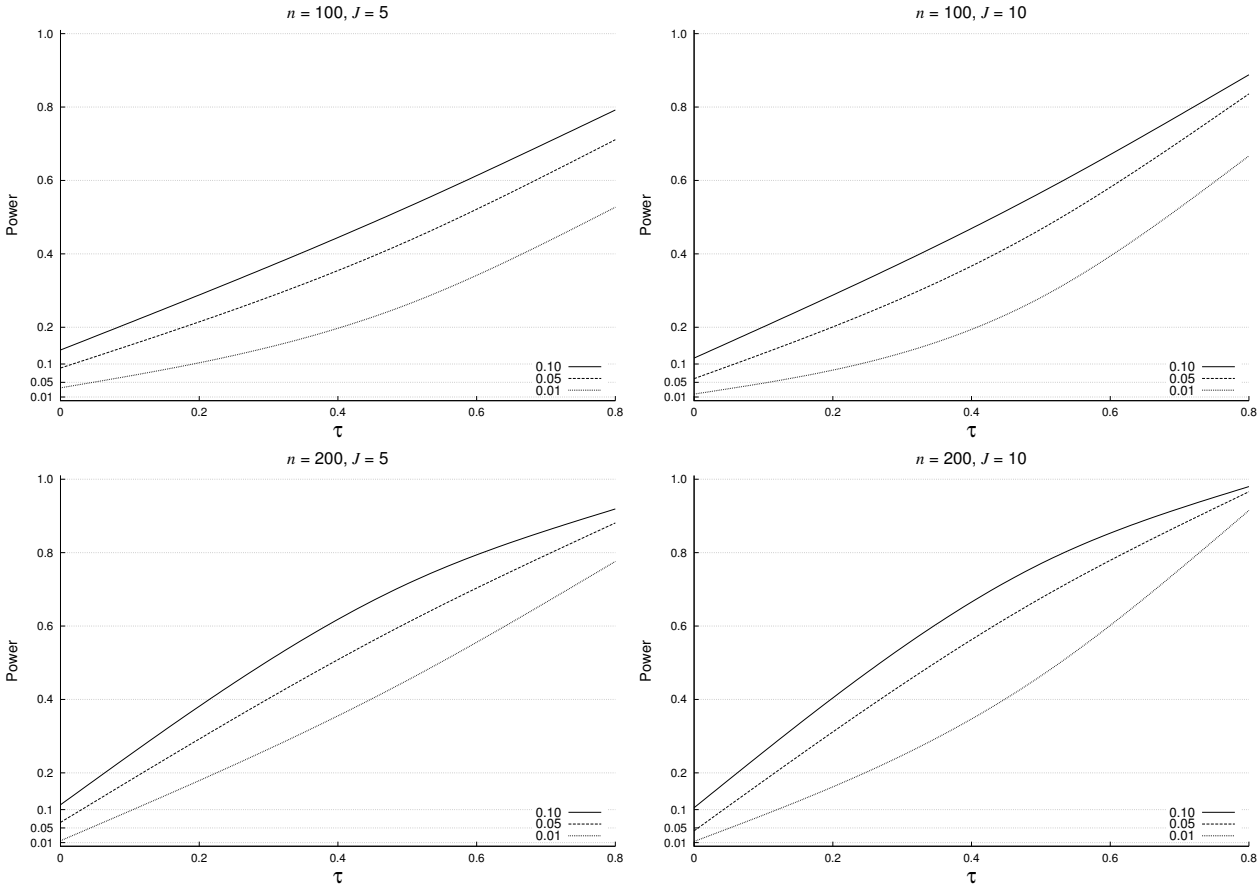
Tables 7- 12 summarize the values of the empirical size of test for binary responses models with longitudinal data and for IRT models considering the three different discrete distributions for α_i described in the previous section. Each experiment is repeated for $n = 500, 1000$ and $J = 5, 10$. The Hausman test compares $\hat{\boldsymbol{\gamma}}_M$, obtained by estimating a finite-mixture model for $k = 1, \ldots, 6$, with $\hat{\boldsymbol{\gamma}}_C$ obtained by CML. For each experiment, the tables also report the number of times (out of 1000 replications) the Hausman test does not reject the null hypothesis at k and compare these results with the number of times information criteria are minimized in k.

The results of our simulation study suggest that the proposed Hausman-type test behaves quite nicely as a test of correct specification in finite-mixture models. Tables 7-12 show that the empirical size of the proposed test reaches its nominal value with $k = 3$ for all the values of n and J considered. When the finite-mixture model is estimated under the assumption of $k = 1$, the proposed test exhibits high rejection rates increasing in both n and J. However, when symmetric discrete distributions for α_i are considered, the rejection rate for $k = 2$ is rather low and slowly increasing in n and J (Tables 7-10). Nevertheless, the standard selection criteria also seem to favor specifications with $k = 2$ support points. In contrast, Tables 11 and 12 show that the rejection rate with $k = 1, 2$ is almost 100% in every scenario, while the empirical size attains its nominal value when $k = 3$.

Tables 13 and 14 report the simulation results for a different misspecification of α_i, namely when the random effects are continuous. Table 13 shows that, with $n = 500$, $J = 5$, and $k = 2$, the Hausman test does not reject the null hypothesis in a considerable number of replications from 415 for the nominal size 10% to 841 for 1% and for most of the remaining replications it selects $k = 3$. With the exception of AIC_c, selection criteria tend instead to select a model with $k = 3$. Notice that BIC and the Hausman test at 1% present almost the opposite behavior. A similar situation occurs when $n = 1000$ and $J = 5$, while with $n = 500, 1000$ and $J = 10$ the proposed test tends to select $k = 3$, while the information criteria lean towards $k = 4$. In the case of the IRT parametrization, the same pattern is even clearer: with $n = 500$ and $J = 5$ the Hausman test selects a model with $k = 2$ in the majority of cases, while selection criteria tend to select models with a higher k (Table 14). For greater values of n and J selection criteria, with exception of BIC, select 4 or even 5 support points, while the proposed test favors a more parsimonious specification of the number of classes.

The major implication of these results is that, when the random effects are a continuous random variable, a situation that is likely to occur with real data, the Hausman test selects a model that produces a consistent estimator of $\boldsymbol{\gamma}$ regardless of the misspec-

Figure 1: Hausman-type specification test for endogeneity: longitudinal data scenario



ification in the distribution of the random effects. In addition, compared to standard model selection criteria, the proposed test chooses a more parsimonious specification of the number of support points especially for large values of J, which usually leads to a clearer interpretation of the results, especially when the aim is data classification or when the interest in on the regression parameters.

The last part of our Monte Carlo study investigates the power properties of the test when the hypothesis of independence between the random effects and the regression covariates is violated. The correlation is controlled by means of the parameter τ which varies between 0 and 0.8 (Section 7). The simulation results are displayed in Figure 1 for finite-mixture models for longitudinal binary data and in Figure 2 for the Rasch model. In both cases, we limit the analyses to sample sizes of $n = 100, 200$. With $\tau = 0$, the proposed test maintains its size properties in both models. The power of the test increases with the correlation τ and in the sample size, while increasing number of occasions J seems to only slightly affect the rejection rates.

Table 7: Hausman-type specification test: number of selections of k classes and empirical size, $k_0 = 3$ symmetric zero mean, binary data

							$n = 500$		$J = 5$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	298	0.631	445	0.511	748	0.233	0	70	1	113	0	0	1	10
2	574	0.100	484	0.052	229	0.013	901	927	953	885	905	379	958	974
3	22	0.103	10	0.057	4	0.018	99	3	46	2	95	510	41	16
4	4	0.104	7	0.059	3	0.020	0	0	0	0	0	71	0	0
5	20	0.105	10	0.059	4	0.020	0	0	0	0	0	29	0	0
6	52	0.107	32	0.059	9	0.021	0	0	0	0	0	11	0	0
							$n = 500$		$J = 10$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	1	0.991	9	0.978	77	0.916	0	0	0	0	0	0	0	0
2	809	0.123	900	0.055	899	0.013	541	952	699	974	552	45	703	823
3	76	0.101	38	0.049	8	0.012	456	48	299	26	445	744	295	177
4	10	0.101	3	0.049	4	0.012	3	0	2	0	3	143	2	0
5	13	0.108	11	0.052	5	0.016	0	0	0	0	0	50	0	0
6	51	0.114	25	0.055	3	0.017	0	0	0	0	0	18	0	0
							$n = 1000$		$J = 5$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	42	0.908	138	0.822	379	0.603	0	0	0	0	0	0	0	0
2	821	0.108	787	0.060	596	0.012	812	997	902	998	817	250	953	979
3	37	0.097	14	0.060	6	0.018	187	3	98	2	182	598	47	21
4	2	0.100	1	0.061	1	0.019	1	0	0	0	1	69	0	0
5	12	0.099	7	0.062	9	0.020	0	0	0	0	0	48	0	0
6	49	0.104	35	0.064	6	0.023	0	0	0	0	0	35	0	0
							$n = 1000$		$J = 10$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	754	0.181	876	0.096	961	0.024	286	873	444	914	290	9	559	692
3	135	0.099	68	0.052	24	0.011	706	127	554	86	702	688	441	308
4	9	0.101	4	0.052	3	0.013	8	0	2	0	8	202	0	0
5	8	0.106	2	0.054	1	0.014	0	0	0	0	0	64	0	0
6	50	0.110	24	0.054	10	0.016	0	0	0	0	0	37	0	0

The random effects $\alpha_i \in \left[-\sqrt{3/2}, 0, \sqrt{3/2}\right]$ with probabilities 0.25, 0.50, and 0.25 respectively. The test statistic $T_2 \xrightarrow{d} \chi^2_1$. The number of replications is 1000.

Table 8: Hausman-type specification test: number of selections of k classes and empirical size, $k_0 = 3$ symmetric zero mean, binary data, IRT model

							$n = 500$	$J = 5$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	565	0.392	727	0.241	926	0.049	8	193	29	281	9	0	30	65
2	315	0.094	204	0.045	43	0.013	929	806	943	719	936	491	943	925
3	4	0.111	5	0.060	3	0.026	63	1	28	0	55	404	27	10
4	2	0.114	4	0.063	2	0.028	0	0	0	0	0	63	0	0
5	19	0.119	14	0.066	13	0.031	0	0	0	0	0	27	0	0
6	25	0.120	18	0.065	7	0.031	0	0	0	0	0	15	0	0
							$n = 500$	$J = 10$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	147	0.839	290	0.700	597	0.396	0	0	0	0	0	0	0	0
2	735	0.083	649	0.038	391	0.003	668	986	815	994	692	164	818	913
3	11	0.100	10	0.047	2	0.009	331	14	184	6	307	667	181	86
4	3	0.105	2	0.049	2	0.010	1	0	1	0	1	113	1	1
5	12	0.105	7	0.051	3	0.010	0	0	0	0	0	42	0	0
6	25	0.109	16	0.054	3	0.012	0	0	0	0	0	14	0	0
							$n = 1000$	$J = 5$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	185	0.799	312	0.667	627	0.356	0	7	1	14	0	0	1	5
2	691	0.099	615	0.051	352	0.012	851	990	931	986	856	315	958	975
3	16	0.105	12	0.058	1	0.019	146	3	68	0	142	509	41	20
4	1	0.108	2	0.059	0	0.020	3	0	0	0	2	104	0	0
5	5	0.110	7	0.063	8	0.021	0	0	0	0	0	46	0	0
6	23	0.114	19	0.068	4	0.026	0	0	0	0	0	26	0	0
							$n = 1000$	$J = 10$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	1	0.999	1	0.998	21	0.978	0	0	0	0	0	0	0	0
2	887	0.086	935	0.038	960	0.010	422	947	613	965	435	41	710	807
3	18	0.090	11	0.046	2	0.012	568	53	385	35	555	695	290	193
4	3	0.092	4	0.050	3	0.015	10	0	2	0	10	192	0	0
5	3	0.095	4	0.053	1	0.018	0	0	0	0	0	53	0	0
6	17	0.093	17	0.053	5	0.017	0	0	0	0	0	19	0	0

The random effects $\alpha_i \in \left[-\sqrt{3/2}, 0, \sqrt{3/2}\right]$ with probabilities 0.25, 0.50, and 0.25 respectively. The test statistic $T_2 \xrightarrow{d} \chi^2_{J-1}$. The number of replications is 1000.

Table 9: Hausman-type specification test: number of selections of k classes and empirical size, $k_0 = 3$ symmetric unit mean, binary data

							$n = 500$		$J = 5$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	304	0.625	483	0.489	793	0.203	2	76	6	113	2	0	6	18
2	587	0.071	470	0.036	196	0.008	900	923	955	886	909	401	956	968
3	14	0.090	9	0.036	1	0.009	95	1	38	1	87	494	37	14
4	6	0.094	2	0.038	1	0.009	3	0	1	0	2	65	1	0
5	30	0.096	7	0.040	1	0.011	0	0	0	0	0	28	0	0
6	36	0.095	19	0.038	4	0.009	0	0	0	0	0	12	0	0
							$n = 500$		$J = 10$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	3	0.991	12	0.977	102	0.890	0	0	0	0	0	0	0	0
2	808	0.127	882	0.076	878	0.009	570	965	748	983	579	74	755	869
3	77	0.106	50	0.054	9	0.009	429	35	251	17	420	721	244	130
4	6	0.107	3	0.054	2	0.010	1	0	1	0	1	147	1	1
5	6	0.108	5	0.054	7	0.010	0	0	0	0	0	42	0	0
6	53	0.112	27	0.057	1	0.013	0	0	0	0	0	16	0	0
							$n = 1000$		$J = 5$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	53	0.893	150	0.814	421	0.564	0	0	0	0	0	0	0	0
2	819	0.102	779	0.054	560	0.010	840	999	913	1000	841	240	958	979
3	21	0.106	16	0.053	5	0.014	160	1	87	0	159	585	42	21
4	4	0.103	3	0.054	0	0.014	0	0	0	0	0	80	0	0
5	17	0.103	14	0.051	5	0.016	0	0	0	0	0	52	0	0
6	45	0.105	22	0.056	8	0.018	0	0	0	0	0	43	0	0
							$n = 1000$		$J = 10$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	0.999	0	0	0	0	0	0	0	0
2	802	0.145	887	0.079	973	0.017	284	876	450	922	288	13	554	700
3	101	0.090	61	0.047	14	0.010	698	124	547	78	697	664	444	300
4	5	0.092	3	0.049	0	0.012	18	0	3	0	15	222	2	0
5	7	0.093	5	0.050	4	0.014	0	0	0	0	0	65	0	0
6	45	0.092	28	0.049	6	0.013	0	0	0	0	0	36	0	0

The random effects $\alpha_i \in \left[1 - \sqrt{3/2}, 1, 1 + \sqrt{3/2}\right]$ with probabilities 0.25, 0.50, and 0.25 respectively. The test statistic $T_2 \xrightarrow{d} \chi^2_1$. The number of replications is 1000.

Table 10: Hausman-type specification test: number of selections of k classes and empirical size, $k_0 = 3$ symmetric unit mean, binary data, IRT model

							$n = 500$		$J = 5$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	572	0.391	743	0.229	942	0.039	7	202	26	307	7	0	28	70
2	322	0.081	196	0.049	36	0.010	924	796	940	693	934	483	939	915
3	1	0.100	5	0.053	2	0.019	69	2	34	0	59	420	33	15
4	2	0.104	4	0.055	1	0.020	0	0	0	0	0	58	0	0
5	22	0.106	7	0.057	9	0.022	0	0	0	0	0	25	0	0
6	16	0.106	17	0.056	7	0.022	0	0	0	0	0	14	0	0
							$n = 500$		$J = 10$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	73	0.921	155	0.839	436	0.560	0	0	0	0	0	0	0	0
2	821	0.078	796	0.031	551	0.004	645	967	789	988	674	127	794	892
3	8	0.089	2	0.043	3	0.007	352	33	210	12	324	718	205	108
4	4	0.093	3	0.045	2	0.010	3	0	1	0	2	119	1	0
5	9	0.097	6	0.046	2	0.011	0	0	0	0	0	27	0	0
6	22	0.095	12	0.045	4	0.011	0	0	0	0	0	9	0	0
							$n = 1000$		$J = 5$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	132	0.852	255	0.729	563	0.427	0	6	0	8	0	0	0	1
2	756	0.083	688	0.041	419	0.007	862	992	929	991	866	309	958	981
3	7	0.101	6	0.049	1	0.015	137	2	71	1	134	517	42	18
4	1	0.104	1	0.050	2	0.017	1	0	0	0	0	99	0	0
5	9	0.105	9	0.050	4	0.017	0	0	0	0	0	54	0	0
6	28	0.106	13	0.053	6	0.020	0	0	0	0	0	21	0	0
							$n = 1000$		$J = 10$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	1	0.998	8	0.992	0	0	0	0	0	0	0	0
2	876	0.092	928	0.043	971	0.009	382	929	556	959	398	25	686	789
3	34	0.085	15	0.050	8	0.008	609	71	442	41	593	690	314	211
4	5	0.088	5	0.053	5	0.010	8	0	2	0	8	214	0	0
5	2	0.088	2	0.053	0	0.012	1	0	0	0	1	54	0	0
6	20	0.092	20	0.058	2	0.017	0	0	0	0	0	17	0	0

The random effects $\alpha_i \in \left[1 - \sqrt{3/2}, 1, 1 + \sqrt{3/2}\right]$ with probabilities 0.25, 0.50, and 0.25 respectively. The test statistic $T_2 \xrightarrow{d} \chi^2_{J-1}$. The number of replications is 1000.

Table 11: Hausman-type specification test: number of selections of k classes and empirical size, $k_0 = 3$ asymmetric, binary data

							$n = 500$	$J = 5$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC^*	$CAIC^*$
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	2	0.993	4	0.992	9	0.987	0	0	0	0	0	0	0	0
3	731	0.094	792	0.047	839	0.013	958	1000	984	1000	961	458	985	993
4	144	0.144	131	0.103	121	0.071	39	0	15	0	36	111	14	6
5	42	0.153	30	0.108	18	0.072	3	0	1	0	3	205	1	1
6	20	0.146	9	0.100	2	0.055	0	0	0	0	0	226	0	0
							$n = 500$	$J = 10$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC^*	$CAIC^*$
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	0	1.000	1	0.999	1	0.999	0	0	0	0	0	0	0	0
3	719	0.120	793	0.049	847	0.011	972	999	995	1000	976	451	995	998
4	137	0.150	129	0.098	131	0.061	22	1	5	0	18	113	5	2
5	43	0.169	29	0.110	9	0.060	6	0	0	0	6	224	0	0
6	17	0.171	14	0.108	4	0.061	0	0	0	0	0	212	0	0
							$n = 1000$	$J = 5$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC^*	$CAIC^*$
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	1	0.998	4	0.996	11	0.987	0	0	0	0	0	0	0	0
3	769	0.096	826	0.057	886	0.014	979	1000	992	1000	980	413	998	1000
4	86	0.145	97	0.087	80	0.044	21	0	8	0	20	111	2	0
5	36	0.159	17	0.099	7	0.048	0	0	0	0	0	212	0	0
6	33	0.158	11	0.101	3	0.051	0	0	0	0	0	264	0	0
							$n = 1000$	$J = 10$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC^*	$CAIC^*$
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
3	728	0.106	786	0.061	851	0.011	970	1000	994	1000	973	365	998	1000
4	144	0.163	137	0.112	133	0.059	23	0	5	0	21	100	2	0
5	22	0.161	14	0.117	4	0.059	5	0	1	0	4	271	0	0
6	29	0.153	19	0.108	5	0.055	2	0	0	0	2	264	0	0

The random effects $\alpha_i \in [-5, 0, 25]$ with probabilities 0.33, 0.50, and 0.17 respectively. The test statistic $T_2 \xrightarrow{d} \chi^2_1$. The number of replications is 1000.

Table 12: Hausman-type specification test: number of selections of k classes and empirical size, $k_0 = 3$ asymmetric, binary data, IRT model

							$n = 500$		$J = 5$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC^*	$CAIC^*$
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	0	1.000	1	0.999	10	0.982	0	0	0	0	0	0	0	0
3	593	0.102	636	0.066	673	0.018	959	999	984	1000	963	579	984	991
4	258	0.184	266	0.149	263	0.107	35	1	16	0	34	145	16	9
5	48	0.201	36	0.149	37	0.106	5	0	0	0	3	134	0	0
6	10	0.261	4	0.214	7	0.175	1	0	0	0	0	142	0	0
							$n = 500$		$J = 10$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC^*	$CAIC^*$
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
3	703	0.108	752	0.063	795	0.020	975	1000	992	1000	979	585	992	999
4	158	0.180	165	0.132	169	0.091	24	0	8	0	21	151	8	1
5	20	0.183	15	0.130	12	0.087	1	0	0	0	0	115	0	0
6	12	0.191	6	0.136	4	0.087	0	0	0	0	0	149	0	0
							$n = 1000$		$J = 5$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC^*	$CAIC^*$
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
3	662	0.114	711	0.058	754	0.016	979	1000	991	1000	980	520	996	997
4	196	0.179	201	0.126	208	0.086	20	0	9	0	19	131	4	3
5	25	0.169	27	0.118	17	0.076	1	0	0	0	1	144	0	0
6	13	0.239	7	0.190	6	0.142	0	0	0	0	0	205	0	0
							$n = 1000$		$J = 10$					
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC^*	$CAIC^*$
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
3	756	0.103	797	0.052	846	0.018	974	1000	990	1000	975	525	998	999
4	120	0.159	124	0.110	125	0.072	24	0	10	0	23	150	2	1
5	13	0.160	19	0.118	11	0.067	2	0	0	0	2	118	0	0
6	12	0.156	10	0.112	0	0.062	0	0	0	0	0	207	0	0

The random effects $\alpha_i \in [-5, 0, 25]$ with probabilities 0.33, 0.50, and 0.17 respectively. The test statistic $T_2 \xrightarrow{d} \chi^2_{J-1}$. The number of replications is 1000.

Table 13: Hausman-type specification test: number of selections of k classes and empirical size, continuous random effects, binary data

							$n = 500$	$J = 5$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC^*	$CAIC^*$
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	2	0.995	0	0	0	0	0	0	0	0
2	415	0.467	597	0.328	841	0.115	5	152	17	223	5	0	17	45
3	398	0.101	297	0.048	122	0.010	627	815	764	757	635	124	774	831
4	59	0.130	40	0.069	19	0.022	355	33	216	20	347	550	206	124
5	64	0.145	35	0.087	11	0.036	13	0	3	0	13	249	3	0
6	40	0.145	20	0.085	5	0.033	0	0	0	0	0	77	0	0
							$n = 500$	$J = 10$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC^*	$CAIC^*$
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	4	0.987	18	0.967	143	0.850	0	0	0	0	0	0	0	0
3	572	0.352	703	0.233	772	0.071	21	295	66	391	26	0	67	129
4	298	0.103	219	0.047	75	0.008	593	663	710	582	608	134	713	740
5	70	0.095	36	0.053	7	0.014	360	42	211	27	343	543	207	128
6	25	0.096	13	0.056	0	0.015	26	0	13	0	23	323	13	3
							$n = 1000$	$J = 5$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC^*	$CAIC^*$
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	156	0.786	292	0.671	603	0.379	0	4	0	11	0	0	0	1
3	636	0.145	603	0.065	362	0.012	257	859	426	896	266	12	550	690
4	106	0.089	51	0.043	22	0.010	687	137	557	93	681	444	442	306
5	56	0.100	32	0.058	9	0.023	53	0	17	0	52	377	8	3
6	22	0.104	10	0.064	2	0.027	3	0	0	0	1	167	0	0
							$n = 1000$	$J = 10$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC^*	$CAIC^*$
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	0	1.000	0	1.000	2	0.998	0	0	0	0	0	0	0	0
3	265	0.673	447	0.527	720	0.269	0	37	1	70	0	0	2	7
4	547	0.154	465	0.072	258	0.018	221	784	378	808	228	23	501	612
5	125	0.091	61	0.039	15	0.007	647	176	563	119	646	415	462	366
6	31	0.100	16	0.041	2	0.012	132	3	58	3	126	562	35	15

The random effects are $\alpha_i \sim N(0, 3)$. The test statistic $T_2 \xrightarrow{d} \chi^2_1$. The number of replications is 1000.

Table 14: Hausman-type specification test: number of selections of k classes and empirical size, continuous random effect, binary data, IRT model

							$n = 500$	$J = 5$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	9	0.991	0	0	0	0	0	0	0	0
2	730	0.194	831	0.100	931	0.009	33	328	86	420	38	0	87	151
3	137	0.085	81	0.044	26	0.005	699	654	783	570	714	201	788	782
4	22	0.111	19	0.073	10	0.028	260	18	131	10	241	551	125	67
5	31	0.119	26	0.080	19	0.035	7	0	0	0	6	189	0	0
6	33	0.132	24	0.091	5	0.046	1	0	0	0	1	59	0	0
							$n = 500$	$J = 10$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	299	0.677	471	0.511	777	0.215	0	0	0	1	0	0	0	0
3	561	0.091	466	0.037	212	0.004	55	468	124	554	61	2	129	228
4	59	0.068	29	0.031	7	0.004	680	517	734	437	694	221	736	704
5	26	0.087	9	0.040	0	0.010	245	15	140	8	231	547	133	68
6	17	0.092	13	0.044	2	0.011	20	0	2	0	14	230	2	0
							$n = 1000$	$J = 5$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	460	0.497	635	0.324	875	0.093	1	38	3	61	1	0	5	11
3	407	0.088	292	0.034	92	0.006	449	892	606	891	457	47	714	820
4	32	0.095	17	0.050	13	0.019	523	70	383	48	517	515	277	168
5	23	0.109	25	0.063	15	0.030	26	0	8	0	24	317	4	1
6	25	0.116	15	0.070	4	0.036	1	0	0	0	1	121	0	0
							$n = 1000$	$J = 10$						
	Hausman test						AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
	10%		5%		1%									
k	sel.	e. s.	sel.	e. s.	sel.	e. s.								
1	0	1.000	0	1.000	0	1.000	0	0	0	0	0	0	0	0
2	6	0.992	20	0.980	120	0.880	0	0	0	0	0	0	0	0
3	756	0.192	846	0.110	848	0.028	0	90	4	146	0	0	14	20
4	142	0.072	80	0.034	26	0.006	408	838	584	810	424	38	684	764
5	34	0.084	20	0.048	2	0.004	532	72	392	44	518	526	292	212
6	20	0.092	8	0.056	0	0.008	60	0	20	0	58	436	10	4

The random effects are $\alpha_i \sim N(0, 3)$. The test statistic $T_2 \xrightarrow{d} \chi^2_{J-1}$. The number of replications is 1000.

Figure 2: Hausman-type specification test for endogeneity: Rasch model scenario