Post-print version of the paper published in Multivariate Behavioral Research;
https://doi.org/10.1080/00273171.2017.1361803

Evaluation of student performance through a multidimensional finite mixture IRT model

Silvia Bacci* Francesco Bartolucci* Leonardo Grilli†

Carla Rampichini†

Abstract

In the Italian academic system, a student can enroll for an exam immediately after the end of the teaching period or can postpone it; in this second case the exam result is missing. We propose an approach for the evaluation of a student performance *along the course of study*, accounting also for non-attempted exams. The approach is based on an Item Response Theory model that includes two discrete latent variables representing student performance and priority in selecting the exams to take. We explicitly account for non-ignorable missing observations as the indicators of attempted exams also contribute to measure the performance (within-item multidimensionality). The model also allows for individual covariates in its structural part.

Keywords: Education, Latent class model, Non-ignorable missing responses, Ordinal responses, Within-item multidimensionality

*Department of Economics, University of Perugia (IT).

†Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence (IT).

1 Introduction

The Italian university system suffers from high drop-out rates and slow progression (OECD, 2016; Gitto et al., 2015). An in-depth analysis of university student careers is necessary to improve the system through planning courses and student guidance. In particular, the performance of students at the end of the first academic year is highly predictive of the final outcome; its evaluation can consider an overall measure, such as the total number of gained credits (Grilli et al., 2016), or it can exploit a multivariate analysis of first-year exams (Bertaccini et al., 2013). Specifically, Bertaccini et al. (2013) proposed an item response multiple indicators multiple causes model, where each compulsory first-year exam corresponds to a binary item. The success corresponds to passing the exam within the first year. However, this model does not account for two aspects which are common in the Italian university system: (i) courses with a large number of students are divided into parallel groups; (ii) a student can take first-year exams in any order and not necessarily during the first year. The issue of parallel groups may be addressed by introducing a distinct item for each group. The second aspect requires to extend the model to account for student decision to take an exam during the first year or to postpone it. Thus, for a certain student at a given time point the result of an exam can be missing for two reasons: (i) the item corresponding to that exam is not due since the student belongs to another group; (ii) the exam is due, but the student decided to postpone it after the first year of study. The first kind of missing data is structural and thus it can be assumed to be ignorable, whereas the second kind of missing data is potentially informative, as it could be related to student performance. Indeed, the data analyzed in this paper show

that the number of postponed exams is highly (and negatively) correlated with student performance on the attempted exams, as measured by the rate of passed exams and their average grade.

In the context at issue, we aim at developing a model-based approach that can help decision makers in planning the degree programs and organizing student tutoring. In particular, the model should be useful for: (i) evaluating student performance by considering, for first-year exams, both the decision to take the exam and the result if the exam is taken; (ii) characterizing first-year exams in terms of their difficulty and discrimination power; (iii) comparing parallel groups of each exam to check whether they behave similarly; (iv) clustering students into homogeneous classes of performance, controlling for observed student characteristics.

We illustrate the proposed model through the analysis of student careers at the University of Florence, considering freshmen of the academic year 2013/2014 who are enrolled in the degree programs *Business* and *Economics*. These programs share six compulsory courses in the first year. Given the large number of freshmen, each course is organized in four parallel groups on the basis of the first letter of the student's surname. This entails a set of $6 \times 4 = 24$ items, thus generating the structural missing values mentioned above. A student can take exams in any order during the examination sessions of the academic year (January, February, June, July, September, December). To take an exam in a given examination session, the student has to enroll via a web procedure, which is also used to record the exam result. Most freshmen cannot manage the entire workload, so they decide to postpone one or more exams to the following academic years. A student postponing an exam never enrolls for that exam during the first year, thus generating a missing value

that is potentially informative. Therefore, postponing an exam out of the observation period generates Missing Not At Random (MNAR) values that, by definition, are not ignorable (Little & Rubin, 2002; Mealli & Rubin, 2015).

The model we develop is of Item Response Theory (IRT) type (Hambleton & Swaminathan, 1985; Van der Linden & Hambleton, 1997; Bartolucci et al., 2015) and accounts for both structural and non-ignorable missing values. There exist several approaches in the IRT literature to handle non-ignorable missingness; for a detailed review see Rose et al. (2010). Common practice consists in ignoring omitted item responses or in recoding them as wrong or partially correct responses. Rose et al. (2010), through a complex simulation study, concluded that both approaches are quite simplistic and give rise to strongly biased item parameter estimates. Valid alternatives explicitly account for the non-ignorability of the missingness process and rely on suitable extensions of IRT models, such as latent regression models (Mislevy, 1987; Zwinderman, 1991) and multidimensional IRT models (Reckase, 2010). In the latent regression IRT approach, the latent trait is regressed on the observed response rate, which is used as a proxy for the propensity to answer. On the contrary, in the multidimensional IRT approach a new latent trait for the propensity to answer and a response indicator (observed/missing) for each item are introduced. This new latent trait is usually correlated with that measured by the test items and both latent traits may affect the values of the item response indicators. Therefore, in the latent regression approach the non-ignorable missingness is represented through a covariate contributing *to explain* the latent trait, whereas in the multidimensional approach the presence of missing item responses provides additional information *to measure* the latent trait. Despite the two approaches have different perspectives, they typically yield

similar parameter estimates, as illustrated by Rose et al. (2010).

In our setting, we consider the decision to take an exam as an indicator of student performance, thus as an important outcome *per se*. Therefore, we adopt the multidimensional IRT approach, whose use for the treatment of MNAR responses was originally proposed by Lord (1983) and further developed in the parametric setting by Holman & Glas (2005). In particular, we assume that student overall behavior is driven by two latent variables. The first latent variable affects both the enrollment decision and the exam result, thus representing student performance that is of main interest; the second latent variable only affects the enrollment decision, thus representing student priority in taking the exams. Exams are treated as items with ordinal responses measuring the latent variable representing student performance, whereas binary indicators of exam enrollment also measure the latent variable for the exam priority. This structure, where a set of items contributes to measure more latent variables, is known as *within-item multidimensionality* (Adams et al., 1997). For a review of potentialities of within-item multidimensional IRT models see Cai (2010) and the references therein.

The resulting multidimensional IRT model with continuous latent traits is also known as *two-tier model*, which represents a particular within-item multidimensional formulation with each item loading on at most two latent variables (Gibbons & Hedeker, 1992; Cai, 2010). Differently from the usual approach, we adopt a finite mixture specification, namely we assume that the latent traits have a discrete rather than continuous distribution. This choice increases the flexibility of the model as it avoids parametric assumptions on the distribution of such traits, and it results in a semi-parametric maximum likelihood estimator in the sense of Lindsay et al. (1991). We denote the proposed model as LC-IRT

model, where LC stands for Latent Class (Lazarsfeld & Henry, 1968; Goodman, 1974) as we interpret the support points of the discrete distribution as latent classes. This approach allows us to cluster individuals into homogeneous groups and it has a practical utility in all those applications where decision-makers are interested in addressing individuals sharing the same characteristics to the same treatment. The weights of the mixture correspond to the class membership probabilities. We let these probabilities depend on individual characteristics.

The proposed model extends the multidimensional finite mixture model of Bacci & Bartolucci (2015) in two directions. Firstly, it allows for mixed items (i.e., binary and ordinal) instead of just binary items: such an extension is needed to fully exploit items with a varying number of ordered categories used in many fields, like customer satisfaction and health assessment; see, for instance, the Short Form-36 (SF-36) and Short Form-12 (SF-12) questionnaires used to measure the health-related quality of life (Stewart & Ware, 1992; Ware & Sherbourne, 1992). The second extension accounts for structural missing values, in addition to informative ones: this feature allows us to face situations where unequal sets of items are proposed to different groups of subjects or to the same subjects at different occasions. These situations are common in educational large-scale assessment surveys, like the Programme for International Student Assessment (PISA; OECD, 2014), as well as in the longitudinal health assessment setting, like the repeated mild cognitive impairment assessment through original Montreal Cognitive Assessment (MoCA) test and its alternate versions (Smith et al., 2007; Lebedeva et al., 2016), where items are administered according to a rotating scheme to avoid cheating and memory bias, respectively.

The general version of the proposed model can handle a wide range of datasets involving the measurement of multiple latent traits. For example, it may properly manage longitudinal item response data, when a given set of items is administered to the same individuals at several time occasions, as well as the measurement of latent traits representing specific sub-domains, as often happens in the educational and health settings (e.g., quality of life is usually decomposed in physical domains and psychological domains).

The procedures to estimate the proposed multidimensional LC-IRT model are implemented in the `R` package `MLCIRTwithin` (Bartolucci & Bacci, 2016; Bacci & Bartolucci, 2016), freely downloadable from `http://CRAN.R-project.org/package=MLCIRTwithin`. The package has been recently updated to account for the new features of the proposed model, namely ordinal responses and structural missing values.

The remainder of the paper is organized as follows. Section 2 describes the data about freshmen at the University of Florence. Section 3 illustrates the general multidimensional LC-IRT model and Section 4 provides details on its estimation. Section 5 is devoted to the application; in particular, this section describes model selection, including tests for the ignorability of the missing data mechanism and for the homogeneity of groups of the same academic course. Finally, in Section 6 we discuss the main features of the proposed approach and possible developments.

2 Data description

The data set used for the proposed analysis is obtained from the administrative archive on student careers. This archive is referred to the freshmen of the academic year 2013/2014 enrolled in the degree programs *Business* ("Economia Aziendale") and *Economics* ("Econo-

mia e Commercio") of the University of Florence.

The data set includes background characteristics of students and their careers until December 2014. The first year entails six compulsory courses, three in the first semester and three in the second semester. All courses have parallel classes with distinct teachers, according to the first letter of the student's surname. Five courses are common to the two degree programs (Business and Economics), and are divided in four groups (A-C, D-L, M-P, Q-Z), while the course Management differs between the two degree programs, and students are split in two groups for each program (A-L, M-Z).

Students can take the exams in any order, not necessarily at the end of the corresponding course. The exams of the first-semester courses can be taken in any of the six sessions from January to December, while the exams of the second-semester courses can be taken in any of the four sessions from June to December. In order to take the exam in the chosen session, the student has to enroll online. If a student decides to postpone an exam to the next academic year, the enrollment record for that exam is empty.

In the analysis we consider the 861 freshmen who enrolled for at least one exam until December 2014 (89% of the freshmen). For each student, the data set contains information on the number of enrollments for each of the six exams, alongside with their outcomes. Passed exams are scored with integer grades ranging from 18 to 30, plus "30 with honors". For each exam (merging groups), Table 1 reports the percentage of students who enrolled for at least one of the six sessions of 2014 (*enrollment rate*), the distribution of the outcome for students enrolled at least once, considering the best outcome if the exam is repeated, and the percentage of students who passed the exam by December 2014 (*passing rate*), both conditional on enrollment and overall. The *overall passing rate* is obtained as the

product of the *enrollment rate* by the *conditional passing rate*.

[Table 1 about here.]

Table 1 highlights the large variability of student performance across the courses. It is worth noting that the overall passing rate may result from markedly different patterns. For example, the overall passing rate for Accounting is higher than for Law (53.8% versus 25.6%), which is mainly due to different enrollment rates (93.5% versus 48.3%), whereas the conditional passing rates are similar (57.5% versus 52.9%). On the contrary, the higher overall passing rate for Statistics with respect to Mathematics (40.4% versus 21.1%) is mainly due to different conditional passing rates (60.3% versus 34.2%), while being the enrollment rates similar (67.0% versus 67.8%).

Table 2 shows the performance of freshmen by gender, High School (HS) type, HS grade (ranging from 60 to 100), late matriculation, degree program, and course group. The table reports the average number of attempted exams (enrolled at least once) and the number of passed exams.

[Table 2 about here.]

Considering the six compulsory courses, on average students attempted 3.8 exams and passed 2.2 exams, with students from Scientific high schools or with a high HS grade (greater than 80 out of 100) performing better. On the contrary, late matriculated students perform worse in terms of both attempted and passed exams.

3 Model formulation

A distinctive feature of the case study under consideration is represented by missing observations on exam results, which could reflect student performance. Indeed, we expect that the tendency to attempt a certain exam in a given session is higher for students with better performance so that exam results are not missing at random.

In general, data are Missing At Random (MAR) if the conditional distribution of the response indicator, given observed and unobserved data, is the same whatever the unobserved data for all the parameter values (see Mealli & Rubin, 2015, Definition 1). If this condition does not hold, data are MNAR (Mealli & Rubin, 2015, Definition 5), thus the missingness mechanism is non-ignorable and it should be explicitly modeled to avoid incorrect inferential conclusions.

In the statistical literature, different approaches have been proposed to deal with situations of MNAR data, including: (i) the *selection approach* (Diggle & Kenward, 1994), in which a model is specified for the distribution of the complete (i.e., observed and unobserved) data and the conditional distribution of the missing indicators, given these data; (ii) the *pattern-mixture approach* (Little, 1993), in which a model is formulated for the marginal distribution of the missing indicators and the conditional distribution of the complete data, given these indicators; (iii) the *shared-parameter approach* (Wu & Carroll, 1988), which introduces a latent variable to capture the association between the observed responses and the missing process. As mentioned in Section 1, the multidimensional IRT approach may be exploited to formulate a shared-parameter model that may also be seen as a finite mixture Structural Equation Model (SEM); this approach have been developed

by Bacci & Bartolucci (2015) in particular. See Jedidi et al. (1997), Dolan & van der Maas (1998), and Arminger et al. (1999) for details on finite mixture SEMs.

It is important to recall that the model of Bacci & Bartolucci (2015) is characterized by a set of multiple equations that define the relationships among latent variables and observed item responses, missingness indicators, and individual covariates. The resulting model is a multidimensional LC-IRT model (Bartolucci, 2007; von Davier, 2008; Bacci et al., 2014) allowing for within-item multidimensionality (Adams et al., 1997) in which certain items measure more latent traits. Here we propose an extension of this model accounting for ordinal item responses and structural missingness (in addition to potentially non-ignorable missingness).

Considering an individual randomly drawn from the population of interest, let Y_j be the ordinal response to item $j = 1, \ldots, m$, where the response categories are labelled from 1 to L. A missing response is denoted with $Y_j = \text{NA}$ (NA stands for "Not Available"); we recall that two types of missing response are possible: (i) item j is not due by design (structural missing, thus ignorable); (ii) item j is due but it is skipped (potentially non-ignorable missing). The two types of missing data are distinguished by the response indicator R_j, assuming value NA if item j is not due, value 0 if item j is skipped; moreover, $R_j = 1$ if item j is answered. Note that the total number of items is $2m$, namely m test items further to m response indicators.

The item responses Y_j, along with the response indicators R_j, contribute to measure two latent traits, assumed to be independent given a set of exogenous individual covariates denoted by $\boldsymbol{X} = (X_1, \ldots, X_C)'$. The first latent trait is described by a multidimensional latent variable $\boldsymbol{U} = (U_1, \ldots, U_S)'$, representing the abilities measured by the

test items. The second latent trait, corresponding to the multidimensional latent variable $\boldsymbol{V} = (V_1, \ldots, V_T)'$, represents individual preferences in choosing the test items to answer (i.e., the exam to take) or to skip. This structure is represented in the path diagram of Figure 1, which refers to the special case of our application (Section 5), where both latent traits have a single component ($S = T = 1$).

[Figure 1 about here.]

In the following, we assume that $\boldsymbol{U}$ and $\boldsymbol{V}$ have a discrete distribution with k_U vectors of support points $\boldsymbol{u}_{h_U} = (u_{1h_U}, \ldots, u_{Sh_U})'$, $h_U = 1, \ldots, k_U$, and k_V vectors of support points $\boldsymbol{v}_{h_V} = (v_{1h_V}, \ldots, v_{Th_V})'$, $h_V = 1, \ldots, k_V$, respectively. This specification allows us to cluster individuals into latent classes that are homogeneous with respect to the latent traits. Note that if $k_V = 1$ then latent trait $\boldsymbol{V}$ is ruled out and response indicators R_j contribute only to the assessment of $\boldsymbol{U}$, as in Harel & Schafer (2009). Besides, in the spirit of concomitant variable LC models (Dayton & Macready, 1988; Formann, 2007), we allow the membership probabilities of the latent classes to depend on observed covariates through a multinomial logit model (see also Bacci & Bartolucci, 2015):

$$\log \frac{\lambda_{h_U}(\boldsymbol{x})}{\lambda_1(\boldsymbol{x})} = \boldsymbol{x}'\boldsymbol{\phi}_{h_U}, \quad h_U = 2, \ldots, k_U, \tag{1}$$

$$\log \frac{\pi_{h_V}(\boldsymbol{x})}{\pi_1(\boldsymbol{x})} = \boldsymbol{x}'\boldsymbol{\psi}_{h_V}, \quad h_V = 2, \ldots, k_V, \tag{2}$$

with $\lambda_{h_U}(\boldsymbol{x}) = Pr(\boldsymbol{U} = \boldsymbol{u}_{h_U} | \boldsymbol{X} = \boldsymbol{x})$ and $\pi_{h_V}(\boldsymbol{x}) = Pr(\boldsymbol{V} = \boldsymbol{v}_{h_V} | \boldsymbol{X} = \boldsymbol{x})$, where the covariate vector $\boldsymbol{x}$ includes a constant term, that is, $\boldsymbol{x} = (1, x_1, \ldots, x_C)'$. The vectors of coefficients $\boldsymbol{\phi}_{h_U} = (\phi_{h_U 1}, \ldots, \phi_{h_U C})'$ and $\boldsymbol{\psi}_{h_V} = (\psi_{h_V 1}, \ldots, \psi_{h_V C})'$ represent the effects of the covariates on the reference category logits.

The relationships among the latent variables in $\boldsymbol{U}$ and $\boldsymbol{V}$ and the item responses $Y_1, \ldots, Y_m$ and the response indicators $R_1, \ldots, R_m$ are described by the measurement part of the model, specified as a multidimensional LC-IRT model (for details, see Bacci et al., 2014). The proposed model is an extension of the model of Bacci et al. (2014), in that the multidimensional model structure is completely general, in the sense that it allows each indicator to measure one component in $\boldsymbol{U}$ and one component in $\boldsymbol{V}$ (within-item multidimensionality).

Let $q_{h_U h_V, j} = Pr(R_j = 1 | \boldsymbol{U} = \boldsymbol{u}_{h_U}, \boldsymbol{V} = \boldsymbol{v}_{h_V})$ denote the probability of answering item j conditionally on $\boldsymbol{U}$ and $\boldsymbol{V}$ and let $p_{h_U, jy} = Pr(Y_j \geq y | \boldsymbol{U} = \boldsymbol{u}_{h_U})$ denote the probability that the answer to item Y_j is y or higher ($y = 2, \ldots, L$), conditionally on the latent trait $\boldsymbol{U}$. To select the components of $\boldsymbol{U}$ and $\boldsymbol{V}$ entering the probabilities $q_{h_U h_V, j}$ and $p_{h_U, jy}$, we introduce two sets of indicators z_{Usj} and z_{Vtj}, equal to 1 if item j measures components U_s and V_t, respectively. Then, we specify a multidimensional LC Two-Parameter Logistic (2PL) model (Bartolucci, 2007) for the probability of answering item j:

$$\log \frac{q_{h_U h_V, j}}{1 - q_{h_U h_V, j}} = \gamma_{Uj} \sum_{s=1}^{S} z_{Usj} u_{s h_U} + \gamma_{Vj} \sum_{t=1}^{T} z_{Vtj} v_{t h_V} - \delta_j, \tag{3}$$

where δ_j can be interpreted as the difficulty to answer item j, as higher values of δ_j reduces the probability to answer the item; moreover, γ_{Uj} and γ_{Vj} are discrimination parameters, measuring the effects of the latent traits $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively, on the probability to answer the item. Note that, when $\gamma_{Uj} = 0$ for all items, the probabilities of answering the items do not depend on the latent variables in $\boldsymbol{U}$, thus the missingness process is ignorable (see the ignorability test described in Section 5.3). Moreover, the ordinal item

responses Y_j are modeled through a graded response parameterization (Samejima, 1969):

$$\log \frac{p_{h_U,jy}}{1-p_{h_U,jy}} = \alpha_j \sum_{s=1}^{S} z_{Usj} u_{sh_U} - \beta_{jy}, \quad y = 2, \ldots, L, \tag{4}$$

where β_{jy} is specific of item j and category y and it may be interpreted as a difficulty parameter, since higher values of β_{jy} ($y = 2, \ldots, L$) push the probability distribution of the item towards the bottom of the scale. On the other hand, α_j is a discrimination parameter, measuring the effect of variables in $\boldsymbol{U}$ on the probability distribution of the item.

In order to ensure the identification of the proposed within-item multidimensional model, two necessary conditions must be satisfied. The first one requires that at least one item loads only on one of the components of $\boldsymbol{U}$ or only on one of the components of $\boldsymbol{V}$. In our specific context related to the treatment of non-ignorable missingness, this condition is always satisfied, as item responses $Y_1, \ldots, Y_m$ measure only latent vector $\boldsymbol{U}$. Second, suitable constraints on the item parameters are required. In particular, we constrain one of the discrimination parameters (γ_{Uj}, γ_{Vj}, α_j) to be equal to 1 and one difficulty parameter (δ_j, β_{jy}) to be equal to 0 for every component of each latent variable. Generally speaking, any item may be chosen to be constrained, paying attention to select a different item for each dimension. In equation (3) we constrain $\gamma_{Vj_t} = 1$ and $\delta_{j_t} = 0$, whereas in equation (4) we constrain $\alpha_{j_s} = 1$ and $\beta_{j_s 1} = 0$, with j_s and j_t ($j_s, j_t = 1, \ldots, m$) denoting a specific item, say the first one, which measure components s ($s = 1, \ldots, S$) and t ($t = 1, \ldots, T$) of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively.

The proposed model can be used to predict probabilities for the item result Y_j and response indicator R_j conditionally on specific values of the latent traits $\boldsymbol{U}$ and $\boldsymbol{V}$. For

instance, from equation (4) the probability that item j results in category y is:

$$Pr(Y_j = y|\boldsymbol{U} = \boldsymbol{u}_{h_U}) =$$

$$= Pr(Y_j \geq y+1|\boldsymbol{U} = \boldsymbol{u}_{h_U}) - Pr(Y_j \geq y|\boldsymbol{U} = \boldsymbol{u}_{h_U}) =$$

$$= \frac{1}{1+\exp[-(\alpha_j \sum_{s=1}^{S} z_{Usj} u_{sh_U} - \beta_{j,y+1})]} - \frac{1}{1+\exp[-(\alpha_j \sum_{s=1}^{S} z_{Vsj} u_{sh_U} - \beta_{jy})]}.$$

As another example, from equation (3) the probability that item j is answered turns out to be:

$$Pr(R_j = 1|\boldsymbol{U} = \boldsymbol{u}_{h_U}, \boldsymbol{V} = \boldsymbol{v}_{h_V}) =$$

$$= \frac{1}{1+\exp[-(\gamma_{Uj} \sum_{s=1}^{S} z_{Usj} u_{sh_U} + \gamma_{Vj} \sum_{t=1}^{T} z_{Vtj} v_{th_V} - \delta_j)]}.$$

4 Likelihood inference

The proposed LC-IRT model under within-item multidimensionality can be estimated through the maximization of the discrete marginal log-likelihood

$$\ell(\boldsymbol{\eta}) = \sum_{i=1}^{n} \log L_i(\boldsymbol{y}_{i,obs}, \boldsymbol{r}_i|\boldsymbol{x}_i), \tag{5}$$

where $\boldsymbol{\eta}$ is the vector of model parameters involved in assumptions (1) to (4) further to the support points of $\boldsymbol{U}$ and $\boldsymbol{V}$, $\boldsymbol{y}_{i,obs} = (y_{i1}, \ldots, y_{im})'$ is the vector of observed item responses for subject i, $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{im})'$ is the vector of response indicators for subject i, and $\boldsymbol{x}_i$ is the vector of covariates for subject i. The joint marginal likelihood $L_i(\boldsymbol{y}_{i,obs}, \boldsymbol{r}_i|\boldsymbol{x}_i)$ of subject i in equation (5) is given by:

$$L_i(\boldsymbol{y}_{i,obs}, \boldsymbol{r}_i|\boldsymbol{x}_i) = \sum_{h_U=1}^{k_U} \sum_{h_V=1}^{k_V} \lambda_{h_U}(\boldsymbol{x}_i) \pi_{h_V}(\boldsymbol{x}_i) p_{h_U h_V}(\boldsymbol{y}_{i,obs}, \boldsymbol{r}_i),$$

where, given the local independence assumption,

$$p_{h_U h_V}(\boldsymbol{y}_{i,obs}, \boldsymbol{r}_i) = \prod_{j=1\,(r_j=1)}^{m} p_{h_U,jy} \prod_{j=1\,(r_j \neq \text{NA})}^{m} q_{h_U h_V,j}^{r_j} (1 - q_{h_U h_V,j})^{1-r_j}. \tag{6}$$

Note that if item j is not due, that is, it is missing by design (r_j = NA), it does not contribute to equation (6); on the other hand, if item j is due but it is skipped ($r_j = 0$) it contributes to equation (6) only through the term $(1 - q_{h_U h_V, j})$.

The estimation of the proposed model can be performed by the specific `R` package `MLCIRTwithin` (Bartolucci & Bacci, 2016), which maximizes the marginal log-likelihood in (5) through the EM algorithm (Dempster et al., 1977), along the same lines as in Bacci & Bartolucci (2015). Moreover, it allows for several options, such as: (*i*) different number of latent classes for the two latent variables, (*ii*) binary or ordinal responses for both the item response process and the missingness process, (*iii*) Rasch or 2PL parameterization for binary items and graded response or partial credit (Masters, 1982) parameterization for ordinal items, (*iv*) multinomial logit or global logit parameterization (Agresti, 2013) for the sub-model that explains the effect of covariates on the probabilities, (*v*) user-specific constraints on support points and item parameters. For more details on the functioning of the package and for a comparison with alternative softwares devoted to the estimation of within-item multidimensional IRT models see Bacci & Bartolucci (2016).

For model selection we rely on information criteria to compare non-nested models (mainly, for the choice of the number of support points k_U and k_V), while we use the likelihood-ratio test to compare nested models. As concerns information criteria, it is worth reminding that the issue of the relative fit is deeply discussed in the literature about mixture models. It is well known that the two most commonly used criteria, that is, the Akaike's Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978), may lead to select different models (McLachlan & Peel, 2000), in particular with large sample sizes. However, several comparative studies (Dias,

2006; Nylund et al., 2007; Yang & Yang, 2007) provide evidence in favor of BIC that typically penalizes the log-likelihood much more than AIC.

In order to simplify the interpretation of the results and the comparison of models with different specifications, we standardize the support points as

$$\hat{u}^*_{sh_U} = \frac{\hat{u}_{sh_U} - \hat{\mu}_{U_s}}{\hat{\sigma}_{U_s}}, \quad s = 1, \ldots, S, \tag{7}$$

$$\hat{v}^*_{th_V} = \frac{\hat{v}_{th_V} - \hat{\mu}_{V_t}}{\hat{\sigma}_{V_t}}, \quad t = 1, \ldots, T, \tag{8}$$

where $\hat{\mu}_{U_s}$ and $\hat{\sigma}_{U_s}$ are the mean and the standard deviation of $\hat{u}_{s1}, \ldots, \hat{u}_{sk_U}$, whereas $\hat{\mu}_{V_t}$ and $\hat{\sigma}_{V_t}$ are the mean and the standard deviation of $\hat{v}_{t1}, \ldots, \hat{v}_{tk_V}$. The item parameters in equations (3) and (4) must be transformed coherently as follows:

$$\hat{\alpha}^*_j = \hat{\alpha}_j \sum_{s=1}^{S} z_{Usj} \hat{\sigma}_{U_s}, \tag{9}$$

$$\hat{\beta}^*_{jy} = \hat{\beta}_{jy} - \hat{\alpha}_j \sum_{s=1}^{S} z_{Usj} \hat{\mu}_{U_s}, \tag{10}$$

$$\hat{\gamma}^*_{1j} = \hat{\gamma}_{1j} \sum_{s=1}^{S} z_{Usj} \hat{\sigma}_{U_s}, \tag{11}$$

$$\hat{\gamma}^*_{2j} = \hat{\gamma}_{2j} \sum_{t=1}^{T} z_{Vtj} \hat{\sigma}_{V_t}, \tag{12}$$

$$\hat{\delta}^*_j = \hat{\delta}_j - \hat{\gamma}_{1j} \sum_{s=1}^{S} z_{Usj} \hat{\mu}_{U_s} - \hat{\gamma}_{2j} \sum_{t=1}^{T} z_{Vtj} \hat{\mu}_{V_t}. \tag{13}$$

The standard errors of the transformed item parameters are obtained through the delta method (Casella & Berger, 2006).

5 Analysis of student careers

We applied the LC-IRT model described in Section 3 to the analysis of the performance of university students described in Section 2. In the following we illustrate model specification and fitting, including a test for the hypothesis of ignorability of the missing data

mechanism, and we report estimates of model parameters. We also discuss the main results, focusing on the discrimination and difficulty of the exams and the interpretation of the latent structure.

5.1 Model specification

For the analysis of the performance of university students we assumed $S = 1$ and $T = 1$, that is, all exams measure the same latent perfomance ($\boldsymbol{U} = U$), and there is one latent tendency to take an exam ($\boldsymbol{V} = V$). The unidimensionality of the latent variable U is in line with our main purpose of clustering freshmen into homogenous groups on the basis of their outcome during the first year of study. This latent performance has a case-specific meaning, since it is intended to summarize the motivation and the multidisciplinary skills needed to tackle the six exams of the first year of the two considered degree programs in academic year 2013/14.

We took explicitly into account that, for each of the six exams, there are four teachers and each of them defines a group, whose assignment to students depends on the first letter of the surname. Consequently, each combination exam-by-group defines a different item and the total number of items is therefore $m = 24$.

The tendency to take an exam, corresponding to V, is measured by the binary variable R_j for $j = 1, \ldots, 24$ that is observed for a given student when item j corresponds to his/her group; R_j is missing by design otherwise ($r_j = \text{NA}$). Given that R_j is observed ($r_j \neq \text{NA}$), it equals 1 if the student enrolls for the corresponding exam at least once during the year and it equals 0 if the student skips the exam. Skipping the exam may depend both on the tendency V to take an exam and on the performance U that university exams contribute

to measure. The structure of the proposed model is illustrated by the path diagram in Figure 1. Conditional on the enrollment, the student can fail or pass the exam with a given grade, ranging from 18 to 30, plus 30 with honors. Our data set includes $3,314$ exams, of which 44% are failures. The distribution of grades of passed exams is quite "irregular" in the sense that there are many accumulation points (see Figure 2): for example, grade 18 has a peak, while grade 29 is rarely used as compared to grades 28 and 30. Moreover, the maximum grade, namely 30 with honors, is out of the quantitative scale.

[Figure 2 about here.]

Therefore, we treated the grade as an ordinal variable. Specifically, we coded the result on exam j by the ordinal variable Y_j with the following categories:

$$\begin{cases} Y_j = \text{NA} & \text{if} \quad R_j = 0, \\ Y_j = 0 & \text{if} \quad R_j = 1 \text{ and } Z_j = \text{NA}, \\ Y_j = 1 & \text{if} \quad 18 \leq Z_j \leq 21, \\ Y_j = 2 & \text{if} \quad 22 \leq Z_j \leq 24, \\ Y_j = 3 & \text{if} \quad 25 \leq Z_j \leq 27, \\ Y_j = 4 & \text{if} \quad Z_j \geq 28, \end{cases}$$

where Z_j is the exam grade in the original scale, with $Z_j \geq 18$ if the exam is passed, and $Z_j = \text{NA}$ if the exam is failed.

We specified a multinomial logit model for the effect of the observed student characteristics (i.e., degree program, gender, HS grade, HS type, and late matriculation) on the probabilities of the latent variables U in equation (1) and V in equation (2). Moreover, we specified a graded response model as in equation (4) for the exam result Y_j, and a 2PL model as in equation (3) for the enrollment R_j.

5.2 Model fitting

In order to select the number of latent classes for U and V, we fitted a series of models with covariates, making comparisons through the BIC. These models are fitted as described in Section 4. As a first step, we considered values of k_U and k_V equal or greater than 2, so as to faithfully reflect the latent structure described by the path diagram in Figure 1. According to the results reported in Table 3, we selected $k_U = 4$ latent classes for U, and $k_V = 2$ latent classes for V.

[Table 3 about here.]

In order to check for local maxima, we repeated the model estimation process with different random starting values of the parameters.

5.3 Testing for the ignorability of the missing data mechanism

In our setting, missing data with respect to variable Y_j arise from the decision to skip exam j in the first year of study. As is already clear, the specified model assumes that the choice to take exam j depends on two latent variables, one representing student performance, and the other one representing student priority in selecting the exams to take. Since the latent performance level affects both the decision to take exam j and its result Y_j, the missing data mechanism is treated as non-ignorable (see Section 3).

To test for the ignorability assumption we compared the proposed multidimensional LC-IRT model with a restricted model where exam enrollment does not depend on the performance U, namely we tested the hypothesis $\gamma_{Uj} = 0, \forall j = 1, \ldots, 24$.

The Likelihood-Ratio Test (LRT) statistic is $LRT = 2 \times (6,533.720 - 6,338.268) =$

390.904, with 24 degrees of freedom yielding a very low p-value. Therefore we proceeded with the proposed multidimensional LC-IRT model accounting for the non-ignorable missing mechanism.

5.4 Discrimination parameters

Table 4 reports the discrimination parameters of equation (4) for the exam outcome Y_j, and the discrimination parameters of equation (3) for exam enrollment R_j. In order to increase the interpretability of the results, all the parameters reported in Table 4 are scaled according to equations (9) to (13) .

[Table 4 about here.]

Note that all the discrimination parameters $\hat{\alpha}_j^*$ relating exam results Y_j to the performance U are significantly different from zero, namely all the exams contribute to measure the latent performance. Accounting, Mathematics, and Statistics tend to have a higher discrimination power, that is, the results of these exams are more sensitive to variations in student performance. However, there are differences across groups of the same course, especially for Law and Management.

According to equation (3), enrollment for an exam R_j is affected by student performance U through the γ_{Uj}^* parameters, and by the latent variable V through the γ_{Vj}^* parameters. The effect of the latent variable V is positive for Mathematics and Statistics, and negative for Law; thus V can be interpreted as the tendency of the student to take exams in quantitative subjects as opposed to exams in qualitative subjects. Considering statistical significance at 5%, student performance U significantly affects the enrollment

for most exams, whereas student tendency V has a significant effect for just around one-third of the items.

The dependence of the R_j variables for the exam enrollment on the performance U suggests that a model for evaluating student performance should account for enrollment decisions; in statistical terms, this provides evidence that the enrollment process generating missing exam grades is not ignorable, as confirmed by the likelihood-ratio test described in Section 5.3.

5.5 Predicted probabilities of exam enrollment and exam result

Equations (3) and (4) include five difficulty parameters for each item j, specifically four parameters $\beta_{1j}^* \dots \beta_{4j}^*$ (after standardization) for each exam result Y_j, and parameter δ_j^* (after standardization) for each exam enrollment variable R_j. The estimates of the difficulty parameters (shown in the online Supplementary Material, Tables 1-2) are not easily interpretable, thus we converted such parameters into response probabilities, conditionally on given values of latent traits $(U = u, V = v)$. In particular, Table 5 reports the probabilities of exam results for an average performance student, that is, $U = 0$, those probabilities depend only on the difficulty parameters. In addition, the right part of Table 5 reports the conditional probability of passing the exam $Pr(Y_j > 0|U)$ for certain values of student performance, that is, $U = -\sigma_U$, $U = 0$, and $U = +\sigma_U$, with σ_U denoting the estimated standard deviation of performance U.

[Table 5 about here.]

We note a large variability among courses and, in some cases, also across groups of the same course. For the majority of items, the most likely result is a failure. Moreover, for

the majority of the courses the modal grade of passed exams is $18-21$, with some notable exceptions, such as Management Econ M-Z. The values of the discrimination parameters $\hat{\alpha}_j^*$ imply that the probability to pass the exam depends on student performance: the range reported in the last column of Table 5 is large, with relevant differences both within and between courses.

The probabilities reported in Table 5 can be used to predict the performance of a student for given values of the latent variable U, depending on the chosen degree program and the group assigned on the basis on the first letter of the surname. For example, for a student with a high level of performance (say, $+\sigma_U$), enrolled in the degree program *Business* and belonging to the A-C group, the probability to pass all the exams can be obtained by multiplying the six probabilities $Pr(Y_j > 0 \mid U = +\sigma_U)$ corresponding to the A-C group, that is, $0.940 \times 0.643 \times \ldots \times 0.942 = 0.191$. It is worth noting that the same probability rises to 0.362 for a student belonging to the Q-Z group, whereas it drops to 0.173 for a student enrolled in the degree program *Economics* and belonging to the D-L group. Similar computations show that the probability to pass all six exams is less than 0.01 for students with average performance ($U = 0$).

In a similar way we can compute the probability of other patterns. For example, the probability to pass only Accounting for a student with performance level equal to the average, who is enrolled in the degree program *Business* and belongs to the D-L group, is $0.352 \times (1 - 0.197) \times \ldots \times (1 - 0.453) = 0.015$; it rises to 0.025 for a colleague belonging to the same group but enrolled in the degree program *Economics* and to 0.070 for a colleague enrolled in the same degree program but belonging to group M-P. Moreover, for a student with a low performance level (say, $-\sigma_U$), enrolled in the degree program

Business and belonging to the D-L group, the probability to pass one exam out of six is 0.306, obtained by adding the probability to pass only Accounting, the probability to pass only Mathematics, and so on. For a similar student belonging to group M-P the probability to pass only one exam rises to 0.349.

Table 6 reports the probability to enroll for an exam for some values of U and V. The first column of Table 6 reports the probabilities for a student with average values for both latent variables ($U = 0, V = 0$), thus depending only on the estimated difficulty parameters $\hat{\delta}_j^*$.

[Table 6 about here.]

Similarly to the exam result Y_j, the enrollment for the exam R_j shows a large variability among courses and, in some cases, also across groups of the same course. The enrollment rate is high for Accounting and low for Microeconomics and Law. Moreover, Microeconomics shows large differences between groups, ranging from 0.14 to 0.60. Looking at the last two columns of Table 6, we see that the probability to enroll in the exam depends more on student performance U than on tendency V, with the exception of Mathematics. The effect of V is relevant and positive for Mathematics and Statistics and negative for Law, confirming the interpretation of V in terms of tendency to take quantitative exams.

5.6 Estimated latent structure and effects of the covariates

Table 7 reports the estimated support points and corresponding average probabilities for the latent classes of performance U and tendency V. The support points $\hat{u}_{h_U}^*$ and $\hat{v}_{h_V}^*$ are standardized as described by equations (7) and (8).

[Table 7 about here.]

Table 8 reports the estimated coefficients $\hat{\phi}_{h_U c}$ ($h_U = 2, 3, 4$) of the multinomial logit model (1) for the probabilities λ_{h_U} of the latent performance U, and the estimated coefficients $\hat{\psi}_{h_V c}$ ($h_V = 2$) of the multinomial logit model (2) for the probabilities π_{h_V} of the latent tendency V. Note that, since V has only two components, the multinomial logit model (2) reduces to a binary logit model.

[Table 8 about here.]

In the model selection process, we retained covariates with a p-value< 0.05 on at least one equation, except for gender, which was retained even if not significant in any equation, since the model enclosing gender has a better fit than the model without gender (LRT statistic equal to 15.373 with 4 degrees of freedom and p-value$= 0.004$); we also outline that gender is recognized as a key variable in several educational studies (e.g. Conger & Long, 2010). The HS type covariate has four categories: we retained these four categories to avoid merging very different types of school.

The distribution of the performance U has four support points and it is right skewed. The average probability of class 1 ($\overline{\lambda}_1 = 0.228$) is close to the observed proportion of students who did not pass any exam (0.243). Class 4 is the smallest one ($\overline{\lambda}_4 = 0.083$) and it includes very good students, with a performance equal to about two standard deviations above the mean. Some student characteristics have a significant effect on the performance (Table 8): students with a higher grade and students with a scientific HS degree tend to belong to latent classes of better performance (i.e., class 3 and class 4), while the reverse holds for late matriculated students.

The distribution of the tendency V gives rise to two latent classes of similar size, with support points -0.949 and 1.054. As noted in Section 5.5, students belonging to class 2 prefer to take quantitative exams. For a baseline student (male, with degree in *Business*, HS grade at a mid-point, HS type technical, no late matriculation), the probability to belong to class 2 is 0.295. This probability raises to 0.462 for a baseline student but enrolled in the degree of *Economics* and to 0.749 for a baseline student but with a scientific HS degree. On the contrary, this probability decreases to 0.166 for a late matriculated student.

In order to characterize the latent classes of performance U, we allocated the students to these classes according to the maximum posterior probability as in Bacci & Bartolucci (2015). Figure 3 reports enrollment rates and passing rates (conditional on enrollment and overall) for all the exams, considering the students assigned to the corresponding class. All the rates increase moving from class 1 to class 4, though the conditional passing rate grows faster then the enrollment rate. Coherently, the average number of enrolled exams increases from 2.83 for class 1 through 3.54 for class 2 and 4.42 for class 3, up to 4.89 for class 4, whereas the average number of passed exams is equal to 0.80, 1.61, 3.06, and 4.16 for each latent class, respectively. Figure 4 represents the enrollment rate (left panel) and the conditional passing rate (right panel) separately for the six exams. We notice different patterns. For example, Accounting shows a nearly flat enrollment rate, and a steep passing rate, while Microeconomics has an opposite pattern. The rate with the highest variation among the latent classes is the conditional passing rate of Math. The average grades of the exams passed by the students assigned to each class are 22.05, 22.79, 24.80, and 26.68, respectively. Therefore, moving from class 1 to class 4, students

show better outcomes in terms of all the considered aspects, namely number of enrolled exams, number of passed exams and average grade. This pattern allows us to interpret U as an overall performance level.

[Figure 3 about here.]

[Figure 4 about here.]

5.7 Testing differences across groups of the same course

The results of Sections 5.4 and 5.5 show that, for some courses, discrimination and difficulty are markedly different across the four groups. In the model described by equations (3) and (4), an item j corresponds to a group of a given course, for example $j = 1$ corresponds to group A-C of Accounting. The model has eight parameters for each item j: three discrimination parameters $(\alpha_j, \gamma_{Uj}, \gamma_{Vj})$ and five difficulty parameters $(\beta_{1j}, \beta_{2j}, \beta_{3j}, \beta_{4j}, \delta_j)$.

A test of homogeneity for the four groups of a given course can be performed comparing the full model with a restricted model, where the items corresponding to the four groups have the same set of parameters. For example, for the course of Accounting the restricted model assumes $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$, and similarly for the other parameters, for a total of $3 \times 8 = 24$ restrictions.

Table 9 reports the LRT statistics comparing the full model with a restricted model for each course, collapsing the items corresponding to different groups. These statistics can be interpreted in terms of differences among groups of a given course.

[Table 9 about here.]

Mathematics shows the lowest LRT value, followed by Statistics, while Management has the highest value. All test statistics deal to reject the homogeneity assumption, except for Mathematics, thus confirming the appropriateness of a model treating the groups as distinct items with certain structural missing values.

5.8 Sensitivity analysis

The results of Section 5.4 suggest a weak role of the latent variable V on each exam enrollment variable R_j; see equation (3) for the specification of this relation. Indeed, the discrimination parameters of V ($\hat{\gamma}^*_{2j}$) in Table 4 are significant (at 5%) for only one-third of the items, and they are lower in absolute value than the discrimination parameters of U ($\hat{\gamma}^*_{1j}$). Moreover, the standard errors for the parameters of Mathematics courses are abnormally high (for details see standard errors $\hat{\text{se}}_{\gamma^*_{Vj}}$ for group Q-Z in Table 4 and also standard errors $\hat{\text{se}}_{\delta^*_j}$ referred to all groups of Mathematics shown in the online Supplementary Material, Table 2).

The latent tendency V is related to non-response patterns, but its deletion does not alter the feature of the model to account for informative missing data. In particular, to evaluate the role of V in the current application, we compared the selected model ($k_U = 4$, $k_V = 2$ in Table 3) against a restricted version without V ($k_U = 4$, $k_V = 1$), entailing a reduction of model parameters from 226 to 194. In the restricted version of the model, the standard errors for Mathematics are no more problematic (see estimates shown in the online Supplementary Material, Tables 3-5), anyway the main findings about the effects of the latent performance U are unchanged (online Supplementary Material, Tables 6-9).

The additional complexity associated to the inclusion of the latent tendency V is not

worthwhile in terms of model fit as measured by the BIC (the restricted model has a BIC value of 14,166.19, as compared to 14,203.86 for the full model). However, we decided to carry out the analysis with a model including V, since it gives additional insights into the student decision process. Moreover, this specification is useful to illustrate the potentialities of the proposed approach, which allows for more flexibility in modeling non-response patterns.

6 Conclusions

In order to evaluate university student performance during the first year of study, we proposed an IRT approach that jointly account for the observed exam results and for the information on the exam enrollment. Indeed, in our case it often happens that a student does not enroll for a given exam during the first year, thus the corresponding exam result is missing. To take into account such informative missing mechanism, the proposed multidimensional latent class IRT model has a latent variable U, representing student performance, which affects both the enrollment decision and the exam result. Another latent variable V accounts for student preferences in choosing which exams to take during the first year. Our model is an extension of the finite mixture model of Bacci & Bartolucci (2015), accounting for mixed items (i.e., binary and ordered) and two kinds of missing data (i.e., structural and informative). It is worth noting that the IRT approach exploits the multivariate nature of the observed outcomes as each first-year exam is treated as an item. This is an advantage over usual approaches based on a summary measure of student outcome, such as the total number of gained credits.

The analysis showed that, even controlling for student latent performance U and

preference V, the enrollment rates vary both between disciplines and between groups of the same discipline. Such differences may stem from several factors, including the teacher ability to motivate the student and organizational issues. Moreover, we found that the enrollment depends mainly on student latent performance U, so that the enrollment mechanism is not ignorable. Therefore, the prediction of student outcome requires to jointly model the enrollment decision and exam result.

The probabilities of passing the exams and of exam grades for a student with a given level of performance U are remarkably different among disciplines. As expected, Mathematics is the hardest exam, with low probability of passing the exam and low grades, highlighting problems with either the course content or the use of the grading scale. Moreover, some courses show worrying differences between groups concerning both enrollment rates and passing rates. This fact poses a serious issue of fairness, given that the assignment of students to groups is based on surname, thus groups are expected to be homogeneous with respect to student performance.

The model was used to cluster students into four latent classes of performance, corresponding to widely different outcomes. In particular, assigning students to latent classes on the basis of the maximum posterior probability, we observed that moving from the first to the last latent class of U, students show better outcomes in terms of all considered aspects, namely number of enrolled exams, number of passed exams, and average grade. This pattern allows us to interpret the latent variable U as an overall performance level.

The structural part of the model relates class membership to observed characteristics. It turned out that the probability to belong to classes of better performance is higher for students coming from a scientific high school, students with a good school grade, and

students beginning university in the academic year following the end of high school. Such information can be used by potential freshmen and by the university management for planning guidance and tutoring activities.

Beyond the case study here illustrated, the proposed LC-IRT model based on within-item multidimensionality can be useful in a wide range of applications characterized by binary and ordinal items with structural as well as informative missing item responses. Examples include assessment in education, customer satisfaction, quality of life, and the assessment of physical or psychological disabilities.

The model and the corresponding software can be extended to let the covariates affect the item responses, so as to allow for differential item functioning (Thissen et al., 1988). As for latent class models in general, the optimal number of latent classes is a controversial issue (McLachlan & Peel, 2000) calling for further theoretical and simulation work.

References

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

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). New York: Wiley.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium of information theory* (p. 267-281). Budapest: Akademiai Kiado.

Arminger, G., Stein, P., & Wittenberg, J. (1999). Mixtures of conditional mean- and covariance-structure models. *Psychometrika*, *64*, 475-494.

Bacci, S., & Bartolucci, F. (2015). A multidimensional finite mixture structural equation model for nonignorable missing responses to test items. *Structural Equation Modeling*, *22*, 352–365.

Bacci, S., & Bartolucci, F. (2016). Two-tier latent class IRT models in R. *The R Journal*, *8*, 139–166.

Bacci, S., Bartolucci, F., & 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.

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

Bartolucci, F., & Bacci, S. (2016). `MLCIRTwithin`*: latent class item response theory models under "within-item multidimensionality".* R package version 2.1, URL http://CRAN.R-project.org/package=MLCIRTwithin.

Bartolucci, F., Bacci, S., & Gnaldi, M. (2015). *Statistical Analysis of Questionnaires: A Unified Approach Based on R and Stata*. Boca Raton, FL: Chapman & Hall, CRC Press.

Bertaccini, B., Grilli, L., & Rampichini, C. (2013). An IRT-MIMIC model for the analysis of university student careers. *Quaderni di Statistica*, *15*, 95 - 110.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

Casella, G., & Berger, R. L. (2006). *Statistical inference*. Pacific Grove, CA: Thomson Learning.

Conger, D., & Long, M. C. (2010). Why are men falling behind? Gender gaps in college performance and persistence. *The Annals of the American Academy of Political and Social Science*, *627*, 184-214.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*, 173–178.

Dempster, A. P., Laird, N. M., & 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. G. (2006). Model selection for the binary latent class model: A Monte Carlo simulation. In V. Batagelj, H. H. Bock, A. Ferligoj, & A. Ziberna (Eds.), *Data Science and Classification* (p. 91-99). Berlin, Heidelberg: Springer-Verlag.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics*, *43*, 49-93.

Dolan, C. V., & van der Maas, H. L. J. (1998). Fitting multivariage normal finite mixtures subject to structural equation modeling. *Psychometrika*, *63*, 227-253.

Formann, A. K. (2007). Mixture analysis of multivariate categorical data with covariates and missing entries. *Computational Statistics & Data Analysis*, *51*, 5236-5246.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Gitto, L., Minervini, L. F., & Monaco, L. (2015). University dropout rates in Italy. In G. Coppola & N. O'Higgins (Eds.), *Youth and the Crisis: Unemployment, Education and Health in Europe* (pp. 75–88). Paignton, UK: Routledge.

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

Grilli, L., Rampichini, C., & Varriale, R. (2016). Statistical modelling of gained university credits to evaluate the role of pre-enrolment assessment tests: An approach based on quantile regression for counts. *Statistical Modelling*, *16*, 47–66.

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

Harel, O., & Schafer, J. L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika*, *96*, 37–50.

Holman, R., & Glas, A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*, 1-17.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Stemm: A general finite mixture structural equation model. *Journal of Classification*, *14*, 23-50.

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

Lebedeva, E., Huang, M., & Koski, L. (2016). Comparison of alternate and original items on the Montreal Cognitive Assessment. *Canadian Geriatrics Journal*, *19*, 15-18.

Lindsay, B., Clogg, C. C., & Greco, 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.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*, 125 - 134.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley and Sons.

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, *48*, 477-482.

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

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

Mealli, F., & Rubin, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, *102*, 995-1000.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, *11*, 81–91.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (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.

OECD. (2014). *PISA 2012 Technical Report*. Paris: OECD Publisher.

OECD. (2016). *Education at a Glance 2016: OECD Indicators*. Paris: OECD Publisher.

Reckase, M. (2010). *Multidimensional Item Response Theory*. New York: Springer-Verlag.

Rose, N., Von Davier, M., & Xu, X. (2010). *Modeling Nonignorable Missing Data with Item Response Theory (IRT)* (Tech. Rep.). ETS Research Report No. RR-10-11, Princeton.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*, 100.

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

Smith, T., Gildeh, N., & Holmes, C. (2007). The Montreal Cognitive Assessment: validity and utility in a memory clinic setting. *Canadian Journal of Psychiatry*, *52*, 329-332.

Stewart, A. L., & Ware, J. E. (1992). *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach*. Durham: Duke University Press.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New-York: Springer-Verlag.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287-307.

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36): I. conceptual framework and item selection. *Medical Care*, *30*, 473-483.

Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, *44*, 175–188.

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

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*, 589–600.

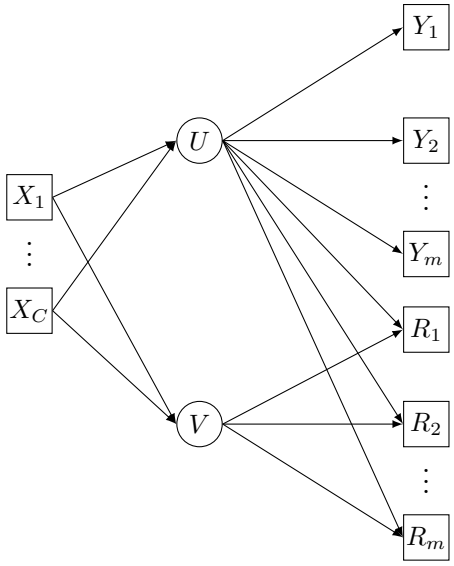

Figure 1: *Path diagram of the LC-IRT model with latent traits (U and V: latent traits; $Y_1, \ldots, Y_m$: observed item responses; $R_1, \ldots, R_m$: item response indicators; $X_1, \ldots, X_C$: exogenous individual covariates).*

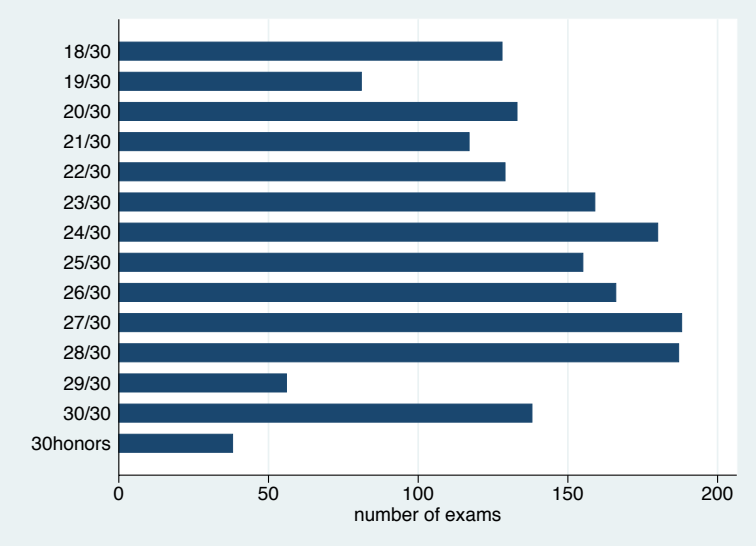

Figure 2: *Distribution of grades, passed exams.*

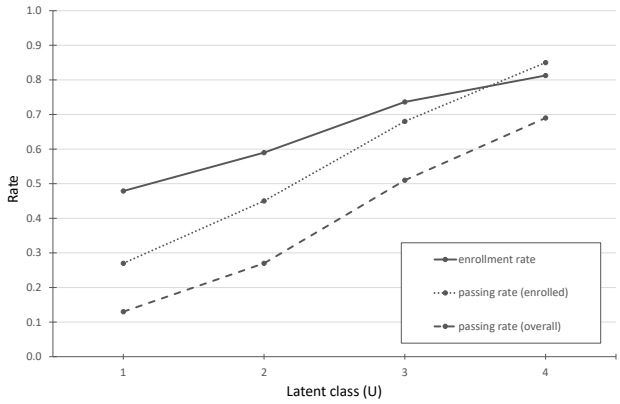

Figure 3: *Enrollment and passing rates for all the exams.*

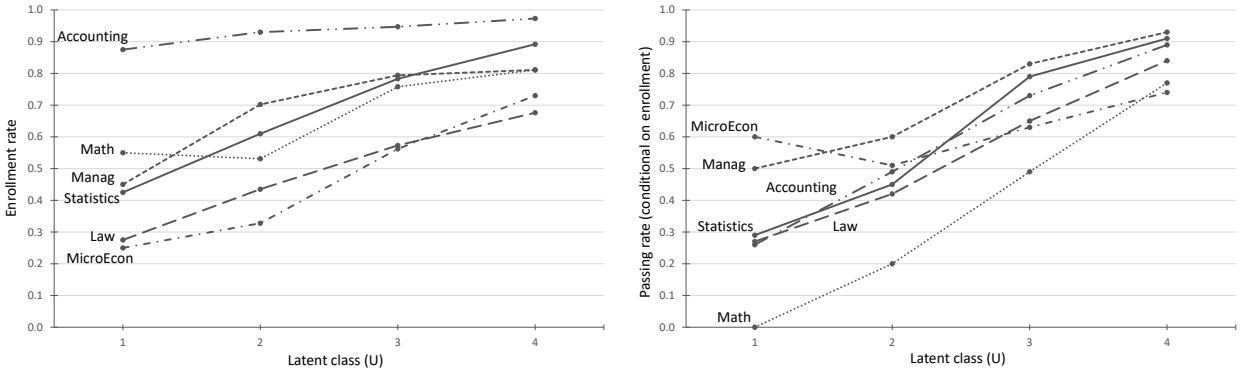

Figure 4: *Enrollment rates (left panel) and passing rates conditionally on enrollment (right panel), for each exam.*

Table 1: *Enrollment rates and exam results of first-year exams by course (freshmen 2013/2014, University of Florence, degree programs* Business *and* Economics*, examination sessions from January to December 2014).*

Course (semester)		Enrollment rate (%)	Exam grade (%)					Passing rate (%)	
			failed	18-21	22-24	25-27	≥ 28	enrolled	overall
Accounting	(I)	93.5	42.5	15.9	17.3	17.0	7.3	57.5	53.8
Mathematics	(I)	67.8	65.8	16.2	7.3	6.8	4.0	34.2	21.1
Law	(I)	48.3	47.1	14.2	16.1	14.4	8.2	52.9	25.6
Management	(II)	72.5	30.6	8.2	16.7	23.2	21.3	69.4	50.3
MicroEcon	(II)	41.8	41.9	10.6	11.4	18.1	18.1	58.1	24.3
Statistics	(II)	67.0	39.7	16.8	13.5	11.4	18.5	60.3	40.4

Table 2: *Average number of attempted and passed first-year exams by student characteristics (freshmen 2013/2014, University of Florence, degree programs* Business *and* Economics*, examination sessions from January to December 2014).*

	N	Average number of exams	
		enrolled to	passed
All freshmen	861	3.8	2.2
Gender			
Male	502	3.8	2.1
Female	359	3.9	2.2
HS type			
Technical	765	3.8	2.1
Humanities	201	3.7	1.9
Scientific	321	4.1	2.4
Other	284	3.6	1.8
HS grade			
< 80	596	3.6	1.7
≥ 80	265	4.4	3.3
Late matriculation			
No	759	4.0	2.3
Yes	102	3.0	1.3
Degree program			
Business	588	3.8	2.1
Economics	273	3.9	2.3
Course group			
A-C	257	3.7	2.2
D-L	240	3.8	2.2
M-P	204	4.0	2.1
Q-Z	160	3.9	2.3

Table 3: *Selection of the number of latent classes.*

k_U	k_V	$\hat{\ell}$	# par	BIC
2	2	-6520.37	208	14446.41
2	3	-6505.63	217	14477.76
3	2	-6387.18	217	14240.87
3	3	-6364.32	226	14255.96
4	2	-6338.27	226	14203.86
4	3	-6325.72	235	14239.58
5	2	-6323.84	235	14235.84
5	3	-6304.16	244	14257.30

Table 4: *Estimated scaled discrimination item parameters.*

Item		$U \to Y_j$			$U \to R_j$			$V \to R_j$		
Course	Group	$\hat{\alpha}^*_j$	$\hat{\text{se}}_{\alpha^*_j}$	p-value	$\hat{\gamma}^*_{Uj}$	$\hat{\text{se}}_{\gamma^*_{Uj}}$	p-value	$\hat{\gamma}^*_{Vj}$	$\hat{\text{se}}_{\gamma^*_{Vj}}$	p-value
Accounting	A-C	2.127	0.285	<0.001	0.904	0.375	0.016	0.401	0.274	0.144
	D-L	1.795	0.259	<0.001	0.535	0.349	0.125	0.594	0.433	0.170
	M-P	2.527	0.380	<0.001	1.172	0.533	0.028	-0.503	0.525	0.338
	Q-Z	2.337	0.357	<0.001	0.433	0.318	0.173	0.522	0.311	0.093
Mathematics	A-C	2.241	0.410	<0.001	1.918	0.893	0.032	2.448	1.176	0.037
	D-L	2.134	0.386	<0.001	1.278	0.440	0.004	2.251	0.772	0.004
	M-P	1.731	0.402	<0.001	1.700	0.503	0.001	1.782	0.587	0.002
	Q-Z	2.963	0.707	<0.001	5.050	4.605	0.273	6.783	5.266	0.198
Law	A-C	1.849	0.427	<0.001	0.903	0.196	<0.001	-0.380	0.221	0.085
	D-L	3.016	0.506	<0.001	1.303	0.254	<0.001	-0.390	0.267	0.144
	M-P	1.391	0.306	<0.001	1.144	0.267	<0.001	-0.804	0.314	0.011
	Q-Z	1.783	0.425	<0.001	1.033	0.241	<0.001	-0.279	0.214	0.192
Management	Busi A-L	2.990	0.530	<0.001	2.287	0.389	<0.001	0.675	0.315	0.032
	Busi M-Z	3.163	0.484	<0.001	1.339	0.249	<0.001	-0.304	0.221	0.169
	Econ A-L	1.976	0.389	<0.001	1.106	0.250	<0.001	0.539	0.296	0.068
	Econ M-Z	0.662	0.306	0.030	2.212	0.500	<0.001	0.605	0.455	0.184
MicroEcon	A-C	1.249	0.325	<0.001	1.429	0.247	<0.001	0.310	0.238	0.193
	D-L	1.130	0.402	0.005	3.114	0.598	<0.001	0.450	0.319	0.159
	M-P	1.822	0.366	<0.001	1.889	0.353	<0.001	-0.447	0.294	0.128
	Q-Z	2.350	0.588	<0.001	2.202	0.440	<0.001	0.926	0.282	0.001
Statistics	A-C	2.787	0.445	<0.001	2.333	0.466	<0.001	0.946	0.387	0.014
	D-L	2.496	0.389	<0.001	1.567	0.290	<0.001	0.808	0.295	0.006
	M-P	2.867	0.497	<0.001	1.772	0.319	<0.001	0.104	0.292	0.722
	Q-Z	2.258	0.468	<0.001	1.998	0.469	<0.001	1.020	0.347	0.003

Table 5: *Predicted probabilities of exam result Y_j for some values of latent performance U.*

Item		$Pr(Y_j = y \mid U = 0)$					Passing rate $Pr(Y_j > 0 \mid U)$			
Course	Group	0	1	2	3	4	$U = u$			
		Failed	18-21	22-24	25-27	≥ 28	$-\sigma_U$	0	$+\sigma_U$	range
Accounting	A-C	0.348	0.328	0.172	0.124	0.028	0.183	0.652	0.940	0.758
	D-L	0.648	0.168	0.126	0.052	0.006	0.083	0.352	0.766	0.683
	M-P	0.378	0.219	0.305	0.094	0.005	0.116	0.622	0.954	0.837
	Q-Z	0.141	0.295	0.355	0.176	0.033	0.371	0.859	0.984	0.614
Mathematics	A-C	0.839	0.114	0.029	0.012	0.005	0.020	0.161	0.643	0.623
	D-L	0.803	0.136	0.037	0.019	0.005	0.028	0.197	0.674	0.646
	M-P	0.872	0.069	0.036	0.016	0.007	0.025	0.128	0.453	0.427
	Q-Z	0.906	0.085	0.005	0.004	0.000	0.005	0.094	0.667	0.662
Law	A-C	0.824	0.105	0.060	0.008	0.004	0.033	0.176	0.576	0.544
	D-L	0.530	0.168	0.237	0.057	0.008	0.042	0.470	0.948	0.906
	M-P	0.709	0.142	0.042	0.092	0.016	0.093	0.291	0.623	0.530
	Q-Z	0.480	0.259	0.177	0.070	0.014	0.154	0.520	0.866	0.711
Management	Busi A-L	0.222	0.187	0.350	0.181	0.061	0.150	0.778	0.986	0.836
	Busi M-Z	0.761	0.063	0.131	0.037	0.008	0.013	0.239	0.881	0.868
	Econ A-L	0.381	0.204	0.209	0.163	0.043	0.184	0.619	0.921	0.738
	Econ M-Z	0.101	0.026	0.093	0.571	0.209	0.821	0.899	0.945	0.124
MicroEcon	A-C	0.707	0.124	0.102	0.049	0.018	0.106	0.293	0.591	0.485
	D-L	0.809	0.046	0.046	0.063	0.036	0.071	0.191	0.422	0.351
	M-P	0.355	0.157	0.161	0.248	0.078	0.227	0.645	0.918	0.691
	Q-Z	0.650	0.117	0.067	0.081	0.085	0.049	0.350	0.850	0.801
Statistics	A-C	0.501	0.292	0.120	0.055	0.032	0.058	0.499	0.942	0.884
	D-L	0.547	0.231	0.157	0.041	0.024	0.064	0.453	0.909	0.846
	M-P	0.673	0.197	0.076	0.038	0.017	0.027	0.327	0.895	0.868
	Q-Z	0.625	0.197	0.057	0.073	0.048	0.059	0.375	0.852	0.793

Table 6: *Predicted probabilities of enrollment R_j for some values of latent performance U and latent tendency V.*

Item		$Pr(R_j = 1 \mid U = u, V = v)$						
Course	Class	u	0	$-\sigma_U$	$+\sigma_U$	0	0	Range	
		v	0	0	0	$-\sigma_V$	$+\sigma_V$	$\pm\sigma_U$	$\pm\sigma_V$
Accounting	A-C		0.962	0.910	0.984	0.944	0.974	0.074	0.030
	D-L		0.952	0.921	0.971	0.916	0.973	0.050	0.056
	M-P		0.973	0.919	0.992	0.984	0.957	0.073	-0.027
	Q-Z		0.913	0.872	0.942	0.861	0.946	0.070	0.085
Mathematics	A-C		0.797	0.367	0.964	0.254	0.979	0.597	0.724
	D-L		0.785	0.505	0.929	0.278	0.972	0.425	0.694
	M-P		0.768	0.376	0.948	0.357	0.952	0.571	0.594
	Q-Z		0.905	0.058	0.999	0.011	1.000	0.942	0.989
Law	A-C		0.323	0.162	0.541	0.411	0.246	0.379	-0.165
	D-L		0.506	0.218	0.790	0.602	0.409	0.572	-0.192
	M-P		0.580	0.306	0.813	0.755	0.382	0.507	-0.373
	Q-Z		0.564	0.315	0.784	0.631	0.495	0.469	-0.137
Management	Busi A-L		0.879	0.424	0.986	0.787	0.934	0.562	0.148
	Busi M-Z		0.792	0.499	0.936	0.837	0.737	0.437	-0.100
	Econ A-L		0.644	0.374	0.845	0.513	0.756	0.471	0.243
	Econ M-Z		0.885	0.456	0.986	0.807	0.933	0.530	0.126
MicroEcon	A-C		0.268	0.081	0.605	0.212	0.333	0.524	0.121
	D-L		0.144	0.007	0.791	0.097	0.209	0.784	0.112
	M-P		0.599	0.184	0.908	0.700	0.488	0.724	-0.212
	Q-Z		0.526	0.109	0.909	0.305	0.737	0.800	0.431
Statistics	A-C		0.852	0.358	0.983	0.691	0.937	0.625	0.246
	D-L		0.679	0.306	0.910	0.485	0.826	0.604	0.341
	M-P		0.690	0.275	0.929	0.667	0.712	0.654	0.044
	Q-Z		0.866	0.466	0.979	0.699	0.947	0.513	0.248

Table 7: *Estimated support points and corresponding average probabilities for the latent classes of performance U and tendency V.*

	Performance U latent class h_U				Tendency V latent class h_V	
	1	2	3	4	1	2
Support points ($\hat{u}^*_{h_U}$, $\hat{v}^*_{h_V}$)	-1.485	-0.129	0.784	1.937	-0.949	1.054
Average probabilities ($\overline{\lambda}_{h_U}$, $\overline{\pi}_{h_V}$)	0.228	0.395	0.294	0.083	0.526	0.474

Table 8: *Estimated coefficients of the multinomial logit models for the probabilities of the latent classes of performance U and tendency V.*

	Performance U			Tendency V
	$\hat{\phi}_{2c}$	$\hat{\phi}_{3c}$	$\hat{\phi}_{4c}$	$\hat{\psi}_{2c}$
Constant	0.748*	0.568	-1.956*	-0.869*
Degree Economics	-0.434	-0.030	0.045	0.716*
Gender (ref: male)	0.434	0.059	-0.487	-0.147
HS grade ≥ 80	0.014	0.118*	0.265*	-0.020
HS type (ref:technical)				
HS humanities	0.138	0.148	0.691	-0.240
HS scientific	-0.061	1.020*	2.219*	1.963*
HS other	-0.163	-0.163	-0.282	-0.336
Late matriculation	-0.191	-1.415*	-1.834*	-0.744*

Parameters with * have p-value< 0.05 .

Table 9: *Likelihood-ratio test comparing full model with models collapsing groups.*

Exam	$logL$	# par	LRT stat.	df	p-value
Full model	-6338.27	226	–	–	–
Accounting	-6389.59	202	102.64	24	0.000
Mathematics	-6349.62	202	22.70	24	0.538
Law	-6397.06	202	117.59	24	0.000
Management	-6464.91	202	253.29	24	0.000
MicroEcon	-6411.62	202	146.71	24	0.000
Statistics	-6358.22	202	39.90	24	0.022