

A multilevel finite mixture item response model to cluster examinees and schools

Michela Gnaldi · Silvia Bacci · Francesco Bartolucci

Received: 5 December 2013 / Revised: 28 November 2014 / Accepted: 27 December 2014 /
Published online: 24 January 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Within the educational context, a key goal is to assess students' acquired skills and to cluster students according to their ability level. In this regard, a relevant element to be accounted for is the possible effect of the school students come from. For this aim, we provide a methodological tool which takes into account the multilevel structure of the data (i.e., students in schools) and allows us to cluster both students and schools into homogeneous classes of ability and effectiveness, and to assess the effect of certain students' and school characteristics on the probability to belong to such classes. The proposed approach relies on an extended class of multidimensional latent class IRT models characterised by: (i) latent traits defined at student and school level, (ii) latent traits represented through random vectors with a discrete distribution, (iii) the inclusion of covariates at student and school level, and (iv) a two-parameter logistic parametrisation for the conditional probability of a correct response given the ability. The approach is applied for the analysis of data collected by two national tests administered in Italy to middle school students in June 2009: the INVALSI Language Test and the Mathematics Test.

Keywords EM algorithm · INVALSI Tests · Latent class model · Multilevel multidimensional item response models · Two-parameter logistic model

Mathematics Subject Classification 62-07 · 62H30 · 62H12 · 62F99

M. Gnaldi (✉)
Department of Political Sciences, University of Perugia,
via A. Pascoli, 20, 06123 Perugia, Italy
e-mail: michela.gnaldi@unipg.it

S. Bacci · F. Bartolucci
Department of Economics, University of Perugia, via A. Pascoli, 20, 06123 Perugia, Italy

1 Introduction

According to the assumption of unidimensionality, which characterises classical Item Response Theory (IRT) models, responses to a set of items only depend on a single latent trait which, in the educational setting, can be interpreted as the students ability. If unidimensionality is not met, summarising students ability through a single score may be misleading as test items indeed measure more than one ability, and the proper evaluation of a student's ability requires the development of multidimensional methods. The present article is focused on these methods, in particular when the sample has a multilevel structure and the interest is in clustering subjects in homogenous classes according to the latent trait measured by the test items.

Clustering individuals in homogenous classes is possible relying on the Latent Class (LC) analysis (Lazarsfeld and Henry 1968; Goodman 1974), which is based on the idea that the population under study is formed by a finite number of latent (i.e., unobservable) classes of subjects with homogeneous ability levels. In this regard, finite mixture IRT models represent a combination of IRT and LC analysis based on the idea that the same IRT model holds for the subjects in the same class, whereas possibly different IRT models hold among different classes. Thus, finite mixture (or LC) IRT models allow for the discreteness of the ability distribution with a number of support points equal to the number of latent classes. We consider, in particular, the class of LC-IRT models for binary items and ordinal polytomous items proposed in Bartolucci (2007) and Bacci et al. (2014), which assume fixed (rather than class-specific) item parameters among latent classes. Moreover, these models take into account multidimensional latent traits (Reckase 2009) and more general item parameterisations than those of Rasch-type models (Rasch 1961), such as the Two-Parameter Logistic (2PL) model introduced by Birnbaum (1968).

An approach related to that of Bartolucci (2007) is due to von Davier (2008), who proposed a diagnostic model based on fixed rather than free abilities. For further examples of LC versions of IRT models we also recall Lindsay et al. (1991), Formann (1992, 1995), Hoijtink and Molenaar (1997), Vermunt (2001), and Smit et al. (2003). See also Masters (1985), Langheine and Rost (1988), Heinen (1996), Christensen et al. (2002), and Formann (2007a), which outline the greater flexibility of IRT models based on the assumption that the latent traits follow a discrete rather than a continuous distribution. Among others, one of the most known variants of IRT models based on a discrete ability distribution is given by the mixed Rasch model for binary and ordinal polytomous data (Rost 1990, 1991; von Davier and Rost 1995), built as a mixture of separate Rasch models characterised by class-specific person and item parameters. Besides, from an applied point of view, mixture IRT models have been used to set proficiency standards (Jiao et al. 2012), identify solution strategies (Mislevy and Verhelst 1990), study the effects of test speediness (Bolt et al. 2002), and identify latent classes which differ with regards to the use of response scales (see, for example, Majj-de Meij et al. 2008).

Another relevant element, which is ignored by classical IRT models, concerns that part of unobserved heterogeneity of item responses is due to multilevel structures of data, where individuals are nested in groups, such as students within schools. In this case, it is reasonable to assume that students sharing the same school context are more

similar than their colleagues belonging to different schools in terms of latent trait levels and, consequently, the corresponding item responses cannot be assumed to be independent. A wide class of models which accounts for the group effect and also encloses the discreteness and multidimensionality of the latent traits is represented by multilevel finite mixture models and its variants for longitudinal data (Vermunt 2003; Skrondal and Rabe-Hesketh 2004; Vermunt 2008; Grilli and Rampichini 2007; Cho and Cohen 2010; Bartolucci et al. 2011); see also Kamata (2001), Maier (2001), and Fox (2005) for multilevel IRT models under the assumption of normality of latent traits. Overall, these types of model are based on specifying different latent traits at each hierarchical level. Latent traits at individual level (first level) affect the test item responses and, in turn, are affected by one or more latent variables at group level (second level) that allow us to represent unobserved group effects. Besides, as a consequence of the discreteness assumption, a multilevel finite mixture model allows us to cluster both individuals and groups in classes, which are homogeneous with respect to the corresponding latent traits and variables.

A further element which is of interest, other than the measurement of the latent traits, is represented by the possible effect of covariates on the level of these traits. In principle, a simple procedure to account for the effect of such covariates involves three consecutive steps: (i) building a mixture IRT model for a set of response variables, (ii) assigning subjects to the latent classes based on their posterior probabilities, and (iii) evaluating the association between class membership and external variables of interest (i.e., gender, age, etc.) using cross-tabulations or a multinomial logistic regression analysis (Vermunt 2010). However, it has been demonstrated (Bolck et al. 2004) that this procedure substantially underestimates the association between class membership and covariates. Subsequent studies introduced correction methods that involve modifying the third step (Vermunt 2010). Moreover, a number of studies (see, for instance, Smit et al. 1999, 2000) showed that incorporating collateral information into various mixture IRT models considerably reduces the standard errors in the item parameter estimates. These authors also demonstrated through simulation studies that latent class assignment can benefit substantially from incorporating external variables associated with the latent classes, especially when the sample size is large. More recently, within a differential item functioning (DIF) framework, further studies (see, for instance, Tay et al. 2011, 2013) proposed to overcome such a three step procedure through an Item Response Theory with Covariates (IRT-C) procedure which allows us to evaluate the effect of class membership and external variables simultaneously. In multilevel settings, covariates may express examinees' characteristics (e.g., gender) as well as schools' characteristics (e.g., geographic area), so that we are dealing with first and second level covariates, respectively.

Motivated by an application in education, in this paper we introduce a multilevel extension of the class of LC-IRT models developed by Bartolucci (2007) to include the group effect due to the aggregation of examinees in different schools, other than the effect of covariates. Specifically, we assume the presence of a single latent trait at school level which affects the abilities at student level. Such individual abilities are then measured through a multidimensional version of LC-IRT models based on a 2PL parameterisation for the conditional probability of a certain response given the underlying ability, in the case of binary items. The proposed model is estimated

through the maximum marginal likelihood method making use of the Expectation–Maximisation (EM) algorithm (Dempster et al. 1977), avoiding in this way the use of three steps methods. The described estimation method is implemented in the R package `MultiLCIRT`;¹ see Bartolucci et al. (2014) for a detailed description of this package in simpler contexts.

The proposed multilevel LC-IRT model with covariates is applied for the analysis of data deriving from two Italian National Tests for the assessment of primary, lower middle, and high-school students, which are developed and yearly collected by the National Institute for the Evaluation of the Education System (INVALSI). Here we focus on the INVALSI Tests administered to middle school students as they are having an increasing relevance in the Italian educational context and their collection will become compulsory in the near future. These data are based on a nationally representative sample of 27,592 students within 1,305 schools and refer to students' results at the Language and Mathematics Tests, administered in June 2009.

With reference to the data mentioned above, we are interested in detecting heterogeneity between examinees and schools, studying the latent score distribution and the size of the latent classes, and examining the relationship between observed covariates and latent traits standing within each latent class, at both hierarchical levels. The aim of this work is twofold. First of all, we aim at clustering students and schools into homogeneous classes of latent traits, evaluating, on one side, the degree to which latent subgroups of examinees show distinct response strategies and, on the other side, the degree to which latent subgroups (or types) of schools differently characterise the expected abilities of their students. Moreover, we aim at assessing if and how examinees' and school covariates (i.e., gender and geographic area, respectively) affect the probability for a student or for a school to belong to each of these latent classes.

The remainder of this paper is organised as follows. In the next section we describe the INVALSI data used in our analysis. The statistical methodological approach employed to investigate the structure of the questionnaires is described in Sect. 3. First, we describe the basic assumptions for the model adopted in our study; then, we illustrate the extension to take into account the multilevel structure of the data and the effect of covariates. Details about the estimation algorithm are given in Sect. 4. Finally, in Sect. 5 we illustrate the main results obtained by applying the proposed approach to the INVALSI datasets and in Sect. 6 we draw the main conclusions of the study.

2 The INVALSI data

The INVALSI Language and Mathematics Tests were administered in June 2009, at the end of the pupils' compulsory educational period. Afterwards, a nationally representative sample made of 27,592 students was drawn (INVALSI 2009a). From each of the 20 strata, corresponding to the 20 Italian geographic regions, a random sample of schools was drawn; allocation of sample units within each stratum was chosen propor-

¹ Downloadable from <http://www.CRAN.R-project.org/package=MultiLCIRT>

tional to an indicator based on the standard deviations of certain variables, such as the school size and the stratum size. Classes within schools were then sampled through a random procedure, with one class sampled in each school. Overall, 1,305 schools (and classes) were sampled.

The INVALSI Language Test includes two sections, a Reading Comprehension section and a Grammar section. The first section is based on two texts: a narrative type text (where readers engage with imagined events and actions) and an informational text (where readers engage with real settings); see [INVALSI \(2009b\)](#). The comprehension processes are measured by 30 items, which require students to demonstrate a range of abilities and skills in constructing meaning from the two written texts. Two main types of comprehension process are considered in developing the items: Lexical Competency, which covers the ability to make sense of words in the text and to recognise meaning connections among them, and Textual Competency, which relates to the ability to: (i) retrieve or locate information in the text, (ii) make inferences, connecting two or more ideas or pieces of information and recognising their relationship, and (iii) interpret and integrate ideas and information, focusing on local or global meanings. The Grammar section is made of ten items, which measure the ability of understanding the morphological and syntactic structure of sentences within a given text.

The INVALSI Mathematics Test consists of 27 items covering four main content domains: Numbers, Shapes and Figures, Algebra, and Data and Predictions ([INVALSI 2009c](#)). The Number content domain consists of understanding (and operation with) whole numbers, fractions and decimals, proportions, and percentage values. The Algebra domain requires students the ability to understand, among others, patterns, expressions and first order equations, and to represent them through words, tables, and graphs. Shapes and Figures domain covers topics such as geometric shapes, measurement, location, and movement. It entails the ability to understand coordinate representations, to use spatial visualisation skills in order to move between two and three dimensional shapes, draw symmetrical figures, and understand and being able to describe rotations, translations, and reflections in mathematical terms. The Data and Previsions domain includes three main topic areas: data organisation and representation (e.g., read, organise and display data using tables and graphs), data interpretation (e.g., identify, calculate and compare characteristics of datasets, including mean, median, mode), and chance (e.g., judge the chance of an outcome, use data to estimate the chance of future outcomes).

All items included in the Language Test are of multiple choice type, with one correct answer and three distractors, and are dichotomously scored (assigning 1 point to correct answers and 0 otherwise). The Mathematics Test is also made of multiple choice items, but it also contains two open questions for which a score of 1 was assigned to correct answers and a score of 0 to incorrect or partially correct answers.

A preliminary analysis (see [Table 1](#)) shows that students' scores are affected by students' gender [male (M), female (F)] and school geographic area [North–West (NW), North–East (NE), Centre, South, and Islands]. Overall, females performed better than males at the Language Test, but worse than males at the Mathematics Test. In both Tests, average percentage scores per geographic area show very different levels of attainment.

Table 1 Average relative score per gender and geographic area for the three dimensions of the INVALSI Tests

	Area					Total
	NW (17.70%)	NE (20.95%)	Centre (20.39%)	South (20.91%)	Islands (20.05%)	
M (50.28%)						
Reading (V1)	0.730	0.711	0.731	0.725	0.701	0.721
Grammar (V2)	0.781	0.743	0.780	0.815	0.781	0.780
Mathematics (V3)	0.770	0.760	0.783	0.795	0.789	0.779
F (49.72%)						
Reading (V1)	0.744	0.737	0.755	0.744	0.732	0.743
Grammar (V2)	0.810	0.778	0.816	0.836	0.812	0.811
Mathematics (V3)	0.744	0.732	0.769	0.792	0.781	0.764
Total	0.763	0.743	0.772	0.785	0.767	0.766

3 Methodological approach

In this section, we illustrate the methodological approach adopted to investigate the students' abilities. First, we review the basic model proposed by [Bartolucci \(2007\)](#) and then we extend it to the multilevel setting.

3.1 Preliminaries

The class of multidimensional LC-IRT models developed by [Bartolucci \(2007\)](#) presents two main differences with respect to classic IRT models: (i) the latent structure is multidimensional and (ii) it is based on latent variables that have a discrete distribution; see [Bacci et al. \(2014\)](#) for a more general formulation for polytomously-scored items. We consider in particular the version of these models based on the 2PL logistic parameterisation of the conditional response probabilities ([Birnbaum 1968](#)).

Let n denote the number of subjects in the sample and suppose that these subjects answer r dichotomous test items that measure s different latent traits or dimensions. For the moment, possible multilevel structures are ignored. Also let $\mathcal{J}_d, d = 1, \dots, s$, be the subset of $\mathcal{J} = \{1, \dots, r\}$ containing the indices of the items measuring the latent trait of type d and let r_d denoting the cardinality of this subset, so that $r = \sum_{d=1}^s r_d$. Since we assume that each item measures only one latent trait, the subsets \mathcal{J}_d are disjoint; on the other hand, these latent traits may be correlated. Moreover, adopting a 2PL parameterisation, it is assumed that

$$\text{logit}[p(Y_{ij} = 1 \mid V_i = v)] = \gamma_j \left(\sum_{d=1}^s \delta_{jd} \xi_{vd}^{(V)} - \beta_j \right),$$

$$i = 1, \dots, n, \quad j = 1, \dots, r. \quad (1)$$

In the above expression, Y_{ij} is the random variable corresponding to the response to item j provided by subject i ($Y_{ij} = 0, 1$ for wrong or right response, respectively), β_j is the difficulty level of item j and γ_j is its discriminating level. Moreover, V_i is a latent variable indicating the latent class of the subject, v denotes one of the possible realisations of V_i , and δ_{jd} is a dummy variable equal to 1 if index j belongs to \mathcal{J}_d (and then item j measures the d th latent trait) and to 0 otherwise. Finally, a crucial assumption is that each random variable V_i has a discrete distribution with support $1, \dots, k_V$, which correspond to k_V latent classes in the population. Associated to subjects in latent class v there is a vector $\xi_v^{(V)}$ with elements $\xi_{vd}^{(V)}$ corresponding to the ability level of subjects in the class with respect to dimension d . Note that, when $\gamma_j = 1$ for all j , then the above 2PL parameterisation reduces to a multidimensional Rasch parameterisation. At the same time, when the elements of each support vector $\xi_v^{(V)}$ are obtained by the same linear transformation of the first element, the model is indeed unidimensional even when $s > 1$.

As for the conventional LC model (Lazarsfeld and Henry 1968; Goodman 1974), the assumption that the latent variables have a discrete distribution implies the following *manifest distribution* of the full response vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ir})'$:

$$p(\mathbf{y}_i) = p(\mathbf{Y}_i = \mathbf{y}_i) = \sum_{v=1}^k p_v(\mathbf{y}_i)\pi_v^{(V)}, \tag{2}$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{ir})'$ denotes a realisation of \mathbf{Y}_i , $\pi_v^{(V)} = p(V_i = v)$ is the weight or a priori probability of the v th latent class, with $\sum_v \pi_v^{(V)} = 1$ and $\pi_v^{(V)} > 0$ for $v = 1, \dots, k_V$. Furthermore, the *local independence assumption* which characterises all IRT models, implies that

$$p_v(\mathbf{y}_i) = p(\mathbf{Y}_i = \mathbf{y}_i \mid V_i = v) = \prod_{j=1}^r p(Y_{ij} = y_{ij} \mid V_i = v), \quad v = 1, \dots, k_V.$$

The specification of the multidimensional LC-2PL model, based on the assumptions illustrated above, univocally depends on: (i) the number of latent classes (k_V), (ii) the number of the dimensions (s), and (iii) the way items are associated to the different dimensions. The last feature is related to the definition of the subsets $\mathcal{J}_d, d = 1, \dots, s$.

3.2 Extension to multilevel setting

The INVALSI data have a hierarchical structure, with students nested into schools, so that students responses are not independent. These groups of students are known, in contrast with the unknown ability classes.

The model illustrated above does not take into account the dependence between item responses of individuals belonging to the same group, which typically arises in the presence of multilevel data, as well as the possible effect of one or more covariates. More generally, it is reasonable to suppose that the ability of a subject coming from

a given latent class of ability is influenced by some unobserved characteristics of the group which she/he belongs to. Such unobserved characteristics define a new latent trait that can be called “group effect”, which adds up to the effect of individual covariates and which may be explained by some group covariates.

In the multilevel context, let Y_{hij} denote the response provided by subject i within group h to item j , with possible values 0 and 1, where $h = 1, \dots, H$, $i = 1, \dots, n_h$, and $j = 1, \dots, r$, with H denoting the number of groups and n_h denoting the size of group h , so that $n = \sum_{h=1}^H n_h$. Note that such a notation is somehow different from that usually adopted in the multilevel model setting (Goldstein 2011), where the first subscript is referred to individuals and the second to groups. However, the reversed order of subscripts allows us to accommodate the third subscript j , which refers to item, coherently with the notation typically adopted in the IRT literature.

Now let $\mathbf{W}_h = (W_{h1}, \dots, W_{hm_U})'$ be a vector of m_U covariates (group level covariates) related to group h and let $\mathbf{X}_{hi} = (X_{hi1}, \dots, X_{him_V})'$ denote a vector of m_V covariates (individual level covariates) for subject i in group h . Besides, according to the definitions given in the previous section, the distribution of the latent traits measured by the questionnaire is described by a latent variable V_{hi} with k_V support points, whereas the group effect is denoted by a discrete latent variable U_h with k_U support points, from 1 to k_U .

The k_V and k_U support points define as many latent classes of individuals and groups, respectively. To avoid any misunderstanding, hereafter we use the term “type” as a synonymous of latent class when we refer to the latent variables U_h at group level.

The relation between V_{hi} and Y_{hij} is based on the formulation illustrated in Sect. 3.1, Eq. (1), where subscript h is added to account for level 2 units (schools):

$$\text{logit}[p(Y_{hij} = 1 \mid V_{hi} = v)] = \gamma_j \left(\sum_{d=1}^s \delta_{jd} \xi_{vd}^{(V)} - \beta_j \right), \quad h = 1, \dots, H, \quad i = 1, \dots, n_h, \quad j = 1, \dots, r. \quad (3)$$

Moreover, since now each V_{hi} depends on U_h and \mathbf{X}_{hi} , then in Eq. (2) there is not any more a constant weight $\pi_v^{(V)} = p(V_{hi} = v)$ for each latent class, but a weight $\pi_{hi,v|u}^{(V)} = p(V_{hi} = v \mid U_h = u, \mathbf{X}_{hi} = \mathbf{x}_{hi})$ depending on U_h and on the individual configuration of \mathbf{X}_{hi} .

The above dependence is represented by a multinomial logit parameterisation (Dayton and Macready 1988; Formann 2007b) for the weights $\pi_{hi,v|u}^{(V)}$, $v = 2, \dots, k_V$, with respect to $\pi_{hi,1|u}$ (or another weight), as follows:

$$\log \frac{\pi_{hi,v|u}^{(V)}}{\pi_{hi,1|u}^{(V)}} = \psi_{0uv}^{(V)} + \mathbf{x}'_{hi} \boldsymbol{\psi}_{1v}^{(V)}, \quad v = 2, \dots, k_V. \quad (4)$$

Each regression parameter in the vector $\boldsymbol{\psi}_{1v}^{(V)}$ corresponds to the effect of individual covariates \mathbf{X}_{hi} on the logit of $\pi_{hi,v|u}^{(V)}$ with respect to $\pi_{hi,1|u}^{(V)}$, whereas $\psi_{0uv}^{(V)}$ is the intercept which is specific for examinees of class v that belong to a school in class (or of type) u .

Let $\pi_{hu}^{(U)} = p(U_h = u | \mathbf{W}_h = \mathbf{w}_h)$ denote the weights associated to the support points for U_h that depend on the group covariate configuration $\mathbf{W}_h = \mathbf{w}_h$. Then, a similar multinomial logit parameterisation is adopted for the conditional distribution of U_h given \mathbf{W}_h :

$$\log \frac{\pi_{hu}^{(U)}}{\pi_{h1}^{(U)}} = \psi_{0u}^{(U)} + \mathbf{w}'_h \boldsymbol{\psi}_{1u}^{(U)}, \quad u = 2, \dots, k_U, \tag{5}$$

where elements of vector $\boldsymbol{\psi}_{1u}^{(U)}$ denote the effect of group covariates \mathbf{W}_h on the logit of $\pi_{hu}^{(U)}$ with respect to $\pi_{h1}^{(U)}$ and $\psi_{0u}^{(U)}$ is the intercept specific for schools within type u .

4 Likelihood based inference

In this section, we deal with maximum likelihood of the extended model based on assumptions (4) and (5). Because one assumes conditional independence (i.e. responses of level 1 units belonging to the same level 2 units are conditionally independent given the corresponding latent variable U_h), the likelihood for a level 2 unit can be obtained as a product of individual likelihoods.

For given k_U and k_V , the parameters of the proposed model may be estimated by maximising the log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{h=1}^H \log \sum_{u=1}^{k_U} \pi_{hu}^{(U)} \rho_h(u), \tag{6}$$

where $\boldsymbol{\theta}$ is the vector containing all the free parameters, and

$$\rho_h(u) = \prod_{i=1}^{n_h} \sum_{v=1}^{k_V} \pi_{hi,v|u}^{(V)} \prod_{j=1}^r p(y_{hij} | V_{hi} = v),$$

with $p(y_{hij} | V_{hi} = v)$ defined as in (3).

The vector $\boldsymbol{\theta}$ contains item parameters β_j (difficulty) and γ_j (discriminating index), covariate parameters $\boldsymbol{\psi}_{1v}^{(V)}$ and $\boldsymbol{\psi}_{1u}^{(U)}$, ability levels $\boldsymbol{\xi}_v^{(V)}$ affecting the individual and group weights $\pi_{hi,v|u}^{(V)}$ and $\pi_{hu}^{(U)}$. However, to make the model identifiable, we adopt the constraints

$$\beta_{j_d} = 0, \quad \gamma_{j_d} = 1, \quad d = 1, \dots, s,$$

with j_d denoting a reference item for the d -th dimension (usually, but not necessarily, the first item for each latent trait); see also Bartolucci (2007). In this way, for each item j , with $j \in \mathcal{J}_d \setminus \{j_d\}$, the parameter β_j is interpreted in terms of differential difficulty level of this item with respect to item j_d ; similarly, γ_j , is interpreted in terms of ratio between the discriminant index of item j and that of item j_d .

Considering the above identifiability constraints, the number of free parameters collected in θ is equal to

$$\#par = (k_U - 1)(m_U + 1) + (k_V - 1)(m_V + k_U) + k_V s + 2(r - s).$$

In fact, there are $(k_U - 1)(m_U + 1) + (k_V - 1)(m_V + k_U)$ regression coefficients for the latent classes, $k_V s$ ability parameters, $r - s$ free discriminating parameters, and $r - s$ free difficulty parameters. Under the Rasch parameterisation, the number of parameters decreases by $r - s$ as the discriminating indices have not to be estimated, as they are all set equal to 1. Note that only the ability parameters $\xi_v^{(V)}$ are estimated, whereas parameters $\xi_u^{(U)}$, which are interpretable as the effect on the students' abilities for schools of type u , are estimated as average of the $\xi_v^{(V)}$ with suitable weights, that is,

$$\hat{\xi}_u^{(U)} = \frac{1}{ns} \sum_{d=1}^s \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{v=1}^{k_V} \hat{\xi}_{dv} \hat{\pi}_{hi,v|u}.$$

In order to maximise the log-likelihood $\ell(\theta)$, we make use of the EM algorithm of [Dempster et al. \(1977\)](#), which is developed along the same lines as in [Bartolucci \(2007\)](#); see also [Bartolucci et al. \(2011\)](#) for a version for longitudinal data.

The *complete log-likelihood*, on which the EM algorithm is based, may be expressed as

$$\ell^*(\theta) = \sum_{h=1}^H \ell_{1h}^*(\theta) + \ell_{2h}^*(\theta) + \ell_{3h}^*(\theta),$$

with

$$\begin{aligned} \ell_{1h}^*(\theta) &= \sum_{i=1}^{n_h} \sum_{j=1}^r \sum_{v=1}^{k_V} z_{hiv} \log p(y_{hij} | V_{hi} = v), \\ \ell_{2h}^*(\theta) &= \sum_{i=1}^{n_h} \sum_{u=1}^{k_U} \sum_{v=1}^{k_V} z_{hu} z_{hiv} \log \pi_{hi,v|u}^{(V)}, \\ \ell_{3h}^*(\theta) &= \sum_{u=1}^{k_U} z_{hu} \log \pi_{hu}^{(U)}, \end{aligned}$$

which is directly related to the *incomplete log-likelihood* defined in (6). In the above expression, z_{hiv} is the indicator function for subject i being in latent class v ($V_{hi} = v$) and z_{hu} is the indicator function for cluster h being of typology u ($U_h = u$). Consequently, $z_{hu} z_{hiv}$ is equal to 1 if both conditions are satisfied and to 0 otherwise.

Obviously, $\ell^*(\theta)$ is much easier to maximise with respect of $\ell(\theta)$, provided that the indicator variables z_{hiv} and z_{hu} are known. Based on function $\ell^*(\theta)$, the EM algorithm alternates the following two steps until convergence in $\ell(\theta)$:

- **E-step.** It consists of computing the expected value of the complete log-likelihood $\ell^*(\theta)$. In practice, this is equivalent to computing the posterior expected values of the indicator variables. In particular, we have that

$$\hat{z}_{hiv} = p(V_{hi} = v | \mathbf{D}) = \sum_{u=1}^{k_U} (\widehat{z_{hu} z_{hiv}}). \tag{7}$$

where \mathbf{D} is a short-hand notation for the observed data. Moreover, we have

$$\begin{aligned} (\widehat{z_{hu} z_{hiv}}) &= p(U_h = u, V_{hi} = v | \mathbf{D}) = p(V_{hi} = v | U_h = u, \mathbf{D}) p(U_h = u | \mathbf{D}) \\ &= \frac{\pi_{hi,v|u}^{(V)} \prod_{j=1}^r p(y_{hij} | V_{hi} = v)}{\sum_{v'=1}^{k_V} \pi_{hi,v'|u}^{(V)} \prod_{j=1}^r p(y_{hij} | V_{hi} = v')} \hat{z}_{hu} \end{aligned}$$

and

$$\hat{z}_{hu} = p(U_h = u | \mathbf{D}) = \frac{\pi_{hu}^{(U)} \rho_h(u)}{\sum_{u'=1}^{k_U} \pi_{hu'}^{(U)} \rho_h(u')}. \tag{8}$$

- **M-step.** It consists of updating the model parameters by maximising the expected value of $\ell^*(\theta)$ obtained at the E-step. As an explicit solution does not exist for the model parameters, iterative optimisation algorithms of Newton–Raphson type are used, but they are of simple implementation. The resulting estimate of θ is used to update the expected value of $\ell^*(\theta)$ at the next E-step and so on.

When the algorithm converges, the last value of θ , denoted by $\hat{\theta}$, corresponds to the maximum of $\ell(\theta)$ and then it is taken as the maximum likelihood estimate of this parameter vector. It is important to highlight that the number of iterations and, in particular, the detection of a global rather than a local maximum point of the target function crucially depend on the initialisation of the EM algorithm. Therefore, following [Bartolucci \(2007\)](#), we recommend to try several different starting values, even randomly chosen, for this algorithm.

The EM algorithm described above is implemented in the R package named `MULTILCIRT` ([Bartolucci et al. 2014](#)). We also clarify that, alternatively, analyses similar to the one here proposed may be performed by means of `Latent GOLD` ([Vermunt and Magidson 2005](#)), `mdltn` ([von Davier 2005](#)), and `Mplus` ([Muthén and Muthén 2012](#)) softwares, which allow for multidimensional IRT models, discrete latent variables, multilevel data structures, and presence of covariates. However, at least to our knowledge, no other software treats multidimensionality and discreteness of latent traits at the same time.

After parameter estimation, each subject i can be allocated to one of the k_V latent classes on the basis of the response pattern y_i she/he provided, her/his covariates x_i , and the typology of group she/he belongs to. Similarly, each group h can be allocated to one of the k_U latent classes. In both cases, the most common approach is to assign the subject and the group to the class with the highest posterior probability, computed as in Eqs. (7) and (8), respectively.

5 Application to the INVALSI dataset

In this section, we illustrate the application of the multilevel finite mixture IRT models to the data collected by the two INVALSI Tests. For the purposes of this analysis, the 30 items which assess reading comprehension within the Language Test are kept distinct from the 10 items which assess grammar competency, as these sections deal with two different competencies. Overall, we consider a model with three dimensions: Reading (V1), Grammar (V2), and Mathematics (V3). Besides, regarding the way of taking the covariate effect into account, we consider subjects classified according to gender, and schools classified according to the geographic area.

In the following, we first deal with the problem of the model selection, regarding in particular the optimal number of latent classes and the item parameterisation. Then, we deal with the analysis of the ability distribution and with the assessment of the covariate effect at both levels of the hierarchy.

5.1 Model selection

In analysing the INVALSI dataset by the model described in Sect. 3, a key point is the choice of the number of latent classes at the individual level and at the group level, that is, k_V and k_U , respectively.

In educational settings, each student's ability may be classified into one of several categories on the basis of cut scores. The setting of cut scores on standardised tests is a composite judgmental process (Loomis and Bourque 2001; Cizek et al. 2004), whose complexities and nuances are well beyond the scope of this work. For the purposes of the analysis described in the following, it is enough to acknowledge that it is possible to select a different number of groups depending on the adopted judgmental criteria. Here, we adopt a widespread classification of students into three groups (i.e., basic, advanced, and proficient), corresponding to $k_V = 3$.

Given the value of k_V , we choose the number of school types relying on the main results reported in the literature about finite mixture models (Biernacki and Govaert 1999; McLachlan and Peel 2000; Fraley and Raftery 2002; Nylund et al. 2007), who suggest the use of the Bayesian Information Criterion (BIC) of Schwarz (1978). On the basis of this criterion, the selected number of school types is the one corresponding to the minimum value of

$$BIC = -2\ell(\hat{\theta}) + \log(n)\#\text{par.}$$

In practice, we fit the model for increasing values of k_U and when BIC starts to increase, the previous value of k_U is taken as the optimal one. Note that, apart from k_V and k_U , the other elements characterising the model, that is, the item parameterisation and the multidimensional structure of items, remain fixed. If one already has some a priori knowledge about the multidimensional structure of the set of items, then it is convenient to adopt it. Otherwise, we suggest to select the number of latent classes taking a very general model based on a different dimension for each item. Similarly, we suggest to adopt a basic LC model in absence of any specific indication about the

Table 2 Log-likelihood, number of parameters and BIC values for $k_U = 1, \dots, 6$ latent classes for the INVALSI test; in boldface is the smallest BIC value

k	$\ell(\hat{\theta})$	#par	BIC
1	-531,346.4	205	1,064,688
2	-530,195.9	212	1,062,455
3	-529,947.7	219	1,062,027
4	-529,829.2	226	1,061,858
5	-529,782.3	233	1,061,833
6	-529,766.5	240	1,061,869

Table 3 Model selection: log-likelihood and BIC values for the Rasch (1PL) model and 2PL model with covariates; in boldface is the smallest BIC value

Model	$\ell(\hat{\theta})$	#par	BIC
1PL	-533,994.6	105	1,069,011
2PL	-530,039.6	169	1,061,724

item parameterisation. For further details about this strategy see, for instance, [Bacci et al. \(2014\)](#).

In the application described in the present paper, for the selection of the school types (i.e., the number of latent classes at the school level) we fit the multilevel LC model with covariates (gender and geographic area) in the three-dimensional version (V1, V2, and V3), in which each item measures just one ability, for values of k_U from 1 to 6. The results of this preliminary fitting are reported in [Table 2](#). On the basis of these results, we choose $k_U = 5$ types of schools, which corresponds to the smallest BIC value.

After the selection of k_V and k_U has been made as described above, alternative models with the given number of classes at the two levels of the hierarchy and the same latent variables are considered. In particular, a Rasch (1PL) model with covariate effects and a 2PL model with covariate effects are fitted (see [Table 3](#)).

The BIC value of the 2PL model with covariates is smaller than that observed for the 1PL model. Therefore, we retain the 2PL model with $k_V = 3$ and $k_U = 5$.

5.2 Distribution of the abilities

In this section, we discuss the estimation of the parameters of the multilevel 2PL model with $k_V = 3$ and three dimensions (V1, V2, and V3) at student level, and $k_U = 5$ at school level.

Ability values are expressed on a standardised scale and class weights are obtained as average values of the estimated individual-specific class weights, denoted by $\hat{\pi}_v^{(V)}$ and obtained as

$$\hat{\pi}_v^{(V)} = \frac{1}{n} \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{u=1}^{k_U} \hat{\pi}_{hi,v|u}^{(V)} \hat{\pi}_{hu}^{(U)}.$$

Table 4 Student level: distribution of the ordered estimated abilities $\hat{\xi}_v^{(V)}$ for the three involved dimensions within classes, together with the average weights ($\hat{\pi}_v^{(V)}$)

	V1	V2	V3	$\hat{\pi}_v^{(V)}$
Class 1	-0.643	1.322	1.214	0.174
Class 2	0.657	2.302	1.671	0.428
Class 3	1.988	3.564	2.214	0.398

Table 5 School level: distribution of the estimated average abilities $\hat{\xi}_u^{(U)}$, for $u = 1, \dots, 5$, and the average weights ($\hat{\pi}^{(U)}$)

	Type 1	Type 2	Type 3	Type 4	Type 5
$\hat{\xi}_u^{(U)}$	1.009	1.531	1.594	2.046	2.451
$\hat{\pi}^{(U)}$	0.081	0.081	0.370	0.351	0.116

Table 4 shows the estimated ability levels and corresponding average weights for the three classes of students and the three involved dimensions. Inspection of these estimates shows that students belonging to class 1 within the two sections (V1 and V2) of the Language Test and the Mathematics Test (V3) tend to have the lowest ability level in relation with the involved dimensions. Overall, the weight of low attainment students grouped in class 1 is quite negligible in terms of class proportions, as they count for slightly more than 17% of the students, overall. Besides, students with the highest ability levels belong to class 3, which counts for a little less than 40%, while class 2 is a class of examinees with intermediate ability levels over the three involved dimensions.

Overall, we observe that predicted abilities over the three dimensions tend to be correlated. In fact, the Spearman correlation coefficients between the three dimensions are always very high (and >0.99), confirming that the three classes tend to group examinees who show consistent levels of ability over the involved dimensions.

At school level, the distribution of the estimated average abilities $\hat{\xi}_u^{(U)}$ for the five chosen types (see Table 5) allows us to qualify the schools from the worst ones, classified in the Type 1, to the best ones, classified in the Type 5. We observe that more than 46 per cent of the schools belong to the best types (Type 4 and 5), whereas only the 16% is classified among the worst ones; the remaining 37% is of intermediate type.

5.3 Effects of level 1 and level 2 covariates

As previously stated, in our multilevel setting, covariates express examinees' characteristics (e.g., gender) as well as school characteristics (e.g., geographic area) and therefore we have first- and second-level covariates. In the following, we discuss estimates of the regression parameters for such covariates, to further study the nature of the classes at both levels of the hierarchy and the substantive differences among them.

Regression parameters for the first-level covariate (i.e., gender) are estimated taking as reference class the one characterised by the worst level of estimated ability over each of the three involved dimensions (Class 1), and category Males as reference category.

Table 6 Estimated conditional probabilities $\hat{\pi}_{h\mu}^{(U)}$ to belong to the five types of schools given the geographic area

	Type 1	Type 2	Type 3	Type 4	Type 5
NE	0.042	0.000	0.498	0.387	0.073
NW	0.036	0.000	0.620	0.329	0.014
Centre	0.031	0.044	0.358	0.472	0.095
South	0.115	0.113	0.239	0.341	0.192
Islands	0.180	0.243	0.140	0.232	0.206

The estimates of the regression parameters ($\psi_{1v}^{(V)}$) used in Eq. (4) over Class 2 and Class 3 (0.117 and 0.175, respectively), and the corresponding standard errors (0.053 and 0.057, respectively) show that females tend to be grouped into these classes, and therefore they tend to score higher than males at the INVALSI Tests.

Similarly, regression parameters for the second-level covariate (i.e., geographic area) are estimated taking Type 1 (i.e., the worst schools) as a reference class, and category NW (North West) as reference category. For easiness of interpretation, the estimated regression parameters and the corresponding standard errors are not shown here and they are replaced by the estimated probabilities to belong to any of the five types of school given the geographic area. Results in Table 6 confirm those of the preliminary descriptive analysis (see Table 1), that is, different levels of attainment according to the school geographic area.

On the whole, the great majority of the Italian schools tend to be classified into average and high attainment schools. However, schools of the North West and North East show a very similar profile, as they display high probability to belong to medium attainment schools and high attainment schools (Type 3 and Type 4 schools, respectively). Finally, schools from the South and the Islands have a relatively higher probability than the schools of the rest of Italy to belong to the best schools (i.e., Type 5) and, at the same time, to the worst schools (Type 1). The latter apparently inconsistent result may be related to the presence in the Southern regions of a few schools with exceptionally positive results (Sani and Grilli 2011).

6 Conclusions

In this article we propose a framework for assessing the relationship between unobserved classes of examinees and schools, and observed characteristics, and establishing the ways observed characteristics are related to unobserved groupings, accounting at the same time for the multilevel structure of our data.

The data analysed by the proposed framework were collected in 2009 by the National Institute for the Evaluation of the Education System (INVALSI) and refer to two assessment Tests—on Italian language competencies (Reading comprehension, Grammar) and Mathematical competencies—administered to middle school students in Italy.

The adopted approach may be seen as an extension of that developed by Bartolucci (2007), by accounting for the multilevel structure of the data and the effects of observed covariates at the students' and school levels. This approach has advantages with respect to other approaches which account for observed variables only, on one side, or on

just latent classes of examinees, on the other side. In fact, at the various levels of the hierarchy, our approach permits the combined use of information derived from observed group membership (i.e., examinees' gender and school geographic area) and unobserved groupings (i.e., latent classes of examinees and type of school) and, thus, to characterise distinct latent classes of examinees and latent classes of schools, which are named "type".

Based on the proposed model, for the data at hand we ascertain the existence of latent classes of examinees who show consistent levels of ability over the involved dimensions, and of a few types of schools, from lowest attainment schools to highest attainment ones. Next, we study the relationship between observed level 1 and level 2 variables, and latent classes. At student level, we find that gender has a significant effect on class membership with females who tend to be grouped in the highest attainment groups of students. At the school level, results reveal how and to what extent factors related to school geographic area affect the probability for a school to be grouped in a certain school type.

Overall, the discussed extension of the latent class IRT model developed by Bartolucci (2007) to account for the multilevel structure of the data and the covariate effect allows us to characterise the classes at the two hierarchical levels in such a way that would not have been detectable through other available models.

References

- Bacci S, Bartolucci F, Gnaldi M (2014) A class of multidimensional latent class IRT models for ordinal polytomous item responses. *Commu Stat Theory 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, Pennoni F, Vittadini G (2011) Assessment of school performance through a multilevel latent Markov Rasch model. *J Educ Behav Stat* 36:491–522
- Bartolucci F, Bacci S, Gnaldi M (2014) MultiLCIRT: an R package for multidimensional latent class item response models. *Comput Stat Data Anal* 71:971–985
- Biernacki C, Govaert G (1999) Choosing models in model-based clustering and discriminant analysis. *J Stat Comput Simul* 64:49–71
- Birnbaum A (1968) Some latent trait models and their use in inferring an examinee's ability. In: Lord FM, Novick MR (eds) *Statistical theories of mental test scores*. Addison-Wesley, Reading, pp 395–479
- Bolck A, Croon M, Hagensars J (2004) Estimating latent structure models with categorical variables: one-step versus three-step estimators. *Polit Anal* 12:3–27
- Bolt D, Cohen A, Wollack J (2002) Item parameter estimation under conditions of test speededness: application of a mixture Rasch model with ordinal constraints. *J Educ Meas* 39:331–348
- Cho SJ, Cohen AS (2010) A multilevel mixture IRT model with an application to DIF. *J Educ Behav Stat* 35:336–370
- Christensen K, Bjorner J, Kreiner S, Petersen J (2002) Testing unidimensionality in polytomous Rasch models. *Psychometrika* 67:563–574
- Cizek G, Bunch M, Koons H (2004) Setting performance standards: contemporary methods. *Educ Meas: Issues Pract* 23:31–50
- Dayton CM, Macready GB (1988) Concomitant-variable latent-class models. *J Am Stat Assoc* 83:173–178
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J R Stat Soc Ser B* 39:1–38
- Formann AK (1992) Linear logistic latent class analysis for polytomous data. *J Am Stat Assoc* 87:476–486
- Formann AK (1995) Linear logistic latent class analysis and the Rasch model. In: Fischer G, Molenaar I (eds) *Rasch models: foundations, recent developments, and applications*. Springer, New York, pp 239–255

- Formann AK (2007a) (Almost) equivalence between conditional and mixture maximum likelihood estimates for some models of the Rasch type. In: von Davier M, Carstensen C (eds) *Multivariate and mixture distribution Rasch models*. Springer, New York, pp 177–189
- Formann AK (2007b) Mixture analysis of multivariate categorical data with covariates and missing entries. *Computat Stat Data Anal* 51:5236–5246
- Fox JP (2005) Multilevel IRT using dichotomous and polytomous response data. *Br J Math Stat Psychol* 58:145–172
- Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* 97:611–631
- Goldstein H (2011) *Multilevel statistical models*. Wiley, Hoboken
- Goodman LA (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61:215–231
- Grilli L, Rampichini C (2007) Multilevel factor models for ordinal variables. *Struct Equ Model* 14:1–25
- Heinen T (1996) *Latent class and discrete latent traits models: similarities and differences*. Sage, Thousand Oaks
- Hojtink H, Molenaar I (1997) A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika* 62:171–190
- INVALSI (2009a) *Esame di stato di primo ciclo. a.s. 2008/2009*. In: INVALSI technical report
- INVALSI (2009b) *Prove invalsi 2009*. In: Report IT (ed) *Quadro di riferimento di Italiano*
- INVALSI (2009c) *Prove invalsi 2009*. In: Report IT (ed) *Quadro di riferimento di Matematica*
- Jiao H, Lissitz R, Macready G, Wang S, Liang S (2012) Exploring levels of performance using the mixture Rasch model for standard setting. *Psychol Test Assess Model* 53:499–522
- Kamata A (2001) Item analysis by the hierarchical generalized linear model. *J Educ Meas* 38:79–93
- Langheine R, Rost J (1988) *Latent trait and latent class models*. Plenum, New York
- Lazarsfeld PF, Henry NW (1968) *Latent structure analysis*. Houghton Mifflin, Boston
- Lindsay B, Clogg C, Greco J (1991) Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J Am Stat Assoc* 86:96–107
- Loomis S, Bourque M (2001) From tradition to innovation: standard setting on the national assessment of educational progress. In: Cizek GJ (ed) *Setting performance standards: concepts methods and perspectives*. Lawrence Erlbaum Associates, Mahwah
- Maier KS (2001) A Rasch hierarchical measurement model. *J Educ Behav Stat* 26:307–330
- Maij-de Meij AM, Kelderman H, van der Flier H (2008) Fitting a mixture item response theory model to personality questionnaire data: characterizing latent classes and investigating possibilities for improving prediction. *Appl Psychol Meas* 32:611–631
- Masters G (1985) A comparison of latent trait and latent class analyses of Likert-type data. *Psychometrika* 50:69–82
- McLachlan G, Peel D (2000) *Finite mixture models*. Wiley, New York
- Mislevy RJ, Verhelst N (1990) Modeling item responses when different subjects employ different solution strategies. *Psychometrika* 55:195–215
- Muthén L, Muthén B (2012) *Mplus user's guide*. Muthén and Muthén edition, Los Angeles
- Nylund KL, Asparouhov T, Muthén BO (2007) Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct Equ Model* 14:535–569
- Rasch G (1961) On general laws and the meaning of measurement in psychology. In: *Proceedings of the IV Berkeley symposium on mathematical statistics and probability*, The Regents of the University of California, pp 321–333
- Reckase MD (2009) *Multidimensional item response theory*. Springer, New York
- Rost J (1990) Rasch models in latent classes: an integration of two approaches to item analysis. *Appl Psychol Meas* 14(3):271–282
- Rost J (1991) A logistic mixture distribution model for polychotomous item responses. *Br J Math Stat Psychol* 44:75–92
- Sani C, Grilli L (2011) Differential variability of test scores among schools: a multilevel analysis of the fifth grade INVALSI test using heteroscedastic random effects. *J Appl Quant Methods* 6:88–99
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
- Skrondal A, Rabe-Hesketh S (2004) *Generalized latent variable modeling. Multilevel, longitudinal and structural equation models*. Chapman and Hall/CRC, London
- Smit A, Kelderman H, van der Flier H (1999) Collateral information and mixed Rasch models. *Methods Psychol Res* 4:19–32

- Smit A, Kelderman H, van der Flier H (2000) The mixed Birnbaum model: estimation using collateral information. *Methods Psychol Res Online* 5:31–43
- Smit A, Kelderman H, van der Flier H (2003) Latent trait latent class analysis of an Eysenck personality questionnaire. *Methods Psychol Res Online* 8:23–50
- Tay L, Vermunt JK, Wang C (2013) Assessing the item response theory with covariate (IRT-C) procedure for ascertaining differential item functioning. *Int J Test* 13:201–222
- Tay L, Newman DA, Vermunt JK (2011) Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organ Res Methods* 14:147–176
- Vermunt JK (2001) The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Appl Psychol Meas* 25:283–294
- Vermunt JK (2003) Multilevel latent class models. *Sociol Methodol* 33:213–239
- Vermunt JK, Magidson J (2005) *Latent GOLD 4.0 user's guide*. Statistical Innovations Inc., Belmont
- Vermunt JK (2008) Multilevel latent variable modeling: an application in education testing. *Austrian J Stat* 37:285–299
- Vermunt JK (2010) Latent class modeling with covariates: two improved three-step approaches. *Polit Anal* 18:450–469
- von Davier M (2005) *mdltm* [computer software]. ETS edn, Princeton
- von Davier M (2008) A general diagnostic model applied to language testing data. *Br J MathStat Psychol* 61:287–307
- von Davier M, Rost J (1995) Polytomous mixed Rasch models. In: Fischer G, Molenaar I (eds) *Rasch models. Foundations, recent developments, and applications*. Springer, New York, pp 371–379