Qual Quant (2017) 51:1167–1182
DOI 10.1007/s11135-016-0323-4

A multidimensional IRT approach for dimensionality assessment of standardised students' tests in mathematics

Michela Gnaldi[1]

Published online: 26 February 2016

Abstract Mathematics proficiency involves several content domains and processes at different levels. This essentially means that mathematics ability is a complex latent variable. In standardised testing, the complex, and unobserved, latent constructs underlying a test are traditionally appraised by expert panels through subjective measures. In the present research, we deal with the issue of dimensionality of the latent structure behind a test measuring the mathematics ability of Italian students from a statistical and objective point of view, within an IRT framework. The data refer to a national standardised test developed and collected by the Italian National Institute for the Evaluation of the Education System (INVALSI), and administered to lower secondary school students (grade 8). The model we apply is based on a class of multidimensional latent class IRT models, which allows us to ascertain the test dimensionality based on an explorative approach, and by concurrently accounting for non-constant item discrimination and a discrete latent variable formulation. Our results show that the latent abilities underlying the INVALSI test mirror the assessment objectives defined at the national level for the mathematics curriculum. We recommend the use of the proposed extended IRT models in the practice of test construction, primarily—but not exclusively—in the educational field, to support the meaningfulness of the inferences made from test scores about students' abilities.

Keywords Standardised national students' tests · INVALSI tests · Mathematics ability · Multidimensional latent class IRT models · Hierarchical clustering

1 Introduction

Students' standardised tests are meant to measure complex latent abilities. Mathematical ability and, within it, mathematical problem solving, are examples of such multifaceted unobservable entities. These attributes are referred alternatively to as latent traits, latent

✉ Michela Gnaldi
michela.gnaldi@unipg.it

[1] Department of Political Sciences, University of Perugia, Via A. Pascoli 20, 06123 Perugia, Italy

variables, latent constructs, as there is no means to directly measure them. In the field of mathematics education, some recent developments (Bartolini Bussi et al. 1999; Douek 2006; Schoenfeld 1992) identify in *understanding*, *problem solving*, and *reasoning* the three main complex constructs involved in mathematics ability. *Understanding* includes the ability to understand and use mathematical notions (and their semiotic representations and properties), operations, and the relationships among them. *Problem solving* implies the ability to deal with a problem, identifying or developing appropriate analyses in relation to the posed problem. *Reasoning* refers to the ability to verify the validity of an assertion or procedure in relation to the given context, also providing an explanation of one's choice.

The assurance that the inferences made from standardised test scores about students' abilities are appropriate for the stated purposes of the test is critical to test validity. In other terms, the key to obtaining valid results from large national surveys is having access to standardised tests which are fit for the intended educational purposes, usually defined at the national level. Thus, test items should align with the requirements of the national framework and cover a wide range of competencies to give students fair opportunity to demonstrate their abilities (Tout and Spithill 2014). National curriculum documents and other legislative texts on expectations for students' learning are the foundations for standardised test content design and specification to measure students' achievement. They serve to communicate to the whole education community (i.e., educators, students, and the public) what is being valued and what students are expected to know and do within a content area and a process at specific points during their formal education (Webb 2006). This is also the case of Italy, where the national *Quadro di Riferimento*, together with the *Indicazioni nazionali per il curricolo della scuola dell'infanzia e del primo ciclo di istruzione* (INVALSI 2012a, b) define the national assessment objectives and the content domains of standardised assessment instruments developed by the National Institute for the Evaluation of the Education System (INVALSI). Such national objectives specifically refer to the three main complex constructs stated above (*understanding*, *problem solving*, and *reasoning*), and are used as a frame of reference for the INVALSI test development.

Typically, during test development, expert panels express judgments about the relevance of the test content domains, the adequacy of the test specifications and the representativeness of the tasks, given the national framework (Kane 2006). A central matter in such a test validation phase is a focus on the abilities covered by the test and their item specifications, which should mirror the stated educational purposes. These abilities are, as previously stated, latent variables, that is, unobservable attributes not directly measurable and appraised by the experts through performances on specific tasks included in standardised tests.

The traditional validation model is liable to a number of criticisms. First, the assurance that the inferences derived from test scores about students' latent abilities are appropriate for the stated purposes of the test is based on experts judgment and, as a consequence, it lacks in objectivity. Second, validity evidence cannot justify conclusions about the interpretation of test scores because it does not involve test scores (Messick 1989; Kane 2006). According to this last key objection, in the traditional approach, validity evidence can play a limited role because it does not provide direct evidence for inferences to be made from test scores.

The aim of this article is to show the utility of the use of multidimensional extensions of IRT models in supporting the validation process of students' standardised tests, by overtaking the typical weaknesses of traditional validity evidence tasks. In fact, as it will be shown in the following, extended IRT models can be used to help identifying latent abilities through objective measures which, also, directly use test scores. Such objective measures produce a clustering of the items of a test, in such a way that the items in the same group are referred to the same ability or latent trait. Besides, as scores are expressed relative to the identified latent abilities, they spread light on students' weaknesses and strengths in relation to these last complex constructs, in this way providing a much reacher information on students' achievement than do single scores on test items. Overall, the outcome of the dimensionality analysis allows us to verify if, and to what extent, the test meets the educational objectives stated at the national level in the relevant content domains.

Specifically, the data we use in this work refer to a national standardised test in mathematics developed and collected by the INVALSI. The INVALSI mathematics test was administered in June 2014 to a national sample of 25348 lower secondary school Italian students (Grade 8). The model we apply is based on a class of multidimensional latent class IRT models, which allows us to test dimensionality by concurrently accounting for non-constant item discrimination (Birnbaum 1968) and a discrete latent variable formulation (Bartolucci 2007; Bartolucci et al. 2014; Gnaldi et al. 2015).

The remainder of this paper is organized as follows. The methodological approaches employed to investigate the latent dimensionality of a test are described in the next section. In Sect. 3, we recall the basics for the model adopted in our study (Bartolucci 2007) and in Sect. 3.1, we describe the clustering algorithm which allows us to cluster items in groups referred to the same latent ability. We describe the data in Sect. 4 and illustrate the main results obtained by applying the proposed approach to the INVALSI data in Sect. 5. Section 6 draws the main conclusions of the study.

2 Statistical approaches for dimensionality assessment of students' standardised tests

Students' standardised tests are intended at measuring latent constructs or content domains. These attributes are referred to as latent traits as there is no means to directly measure them, so that the degree to which a certain latent ability characterizes a student can be merely inferred from overt behaviors, which represent the construct observable manifestation (Bartolucci et al. 2015). In turn, observed behaviors can be seen only as proxies of the associated latent construct (i.e., the responses to the items of a mathematics test can be considered as proxies of the ability under study). Unlike constructs, observable proxies only reflect specific aspects of the construct and, as such, they are not their perfect measures (Raykov and Marcoulides 2011). Another way of thinking about students' abilities is as latent dimensions along which students are positioned and differ one another. As latent dimensions are not directly observable and measurable, students' exact location on the latent traits are not known, and therefore, they are not precisely identified.

When a test measures only a latent ability it is referred to as unidimensional. However, students' standardised tests are often composed by subsets of items measuring different constructs. Tests of the latest type are referred to as multidimensional. The dimensionality assessment of students' standardised tests essentially aims at identifying the number and kind of abilities covered by a test.

The literature has proposed various approaches to evaluate the dimensionality of test items. Item factor analysis (Wirth and Edwards 2007), among parametric approaches, generalizes traditional factor analysis, based on continuous observed responses, to discrete observed responses.

The most known and broadly used form of confirmatory factor analysis is the bi-factor analysis, introduced by Holzinger and Swineford (1937). In bi-factor analysis, the first factor is called general factor and may be seen as the overall latent ability assessed by the test (i.e., the mathematical ability). The remaining factors are called group factors and may be viewed as more specific sub-components of the overall ability assessed by the test (Golay and Lecerf 2011).

The traditional bi-factor analysis has a confirmative nature. Therefore, to apply it, a specific bi-factor structure has to be specified in advance. Such a specification may be difficult to find in practice, and this happens whenever one does not have prior information on the structure of dimensionality of a test. To try to overcome this limitation, Jennrich and Bentler (2011) introduced recently a new method which, however, has not been followed by widespread applications yet, at least to our knowledge. The model is essentially a form of exploratory bi-factor analysis, designed to give a rotated loading matrix with an approximate bi-factor structure. The same authors suggest a possible extension of this procedure, which uses an orthogonal rotation, to account for more general rotation methods (i.e., oblique rotation) so that group factors are allowed to depend one another (Jennrich and Bentler 2012).

Finally, for possible generalization of bi-factor analysis in the context of multidimensional item response theory, we refer the reader to Cai and Hansen (2011). As known, traditional IRT models assume that the associations between the responses of a student are fully accounted for by only one latent trait which, in the educational setting, is the students' ability. The characterization of the latent person space in terms of a single unidimensional ability implies that all items of a test are located on the same scale, contributing to measure a single latent trait (Bartolucci et al. 2015). However, students' assessment tests are often composed by subsets of items measuring different but potentially related constructs or content domains. In such later contexts, the traditional IRT assumption of only one underlying latent variable is inappropriate and restrictive for the data at issue. In fact, the unidimensional approach ignores the differential information on students' ability levels relative to several latent traits, which are confused in the same measurement (Camilli 1992; Embretson 1991; Luecht and Miller 1992).

In the IRT framework, a consecutive approach (Briggs and Wilson 2003) is commonly adopted to assess dimensionality, which consists of modeling each latent trait independently of the others, formulating a specific unidimensional IRT model for each ability. This approach has the advantage of providing person ability estimates and standard errors for each latent variable; however, the possibility that these latent variables are related is ignored. On the other hand, this aspect is accounted for in multidimensional IRT approaches, which enclose the correlation between the latent variables. In fact, in a multidimensional IRT model, each latent trait is assumed to have a direct influence on the responses to a certain subset of items and also an indirect influence on the responses to other items. Hence, the main advantage of using the multidimensional approach is that the structure of the test is explicitly taken into account, so that estimates for the correlation between the latent traits are provided and more accurate parameter estimates are obtained.

The hypothesis of unidimensionality has been tested extensively in the literature, especially in relation to the Rasch model, (Rasch 1961; Glas and Verhelst 1995; Vermunt 2001). Martin-Löf (1973), for example, proposed to test the hypothesis that the Rasch model holds for the whole set of items against the hypothesis that this model holds for two disjoint subsets of items defined in advance. The method has two main limitations, that is, it assumes constant item discrimination and has a confirmative cut. Therefore, whenever the discrimination power of test items is not constant, as it is the case of many standardised test items—such as the INVALSI items at issue, see for instance Gnaldi et al. (2015)—and the structure of dimensionality of a test is not known a priori, the approach cannot be appropriately applied in practice.

To overcome these limitations, Bartolucci (2007) proposed a semi-parametric approach based on a class of multidimensional latent class (LC) IRT models. As it will be further described in the following section, such approach takes into account multidimensional latent traits (Reckase 2009) and more general item parameterisations than those of Rasch-type models (Rasch 1961), that is, the two-parameter logistic (2PL) model introduced by Birnbaum (1968). Moreover, the model at issue represents abilities by a random vector with a discrete distribution common to all subjects. Representing the ability distribution through a discrete latent variable is more flexible than representing it by means of a continuous distribution, as it allows to classify individuals in homogeneous classes having very similar latent characteristics.

Among non parametric approaches developed in the IRT framework, we remind the Mokken Scale analysis (Mokken 1971; Sijtsma and Molenaar 2002) and the Detect method (Stout 1987; Zhang and Stout 1999a, b). Both are based on the computation of the covariance between pairs of items to obtain a synthetic indication of the multidimensionality of a set of items, distinguishing one another for the type of covariance: unconditional in the former case and conditional to the latent variables vector in the latter case. Finally, for multidimensional IRT models applied to the Italian context we refer the reader to Matteucci and Mignani (2015).

3 The multidimensional LC IRT model

The class of multidimensional LC-IRT models developed by Bartolucci (2007), and applied in this paper, 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, meaning that the population under study is made up by a finite number of classes, with subjects in the same class having the same ability level (Lazarsfeld and Henry 1968; Formann 1995; Lindsay et al. 1991); see Bacci et al. (2014) for a more general formulation for polytomously-scored items. In this paper, we consider in particular the version of these models based on the two-parameter (2PL) logistic parameterisation of the conditional response probabilities (Birnbaum 1968).

Let n denote the number of students in the sample and suppose that they answer r dichotomous test items that measure s different latent traits or dimensions. Besides, let $\mathcal{J}_d$, $d = 1, \ldots, s$, be the subset of $\mathcal{J} = \{1, \ldots, r\}$ 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 $r = \sum_{d=1}^{s} r_d$. The subsets $\mathcal{J}_d$ are disjoint as each item measures only a latent trait. On the other hand, we assume that he latent traits may be correlated.

The 2PL parameterisation implies that

$$\operatorname{logit}[p(Y_{ij} = 1 \mid V_i = v)] = \gamma_j \left(\sum_{d=1}^{s} \delta_{jd} \xi_{vd}^{(V)} - \beta_j \right), \quad i = 1, \ldots, n, \quad j = 1, \ldots, r. \tag{1}$$

where Y_{ij} is the response to item j provided by student 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. Besides, V_i is a latent variable indicating the latent class of the subject, v is 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 dth latent trait) and to 0 otherwise. Finally, each random variable V_i has a discrete distribution with support $1, \ldots, k_V$ corresponding to the k_V latent classes in the population.

Associated to subjects in latent class v there is a vector $\boldsymbol{\xi}_v^{(V)}$ with elements $\xi_{vd}^{(V)}$ corresponding to the ability level of subjects in latent class v 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 $\boldsymbol{\xi}_v^{(V)}$ are obtained by the same linear transformation of the first element, the model is indeed unidimensional even when $s > 1$.

The assumption that the latent variables have a discrete distribution implies the following *manifest distribution* of the full response vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ir})'$:

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

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

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

3.1 Dimensionality assessment

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 latent dimensions (s), and (iii) the way items are associated to the different latent dimensions. While latent classes refer to groups of units, latent dimensions refer to groups of variables (i.e., items) in a dataset. As to the number of latent classes (k_V), they can be chosen through a statistical approach—i.e., an information criterium such as the Bayesian Information Criterion (BIC) or the Akaike information criterion (AIC)—or on the basis of subjective choices based on specific objectives or previous research.

Once the number of latent classes k is chosen for the 2PL model expressed by Eq. (1), the next step is the assessment of the dimensionality of the test. For this purpose, we test the hypothesis that the r items of the questionnaire measure $s - 1$ instead of s dimensions: the $s - 1$ dimensions are specified by collapsing two dimensions of the s initial ones into one, and then grouping the corresponding items.

To cluster items into a reduced number of groups, we adopt a hierarchical algorithm which allows us to group items measuring the same ability in the same cluster. The algorithm builds a sequence of nested models. It starts with estimating the most general model, that is, a multidimensional LC 2-PL IRT model with a different dimension for each item—corresponding to the classic LC model—and ends with the most restrictive model, that is, a model with only one common dimension to all items—corresponding to a unidimensional LC 2-PL IRT model. At each step of the procedure, the algorithm estimates a multidimensional LC 2-PL IRT model and reduces the dimensionality of the test of a dimension, by collapsing two items in the same group (or two groups of items in the same group). Specifically, at each step, the following likelihood ratio (LR) test is performed for every pair of possible aggregations of items (or groups of items):

$$LR = 2\sum_{\boldsymbol{y}} n(\boldsymbol{y}) \log\left[\frac{\hat{p}(\boldsymbol{y})}{\hat{p}_0(\boldsymbol{y})}\right], \tag{3}$$

where $\hat{p}(\boldsymbol{y})$ and $\hat{p}_0(\boldsymbol{y})$ are the estimated probability of configuration $\boldsymbol{y}$ under the model with s and $s-1$ dimensions, respectively.

The LR test is thus used to compare models which differ only in terms of their dimensional structure, all the rest keeping constant. This type of statistical test allows us to evaluate the similarity between a general model and a restricted model, i.e., a model which is obtained by the general one by imposing one constraint (i.e., so that the restricted model is nested in the general one). More precisely, the LR test evaluates, at a given significance level, the null hypothesis of equivalence between the two nested models at issue. If the null hypothesis is not rejected, the restricted model is preferred, in the interest of parsimony. If the null hypothesis is rejected, the general model is preferred. We remind that in our framework, the most general model is the one with a dimension for each item, whereas the most restricted model is used when all items belong to the same dimension.

The output of the above clustering algorithm may be displayed through a dendrogram that shows the deviance between the initial (s-dimensional) LC model and the model selected at each step of the clustering procedure. As known, the results of a cluster analysis based on a hierarchical procedure, and the consequent choice of the number of dimensions of a test, depend on the adopted rule to cut the dendrogram, which may be chosen according to several criteria. Bartolucci (2007) proposed a criterion based on the LR test statistic: the dendrogram is cut at the level corresponding to the first aggregation for which the test leads to reject H_0. However, such an approach can be misleading for large samples, because it leads to overestimate the dimensionality of the latent structure. A more suitable rule which may be adopted for large samples, as ours, is based on a suitable information criterion, such as the Bayesian Information Criterion (Schwarz 1978), defined by:

$$BIC = -2\hat{\ell} + (\log n)m,$$

where $\hat{\ell}$ is the maximum of the log-likelihood for a given model, whose number of parameters is equal to m, while n is the sample size. In particular, we select the model (then the number of dimensions) for which the difference between its BIC and that of the LC model becomes positive.

4 The data

The data we refer to in this work is a national standardised test of mathematics developed and collected by INVALSI. The INVALSI mathematics test was administered in June 2014 to a national sample of 25,348 lower secondary school Italian students, at their grade 8. As this is the step where Italian students end their first cycle of instruction, it is a very important stage which laid the foundations for their higher cognitive processes. Our data is derived from a sample of controlled classes, that is, classes where the administration of the INVALSI test has been carried out by inspectors (and not by class teachers) and where, therefore, cheating should be marginal.

The Italian *Quadro di Riferimento* and the *Indicazioni nazionali per il curricolo della scuola dell'infanzia e del primo ciclo di istruzione* (INVALSI 2012a, b) define the national assessment objectives and the content topics of standardised assessment instruments developed by the INVALSI. Such objectives and content topics are used as a beginning point for test content identification. The Italian *Quadro di Riferimento* for mathematics education of the first cycle of instruction is organized in two dimensions: the content dimension, which defines the mathematics content topics test items should cover, and the cognitive dimension, which relates to the processes students activate when answering each test item.

The content dimension is articulated into four content domains: numbers (NU), shapes and figures (SF), relations and functions (RF), and data and previsions (DP). The number content domain consists of understanding, operating with, properties and representations of natural numbers, whole numbers, fractions and decimals, proportions, and percentage values and power and roots. 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. It also includes classification and relationships between mathematical objects; verbal, numeric, symbolic representations of mathematical objects. The shapes and figures domain covers topics such as geometric shapes, measurement, location, and movement, with an emphasis on the ability to represent figures through graphical tools. The data and previsions domain includes three main topic areas: data organization and representation (reading, organizing, and displaying data using tables and graphs), data interpretation (identifying, calculating, and comparing characteristics of datasets, including mean, median, mode), and chance (e.g., judging the chance of an outcome, using data to estimate the chance of future outcomes).

The eight processes refer to the various components of mathematics competence and specifically to:

1. Knowledge and mastery of specific mathematics contents (i.e., mathematics objects, properties, structures)
2. Knowledge and use of simple algorithms and procedures (in the arithmetic, geometric, algebraic, statistics and probabilistic fields) (INVALSI 2012a, b)
3. Knowledge of the different representation forms (verbal, numeric, symbolic, graphical, etc.) and being able to move from one representation to another
4. Solving problems using various strategies in different fields
5. Recognising the measurable nature of objects and phenomena, using measurement tools and quantities, estimating measuring quantities
6. Gradually acquiring typical forms of the mathematics reasoning (i.e., make conjectures, discuss, verify, define, generalize) (INVALSI 2012a)
7. Using mathematical instruments, models and representations to deal with quantitative information in the scientific, technologic, economic and social fields.
8. Recognising shapes and figures in a space e using them to solve geometric and modeling problems.

The 50 items which compose the 2014 INVALSI mathematics test refer together to a content dimension and a process dimension. Table 1 shows the classification of the items of the 2014 INVALSI mathematics test according to the two dimensions. The content dimension and the process dimension covered by each item of the test have to be seen more as prevalent processes than as exclusive processes. In fact, because, in general, any answer to each item involves several levels of knowledge and requires the mastery of several

Table 1 Classification of the items of the 2014 INVALSI mathematics test according to the process dimension and the content dimension involved (*NU* numbers, *SF* shapes and figures, *RF* relations and functions, *DP* data and previsions)

Process	Content				Total
	NU	SF	RF	DP	
1	2	0	0	1	3
2	1	6	1	1	9
3	2	0	0	0	2
4	2	1	1	3	7
5	2	1	0	6	9
6	0	0	0	1	1
7	0	0	14	0	14
8	0	5	0	0	5
Total	9	13	16	12	50

abilities, it is not possible to set a univocal correspondence between a single item and a unique content (in terms of knowledge and ability).

The formats of the items of the 2014 INVALSI mathematics test include multiple choice and open items. Multiple choice items are both simple multiple choice items with one correct answer and three distractors, dichotomously scored (assigning 1 point to correct answers and 0 otherwise) and multiple response (dichotomy) questions, where students are presented with a number of options and are invited to tick or otherwise indicate all those which apply. Open items include both items with univocal answers (i.e., correct answers can be rigidly specified *a priori*), which require to provide a result, or to complete a table or figure, and open items with non-univocal answers, such as those requiring students to describe a procedure or to justify an answer or choice. Overall, 16 items of the 2014 INVALSI mathematics test are of multiple choice type, 21 are multiple response (dichotomy) items, 11 are univocal open-ended items and 2 are non-univocal open items. For the purposes of the analysis described below, we used all the 50 items composing the test, dichotomously re-scored by the INVALSI itself.

5 Application to the INVALSI dataset

In this section, we deal with the dimensionality assessment of the 2014 INVALSI mathematics test which allows us to identify the number and kind of complex abilities covered by the test.

In analysing the INVALSI dataset by the model described in the Sections 3 and 3.1, a key point is the choice of the number of latent classes. In education settings, the ability of each student 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 beyond the aim of this article. 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 $k_V = 3$. Other less subjective criteria for the selection of the number of classes are available, see for example Bartolucci (2007), Bartolucci et al. (2014) and Gnaldi et al. (2015).

The analysis of dimensionality of the 2014 INVALSI mathematics test is carried out through two consecutive steps. In the first step, the clustering algorithm described in Sect. 3.1 has been applied separately to the items which compose each of the four groups defined by the experts on account of their content. Thus, we obtain a clustering of the items which compose the NU (i.e., number) group, a clustering of the items which compose the SP (i.e. shapes and figures) group, and so on. Afterwards, the clusters of items selected at the first step are further clustered in the second step. Therefore, at the second step, the procedure groups clusters of item. This two-steps clustering procedure is necessary as the algorithm cannot be applied to a very high number of items (i.e., the 50 items which overall compose the test).

The output of the clustering algorithm run at the first step is represented by the dendrograms in Figs. 1, 2, 3, and 4, which are referred, respectively, to the numbers (i.e., NU), shapes and figures (i.e., SF), relations and functions (i.e., RF), and data and previsions (i.e., DP) groups.

At this first step of the dimensionality assessment, and differently from step 2, the selection of the number of clusters within each content group is carried out without relying on statistical criteria (i.e., the BIC), as the aim of this first step is to reduce the number of items (i.e., to a smaller number than the initial 50 items) in order to be able to run the procedure in a reasonable time. Table 2 shows the distribution of items in the selected clusters within each group content at Step 1. The cluster selection which results at this step is a compromise between the need to reduce the number of items and the demand to retain as much information as possible. Overall, 16 clusters are selected at this first step.

At the second step, the 16 clusters chosen at the first step are further clustered by running the same procedure a second time. The dendrogram in Fig. 5 and Table 3 show the details of the clusters formed at each step of the procedure.

In particular, Table 3 shows the clustered groups at each step, together with the deviance with respect to the initial model, the degree of freedom, and the values of the increase of BIC with respect to the initial model. Note that the number of steps of the clustering algorithm depends on the number of items (or groups of items, as in our case) minus one. Besides, note that, at each step, the procedure reduces the number of dimensions of a dimension; that is, at the first step, the dimensions are reduced to 15 (from the initial 16), by collapsing together the items in brackets (i.e. 9 and 14), and so on.

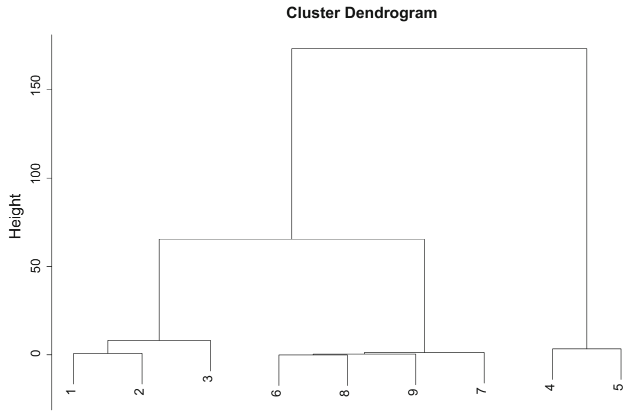

Fig. 1 Step 1: dendrogram for the mathematics test—numbers (NU)

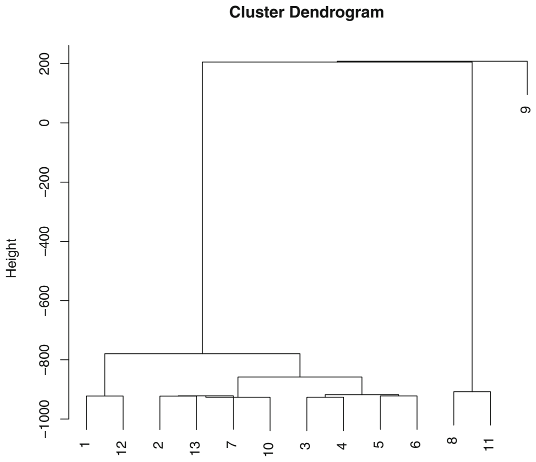

Fig. 2 Step 1: dendrogram for the mathematics test—shapes and figures (SF)

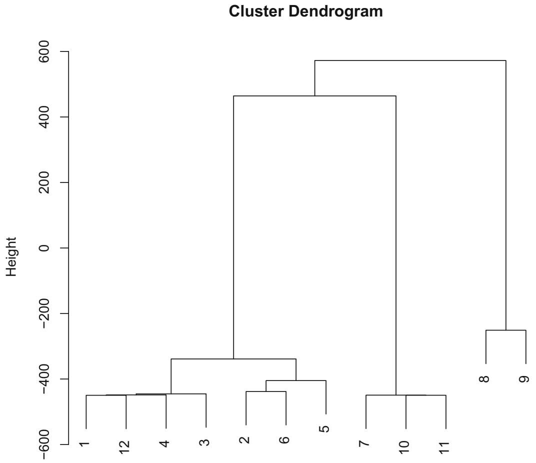

Fig. 3 Step 1: dendrogram for the mathematics test—relations and functions (RF)

Following the method outlined in Sect. 3.1, we adopt as a criterion to cut the dendrogram the one based on BIC. In particular, since BIC tends to select more parsimonious models than other criteria (in particular with large sample sizes), we rely on the increase of BIC with respect to the initial model (i.e., the model with one dimension for each item).

The results in Table 4 show that, with the adopted cut criterion and the chosen number of latent classes, it is reasonable to assume that the test at issue is made up of $s = 2$ groups of items which identify two latent dimensions. The two groups are made of 20 and 30 items, corresponding to different complex latent constructs which may be synthetically characterized as: (i) knowledge and use of mathematical algorithms and procedures and the ability to solve problems; (ii) ability to use mathematical instruments, models and representations to deal with quantitative information in the scientific, technologic, economic and social fields.

The second complex latent construct appears especially typified by items in the *data and previsions* content domain so that, overall, this dimension may be defined as the students' ability to verify and explain the degree of sensibleness of statements and procedures when faced with data representations (i.e., statistical tables and graphs), and

Fig. 4 Step 1: dendrogram for the mathematics test—data and previsions (DP)

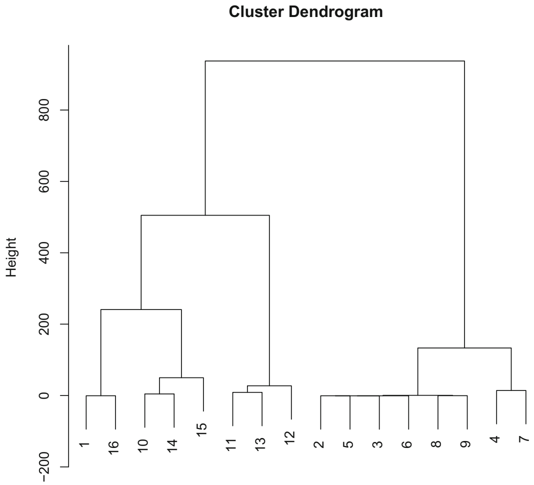

Table 2 Distribution of clusters and items within each group content selected at Step 1

Group content	Cluster labels	Items in cluster (original labels)
NU	10	D4 D6 D11
NU	11	D21 D26 D29 D25
NU	12	D12 D19
SF	13	D2a1 D20
SF	14	D2a2 D22 D14 D17c
SF	15	D2a3 D2b D7 D10
SF	16	D17a D17d D17b
RF	1	D3 D27 D13b D13a
RF	2	D9 D16 D15
RF	3	D18a D18b3 D18b4
RF	4	D18b1 D18b2
DP	5	D1 D28
DP	6	D24a D24b4 D24b5
DP	7	D24b1 D24b3 D24b2
DP	8	D5 D8b2 D8a D8b3 D8b5 D23
DP	9	D8b1 D8b4

chance (e.g., judging the chance of an outcome, using data to estimate the chance of future outcomes). Differently, the first dimension concerns primarily all the content areas but the *data and previsions* one. Therefore, overall, the first complex construct identifies the ability to use numeric, geometric, algebraic algorithms and procedures to solve real problems.

6 Discussion and conclusions

The primary intent of this article is to show the potentialities of extended IRT models in supporting the process of construction and validation of students' standardised tests carried out by national agencies in charge of such tasks.

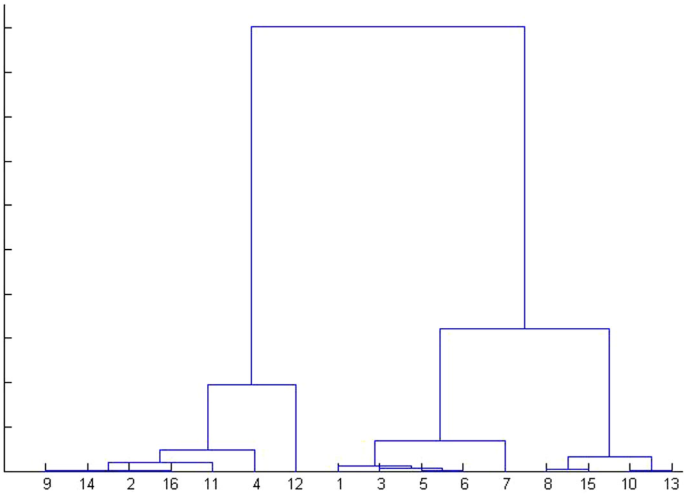

Fig. 5 Step 2: dendrogram for the overall mathematics test

The data we use in this work refer to a national standardised test in mathematics developed and collected by the Italian National Institute for the Evaluation of the Education System (INVALSI). The test was administered in June 2014 to a national sample of 25348 lower secondary school Italian students. We choose to work with these data as this is the stage where Italian students end their primary cycle of instruction, and where the foundations for their higher cognitive processes are laid.

The model we apply is based on a class of multidimensional latent class IRT models (Bartolucci 2007), which allows us to ascertain the test dimensionality based on an explorative approach, and by concurrently accounting for non-constant item discrimination and a discrete latent variable formulation. Relying on this model, we apply a clustering algorithm which identifies clusters of complex and unobserved mathematical abilities underlying the test, which we characterise as the ability to solve mathematical problems, and as the ability to verify and explain the degree of sensibleness of statements and procedures in the data and prevision content domain.

Further, our results can be used to assess the degree of correspondence between the identified latent abilities and the mathematical constructs recognised at the national level as key abilities in the subject. In this sense, we can state that the INVALSI mathematics test adequately covers the three main mathematical abilities (i.e., understanding, problem solving, and reasoning). In fact, the *understanding* and *problem solving* abilities are summarized by the 20 items grouped in the first clustered dimension, and the *reasoning* ability is reflected by the 30 items in the second dimension.

To our knowledge, there are no other attempts to use such extended IRT models in the practice of test construction in the educational field. We recommend their use as they allow national agencies to identify latent abilities through objective measures, providing in this way a non-subjective evidence to support the meaningfulness of the inferences made from test scores about students' abilities. We believe that, without such objective measures, the strength of any inferences made from a standardised test by the experts' opinion is possibly shrunk down by their lack of objectivity.

Table 3 Step 2: diagnostics for the hierarchical clustering algorithm for the mathematics test: number of dimensions (n dim), aggregate groups, deviance with respect to initial model, degrees of freedom, increase BIC with respect to initial model; in boldface is the first BIC positive value

n dim	Aggregate groups	Deviance wrt initial model	df	Increase BIC wrt initial model
15	1,2,3,4,5,6,7,8,10,11,12,13,15,16 (9,14)	00,001	1	−101,403
14	1 3 4 5 6 7 8 10 11 12 13 15 (9,14) (2,16)	02,305	3	−301,909
13	1 3 4 7 8 10 11 12 13 15 (9,14) (2,16) (5, 6)	04,945	5	−502,078
12	1 3 4 7 8 11 12 15 (9,14) (2,16) (5,6) (10,13)	10,612	7	−699,220
11	1 3 4 7 8 11,12,15 (5,6) (10,13) (9,14, 2,16)	22,012	9	−890,629
10	1 3 4 7 11 12 (5,6) (10,13) (9,14,2,16) (8,15)	36,217	11	−1,079,233
9	1 4 7 11 12 (10,13) (9,14,2,16) (8,15) (5,6,3)	54,468	13	−1,263,791
8	4 7 11 12 (10,13) (9,14,2,16) (8,15) (5,6,3,1)	124,695	15	−1,396,373
7	4 7 12 (10,13) (8,15) (5,6,3,1) (9,14,2,16,11)	198,960	17	−1,524,917
6	4 7 12 (5,6,3,1) (9,14,2,16,11) (8,15,10,13)	335,482	19	−1,591,205
5	7 12 (5,6,3,1) (8,15,10,13) (9,14,2,16,11,4)	476,089	21	−1,653,407
4	12 (8,15,10,13) (4,9,14,2,16,11) (5,6,3,1,7)	677,852	23	−1,654,453
3	(8,15,10,13) (5,6,3,1,7) (4,9,14,2,16,11,12)	1,956,491	25	−5,78,622
2	(4,9,14,2,16,11,12) (5,6,3,1,7,8,15,10,13)	3,227,428	27	**489,505**
1	(5,6,3,1,7,8,15,10,13,4,9,14,2,16,11,12)	10,038,189	29	7,097,457

Table 4 Classification of the items of the INVALSI mathematics test into $s = 2$ latent dimensions

Cluster	Items in cluster
Dimension 1	
9	D8b1 D8b4
14	D2a2 D22 D14 D17c
2	D9 D16 D15
16	D17a D17d D17b
11	D21 D26 D29 D25
4	D18b1 D18b2
12	D12 D19
Dimension 2	
1	D3 D27 D13b D13a
3	D18a D18b3 D18b4
5	D1 D28
6	D24a D24b4 D24b5
7	D24b1 D24b3 D24b2
8	D5 D8b2 D8a D8b3 D8b5 D23
15	D2a3 D2b D7 D10
10	D4 D6 D11
13	D2a1 D20

Finally, as the method and processes of validation are central to constructing and evaluating not only performance tests in the educational field, but also other measures in other scientific fields, we advocate the practice to rely on the presented class of extended

IRT models also in the development and validation of measures in the social, health, and human fields.

References

Bacci, S., Bartolucci, F., Gnaldi, M.: A class of multidimensional latent class IRT models for ordinal polytomous item responses. Commun. Stat Theory Methods **43**, 787–800 (2014)

Bartolini Bussi, M.G., Boni, M., Ferri, F., Garuti, R.: Early approach to theoretical thinking: gears in primary school. Educ. Stud. Math. **39**, 67–87 (1999)

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

Bartolucci, F., Bacci, S., Gnaldi, M.: MultiLCIRT: an R package for multidimensional latent class item response models. Comput. Stat. Data Anal. **71**, 971–985 (2014)

Bartolucci, F., Bacci, S., Gnaldi, M.: Statistical Analysis of Questionnaires: A Unified Approach Based on R and Stata. Chapman & Hall/CRCHall, New York (2015)

Birnbaum, A.: Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M., Novick, M.R. (eds.) Statistical Theories of Mental Test Scores, pp. 395–479. Addison-Wesley, Reading, MA (1968)

Briggs, D., Wilson, M.: An introduction to multidimensional measurement using Rasch models. J. Appl. Meas. **4**, 87–100 (2003)

Cai L, Yang JS, Hansen M.: Generalized full-information item bifactor analysis. Psychol. Methods **16**, 221–248 (2011)

Camilli, G.: A conceptual analysis of differential item functioning in terms of a multidimensional item response model. Appl. Psychol. Meas. **16**, 129–147 (1992)

Cizek, G., Bunch, M., Koons, H.: Setting performance standards: contemporary methods. Educ. Meas. **23**(4), 31–50 (2004)

Douek, N.: Some remarks about argumentation and proof. In: Boero, P. (ed.) Theorems in School: From History, Epistemology and Cognition to Classroom Practice. Sense Publishers, Rotterdam (2006)

Embretson, S.E.: A multidimensional latent trait model for measuring learning and change. Psychometrika **56**, 495–515 (1991)

Formann, A.K.: Linear logistic latent class analysis and the Rasch model. In: Fischer, G., Molenaar, I. (eds.) Rasch Models: Foundations, Recent Developments, and Applications, pp. 239–255. Springer, New York (1995)

Glas, C.A.W., Verhelst, N.D.: Testing the rasch model. In: Fischer, G.H., Molenaar, I. (eds.) Rasch Models. Their Foundations, Recent Developments and Applications, pp. 69–95. Springer, New York (1995)

Gnaldi, M., Bartolucci, F., Bacci, S.: A multilevel finite mixture item response model to cluster examinees and schools. Adv. Data Anal. Classif. (2015). doi:10.1007/s11634-014-0196-0

Golay, P., Lecerf, T.: On higher order structure and confirmatory factor analysis of the French Wechsler Adult Intelligence Scale (WAIS-III). Psychol. Assess. **23**, 143–152 (2011)

Holzinger, K., Swineford, S.: The bi-factor method. Psychometrika **47**, 41–54 (1937)

INVALSI: Quadro di riferimento per il primo ciclo di istruzione. Technical report, INVALSI (2012a)

INVALSI: Quadro di riferimento per il secondo ciclo di istruzione. Technical report INVALSI (2012b)

Jennrich, R., Bentler, P.: Exploratory bi-factor analysis. Psychometrika **76**, 537–549 (2011)

Jennrich, R., Bentler, P.: Exploratory bi-factor analysis: the Oblique case. Psychometrika **77**, 442–454 (2012)

Kane, M.: Content-related validity evidence in test development. In: Downing, S.M., Haladyna, T.M. (eds.) Handbook of Test Development. Lawrence Erlbaum Associates, Mahwah, New Jersey (2006)

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

Lindsay, B., Clogg, C., Greco, J.: 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 (1991)

Loomis, S., Bourque, M.: From tradition to innovation: Standard setting on the national assessment of educational progress. In: Cizek, G.J. (ed.) Setting Performance Standards: Concepts Methods and Perspectives. Lawrence Erlbaum Associates, Mahwah, NJ (2001)

Luecht, R.M., Miller, R.: Unidimensional calibrations and interpretations of composite traits for multidimensional tests. Appl. Psychol. Meas. **16**, 279–293 (1992)

Martin-Löf, P.: Statistiska Modeller. Institütet för Försäkringsmatemetik och Matematisk Statistisk vid Stockholms Universitet, Stockholm (1973)

Matteucci M, Mignani S (2015) Multidimensional irt models to analyze learning outcomes of italian students at the end of lower secondary school. In: Millsap R, Bolt D, van der Ark L, Wang W (eds) Quantitative Psychology Research, Springer Proceedings in Mathematics & Statistics, Springer International Publishing Switzerland, vol 89, pp. 91–111

Messick, S.: Validity. In: Linn, R.L. (ed.) Educational Measurement. American Council on Education and Macmillan, New York (1989)

Mokken, R.: A Theory and Procedure of Scale Analysis. De Gruyter, Berlin, Germany (1971)

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, University of California Press, pp. 321–333

Raykov, T., Marcoulides, G.A.: Introduction to Psychometric Theory. Routledge, Taylor & Francis Group, New York (2011)

Reckase, M.: Multidimensional Item Response Theory. Springer, NewYork (2009)

Schoenfeld, A.: Learning to think mathematically: problem solving, metacognition, and sense making in mathematics. In: Grows, D. (ed.) Handbook for Research on Mathematics Teaching and Learning. Macmillan, New York (1992)

Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)

Sijtsma, K., Molenaar, I.: Introduction to Nonparametric Item Response Theory. Sage, Thousand Oaks (2002)

Stout, W.: A non parametric approach for assessing latent trait unidimensionality. Psychometrika **52**(4), 589–617 (1987)

Tout, D., Spithill, J.: The challenges and complexities of writing items to test mathematical literacy. In: Turner, R., Stacey, K. (eds.) Assessing Mathematical Literacy, The PISA Experience. Springer, New York (2014)

Vermunt, J.: The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. Appl. Psychol. Meas. **25**, 283–294 (2001)

Webb, N.L.: Identifying content for student achievement tests. In: Downing, S.M., Haladyna, T.M. (eds.) Handbook of Test Development. Lawrence Erlbaum Associates, Mahwah, New Jersey (2006)

Wirth, R., Edwards, M.: Item factor analysis: current approaches and future directions. Psychol. Methods **12**(1), 58–79 (2007)

Zhang, J., Stout, W.: Conditional covariance structure of generalized compensatory multidimensional item. Psychometrika **64**(2), 129–152 (1999a)

Zhang, J., Stout, W.: The theoretical detect index of dimensionality and its application to approximate simple structure. Psychometrika **64**(2), 213–249 (1999b)