Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione

WORKING PAPER N. 33/2018

A comparison of the goodness-of-fit of three multidimensional IRT models to the INVALSI data.

Simone Del Sarto - INVALSI

https://orcid.org/0000-0003-1102-2881

Michela Gnaldi – Department of Political Science, University of Perugia

https://orcid.org/0000-0002-2785-3279

Collana: Working Papers INVALSI

ISSN: 2611 - 5719

Abstract

The present work aims at comparing the performance of three multidimensional Item Response Theory (IRT) models when applied to the INVALSI data. Such models are extensions of the classical unidimensional IRT models, since they assume that the response process to the test items depends on several, potentially correlated, latent traits (rather than on a unique latent trait), in addition to the specific item characteristics. Among multidimensional models, further issues concern the possibility that each item contributes to measure only a latent trait (between-item multidimensionality), in contrast to the within-item multidimensionality, in which more latent traits can simultaneously affect the item response. In this study, we consider INVALSI data on the mathematics test administrated in 2016 to students at lower secondary school level (Grade 8). Three multidimensional IRT models are applied. In particular, within a between-item multidimensionality context, we consider the Multidimensional Item Response Theory model – assuming a continuous distribution of the latent trait – and the Multidimensional Latent Class IRT model, in which the latent trait distribution is hypothesised to be discrete. In the context of within-item dimensionality, instead, the two-tier Latent Class IRT model is considered, in which we suppose the existence of two underlying multidimensional, but uncorrelated, latent traits, both having a discrete distribution. First results show that the two-tier Latent Class IRT is the model that best fits the INVALSI data at hand.

Keywords: Item Response Theory models; multidimensionality; INVALSI mathematics test

1. Introduction

Students' proficiency (e.g., mathematical ability), like many psychological attributes, is latent by nature, since it is impossible to have a direct manifestation of it. For this reason, the response pattern provided by some students to a set of specific test items may be employed to infer such an unobservable construct. In fact, since there are no ways to directly measure it, the degree to which a certain latent ability characterises a student can be uniquely inferred from overt behaviours, representing the construct observable manifestation (Bartolucci et al., 2015).

Standardised tests can be employed as measurement tool of students' abilities, since the response pattern can be considered as a direct manifestation of the respondents' proficiency. Such tests are characterised by *i.* homogeneity of respondents' working conditions (same test items and same available time) and *ii.* objectivity, that is, test correction is accomplished according to a prespecified protocol, so that the correction is independent by the person who carries out it (INVALSI, 2016b). As such, the tests administrated by the Italian National Institute for the Evaluation of the Education System (INVALSI) are typical cases of standardised test. These tests (on Italian, grammar and mathematics) are annually administrated to school students, with a different content according to the school level the students belong to: grade 2 and 5 (primary school), grade 8 (lower secondary school) and grade 10 (upper-secondary school).

In order to measure students' achievement, standardised test content design and specification have to be founded on the national curriculum documents, the national legislation, and expectations for students' learning. They are communication tools to the entire education community (i.e., teachers, students, the public) about the broad evaluation objectives, i.e., what students are supposed to know and to do within a content area and a process at specific points during their formal education (Webb, 2006). For this purpose, there should be an alignment between the items of a test and the national framework requirements; moreover, the items should cover a broad range of competencies to give students fair opportunity to show their abilities (Tout and Spithill, 2014). The INVALSI test objectives are defined and detailed in the *Quadro Teorico di Riferimento* (INVALSI, 2017), together with the *Indicazioni nazionali per il curricolo della scuola dell'infanzia e del primo ciclo di istruzione.* These documents define the conceptual key points that are fundamental to build the test, the characteristics in terms of cognitive processes requested for solving the tasks and the operational criteria to be used in the test building process along the four school levels (INVALSI, 2016b).

In this paper, the attention is paid on the INVALSI mathematics tests. Mathematical ability is a very complex and multifaceted phenomenon; in fact, when responding to a set of specific items, different but potentially related sub-abilities are involved.

In this regard, some recent developments (Bartolini Bussi et al. 1999; Douek 2006; Gnaldi, 2016, Gnaldi and Del Sarto, 2016) investigate the complex structure of ability in mathematics, highlighting that it cannot be considered a simple and unique construct (i.e., unidimensional), as it engages several content domains and processes at different levels. In other terms, mathematics ability is a multidimensional construct.

Several assessment tools can be employed to investigate the structure of multidimensionality of a test. They can be divided into two main groups: confirmatory and exploratory. The former may be carried out when one knows a priori the multidimensional structure of a test, that is, the groups of items that contribute to measure the specific dimensions are known in advance. Alternatively, exploratory methods can be used when no prior information is available for the test structure. Both can be used for ascertaining the number of dimensions measured by a test and the clusters of items contributing to measure them.

As reported in the technical report (INVALSI, 2016b), the statistical methodology considered by INVALSI to assess the uni/multidimensionality of the data collected within its national survey is the Underlying Variable Approach (UVA; Moustaki, 2000), using the MPLUS software (Muthén and Muthén, 2010). This method assumes that observed variables (i.e., the dichotomous responses to test items) are partial realisations of continuous latent variables with Normal distribution, and it is appropriate for the context at issue as the INVALSI data consist in a data matrix of dichotomous variables (i.e., the wrong/correct responses of students to the test items). In the UVA, the tetrachoric correlation is considered to estimate the association between underlying continuous variables. Besides, in order to evaluate the structure of dimensionality of the data, a multi-criteria approach is followed, according to indexes of model goodness-of-fit (Chi-Squared test, Root Mean Square Error of Approximation, Standardized Root Mean Square Residual) and other typical measures of Factor Analysis (ratio between the first two eigenvalues, eigenvalues scree test, range of the factorial saturations). Our proposal, which can be considered as a further methodological possibility to the current INVALSI approach for dimensionality assessment, is entirely based on Item Response Theory (IRT) models. Such models are very suitable statistical tools to infer a psychological construct, starting from a test response data matrix. IRT models assume that the response process to a set of items depends on *i.* some item features (e.g., difficulty, discrimination) and *ii.* the personal characteristics of the respondent, generally called “latent ability” or “latent trait”, since it cannot be directly observed. Classical IRT models assume unidimensionality, that is, the latent ability underlying the test response process is unique and can be statistically represented by a univariate latent variable. However, a test – like the INVALSI one – is often composed by subsets of items measuring different but potentially related sub-constructs of the main study object. For this kind of tests, Multidimensional IRT models (Reckase, 2009) are very useful, since they take into account that the underlying

latent ability, which influences the response process, is made of several dimensions: this translates in the presence of a multivariate latent variable in the model specification.

The specific purpose of this paper is to compare three multidimensional IRT models applying them to the INVALSI mathematics test data, in order to understand which of them best fits the data at issue. In particular, to investigate the potential multidimensional nature of the INVALSI mathematics test, we adopt a confirmatory approach. For this purpose, we exploit some prespecified item classifications made within the same INVALSI as possible multidimensional setting. In fact, as specified in the *Quadro Teorico di Riferimento* (INVALSI, 2017), the INVALSI mathematics test is built considering two main types of item classifications, one based on the item contents, divided in four groups and the other referred to seven/eight cognitive processes involved when responding to the questions. Moreover, another item classification recently introduced refers to the goal of the National Indications, which groups the items in three main dimensions.

The three IRT models compared in this study are the Multidimensional IRT (MIRT) model (Reckase, 2009), in which a Normal distribution is supposed for the underlying latent variable, the Latent Class (LC) MIRT (Bartolucci, 2007), which is a discrete version of the first model, and the two-tier LC MIRT (Bacci and Bartolucci, 2016), in which the response process is influenced by two (multidimensional) latent variables. This latter model allows for the so-called "within-item multidimensionality", that is, the possibility that an item contributes to simultaneously measure two dimensions, in contrast to the "between-item multidimensionality", which assumes that the response to an item is affected by only one latent trait. This paper is organised as follows: Section 2 briefly describes the data considered in this work, that is, the INVALSI mathematics test administrated in 2016. In Section 3 the statistical methodologies are described, while the results of our analysis are shown in Section 4. Finally, Section 5 ends with some concluding remarks.

2. The INVALSI mathematics test data

The data we deal with in this paper refer to the INVALSI mathematics test administrated in June 2016 to students of lower secondary school (grade 8). In particular, only the data collected in the so-called "sample classes" are considered: within such classes, the test is administrated in the presence of an external supervisor, whose main tasks are the monitoring of the test administration to ensure the respect of the procedures and to report the students' responses on specific electronic forms made available by INVALSI (INVALSI, 2016a). These data consist in the response pattern provided by 27,955 students to the 43 multiple-choice items making up the test, for which the wrong/correct response is coded with 0 and 1, respectively.

INVALSI develops the mathematics test according to two different schemes, one tied with the mathematical *contents* and the other referred to the *processes* used by the students when responding to the

question. As far as the first scheme is concerned, the test items are classified in four contents, according to the Italian *Quadro Teorico di Riferimento* for mathematics education of the first cycle of instruction (INVALSI, 2017): Numbers (NU), Shapes and Figures (SF), Relations and Functions (RF), and Data and Previsions (DP). Moreover, the second scheme classifies the questions into seven cognitive processes, as follows:

1. knowledge and mastery of specific mathematics contents;
2. knowledge and use of algorithms and procedures;
3. knowledge of the different representation forms and ability to move from one representation to another;
4. solving problems using various strategies in different fields;
5. gradually acquiring typical forms of mathematics reasoning;
6. using mathematical instruments, models and representations to deal with quantitative information in the scientific, technological, economic and social fields;
7. recognising shapes and figures in a space and using them to solve geometric and modelling problems.

Furthermore, another possible classification of the test items is according to the *goals* for the proficiency development, in line with the National Indications of the first cycle of instruction. Each item is connected with a goal of the National Indications, and, in turn, such goals are aggregated in three dimensions: Understanding (UND), Problem Solving (PS), and Reasoning (REAS). A description of the 43 items is provided in Table 1, along with the observed proportion of correct response for each item.

3. Statistical methods

Item Response Theory (IRT) models are broadly used statistical methods to infer the response pattern to a questionnaire/assessment test. In particular, as already outlined in Section 1, in contrast with the classic test theory, such models assume that the response process to a set of items depends on *i.* some item features (e.g., difficulty, discrimination) and *ii.* the personal characteristics of the respondent, generally called “latent ability” or “latent trait”, since it cannot be directly observed. In fact, if the interest is to infer some psychological characteristics of people (for example, the mathematics ability, a service satisfaction, a health status, etc.), since such phenomena are unobservable by nature, it is important to deal with an observable manifestation of them, such as the responses to a set of test items in mathematics, or to a questionnaire investigating a service satisfaction or a health status.

Classic IRT models assume unidimensionality, that is, the latent variable representing the underlying person ability, say U, is univariate with a specific distribution (e.g., Normal). Moreover, most unidimensional IRT

(UIRT) models hypothesise that the probability of a correct response to a test item increases as U increases (monotonicity assumption). Another important assumption of such models is the local independence, that is, the responses are conditionally independent given the level of ability U.

Several UIRT models have been proposed, which differ in the functional form connecting the response process to the item and person's characteristics. In this paper, we focus on the two parameter logistic (2-PL) parametrisation for the conditional response probability (i.e., the probability of a correct response given the ability level), in which two item parameters are assumed to affect the above mentioned probability of response, that is, the difficulty parameter and the discrimination parameter.

In many contexts, and especially in the educational context, the unidimensional assumption does not fit the realty of the data at hand, since psychological and educational attributes are complex, multifaceted and present several aspects: for these reasons, such latent constructs cannot be correctly represented through a single latent variable. Similarly, mathematical proficiency investigated by the INVALSI test considered in the present paper is a multifaceted unobservable construct, which involves several content domains and processes at different levels (Bartolini Bussi et al. 1999; Douek 2006; Gnaldi, 2016, Gnaldi and Del Sarto, 2016).

Given the above, multidimensional IRT models (MIRT) have been proposed: the main change with respect to the UIRT is that now the underlying latent trait is composed by D dimensions, then represented by a D-dimensional random vector $\mathbf{U}$ (instead of the single random variable U) with a multivariate distribution. In particular, in this paper we consider three classes of MIRT:

- the multidimensional IRT model (Reckase, 2009), in which a multivariate Normal distribution is assumed for $\mathbf{U}$ (N-MIRT);
- the Latent Class multidimensional IRT model (Bartolucci, 2007), in which $\mathbf{U}$ has a discrete distribution (LC-MIRT);
- the two-tier Latent Class multidimensional IRT model (Bacci and Bartolucci, 2016), in which the response process depends on two uncorrelated and multidimensional random variables (2T LC-MIRT).

These models are briefly described in the following of this section, when they are applied to the case of dichotomously-scored items, that is, items in which the response can be correct or incorrect.

Let us assume that the response variable of person $i = 1, \ldots, n$ to item $j = 1, \ldots, J$, is denoted by Y_{ij}, with possible values equal to 0 or 1 for wrong or correct response, respectively. The N-MIRT model represents the conditional probability of a correct response as follows, given that subject i has $\boldsymbol{u}_i$ as ability level:

$$\text{logit}\, P\left(Y_{ij} = 1 \middle| \mathbf{U}_i = \boldsymbol{u}_i\right) = d_j + \sum_{l=1}^{D} a_{jl} u_{il}, \qquad (1)$$

where d_j is the item intercept, a_{jl} is the item slope – which measures the item discrimination – with respect to dimension l, and $\boldsymbol{u}_i$ is the latent trait vector with elements u_{il}, $l = 1, \ldots, D$ and a D-variate Normal distribution. In Equation (1), the general 2-PL parameterisation is considered, which implies that different item slopes are admissible for every dimension $l = 1, \ldots, D$. It follows that a generic item j can potentially load on all the D different dimensions of **U** and this, in turn, means that the item response can be affected by more than one dimension (*within-item dimensionality*). However, in this study we consider the N-MIRT model in such a way that each item can load on only one pre-specified dimension (*between-item dimensionality*), hence the a_{jl}'s are constrained to be non-zero only in correspondence of the dimension that it contributes to measure, and 0 otherwise. Furthermore, the item intercept can be interpreted as the item difficulty, employed in the following models. In fact, the item difficulty can be obtained by Equation (1), dividing the intercept of opposite sign by the (unique item) slope (discrimination).

The LC-MIRT differs with respect to the above model in the distribution of the latent trait and has a slightly different parameterisation (difficulty/discrimination, rather than intercept/slope). In fact, it is supposed that the latent vector of abilities has a multivariate discrete distribution with k support points, $\mathbf{u}_1, \ldots, \mathbf{u}_k$, and mass probabilities □1, …, □k. The support points identify classes of individuals (i.e., subgroups) that are homogeneous with respect to the latent trait: the generic element u_{cl} represents the ability level of individuals who belong to class c with respect to dimension l, $c = 1, \ldots, k$ and $l = 1, \ldots, D$. Then, each $\boldsymbol{u}_c$ is again a D-dimensional vector. The LCMIRT is based on the following equation, representing the conditional probability of a correct response, given that subject i belongs to latent class c, thus, with a level of ability represented by the vector $\boldsymbol{u}_c$:

$$\text{logit}\, P\left(Y_{ij} = 1 \middle| \mathbf{U}_i = \boldsymbol{u}_c\right) = \gamma_j\left(\sum_{l=1}^{D} \delta_{jl} u_{cl} - \beta_j\right), \qquad (2)$$

where □jis the discrimination power of item j and □jrepresents its difficulty level. Moreover, □jl is an indicator variable, equal to 1 if item j contributes to measure dimension l, and 0 otherwise, $l = 1,$
$\ldots, D$.

Finally, the 2T LC-MIRT is an extension of the above model, in which two latent variables, say **U** and **V**, affect the item response process. Specifically, such latent variables are supposed to be both multidimensional with dimension D_V and D_U, respectively, but they are uncorrelated. In the two-tier model, it is possible that an

item simultaneously loads on a dimension of **U** and a dimension of **V**, but not on two dimensions of the same latent variable. Moreover, **U** and **V** have a discrete distribution with k_U and k_V support points, each having specific mass probabilities. Like in the LC-MIRT model, these support points identify subgroups of individuals with similar characteristics in terms of latent traits represented by **U** and **V**. Poorly speaking, these two latent variables represent two multidimensional but uncorrelated abilities and are referred to the same latent phenomenon (e.g., mathematical proficiency).

The 2T LC-MIRT assumes that, for $c_U = 1, \ldots, k_U$ and $c_V = 1, \ldots, k_V$:

$$\text{logit}\, P\left(Y_{ij} = 1 \middle| \mathbf{U}_i = \boldsymbol{u}_{c_U}, \boldsymbol{V}_i = \boldsymbol{v}_{c_V}\right) = \gamma_{Uj} \sum_{l_U=1}^{D_U} \delta_{jl_U} u_{c_U l_U} + \gamma_{Vj} \sum_{l_V=1}^{D_V} \delta_{jl_V} u_{c_V l_V} - \beta_j, \qquad (3)$$

where δ_{jl_U} and δ_{jl_V} are again indicator variables, equal to 1 if item j loads on dimension l_U or l_V, respectively, $l_U = 1, \ldots, D_U$ and $l_V = 1, \ldots, D_V$. The item difficulty level is again represented by the parameter β_j, while here we have two discrimination parameters, denoted by γ_{Uj} and γ_{Vj}, since each item response can be affected by both **U** and **V**.

The three above methods have similar estimation methods, based on the maximisation of the model log-likelihood. However, a detailed description of such procedures is out of the scope of this work: the reader can refer to the original papers for further details on the models at issue.

4. Results

As outlined above, a confirmatory approach is adopted to the INVALSI data described in Section 2 to understand which multidimensional model, among those described in the previous section, better fits the data at issue. All the analyses reported in this paper are obtained through specific packages within the R environment (R Core Team, 2017). Specifically, the N-MIRT model is run through the '`mirt`' package (Chalmers, 2012), while the '`MultiLCIRT`' (Bartolucci, 2014) and the '`MLCIRTwithin`' (Bacci and Bartolucci, 2016) packages are used for the LC-MIRT and the 2T LC-MIRT, respectively.

In order to choose "the best" model, Information Criteria are employed. In particular, in this paper we consider two widely-known Information Criteria, that is, the Bayesian Information Criteria (BIC; Schwartz, 1978) and the Akaike Information Criteria (AIC; Akaike, 1973). As it is well known, the best model is that showing the minimum Information Criterion.

The analysis reported in this paper is performed through three steps:

1. a comparison between two possible parameterisations (1-PL vs. 2-PL), under the unidimensional assumption, within the MIRT setting;
2. the assessment of the dimensional structure of the INVALSI test, using the best parameterisation chosen at the previous step, within the MIRT setting; among several multidimensional structures hypothesised for the data at issue, we aim at choosing the one that best fits the data;
3. a comparison among multidimensional models accounting for between-item multidimensionality (i.e., each item contributes to measure only a latent trait), and for within-item multidimensionality (i.e., more latent traits can simultaneously affect the item response).

The first step is conducted to assess which type of parameterisation is preferable for the INVALSI data, that is, the two-parameter logistic (2-PL) parameterisation in contrast to the one-parameter logistic (1-PL) one, also known as Rasch model (Rasch, 1961), in which all the items are supposed to equally discriminate. To this aim, the analysis is performed using either the Normal Unidimensional IRT model (N-UIRT) and its Latent Class version (LC-UIRT), that is, the unidimensional version of the N-MIRT and the LC-MIRT, respectively. Besides, the latter model needs to specify in advance the number of latent classes *k*, that is, the number of groups in which the students can be partitioned according to mathematical ability. Such decision can be made according to statistical methods (comparing models with different values of *k* and selecting the best one), or using subjective criteria, based, for example, on previous knowledge or research on the study object. In this paper, we use a statistical criterion (through the BIC and the AIC) to select the number of latent classes: thus, the LC-MIRT model is run for increasing values of *k* (up to 7) and the best one is chosen according to the BIC and the AIC. We remind that *k* represents the groups of ability in which the students may be clustered: for example, if we consider $k = 3$, we suppose that students may be grouped in three classes of ability, which might be labelled, for example, as “low”, “medium” and “high” ability students. Finally, the decision to not consider beyond 7 groups depends on the fact that the INVALSI test is part of the final exam of Italian lower secondary school students. Since the Italian school marking system is expressed in tenth, the INVALSI test results for each student must be successively converted in a vote expressed in tenth, from 4 to 10: in this way, we have 7 different final grades, hence a maximum of 7 groups of students.

The results for the choice of the type of parameterisation (1-PL vs. 2-PL) is reported in Table 2. Here, the BIC and the AIC are shown for the unidimensional version of the LC-MIRT model, estimated on the data at issue for increasing values of *k* and using the two types of parameterisation. In the last row, the BIC and the AIC of the unidimensional N-MIRT model are also reported. As we can observe, looking at the LC-MIRT, the 2-PL parameterisation is preferable with respect to the 1-PL, since the models using two item parameters

exhibit substantially lower BIC (or AIC) than 1-PL models, regardless of the number of latent classes. This behaviour is also confirmed using the N-MIRT. The next step of the analysis is devoted to evaluate the dimensional structure of the data. Now, the interest is to investigate if the latent construct to measure – i.e., the mathematics proficiency of Italian students – can be considered unidimensional or multidimensional, given the observed data. To this end, three possible multidimensional schemes are considered, obtained specifying three possible item aggregations according to Table 1. Specifically, in the first multidimensional scheme, the number of dimensions is chosen on account of the item content, so a four dimensional structure is employed (Numbers, Shapes and Figures, Relations and Functions, and Data and Previsions), labelled as CONT. The second one considers the processes listed in Section 2, so a 7-dimensional structure is specified (labelled as PROC). The last structure is built assuming three dimensions corresponding to the three goal-tied dimensions (Understanding, Problem Solving, and Reasoning), labelled as GOAL. All the previous multidimensional structures are specified using the between-item multidimensionality assumption, so each item response is supposed to be affected by a unique dimension. Results of this step of the analysis are reported in Table 3a, which shows the BIC (top panel) and the AIC (bottom panel) obtained for these three multidimensional structures, under the LC-MIRT and the N-MIRT settings. Moreover, such table also reports the BIC and the AIC of the unidimensional models, labelled as UNI (just reported in Table 1), for a quick comparison.

The first important result to comment on is about the main question of this research, that is, if mathematical ability, measured by the data at issue, can be considered unidimensional or multidimensional. To this aim, we compare the BIC (and the AIC) of a unidimensional model with respect to its multidimensional version. As far as the N-MIRT is concerned, we can observe that the BIC (and the AIC as well) of the UNI model is greater with respect to each multidimensional model: this is a first evidence that the construct analysed here can be considered multidimensional. Besides, among the three multidimensional schemes mentioned above, the best is the one which considers item contents (BIC = 1,359,439; AIC = 1,358,681), then a 4-dimensional structure for our data.

This result is confirmed by the LC-MIRT model. As specified before, it requires the specification of the number of latent classes *k*, so each multidimensional model is run for increasing values of *k* (up to 7). The LC-MIRT results of Table 3a firstly reveal that the multidimensional setting is again preferable with respect to the unidimensional one, since, for each value of *k*, the unidimensional model has worse performance (i.e., larger BIC or AIC) than its multidimensional counterpart. Moreover, as far as the number of latent classes *k* is concerned, we can observe that the best model (i.e., the one with the minimum BIC or AIC) is obtained in correspondence of $k = 7$, for

all the three hypothesised dimensional schemes (CONT, PROC and GOAL). In addition, among the three models with seven latent classes, the best one according to the BIC is again the CONT model, that is, the model whose multidimensional structure is specified according to the item content (BIC = 1,361,226). However, if we look at the AIC (bottom of Table 3a), the best scheme is that referring to the cognitive process (PROC model, AIC = 1,360,238).

The last step of the present study is carried out to compare the models selected at the previous step – that is, IRT models with a single multidimensional latent variable – and the 2T LCMIRT. We recall that the latter considers two uncorrelated multidimensional latent variables **U** and **V**: since their distribution is supposed to be discrete, this, in turn, requires the specification of the number of latent classes for each latent variable (k_U and k_V). Given the results obtained in the previous step, we chose to use $k_U = k_V = 7$. Moreover, since we are dealing with three possible item classifications (according to the item contents, processes and goal-dimensions), we consider all the possible 2T LC-MIRT models, built on the basis of the possible combinations of the above item classifications: *i.* CONT + PROC, *ii.* CONT + GOAL and *iii.* PROC + GOAL. For example, in the CONT + PROC model, the first latent variable **U** is multidimensional with four dimensions according to the item contents, while **V** is multidimensional with seven dimensions on the basis of the item processes. Again, in order to choose the suitable multidimensional classification, the three 2T LC-MIRT models are compared according to their BIC and AIC, reported in Table 3b. As we can see, the best two-tier model is the CONT + PROC one (BIC = 1,338,228; AIC = 1,336,588). Moreover, comparing the BIC (and the AIC) with the best LC-MIRT or N-MIRT obtained in the previous step – that is, models with a multidimensional structure made by only one latent variable – we can see that the two-tier model is preferable according to both Information Criteria, since we can observe lower values of both indexes for the 2T LC-MIRT than the LC-MIRT or N-MIRT ones.

5. Conclusions

The present work aims at comparing the goodness of fit of three multidimensional IRT models to the INVALSI data: the multidimensional IRT model (Reckase, 2009), the Latent Class multidimensional IRT model (Bartolucci, 2007), and the two-tier Latent Class multidimensional IRT model (Bacci and Bartolucci, 2016). All the previous models assume that the responses to the test items depend on more than a single latent trait or ability. However, while in the first two models a unique latent trait is assumed to affect the item response (between-item multidimensionality), the third model is based on

the within-item multidimensionality assumption, which implies that more latent traits can simultaneously affect the item response.

The data used in this study are the INVALSI data on the mathematics test administrated in 2016 to students at lower secondary school level (Grade 8). First results show that any of the three multidimensional models fit the data better than their unidimensional counterpart, confirming that the mathematics test is made of more than a single component or ability. Further, among the three multidimensional models accounted for in this study, the two-tier Latent Class IRT is the model that best fits the INVALSI data. This implies that, for the INVALSI mathematics test, it is realistic to hypothesise that each item contributes to contextually measure two latent abilities – rather than a single one – defined on account of both its content specifications and the mathematical processes involved to resolve it. It follows that the multidimensional structure of the mathematics test can be conveniently specified according to item content and processes, and that the adoption by the INVALSI of such a two-side item classification – where item contents and processes are uncorrelated – is expressive of the real nature of the data.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory* (pp. 267–281). Akademinai Kiado.

Bacci, S., Bartolucci, F. (2016). Two-Tier Latent Class IRT Models in R. *The R Journal*, 8(2), 139-166.

Bartolini Bussi, M.G., Boni, M., Ferri, F., Garuti, R. (1999). Early approach to theoretical thinking: gears in primary school. *Educational Studies in Mathematics*, 39(1), 67-87.

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

Bartolucci, F., Bacci, S., Gnaldi, M. (2014). MultiLCIRT: An R package for multidimensional latent class item response models. *Computational Statistics and Data Analysis*, 71, 971-985.

Bartolucci, F., Bacci, S., Gnaldi, M. (2015). *Statistical analysis of questionnaires: A unified approach based on R and Stata*. Boca Raton, FL: Chapman & Hall/CRC Press.

Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1-29.

Douek, N. (2006). Some remarks about argumentation and proof. In P. Boero (ed.), *Theorems in school: from history, epistemology and cognition to classroom practice*. Rotterdam: Sense Publishers.

Gnaldi, M. (2017). A multidimensional IRT approach for dimensionality assessment of standardised students' tests in mathematics. *Quality & Quantity*, 51(3), 1167-1182.

Gnaldi, M., Del Sarto, S. (2016). Variable weighting via multidimensional IRT models in Composite Indicators construction. *Social Indicators Research*, DOI: 10.1007/s11205-016-1500-5.

INVALSI (2016a). *Rilevazioni nazionali degli apprendimenti 2015_16. Rapporto risultati*.

INVALSI (2016b). *Rilevazioni nazionali degli apprendimenti 2015_16. Rapporto tecnico*.

INVALSI (2017). *Il quadro di riferimento delle prove di matematica del sistema nazionale di valutazione*. URL https://invalsi-areaprove.cineca.it/docs/file/QdR_2017_def.pdf

Moustaki, I. (2000). A latent variable model for ordinal variables. *Applied Psychological Measurement*, 24 (3), 211-223.

Muthén, L.K., Muthén, B.O. (2010). MPLUS user's guide: Statistical Analysis with Latent Variables. Los Angeles, CA: Muthén & Muthén.

R Core Team (2017). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/

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.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464.

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

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

Tables and figures

Table 1: Description of the 43 items composing the 2016 INVALSI mathematics test.

Item	Original label	Content[1]	Process dimension	Goal- [2]	Prop. of correct
1	D1	NUM	3	UND	0.625
2	D2_a	NUM	7	PS	0.804
3	D2_b	DP	4	PS	0.592
4	D3_a	SF	8	UND	0.523
5	D3_b	SF	8	UND	0.537
6	D4_a	RF	7	UND	0.746
7	D4_b	RF	7	UND	0.721
8	D5_a	NUM	4	PS	0.435
9	D5_b	NUM	4	PS	0.349
10	D6	SF	6	REAS	0.237
11	D7_a	DP	7	PS	0.583
12	D7_b	DP	7	PS	0.487
13	D7_c	DP	7	PS	0.283
14	D8	SF	4	PS	0.281
15	D9_a	SF	1	UND	0.595
16	D9_b	RF	4	PS	0.775
17	D9_c	RF	4	PS	0.620
18	D10	DP	7	PS	0.483
19	D11_a	RF	4	PS	0.554
20	D11_b	RF	4	PS	0.485
21	D12	DP	2	PS	0.541
22	D13_a	DP	7	PS	0.841
23	D13_b	NUM	1	UND	0.374
24	D14	SF	1	UND	0.345
25	D15	NUM	6	REAS	0.371
26	D16	DP	7	PS	0.807
27	D17	SF	2	UND	0.436
28	D18	DP	2	PS	0.417
29	D19	SF	8	UND	0.433
30	D20	NUM	4	PS	0.660
31	D21	DP	3	PS	0.831
32	D22	SF	8	UND	0.487

[1] NU: Numbers; SF: Shapes and Figures; RF: Relations and Functions; DP: Data and Previsions
[2] UND: Understanding; PS: Problem Solving; REAS: Reasoning

33	D23_a	RF	6	REAS	0.357
34	D23_b	RF	4	PS	0.640
35	D24	NUM	2	UND	0.411
36	D25	RF	3	UND	0.494
37	D26_a	RF	2	UND	0.498
38	D26_b	RF	6	PS	0.568
39	D26_c	RF	6	PS	0.525
40	D27	NUM	3	UND	0.512
41	D28	NUM	2	UND	0.472
42	D29	NUM	1	UND	0.639
43	D30	DP	1	UND	0.367

Table 2: Results about the comparison of the parameterisations considered for the unidimensional IRT models (UIRT).

LC-UIRT		BIC		AIC	
	k	1-PL	2-PL	1-PL	2-PL
	3	1,386,484	1,372,288	1,386,097	1,371,554
	5	1,378,916	1,362,991	1,378,496	1,362,225
	7	1,378,592	1,362,467	1,378,139	1,361,668
N-UIRT		1,378,758	1,362,681	1,378,395	1,361,972

Table 3: a) Comparison between unidimensional solution (UNI) and multidimensional ones, according to the item contents (CONT), involved processes (PROC) and goal-tied dimensions (GOAL); b) two-tier LC MIRT performance.

a)

BIC	LC-MIRT (*k*)			N-MIRT
	3	5	7	
CONT	1,372,145	1,362,805	1,361,226	1,359,439
PROC	1,372,209	1,362,881	1,361,285	1,359,571
GOAL	1,372,258	1,362,930	1,362,034	1,361,563
UNI	1,372,288	1,362,991	1,362,467	1,362,681

AIC	LC-MIRT (*k*)			N-MIRT
	3	5	7	
CONT	1,371,387	1,361,965	1,360,303	1,358,681
PROC	1,371,426	1,361,966	1,360,238	1,358,690
GOAL	1,371,508	1,362,114	1,361,153	1,360,830
UNI	1,371,554	1,362,225	1,361,668	1,361,972

b)

2T LC-MIRT	BIC	AIC
CONT + GOAL	1,340,143	1,338,685
CONT + PROC	1,338,228	1,336,588
GOAL + PROC	1,339,560	1,337,962