Testing Unidimensionality and Differential Item Functioning of the INVALSI Students' Data

CLADAG 2011
University of Pavia, 7-9 September 2011

M. Gnaldi, F. Bartolucci, S. Bacci
Dapartment of Economics, Finance and Statistics
University of Perugia

OUTLINE

- FRAMING THE ISSUE
- METHODOLOGICAL ISSUES
- THE DATA AND THE APPLICATION
- MAIN RESULTS
- CONCLUSIONS

FRAMING THE ISSUE

Framing the Issue - Basics

Within the educational context, students' assessment tests are validated through Item Response Theory (IRT) models which assume *unidimensionality* and *absence of Differential Item Functioning (DIF)*.

- **Unidimensionality assumption**: responses to a set of items only depend on a single latent trait (the student's ability)
- **Absence of DIF assumption**: items have the same difficulty for all subjects and, therefore, difficulty does not vary among different groups defined, for instance, by gender or geographical area

If the hypotheses of absence of DIF and unidimensionality are not met, summarizing students' performances through a single score may be misleading.

The main objective

We investigate if the no DIF and Unidimensionality assumptions hold for two national tests administered in Italy to middle school students in June 2009:

- **INVALSI Italian Test**
- **INVALSI Mathematics Test**

It is plausible that, given the complexity of the INVALSI study, these assumptions are not met for the INVALSI Test items as they may not discriminate equally well among subjects and may exhibit differential item functioning (DIF).

The methodological framework

- The hypothesis of *unidimensionality* has been extensively tested in connection with the Rasch model. Most statistical tests proposed in the literature are based on the assumptions that:

  (i) item discrimination power is constant

  (ii) the conditional probability to answer a given item correctly does not vary across different groups

- We illustrate an **extension of the 2PL multidimensional latent class (LC) IRT models developed by Bartolucci (2007) to include DIF effects** (students' gender and geographical area included as *covariates)*.

METHODOLOGICAL ISSUES

Basic notation

- n: number of subjects in the sample
- Y_{ij}: random variable corresponding to the response to item j provided by subject i
- r: number of dichotomous items
- s: number of latent traits/dimensions measured by the items
- k: number of latent classes of individuals (the same for each latent trait)
- $\mathcal{J}_d$, $d = 1, \ldots, s$: the subset of $\mathcal{J} = \{1, \ldots, r\}$ containing the indices of the items measuring the latent trait of type d
- r_d: the cardinality of this subset, so that $r = \sum_{d=1}^{r} s_d$.
- $\boldsymbol{\Theta}_i = (\Theta_{i1}, \ldots, \Theta_{is})'$: the vector of latent traits (or dimensions) measured by the test items
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)'$: one of its possible realizations
- δ_{jd}: a dummy variable equal to 1 if item j belongs to $\mathcal{J}_d$ and to 0 otherwise
- γ_j: discrimination index of item j
- β_j: difficulty parameter of item j

The multidimensional 2PL LC IRT model

- It presents two main differences with respect to the classic IRT models:

  (*i*) the latent structure is **multidimensional**

  (*ii*) it is based on latent variables that have a **discrete distribution**

- We consider in particular the version of these models based on the two-parameter (2PL) logistic parameterisation of the conditional response probabilities (Birnbaum, 1968).

$$\text{logit}[p(Y_{ij} = 1 \mid \mathbf{\Theta}_i = \theta)] = \gamma_j \left(\sum_{d=1}^{D} \delta_{jd} \theta_d - \beta_j \right)$$

Further assumptions of the model:

- The random vector $\boldsymbol{\Theta}$ has a **discrete distribution** with support points $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k\}$
- The manifest distribution of the full response vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ir})'$:

$$p_i(\boldsymbol{y}) = p(\boldsymbol{Y}_i = \boldsymbol{y}) = \sum_{c=1}^{k} p_i(\boldsymbol{y} \mid c)\pi_c,$$

where $p(\boldsymbol{\Theta}_i = \boldsymbol{\xi}_c)$ and (Assumption of Local Independence)

$$p_i(\boldsymbol{y} \mid c) = p(\boldsymbol{Y}_i = \boldsymbol{y} \mid \boldsymbol{\Theta}_i = \boldsymbol{\xi}_c) = \prod_{j=1}^{r} p(Y_{ij} = y_j \mid \boldsymbol{\Theta}_i = \boldsymbol{\xi}_c), \quad c = 1, \ldots, k$$

Extension for Differential Item Functioning (1)

- DIF occurs when subjects belonging to different groups with the same latent trait level have a *different probability of providing a certain answer to a given item* (see, for example, Thissen at al., 1993)
- Even in the presence of a 2PL parameterisation, it reasonable to suppose that the main reason of DIF is due to the item difficulty level, which may depend on the individual characteristics of the respondent
- The presence of DIF in the difficulty level of item j may be represented by **shifted values of β_j for one group of subjects with respect to another**

Extension for Differential Item Functioning (2)

Let z_{gi} be a dummy variable which assumes value 1 if subject i belongs to group g (e.g., that of females) and value 0 otherwise and let h be the number of groups so that $g = 1, \ldots, h$.

CASE 1: *Extension for DIF when assuming a* **Unidimensional structure** *and a classification of pupils according to* **one criterium** *(e.g. gender).*

The 2PL parameterisation can be expressed as:

$$\text{logit}[p(Y_{ij} = 1 \mid \Theta_i = \theta)] = \gamma_j \left[\theta - \left(\beta_j + \sum_{g=1}^{h} \phi_{gj} z_{gi} \right) \right],$$

where ϕ_{gj} is the DIF parameter.

If two subjects have the same ability level θ, but belong to two different groups (g_1 and g_2), the difference between the conditional probabilities of a correct response does not depend on the common latent trait θ (**uniform DIF**).

Extension for Differential Item Functioning (3)

CASE 2: *Extension for DIF when assuming a* **Multidimensional structure** *and a classification of pupils according to* **two criteria** *(e.g. gender and geographic areas).*

The 2PL parameterisation can be expressed as:

$$\text{logit}[p(Y_{ij} = 1 \mid \Theta_i = \theta)] = \gamma_j \left[\sum_{d=1}^{s} \delta_{jd}\theta_d - \left(\beta_j + \sum_{g=1}^{h_1} \phi_{gj}^{(1)} z_{hi}^{(1)} + \sum_{g=1}^{h_2} \phi_{gj}^{(2)} z_{hi}^{(2)} \right) \right],$$

where, $z_{gi}^{(1)}$ is equal to 1 if subject i belongs to group g and to 0 otherwise; $\phi_{gj}^{(1)}$ is the DIF parameter; $z_{gi}^{(2)}$ and $\phi_{gj}^{(2)}$ are defined accordingly.

Choice of the number of latent classes

- We rely on the **Bayesian Information Criterion (BIC)** of Schwarz, 1978.
- The selected number of classes is the one corresponding to the minimum value of

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

- In practice, the model is fitted for increasing values of k until BIC does not start to increase. Then, the previous value of k is taken as the optimal one.

Hypothesis Testing - Absence of DIF

- **The hypothesis of absence of DIF** is:

$$H_0 : \phi_{gj} = 0, \quad g = 2, \ldots, h, \; j = 1, \ldots, r,$$

or

$$H_0 : \phi_{2j}^{(1)} = \cdots = \phi_{h_1 j}^{(1)} = \phi_{2j}^{(2)} = \cdots = \phi_{h_2 j}^{(2)} = 0, \quad j = 1, \ldots, r,$$

- For a hypothesis of type $H_0 : \boldsymbol{f}(\boldsymbol{\eta}) = \mathbf{0}$, the following statistic can be used:

$$D = -2[\ell(\hat{\boldsymbol{\eta}}_0) - \ell(\hat{\boldsymbol{\eta}})],$$

which has null asymptotic distribution of χ^2_m type

- We have to fit the model with and without DIF and compare the corresponding log-likelihoods by ()

- **If the obtained value of test statistic is higher than a suitable percentile of the χ^2_m distribution, with $m = r(h-1)$, we reject H_0 and can state that there is evidence of DIF**.

Hypothesis Testing - Unidimensionality (1)

- We compare a unidimensional model with a multidimensional counterpart relying on a **hierarchical clustering algorithm** proposed by Bartolucci (2007), which builds a **sequence of nested models**
- The clustering procedure performs $s - 1$ steps. At each step, the **Wald test statistic** for unidimensionality is computed for every pair of possible aggregations of items:

$$W = \boldsymbol{f}(\hat{\boldsymbol{\eta}})' \boldsymbol{G}(\hat{\boldsymbol{\eta}}) \boldsymbol{f}(\hat{\boldsymbol{\eta}}),$$

  where $\boldsymbol{G}(\boldsymbol{\eta})$ is a suitable matrix computed on the basis of the Jacobian of $\boldsymbol{f}(\boldsymbol{\eta})$ and the information matrix of the model.
- **The aggregation with the minimum value of the statistic is adopted** and the corresponding model fitted before going to the next step.

Hypothesis Testing - Unidimensionality (2)

- The output can be displayed through a **dendrogram** that shows the deviance between the initial (k-dimensional) model and the model selected at each step of the clustering procedure.

- The results of a cluster analysis depend on the adopted **rule to cut the dendrogram**. A possible rule is based on the increase of a suitable information criterion, such as BIC, with respect to the initial or the previous fitted model.

- The dendrogram is cut in correspondence with the *last step showing a negative increase of BIC*.

THE DATA AND THE APPLICATION

The 2009 INVALSI Tests

ITALIAN TEST

- Reading Comprehension section: **30 items** to assess *Lexical Competency* (e.g. the ability to make sense of words in the text) and *Textual Competency* (e.g. make inferences, interpret and integrate ideas and information)
- Grammar section: **10 items** to assess the ability of understanding the morphological and syntactic structure of sentences within a text

MATHEMATICS TEST

27 items to cover four main content domains: *Number, Shapes and Figures, Algebra, Data and Previsions*.

ITEM TYPE: Multiple choice, dichotomously scored.

27592 pupils and **1305 schools** (and classes).

Choice of the Number of Latent Classes

Table 1: Log-likelihood, number of parameters and BIC values for $k = 1, \ldots, 9$ latent classes for the Reading Comprehension and the Grammar sections of the Italian Test and for the Mathematics Test; in boldface is the smallest BIC value for each type of Test.

k	Reading comprehension			Grammar			Mathematics		
	$\ell(\hat{\boldsymbol{\eta}})$	#par	BIC	$\ell(\hat{\boldsymbol{\eta}})$	#par	BIC	$\ell(\hat{\boldsymbol{\eta}})$	#par	BIC
1	-350,474	180	702,743	-100,842	60	202,282	-242,111	162	485,808
2	-329,109	211	660,323	-95,580	71	192,899	-224,506	190	450,873
3	-326,171	242	654,760	-95,645	82	192,110	-221,976	218	446,090
4	-325,516	273	653,750	-95,580	93	192,090	-220,936	246	444,280
5	-324,970	304	**652,970**	-95,517	104	**192,070**	-220,032	274	442,750
6	-324,863	335	653,070	-95,470	115	192,090	-219,619	302	442,190
7	-324,764	366	653,178	-95,464	126	192,184	-219,248	330	441,730
8	-324,684	397	653,327	-95,454	137	192,274	-218,977	358	**441,460**
9	-324,583	428	653,436	-95,429	148	192,334	-218,846	386	441,470

Testing absence of DIF

Table 2: Deviance of the multidimensional 2PL model with uniform DIF with respect to the multidimensional 2PL model with no DIF for the Italian Test - Reading Comprehension section and Grammar section - and the Mathematics Test.

	Deviance	p-value
Reading Compr.	1579.702	$<$0.001
Grammar	1313.427	$<$0.001
Mathematics	2183.573	$<$0.001

Table 3: Estimates of $\phi_{gj}^{(1)}$ *and* $\phi_{gj}^{(2)}$ *for the* **Reading Comprehension Section - INVALSI Italian Test***; significance at levels 0.001 (***), 0.01 (**), 0.05 (*)*

Item	Females	NorthEast	Centre	South	Islands
R1	-0.018	-0.051	-0.262***	-0.173**	0.032
R2	-0.322***	0.132	-0.005	-0.073	0.170
R3	0.021	0.211*	-0.057	0.212*	0.131
R4	-0.447***	0.253*	0.291**	0.428***	0.613***
R5	-0.377***	0.192*	0.094	0.110	0.227**
R6	0.117**	0.132	0.153*	0.305***	0.457***
R7	-0.332***	0.083	0.040	0.196**	0.433***
R8	-0.072*	0.002	0.127*	0.229***	0.159**
R9	-0.170***	0.046	0.008	0.078	0.141**
R10	-0.340***	0.320*	-0.291*	-0.420**	-0.581***
R11	-0.159***	0.038	0.075	0.279***	0.415***
R12	-0.148***	0.069	-0.038	0.277***	0.227***
R13	-0.057*	0.003	0.003	-0.035	0.111*
R14	0.096	0.019	-0.060	0.176*	0.159*
R15	0.001	-0.026	-0.079	-0.079	0.028
R16	-0.352***	0.270**	-0.387***	-0.682***	-0.661***
R17	-0.074**	0.058	-0.065	-0.017	0.067
R18	0.109**	0.036	0.075	0.232***	0.350**
R19	0.260**	0.044	-0.075	-0.480***	-0.566***
R20	0.029	0.049	0.283***	0.207**	0.236**

Table 3 (follows): Estimates of $\phi_{gj}^{(1)}$ *and* $\phi_{gj}^{(2)}$ *for the* **Reading Comprehension Section - INVALSI Italian Test***; significance at levels 0.001 (***), 0.01 (**), 0.05 (*)*

Item	Females	NorthEast	Centre	South	Islands
R21	-0.195***	-0.068	0.018	0.327***	0.290**
R22	-0.193***	0.020	-0.022	0.216***	0.334***
R23	-0.254***	0.050	0.025	0.431***	0.441***
R24	-0.245*	0.223	-0.282*	-0.216	-0.067
R25	-0.068	-0.043	0.001	-0.173*	-0.056
R26	-0.319***	0.053	-0.106	-0.158**	0.160**
R27	-0.239***	-0.116	-0.079	0.150	0.313***
R28	-0.286***	0.094	-0.105	-0.215**	-0.014
R29	-0.179***	-0.071	-0.117	0.008	0.252 **
R30	-0.405***	0.026	-0.119	0.182*	0.309***

Table 4: Estimates of $\phi_{gj}^{(1)}$ and $\phi_{gj}^{(2)}$ for the **Grammar section - INVALSI Italian Test***; significance at levels 0.001 (***), 0.01 (**), 0.05 (*)*

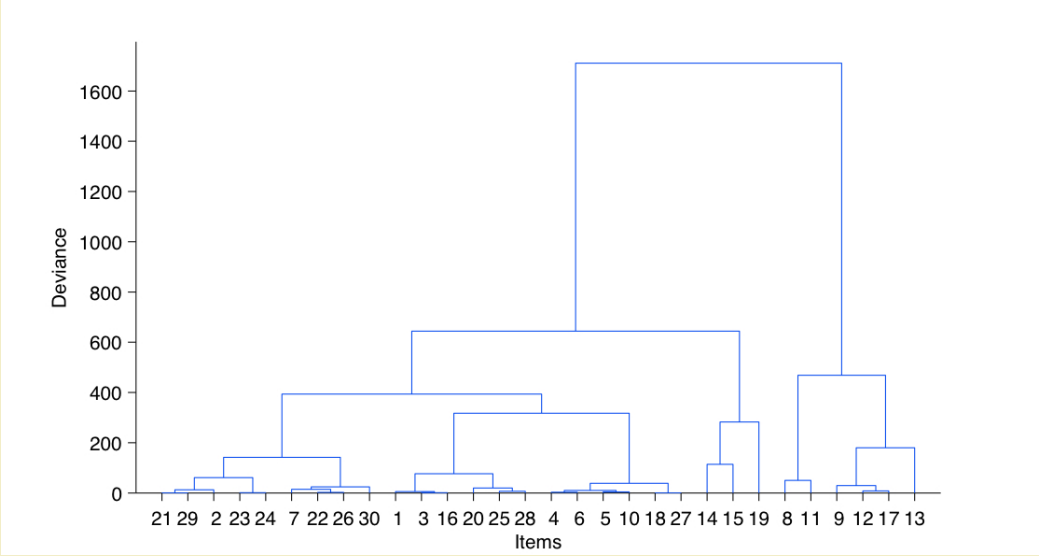
Item	Females	NorthEast	Centre	South	Islands
G1	-0.404***	0.133	0.073	0.129	0.428***
G2	-0.272***	0.156*	0.060	0.051	0.114
G3	-0.137***	0.059	-0.191**	-0.198**	-0.002
G4	-0.052	0.323***	0.004	-0.679***	-0.350***
G5	-0.328***	0.118*	-0.111*	-0.141**	0.126*
G6	-0.261***	0.362***	-0.043	-0.312***	0.072
G7	-0.323***	0.197***	-0.131*	-0.287***	-0.011
G8	-0.309***	0.060	0.144	0.099	0.524***
G9	-0.205*	-0.084	-0.067	0.290*	0.705***
G10	-0.167*	0.269*	-0.280*	-0.504***	-0.324*

Table 5: Estimates of $\phi_{gj}^{(1)}$ and $\phi_{gj}^{(2)}$ for the **INVALSI Mathematics Test**

Item	Females	NorthEast	Centre	South	Islands
M1	0.092	0.013	-0.193**	-0.426***	-0.423***
M2	-0.027	0.058	-0.051	-0.050	0.045
M3	0.023	0.058	-0.044	-0.277***	-0.200***
M4	0.008	0.076*	-0.021	-0.030	0.006
M5	0.024	0.102	0.109	0.057	-0.060
M6	-0.002	0.012	-0.102	0.072	0.076
M7	0.036	-0.097	0.073	-0.090	-0.139***
M8	0.022	-0.058	0.001	0.144	0.148
M9	0.071***	-0.022	-0.056*	-0.078**	-0.027
M10	0.102***	0.023	-0.041	-0.083**	-0.077**
M11	0.029*	0.031	-0.120***	-0.310***	-0.310***
M12	0.087***	0.024	-0.057*	-0.072	0.022
M13	0.055***	0.101***	-0.027	-0.073**	-0.024
M14	0.052**	0.065**	-0.031	-0.166***	-0.193***
M15	0.052**	0.037	-0.012	0.047	-0.002
M16	0.161***	0.030	0.019	-0.008	-0.019
M17	0.164***	-0.001	-0.056	-0.060	0.018
M18	0.049***	-0.001	-0.023	-0.022	-0.025
M19	-0.008	-0.006	0.032	0.103***	0.183***
M20	-0.024	0.029	-0.007	-0.055*	-0.006
M21	0.112***	0.104**	0.023	-0.034	-0.022
M22	0.033	0.078*	-0.036	0.001	-0.060*
M23	0.013	0.049*	-0.049*	-0.062**	-0.057**
M24	-0.012	-0.008	-0.097***	-0.143***	-0.125***
M25	-0.035**	-0.013	0.033	0.032	0.153***
M26	0.087***	0.089**	-0.058	-0.163***	-0.083*
M27	0.010	0.093**	-0.074	-0.269***	-0.312***

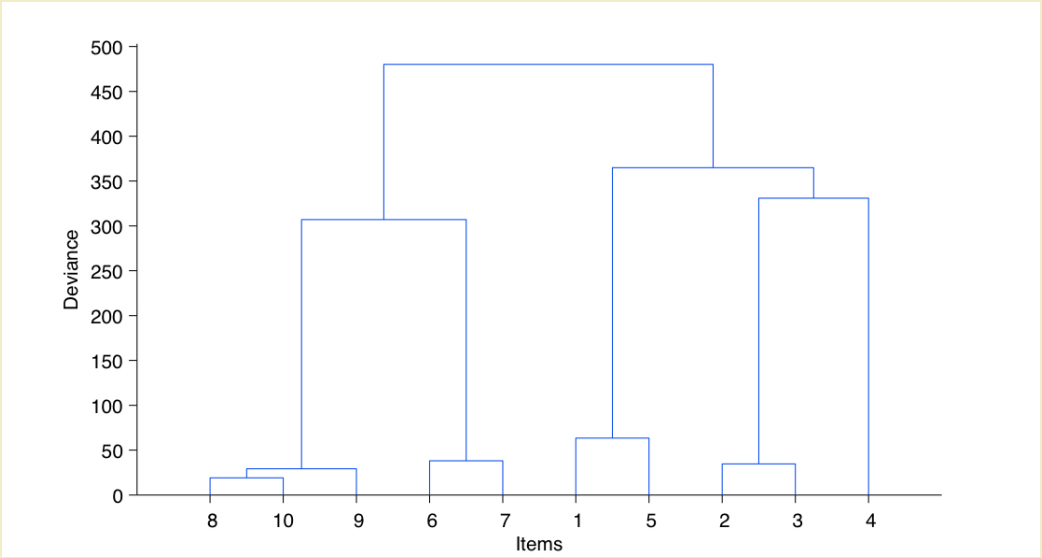
Testing dimensionality

Figure 1: Dendrogram for the Italian Test - Reading Comprehension Section.



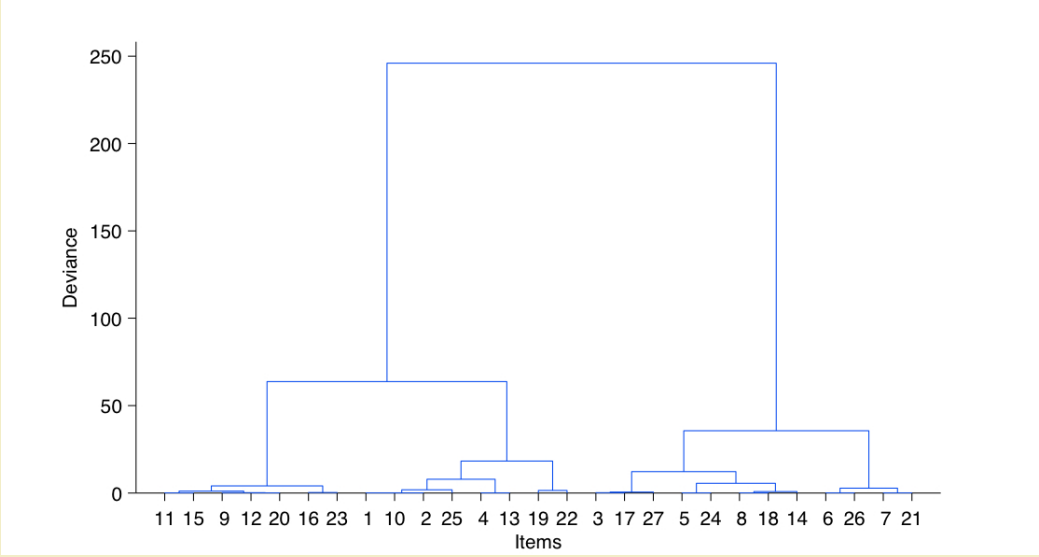
Testing dimensionality

Figure 2: Dendrogram for the Italian Test - Grammar Section.



Testing dimensionality

Figure 3: Dendrogram for the Mathematics Test.



h	s	**Increase BIC wrt initial model**		
		Reading compr.	Grammar	Maths
1	29	-29.8	-10.755	-9.791
2	28	-59.6	-30.642	-19.581
3	27	-89.0	-55.184	-29.371
4	26	-118.4	-81.534	-39.158
5	25	-147.4	-86.135	-48.942
6	24	-176.4	**127.498**	-58.723
7	23	-205.4	121.555	-68.499
8	22	-234.0	125.539	-78.275
9	21	-262.6	210.922	-88.043
10	20	-291.0	–	-97.805
11	19	-319.1	–	-107.550
12	18	-346.7	–	-117.241
13	17	-374.0	–	-126.792
14	16	-399.4	–	-136.315
15	15	-424.5	–	-145.827
16	14	-449.3	–	-155.235
17	13	-470.0	–	-164.625
18	12	-488.7	–	-173.480
19	11	-507.3	–	-181.965
20	10	-522.0	–	-190.288
21	9	-514.0	–	-197.723
22	8	-516.5	–	-203.161
23	7	-508.3	–	-206.950
24	6	-435.6	–	-199.422
25	5	-430.4	–	-181.022
26	4	-384.1	–	-8.600
27	3	-339.4	–	–
28	2	-193.6	–	–
29	1	**843.4**	–	–

Table 7: Support points estimates for the Italian Test - Reading Comprehension section and Grammar section - and the Mathematics Test

	c				
	1	2	3	4	5
Reading Comprehension					
Dimension 1	-1.193	0.221	-0.329	1.012	2.776
Dimension 2	-1.404	-0.859	-0.049	0.646	1.378
Grammar					
Dimension 1	-0.334	2.244	2.536	2.948	4.363
Dimension 2	-0.853	-0.786	0.812	0.935	2.807
Dimension 3	-0.827	-0.554	-2.068	0.598	2.384
Dimension 4	0.782	1.224	2.012	2.507	3.735
Dimension 5	-0.616	-1.069	-0.623	-0.056	1.364
Mathematics					
Dimension 1	0.995	1.509	2.060	–	–

Conclusions

1. Assumption of Absence of DIF: strongly rejected

- *Gender*: The Italian Test favours girls, while the Maths Test tend to favour boys
- *Geographic area*: stronger incidence of items affected by DIF for the southern regions

2. Assumption of Unidimensionality: unreasonable for the Italian Test, acceptable for the Maths Test

- *Reading Comprehension Section*: 2 groups of items corresponding to the ability to (*i*) make sense of worlds and sentences in the text and recognize meaning connections among them and (*ii*) interpret and make inferences from a written text
- *Grammar Section*: 5 groups of items corresponding to the ability to (*i*) recognize verb forms, (*ii*) recognize the meaning of connectives within a sentence, (*iii*) recognize grammatical categories, (*iv*) make a difference between clauses within a sentence, (*v*) recognize the meaning of punctuation marks

3. Support point estimates: students' belonging to the higher latent classes is linked with increasing ability levels.

References

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees ability. In Lord, I. F. M. and M.R. Novick (eds.), Reading, M. A.-W., editors, Statistical theories of mental test scores, pages 395-479.

Formann, A. (1995). Linear logistic latent class analysis and the rasch model. In Molenaar, G. F. . I., editor, Rasch models: Foundations, recent developments, and applications, pages 239-255. Springer-Verlag, New York.

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

INVALSI (2009). Esame di stato di primo ciclo. a.s. 2008/2009. In INVALSI Technical Report.

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

Thissen, D., Steinberg, L., and Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In Holland PW, Wainer H. Hillsdale, N. J. L. E. A., editor, In Differential Item Functioning, pages 67-11.

Verhelst, N. D. (2001). Testing the unidimensionality assumption of the rasch model. Methods of Psychological Research Online, 6:231-271.