SURVEYS ON STUDENTS: INVALSI NATIONAL AND INTERNATIONAL TESTS

VI Seminar "INVALSI data: a tool for teaching and scientific research"

edited by
Patrizia Falzetti

FrancoAngeli

INVALSI PER LA RICERCA
STUDI E RICERCHE

INVALSI

INVALSI PER LA RICERCA

La collana Open Access INVALSI PER LA RICERCA si pone come obiettivo la diffusione degli esiti delle attività di ricerca promosse dall'Istituto, favorendo lo scambio di esperienze e conoscenze con il mondo accademico e scolastico.
La collana è articolata in tre sezioni: "Studi e ricerche", i cui contributi sono sottoposti a revisione in doppio cieco, "Percorsi e strumenti", di taglio più divulgativo o di approfondimento, sottoposta a singolo referaggio, e "Rapporti di ricerca e sperimentazioni", le cui pubblicazioni riguardano le attività di ricerca e sperimentazione dell'Istituto e non sono sottoposte a revisione.

SURVEYS ON STUDENTS: INVALSI NATIONAL AND INTERNATIONAL TESTS

VI Seminar "INVALSI data: a tool for teaching and scientific research"

edited by
Patrizia Falzetti

INVALSI

FrancoAngeli

The opinions expressed are solely of the author(s). In no case should they be considered or construed as representing an official position of INVALSI.

Index

3. A quantitative model for gender gap in G8 standardized Mathematics tests in Italian schools

by Riccardo Orlando, Ottavio G. Rizzo

The gender gap in Mathematics, i.e. the different performances of male and female students, is a well-known and well-documented phenomenon. Testing from OCSE-PISA, in particular, highlights how the gap in Italy is much larger than the international average. The didactic component of this gap has been investigated in the literature through one of two broad strategies: either large-scale, statistical analysis of test results, or item-level analysis of very few selected items with the theory of the didactic contract.

This work is an explorative analysis which aims to determine if it is possible to define specific item categories in which the gender gap is particularly notable, and to interpret these categories with the theory of the didactic contract.

Il divario di genere in Matematica, ovvero la differenza di performance tra studenti maschi e femmine, è un fenomeno ben noto e ben documentato. In particolare, le prove di OCSE-PISA evidenziano il gap italiano, molto maggiore della media internazionale. La componente didattica di questo divario è stata investigata in letteratura mediante una di due principali strategie: l'analisi statistica di un grande numero di prove, oppure un'analisi di alcuni item selezionati basata sulla teoria del contratto didattico.

Questo lavoro è un'analisi esplorativa con l'obiettivo di determinare se sia possibile definire specifiche categorie di item nelle quali il divario di genere è particolarmente notevole, e interpretare queste categorie con la teoria del contratto didattico.

1. Introduction

The gender gap in academic performance is a well-studied topic both at a national and international level (Leder and Forgasz, 2008), in particular thanks to the results of large scale international testings showing that on average, women fare better in language tests while men fare better in Mathematics tests (OECD, 2016).

We know that "[n]o significant differences between boys' and girls' Mathematics achievement [is] found before boys and girls [enter] elementary school or during early elementary years" (Fennema, 1974, mainly referring to the US context) while the gap "is large and significant in the middle school years and beyond" (Fryer and Levitt, 2010). Contini *et al.* (2017) show that in Italian context the gap appears in primary school and widens between grade 5 and 10. We also know that the size of the gap varies considerably between different educational systems (OECD, 2015, 2016), and this suggests that biological or physiological differences (Gallagher *et al.*, 2000) could not be the reason, or at the very least not the only reason.

Overall emancipation of women, as measured by the World Economic Forum Gender Gap Index (Guiso *et al.*, 2008) explains partially the gender gap in Mathematics as a result of social and cultural factors. Indeed, the World Economic Forum Gender Gap Index ranks Italy as 72nd in the world (Hausmann, Tyson and Zahidi, 2009), while Italy presents one of the highest gender gaps in Mathematics in the OCSE-PISA standardized tests (Contini *et al.*, 2017).

On the other hand, once Math anxiety and Math self-beliefs are taken in account, the Mathematics gender gap disappears (OECD, 2015); and we know that men and women utilize different strategies in problem solving (Giberti, 2019; Gallagher *et al.*, 2000; Fennema and Carpenter, 1998); hence, results could vary according to which strategies are activated by a given problem.

In this work, we aim to investigate the relationship between the characteristics of items in standardized testing and gender gap, by constructing a model that highlights item characteristics that produce different results in male and female students.

2. Theoretical Framework

This model assumes the existence of certain categories of items that independently cause discrimination. Each category is associated with a discrimination score.

Each item belongs to zero or more categories, and we expect the item discrimination to be the sum of the discrimination scores of the categories to which it belongs.

That is, given n categories, let **c** be the vector of their discrimination scores; given an item, let **m** be the vector that marks the categories to which it belongs: $\mathbf{m}_i$ is 1 if the item belongs to category *i*, 0 otherwise.

Therefore, an item's total discrimination is given by the scalar product $\mathbf{c}\cdot\mathbf{m}$.

Considering now *N* items, and their discrimination vector **d**, their classification is given by a matrix *M* and we expect $\mathbf{d} = M\mathbf{c}$.

This model formulation treats the item discrimination as unknown, and the category discrimination as known. Of course, in practice the opposite is the case and we obtain c with a least-squares method.

3. Methodology

To obtain each item's discrimination, we first compute the uniform Differential Item Functioning score, using as reference and focal groups male and female students respectively (Meyer, 2014).

This yields a value E on a multiplicative scale, from 0 to ∞, with a score of 1 indicating no discrimination. Therefore, we transform this score to D = -100 log(E), in order to obtain values on an additive scale such that positive values indicate discrimination in favor of female students, and a value of 0 indicates no discrimination.

The figure below shows item discrimination values computed for four INVALSI tests, and their approximate distribution.

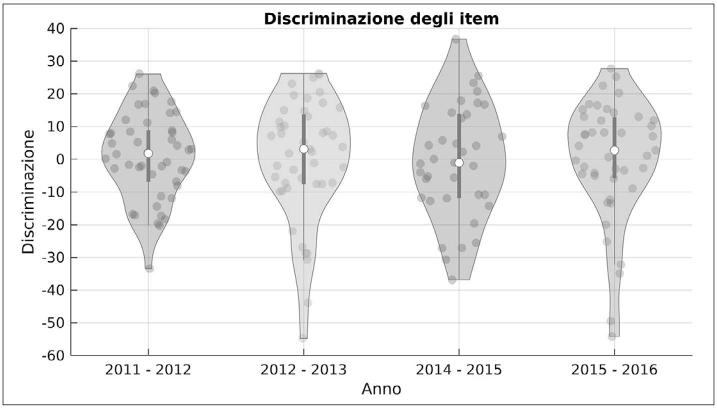

Fig. 1 – Item discrimination distribution for four INVALSI tests

The categories are constructed "a posteriori", according to our analysis of *half* of the available data, namely the tests from school years 2011/12, 2012/13, 2014/15 and 2015/16. We refer to this data as the construction set, and the remaining data (from years 2009/10, 2010/11, 2013/14 and 2016/17) as the validation set.

We compute the E score described above for each item in the construction set, and select the items with highest score (in absolute value).

We then examine these items for common features, with respect to the tested skills or the presentation, and use these commonalities to define a set of possibly discriminating categories.

In particular, these are the main features that we use to construct the categories:

Topic: the Mathematical topic of the question, with reference to the pre-existing INVALSI classification known as "ambiti di contenuto". These group grade 8 items in four "content domains", which are:

– Numbers ("Numeri");
– Relationships and functions ("Relazioni e funzioni");
– Space and Geometric Figures ("Spazio e figure");
– Data and Forecasts ("Dati e previsioni").

We also consider the specific Mathematical skills required to answer the question correctly.

Item type: we consider the type of the item, such as open or multiple choice, as well as the type of the answer: numeric, text, ...

Information density and accessibility: we consider the language and reading comprehension skills required to understand the question, as well as other comprehension skills (such as estimation, or reading a plot).

After drafting the categories, we classify all items of the validation set. This process highlights the definitions that need clarification, as well as those that match too many items or too few.

This allows us to clarify the definitions, and discard the categories that don't match enough items.

Finally, we classify every item of both the construction set and the validation set, denoting all the categories to which each item belongs.

This process results in the following categories:

– **Algebra**: The item asks to deduce or manipulate an algebraic expression;
– **Arithmetics – distractor**: An arithmetic item, such that the simplest solution strategy does not yield the correct result;
– **Asymmetric distractors**: A multiple choice item with a numeric answer, with a wide range of possible options and such that the correct answer is an extreme of the range;

– **Estimate**: The item asks to estimate a measure or an amount given an image or a plot, or to estimate the result of an arithmetic operation;
– **Explain your reasoning**: The item asks to justify the answer – simply giving the correct answer is not enough; the student must provide a correct reasoning. This category includes arithmetic items, where a student must give or choose from a list the expression used to obtain the result;
– **Extrapolation**: The item asks to extrapolate a pattern from given information;
– **Geometry – compute**: An analytic geometry item, or an Euclidean geometry item that asks to compute a value without a proof;
– **Geometry – draw**: The item asks to draw or complete a drawing of a geometric figure;
– **Implicit hypothesis**: The item assumes a hypothesis implicitly, without mentioning it in the text;
– **Multiple choice item**: The item has at most four possible answers, or is clearly divided in sub-items that meet this definition;
– **Nonlinear relationship**: The item includes two quantities that are connected in a non-linear relationship;
– **Numeric answer**: The item answer is a single number, eventually including a unit of measure, or the item is clearly divided in sub-items that meet this definiton;
– **Plot**: The item asks to draw, read or complete a plot or chart;
– **Probability – intuition**: The item is a probaility question that does not require computation;
– **Redundant information**: The item text contains much redundant or unnecessary information;
– **Standard arithmetics**: An arithmetic item, that can be solved correctly using only standard procedures;
– **Base**: All items belong to this category.

The *Base* category exists to allow for non-didactic sources of discrimination: we expect those sources to contribute equally to all items' discrimination scores.

In order to assess the model's performance, for each category we consider the model score on the validation set, and we consider as robust the categories that meet the following criteria:

– They contain at least 3 items of the validation set;
– They contain at least 3 items of the construciton set;
– They have a discrimination score of at least 5 (in absolute value).

4. Results

The following table shows the results for each category.

- **Nc** is the number of items in the construction set that belong to that category;
- **Nv** is the number of items in the validation set that belong to that category;
- **Average** is the average discrimination of those items;
- **Model** is the discrimination score computed by the model.

We obtain five robust categories, in bold in the table.

Tab. 1 – Results by cathegory (robust cathegories in bold)

Cathegory	*Nc*	*Nv*	*Average*	*Model*
Algebra	**17**	**18**	**7.7**	**15.1**
Arithmetics – distractor	8	2	17.4	25.8
Standard arithmetics	43	29	-0.2	1.5
Asymmetric distractors	**4**	**4**	**-4.1**	**-7.9**
Extrapolation	**14**	**31**	**-6.8**	**-8.1**
Geometry – compute	25	18	0.9	1.2
Geometry – draw	**5**	**9**	**17.5**	**21.8**
Explain your reasoning	**22**	**40**	**3.9**	**7.8**
Plot	37	36	-2.2	1.1
Redundant information	19	16	-0.1	1.6
Implicit hypothesis	4	0	-	0
Multiple choice item	95	107	0.1	2.9
Probability – intuition	5	8	-2.4	-4.3
Nonlinear relationship	2	5	-18.5	-21.9
Numeric answer	71	56	-3.2	0.6
Estimate	15	13	-2.3	-0.8
Base	171	171	0.2	-4.9

We consider some variations on this model, which did not yield interesting results.

- Including item difficulty as a parameter did not increase the model effectiveness;
- Computing the category discrimination score on the construction set, then computing the expected discrimination on the validation set, did not change significantly the model effectiveness.

5. Conclusions

We find that the categories *Algebra*, *Geometry – draw* and *Explain your reasoning* have positive discrimination, and therefore favor female students.

This result is not surprising, given their greater language skills and sensitivity to the didactic contract. In particular, this matches with the effect known as the *need for formal justification* (Bolondi *et al.*, 2018).

The two robust categories with negative discrimination are *Asymmetric distractors* and *Extrapolate*. These categories favor male students.

In the second category especially, the need for formal justification imposes the necessity of using only the values explicitly mentioned in the text.

In general, we observe that robust categories match up with known effects from the theory of the didactic contract. Therefore, further work may wish to construct new categories based on this framework, rather than on direct item observation.

This model can be applied with no modifications to grade 5 tests, while its application to grade 2 tests will require the construction of new categories.

Grade 10 and 13 tests, however, will require careful handling of student self-selection as Italy has a common curriculum only up to grade 8: high schools (grade 9 to 13) offer different tracks, with a different amount of time spent on Mathematics and many tracks presenting a significant gender imbalance.

Finally, with the advent of computer-based testing, the Mathematical tools will have to adapt to a sparser coverage of item answers.

References

Bolondi G., Ferretti F., Giberti C. (2018), "Didactic Contract as a Key to Interpreting Gender Differences in Maths", *EPCS Journal,* 18, pp. 415-435.

Contini D., Di Tommaso M.L., Mendolia S. (2017), "The gender gap in Mathematics achievement: Evidence from Italian data", *Economics of Education Review*, 58, pp. 32-42.

Fennema E. (1974), "Mathematics Learning and the Sexes: A Review", *Journal for Research in Mathematics Education*, 5 (3), pp. 126-139.

Fennema E., Carpenter T.P. (1998), "New Perspectives on Gender Differences in Mathematics: An Introduction", *Educational Researcher*, 27 (5), pp. 4-5.

Fryer R.G., Levitt S.D. (2010), "An Empirical Analysis of the Gender Gap in Mathematics", *American Economic Journal: Applied Economics*, 2 (2), pp. 210-240.

Gallagher A.M., De Lisi R., Holst P.C., McGillicuddy-De Lisi A.V., Morely M., Cahalan C. (2000), "Gender Differences in Advanced Mathematical Problem Solving", *Journal of Experimental Child Psychology*, 75 (3), pp. 165-190.

Giberti C. (2019), “Gender differences in Mathematics: from the international literature to the Italian situation”, *Didattica della matematica. Dalla ricerca alle pratiche d’aula*, 5, pp. 44-69.
Guiso L., Monte F., Sapienza P., Zingales L. (2008), “Culture, gender, and math”, *Science*, 320, pp. 1164-1165.
Hausmann R., Tyson L.D., Zahidi S. (2009), *The Global Gender Gap Report: 2009*, World Economic Forum, Geneva.
Leder G., Forgasz H. (2008), “Mathematics education: new perspectives on gender”, *ZDM*, 40 (4), pp. 513-518.
Meyer J.P. (2014), *Applied Measurement with JMetrik*, Routledge, New York.
OECD (2015), *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence*, OECD Publishing, Paris.
OECD (2016), *PISA 2015 Results*, Vol. I: *Excellence and Equity in Education*, OECD Publishing, Paris.

INVALSI, as a part of the National System of Evaluation of the Education and Training System (SNV), conducts periodic and systematic tests on students' knowledge and skills. Albeit with some modifications over time, these standardised tests have been objectively measuring for about 20 years students' achieving and learning in some main skills in Italian, Mathematics and English domains.

In addition to conducting the National Survey, INVALSI coordinates and ensures the participation of Italy in certain main international surveys in education promoted by OECD (Organisation for Economic Cooperation and Development) and IEA (International Association for the Evaluation of Educational Achievement) which, both of them, carry out specific tests on some students' literacies and skills.

At the end of each survey, INVALSI makes useful databases available for studying and analysing the Italian education system – with an international comparison as well – and, on the occasion of the VI Seminar "INVALSI data: a tool for teaching and scientific research" (Rome, from 25th to 28th November 2021), the potential of their use became evident. This volume collects some papers presented there.

The book is therefore full of insights on the possible uses of national and international surveys. We hope that from it reading, researchers, teachers and all stakeholders could find further stimuli to better investigate the Italian education system thanks to INVALSI data and beyond.

Patrizia Falzetti, Technologist Director, is the Head of the INVALSI Area of the Evaluation Research, of the SISTAN Statistical Office and of the INVALSI Statistical Service which manages data acquisition, analysis and return about both national and international surveys on learning (OECD and IEA). She coordinates and manages the process about returning data and statistical analysis to every school and to the Ministry of Education and Merit.

FrancoAngeli
La passione per le conoscenze