Unione Europea

FONDI
STRUTTURALI
EUROPEI

PON
2007-2013

MIUR

Ministero dell'Istruzione, dell'Università e della Ricerca
Dipartimento per la Programmazione
D.G. per gli Affari Internazionali - Ufficio IV
Programmazione e gestione dei fondi strutturali europei
e nazionali per lo sviluppo e la coesione sociale

COMPETENZE PER LO SVILUPPO (FSE) - AMBIENTI PER L' APPRENDIMENTO (FESR)

IMPROVING EDUCATION THROUGHT ACCOUNTABILITY AND EVALUATION LESSONS FROM AROUND THE WORLD

October 3-5, 2012

Palazzo Caetani
via Michelangelo Caetani, 32
Rome, Italy

INVALSI

Association for Public Policy Analysis & Management

University of Maryland
SCHOOL OF PUBLIC POLICY
Celebrating 30 Years

CRELL

AIR

THE WORLD BANK

ATLANTIC COUNCIL

www.invalsi.it/invalsi/ri/improving_education/

Teacher Assessment and Students' Performance in OCSE-PISA 2009

Brunella Fiore
Post-doc researcher
Department of Sociology, University of Milano-Bicocca

Isabella Romeo
Post-doc researcher
Department of Statistics, University of Milano-Bicocca

Abstract

Many educational works focus on the analysis of teachers' judgment in order to analyze evaluation variability on students with similar ability levels. The aim of this work is to analyze whether teachers of upper secondary schools of the Lombardy region overestimate or underestimate theirs students, exploiting the OCSE-PISA 2009 data. For this purpose a comparison between a teacher's mark and a standardized test has been made in order to identify both overestimated and underestimated groups of students. In particular, this paper focuses on identifying which characteristics of students and schools have an impact on both overestimation and underestimation. Then a multilevel multinomial logistic model has been performed. Also a measure of student's ability taken from a Rasch analysis is considered. The results showed that the factors connected with cultural and socio-economic status seemed to have a major impact on these phenomena.

Key words: Teachers' Judgment, Multilevel Multinomial Logistic Model, Rasch Analysis

Theoretical Framework and Aims of the Study

In the last decade there has been growing interest in educational accountability. Measures of assessment for schools, teachers and students have become the most recent watchword in education. The undertaking idea is that external accountability can help schools to make a greater effort to improve student's achievement. At the national and international level, the use of standardized tests has become widespread, given that they allow to objectively measure student performance.

In Anglo-Saxon countries standardized tests or external examiners are considered very important in determining students' ability level in the school system. On the contrary, in most countries as in Italy, standardized tests have still less importance over teachers' evaluations[1].

Under certain circumstances the teacher's judgment should not represent an objective evaluation. Firstly, it could be that the teacher calibrates his judgment on the school or class level without considering a national standardized scale[2]. Secondly, the teachers evaluation could be influenced by a variety of sources beyond the effective student's result such as the student's behavior. Thus, the teacher's judgment represents something more than a simple evaluation of the student's ability level despite of standardized tests[3].

Great interest has been shown on the relation between these different measures of student ability since the end of the 1970's. Early studies on the accuracy of the teacher assessment focus on the correlation between the teacher's judgment and the standardized tests. In particular, Hoge and Coladarci in a review of 16 previous studies found a mean correlation of 0.67 [4]. Next studies found instead, a wide range of correlation values between 0.28 and 0.92[5]. The high variability across research studies underlines the need for a definition of a more robust method.

The aim of this study is to examine in upper secondary schools the variability of teachers' judgment on reading marks compared to the OCSE-PISA 2009 standardized test. Given the well known high performance variability across Italy [6,7] only the Lombardy region has been taken into account in order to analyze a more homogeneous context. In particular, our goal consists in understanding which student and school characteristics have an impact on both overestimation and underestimation of students.

Methods and data sources

As explained above the aim of this work is to analyze if upper secondary school teachers overestimate or underestimate their students. Two student evaluations from the same period are necessary to answer this research question: one from the teacher and another one from a standardized evaluation. This type of information is available in the PISA 2009 dataset.

PISA is a comprehensive and rigorous international programme promoted by OCSE to assess student performance and to collect data on students, families and institutional factors that could help explain differences in performances. This survey collects information about reading, mathematics and science results of 15-year-old students. In particular, information coming from this survey focus on how well students are prepared to meet the challenges of life. The Pisa 2009 survey focused on reading literacy. Thus, this paper is centered on this topic.

In the Pisa 2009 dataset, besides the PISA reading score, also the teacher's mark reported in the second year of upper secondary school education is available. Given the time proximity of these marks (they refer respectively to April and January), they can be considered contemporaneous. However, they are not directly comparable since they have two different numeric scales. On one hand the PISA score is expressed by a numeric scale centered in 500 with a standard deviation of 100. On the other hand the mark received in the second year of upper secondary education is expressed on a numeric scale from 1 to 10. In order to allow them to be comparable, a normalization of them is necessary. In this way the new variables considered have a mean zero and a standard deviation equal to one. The variable obtained as the difference between the teacher's mark and the PISA score is taken into consideration.

In all the analysis and computational process it is necessary to take into account the particular structure of the PISA dataset that means to consider both the five plausible values (PVs) for parameter estimation and the replicates for standard error estimation[8]. In this direction, five differences between the teacher's mark and each PVs are calculated in order to identify underestimated students, overestimated students, and students with teacher's judgment coherence with the PISA data. A positive difference detects an overestimation and a negative one detects an underestimation. In particular, students who present all differences as positive have been assigned to the overestimated group, students who present all differences as negative have been assigned to the underestimated group and finally students with both negative and positive differences are collected in the ``coherence group''.

A multilevel approach has been employed given the hierarchical nature of the data. This structure reflects the existence of two different levels of variables: the one related to school characteristics (variables at level 2) and the other related to student characteristics (variables at level 1). In particular, a multilevel multinomial logistic model has been chosen[9,10] since the aim of this work is to understand which student and school characteristics have an impact on teacher underestimation and overestimation. At the student level (level 1) gender, immigration status, cultural and socio-economic status[i], student repeating a year, together with a measure of student ability have been included. The measure of ability is provided as the summary of the final evaluation of the lower secondary education and the

[i] This variable was created on the basis of the occupational and educational level of student's parents, family wealth, home educational and cultural resources by OCSE[11]

marks reported in the first year of upper secondary school education through the implementation of a partial credit model. This model has been taken into account, given the different ordinal scale of the two marks considered. On one hand, the final evaluation of the lower secondary education is expressed by ``Excellent'', ``Good'', ``Discrete'' or ``Sufficient''. On the other hand, the mark received in the first year of upper secondary education is expressed on a numeric scale from 1 to 10. This model converts the raw ordinal data into interval data, placing the two marks on the same common logit scale[12].

All variables considered at the student level have been aggregated and included at the school level (level 2). In order to make parameters more interpretable, all the variables have been centered on the mean value. Furthermore, also type of secondary school, school size and teacher expectation toward students[ii] have been considered.

Results and Discussion

A first measure about the strength of the level of agreement between teacher assessment and student performance is found in the correlation coefficient. In this context it is equal to 0.38 when all students are considered. This index changes considerably if more homogeneous groups are taken into consideration. When the overestimated group of students is not considered, then the correlation between these measures becomes 0.62. Similarly, when the underestimated group of students is not considered, then the correlation becomes 0.65 when overestimated students and underestimated students are not considered then the correlation became respectively equal to 0.62 and 0.65. These results suggest to take into consideration overestimation and underestimation as distinct issues.

In order to better understand differences among the overestimated, underestimated and ``coherence group'' of students a further graphical descriptive analysis has been

[ii] This variable is created by OCSE [13]

performed exploiting boxplots[iii]. In Figure 1 boxplots allow to compare how PISA scores vary across teacher's marks. High variability of PISA results can be explained on one hand by the presence of both underestimated and overestimated students and on the other hand by different aims of the PISA survey and the educational system. It is important to consider that the PISA survey focuses on how well students are prepared to meet the real challenges of life rather than to examine how well they perform a particular curricula specified by the school system. For example, a teacher's overestimation can also be interpreted as good student knowledge of the curricula education despite a bad attitude in solving the problems of life. Nevertheless PISA is considered a good proxy variable of students' ability.

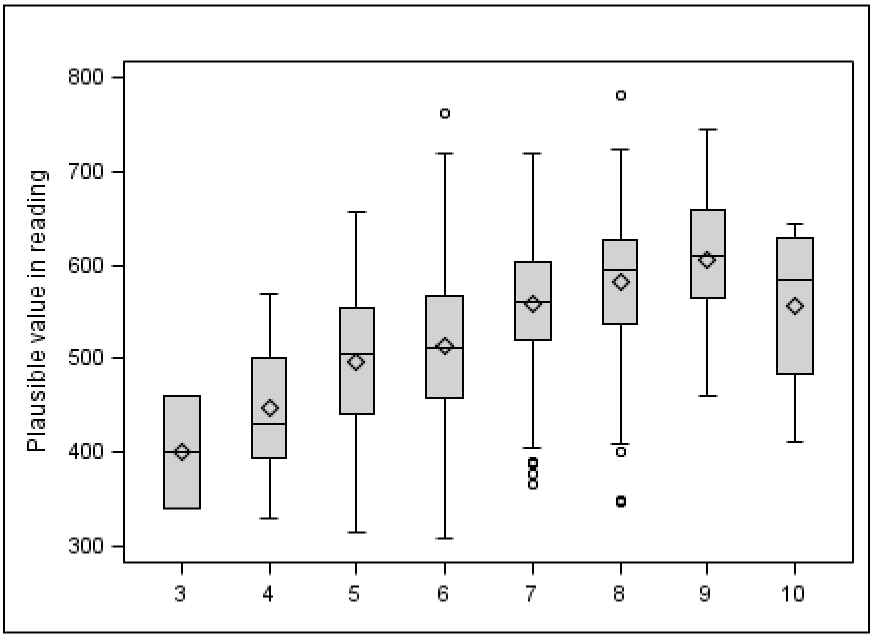

Figure 1 Relation between the PISA score and report mark represented by boxplots

Of great interest is the focus regarding the pass level considered as basic knowledge that is fixed by OCSE to 500 and to 6 in teachers' evaluations in Italy. In general, the

[iii] Boxplots represent a convenient way of graphically displaying differences between groups. They allow to obtain the most important summary statistics of data distribution of each group: the smallest observation, lower quartile, median, upper quartile, and largest observation. The spacings between the different parts of the box help indicate the degree of dispersion and skewness in the data, and which observations might be considered outliers.

variability of PISA scores for students receiving a grade equal to 6 is higher compared to other grades. From a more detailed analysis stratified for groups it has emerged that students belonging to the coherent group with a teacher evaluation equal to 6, on average obtain 487 points on the PISA scale. All overestimated students obtained scores lower then 500 with a mean of 415, and all underestimated students obtained scores greater than 500 with a mean of 578.

Table 1 Significance effects on Overestimated and Underestimated groups

	Variables	**Overestimation**	**Underestimation**
Student	Intercept	0.842	-0.595
	Ability	0.084**	-0.059
	Escs	-0.107	-0.217**
	Student repeating the year (ref. Student not repeating the year)	0.486*	-0.402
	Female (ref. Male)	-0.424*	-0.297
	Immigrant (ref. Italian)	0.004	-0.049
School	Mean Ability	-0.147*	-0.081
	Mean ESCS	-0.734*	1.294***
	Percentage of students not repeating the year	-1.802**	-1.401
	Percentage of girls	0	0.010**
	Percent of immigrants	3.193***	-1.8
	Vocational studies	0.094	1.165*
	Technical institute	0.213	0.263
	School size	-0.310***	0.305***
	Teacher Behavior	0.342**	0.033
	Variance	0.7	0.39

(*) Significance level α=0.1; (**) α =0.05$; (***) α =0.01.

As explained in the previous section, one is interested in understanding which variables have an impact on overestimation and underestimation. For these purposes a multilevel multinomial model has been performed. Table 1 reports the estimates of the model parameters where both groups of overestimated students and underestimated students are compared with the reference group composed of students coherently judged.

For what concerns the overestimated students, it emerges that teachers have a general tendency to overestimate students who show higher ability levels, coherently with the existing literature[1]. Furthermore, males and students repeating a year are more likely to be overestimated. At the school level, factors that have an impact on the probability of being overestimated rather than being coherently judged are the high rate of immigrant students, the low rate of students repeating a year, the low value of school mean ESCS, the low number of students per school and the good teacher's expectation toward students. For what concerns underestimated students, only the ESCS at the student level resulted significant: a lower chance to be underestimated is associated to higher individual ESCS level. At the school level both higher values of school mean ESCS and school size show a great impact on the student chance of being underestimated. From both underestimation and overestimation results, it is possible to make some observations. Firstly, we can get some information from the well known relation[14] between the school mean ESCS and the type of secondary school. On one hand, we can deduce that students attending technical and especially vocational schools (lower mean ESCS) have both less chances of being underestimated and more chance of being overestimated than students attending ``liceo'' [iv] (higher mean ESCS). On the other hand students attending ``liceo'' are the ones most likely to be underestimated and the less likely of being overestimated. Secondly, the high significance of school size both on underestimation and overestimation needs a closer examination. One possible hypothesis is that usually teachers employed outside the metropolitan area, where schools have less students, are more inclined to overestimated students. In conclusion, the analysis indicates that overestimation is a more complex phenomena than underestimation as long as a higher number of significant variables are implied. In particular, it emerges that teachers have a tendency to overestimate students with high abilities and more disadvantaged conditions.

[iv] Scientific, classical, socio-pedagogic high schools.

Notes

1. Feinberg, B., Adam and Edward S., Shapiro, “Teacher accuracy: An examination of teacher-based judgments of students’ reading with differing achievement levels”, *Journal of Educational Research*, 102 (2009): 453-462.

2. Cipollone, Piero, and Paolo Sestito, *Il capitale umano. Come far fruttare i talenti*, Il Mulino, Bologna, 2010.

3. Graney, B., Suzanne, “General education teacher judgments of their low-performing students’ short-term reading progress”. *Psychology in the Schools*, 45(2008): 537-549.

4. Hoge, Robert D. and Theodore Coladarci, “Teacher-based judgments of academic achievement: A review of literature”, *Review of Educational Research*, 59 (1989): 297-313.

5. Ibidem, Feinberg, B.A., Shapiro, E.S.

6. Bratti, M., Checchi, D., Filippin, A., Da dove vengono le competenze degli studenti? I divari territoriali dell'indagine OCSE-PISA 2003, Il Mulino, Bologna (2008).
7. Ibidem, Cipollone, P., Sestito, P.

8. OCSE-PISA, PISA 2009 Technical Report, available at http://www.oecd.org/dataoecd/13/34/48578536.pdf, 2006.

9. Skrondal, Anders and Sophia, Rabe-Hesketh, “Multilevel logistic regression for polytomous data and rankings”, *Psychometrika*, (2003): 267-287.

10. Snijders, Tom and Roel J., Bosker, *Multilevel Analysis. An introduction to basic and advanced multilevel modeling*, Sage Publications Ltd (1999).

11. Ibidem, OCSE-PISA.

12. Bond, T.G., Fox, C.M., Applying the Rasch Model: Fundamental measurement in the human sciences, London: Lawrence Erlbaum Associates Publishers (2001).

13. Ibidem, OCSE-PISA.

14. Martini, Angela e Roberto Ricci, “Un esperimento di misurazione del valore aggiunto delle scuole sulla base dei dati PISA 206 del Veneto”, *Rivista Economica e Statistica del Territorio*, (2010): 78-105.

Bibliography

Bond, G., Trevor and Christine, M., Fox, *Applying the Rasch Model: Fundamental measurement in the human sciences*, London: Lawrence Erlbaum Associates Publishers, 2001.

Bratti, Massimiliano, Daniele, Checchi and Antonio, Filippin, *Da dove vengono le competenze degli studenti? I divari territoriali dell'indagine OCSE-PISA 2003*, Il Mulino, Bologna, 2008.

Cipollone, Piero, and Paolo Sestito, *Il capitale umano. Come far fruttare i talenti*, Il Mulino, Bologna, 2010.

Feinberg, B., Adam and Edward S. , Shapiro, "Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels", *Journal of Educational Research*, 102 (2009): 453-462.

Graney, B., Suzanne, "General education teacher judgments of their low-performing students' short-term reading progress". *Psychology in the Schools*, 45(2008): 537-549.

Hoge, Robert D. and Theodore Coladarci, "Teacher-based judgments of academic achievement: A review of literature", *Review of Educational Research*, 59(1989): 297-313.

OCSE-PISA, PISA 2009 Technical Report, available at http://www.oecd.org/dataoecd/13/34/48578536.pdf, 2006.

Martini, Angela e Roberto Ricci, "Un esperimento di misurazione del valore aggiunto delle scuole sulla base dei dati PISA 206 del Veneto", *Rivista Economica e Statistica del Territorio*, (2010): 78-105.

Skrondal, Anders and Sophia, Rabe-Hesketh, "Multilevel logistic regression for polytomous data and rankings", *Psychometrika*, (2003): 267-287.

Snijders, Tom and Roel J., Bosker, *Multilevel Analysis. An introduction to basic and advanced multilevel modeling*, Sage Publications Ltd (1999).