I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10458

Teacher Assessments versus Standardized Tests: Is Acting "Girly" an Advantage?

Adriana Di Liberto
Laura Casula

DECEMBER 2016

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10458

Teacher Assessments versus Standardized Tests: Is Acting "Girly" an Advantage?

Adriana Di Liberto
University of Cagliari, IZA and CRENoS

Laura Casula
University of Cagliari and CRENoS

DECEMBER 2016

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Teacher Assessments versus Standardized Tests: Is Acting "Girly" an Advantage?*

We study if Italian teachers do apply gender discrimination when judging students. To this aim, we use a difference-in-differences approach that exploits the availability of both teachers (non-blind) and standardized test (blind) scores in math and language that Italian students receive during the school year. Using data for all sixth graders, descriptives show that in both scores girls are better than boys in the language scores, while in math boys perform better than girls in the blind test. Moreover, our analysis suggest that boys are always discriminated by teachers in both subjects. This result holds also when we control for class fixed effects, students noncognitive skills, gender specific-attitude towards cheating and possible cultural differences towards gender attitudes in math or language.

JEL Classification: L2, I2, M1, O32

Keywords: gender stereotypes, discrimination, schooling outcomes

Corresponding author:
Adriana Di Liberto
Dipartimento di Scienze Economiche e Aziendali
University of Cagliari
via S. Ignazio 17
09123, Cagliari
Italy
E-mail: diliberto@unica.it

* We thank the seminar participants at the 2016 AIEL Conference (Trento) and the 2016 Counterfactual Methods for Policy Impact Evaluation (COMPIE) Conference (Milan). We also thank INVALSI and in particular Patrizia Falzetti for providing the data on student outcomes. We are solely responsible for all the remaining errors.

1 Introduction

Are school teachers influenced by students gender when evaluating them at the exams? And, if so, in what way? Do they discriminate against a specific gender and also on specific subjects? These questions are of a great interests as teachers stereotypical perceptions when assessing academic results may have long lasting consequences on students school performance and, through this, on their following labor market outcomes.

The literature suggests the presence of different channels that link gender stereotypical perceptions when evaluating students with their economic and social outcomes. First, evidence shows that the highest performing education systems are those that combine educational quality with equity, and teachers gender biased (mis)judgment may affect many educational outcomes. In particular, dropping out is considered the result of a complex process of student disengagement and teachers gender discrimination is detrimental for misjudged students (OECD, 2012; Lyche, 2010). Data also suggest that there exist a significant difference between boys and girls in educational attainment, with boys more likely to repeat school years than girls and being predominate among early school leavers (Eurydice, 2010). If these results were driven by the presence of teachers gender biased evaluations we should find that at school boys are discriminated against girls.

Second, teachers gender stereotypes could be also differentiated by subject, such as "boys are good in math and science, while girls in literature and poetry" (Lavy and Sands, 2015). This kind of teacher's stereotypes could cause lower/higher grades for girls in math/language and it would result in a misallocation of talents and skills: science-gifted women would invest more in "girly" studies that are also less profitable in terms of labor market outcomes, while the opposite would be true for men. Overall, if this teachers stereotypes behavior represents the rule rather than an exception in an educational system, misallocation processes may significantly affect a nation labor force productivity and they may harm its growth perspectives. Again, international data show that women with the same educational attainment as men are under-represented in many scientific and technical degrees, which typically lead to better paid occupations. Recent evidence also suggests that this gap is narrowing quickly in nations that pursue gender equality policies (Machin and

Pekkarinen 2008, Guiso et al. 2008).

In this study we focus on the Italian school system. With respect to most industrialized countries, the Italian educational system performs poorly: its mean performance at PISA tests in all subjects is below the OECD average. Moreover, boys outperform girls in mathematics by an average of 18 points, and this gap has remained stable since 2003. Conversely, girls outperform boys in reading by an average of 39 score points. The gender gap observed in both subjects is similar to that observed across OECD countries. Finally, considering school dropouts, as in most industrialized countries, girls outperform boys in Italian schools, with dropout rates among boys (17.7%) significantly higher than that observed for girls (12.2%).[1]

In order to test for the existence of gender stereotyping and discrimination by Italian teachers, we follow Lavy (2008) and use a difference-in-differences approach that exploit the presence in the dataset of both blind and non-blind results in two different subjects, math and language. In fact, for each Italian student we have both a measure of the score assigned by math and language teachers and the Invalsi standardized test score results obtained in the same subjects during the same school year. The Invalsi standardized tests are compulsory for all Italian schools and students, both public and private, attending specific grades of schooling. Unlike the score assigned by student's teacher which is a "non-blind" score, given the way it is implemented the Invalsi tests may be considered as a "fair" or "blind" assessment. Thus, here we assume that the blind score may be used as the counterfactual measure to the non-blind score, which may be influenced by the teachers discrimination and other factors related to their culture.[2]

Data are provided by the Invalsi, the Italian institute in charge of evaluating school performance, and include information on both blind (standardized tests carried out by Invalsi) and non-blind (teachers evaluation) students results, together with many additional information on students and school characteristics. That is, the Invalsi dataset provides a very rich set of information on student characteristics that includes not only a full set of demographics, but also information on noncognitive skills, such as students attitude towards learning by subject. This enables us to exploit within country data on all students during the schooling year 2010-11 for sixth graders, a unique

[1] Eurostat, LFS (2014).

[2] On the use of the systematic difference between blind and non-blind tests across groups as a method to underline discrimination see also the work of Blank (1991) and Goldin and Rouse (2000).

advantages over existing analysis, that usually focus on small sample of students, and may suffer from problems of limited external validity and sample selection.

Our main result supports the evidence also found in other studies that teachers assessment always act against male students.[3] This result is also robust to the inclusion of class fixed effects and different model specifications.

First, we control for noncognitive skills using different measures of self-assessed ability and propensity for studying Math or Language. Cornwell et al. (2013) suggest that excluding these skills from the analysis would produce biased results, with teachers gender discrimination vanishing when noncognitive skills are taken into account.

Second, we perform the regression analysis for the subset of classes where external inspectors invigilate students during the blind standardized test. Indeed, cheating is a well-known phenomenon during Italian schooling exams, and girls may have a different attitude towards cheating than boys. Thus, using this subsample we are confident a) that cheating is not an issue and b) since all the steps of the Invalsi testing protocol has been fulfilled, that the blind score is likely to be free of any bias that might be caused by teachers attitude.

Finally, we exploit a specific feature of our Italian sample. In fact, unlike most within-country data sets there exists a deep, persistent duality in Italy between the developed North-Center and the less developed South. This substantial geographical heterogeneity is also present in both education and gender roles. For the former, both quantitative (educational attainments) and qualitative (cognitive skill tests results) educational outcomes stress a large gap between the two areas.[4] For gender roles, as suggested by many labor market outcome indicators, women's traditional role of wife and mother is still more persistent in the South.[5] Thus, in order to take into account for cultural factors and gender roles that may differently affect the choice of how much to invest in studying specific subjects, we have also performed the analysis separately for the Northern and

[3]Together with Lavy (2008), see also Bjorn et al. (2011), Hanna and Linden (2012), Cornwell et al. (2013).

[4]On this see Di Liberto (2008) and Di Liberto et al. (2015).

[5]On gender biased labor market outcomes see Del Boca (2005), and Di Liberto and Sideri (2015) for cultural differences across Italian regions. The importance of cultural bias in teachers' assessment has been also recently stressed by Card and Giuliano (2015). Using data from a large urban school district, they show that underrepresented groups are better under a screening process that places less weight on teachers subjective assessments. In fact, they find that the process for identifying gifted students, through parent and teacher referrals, systematically misses many potentially qualified disadvantaged students and they suggest that factors related to race or culture may play an important role.

Southern regions of the country.

Overall, all our robustness checks confirm that boys who perform equally well as girls on language and math blind tests are graded less favorably by their teachers.

2 Data and descriptives

We constructed a database with rich information on student, school and area characteristics. Our main source of data is the database provided by the National Institute for the Evaluation of the Educational System of Instruction and Training (Invalsi henceforth), a government agency that carries out a yearly evaluation of student attainment in both Mathematics and Language. The Invalsi standardized tests are compulsory for all Italian schools and students, both public and private, attending specific grades of schooling. In our analysis we focus on the 2010-11 school-year data for sixth grade lower secondary school students.[6] Invalsi enforces a protocol for the administration of the tests to reduce discretion and the possibility of teachers manipulations (Invalsi, 2011). First, the type of tasks that students have to complete include multiple choice and closed-format short answer questions. Second, the test is not administered by the class teachers but by other teachers of the school, who in general teach a different subject from the one that is being tested.[7]

Together with the standardized test results, the dataset also include a measure of the score in both language and math assigned by teachers during the first term. Given the way the test is implemented, the Invalsi tests may be considered "blind" or "fair" assessments while, in contrast, the score assigned by student's teacher is a "non-blind" and possibly biased by perceptions (or unfair) assessment.[8]

Moreover, the Invalsi questionnaire is also designed in order to collect detailed information

[6]Tests are carried out also by students attending the second and fifth grade (in primary schools), the sixth and eighth grade (in lower secondary) and the tenth grade (in upper secondary). The Italian school system starts at age six with five years of primary school (grades 1 to 5) followed by three years of lower secondary school (grades 6 to 8). Upper secondary education lasts three to five years depending on the type of school chosen.

[7]Moreover, all the school teachers are simultaneously involved in the transcription process, so that they cross-check each other while the school principal, who is responsible for the correct implementation of the protocol, supervises the whole process. For more on this see also Lucifora and Tonello (2015).

[8]It is difficult to find a comparison of blind and non-blind grading of the exact same tests. One example is in Hinnerich et al. (2010).

about the schools, the student background and family characteristics.[9] In our analysis we include the following additional demographic information about students: gender, citizenship (native, first and second generation immigrant students), if she/he speaks a foreign language at home or an Italian dialect, her/his socio-economic background using the number of books at home, the number of siblings, and parental education.[10] The set of school characteristics includes the number of students per class and school, the proportion of female students per class, and the school-average ESCS index. The latter is an index for student socioeconomic background, analogous to the same one computed by OECD for the PISA test. It is calculated based on the parental occupational status, their educational attainment levels and different measures of household possessions including cultural possessions such as home educational resources and the number of books, and the individual scores of this index are obtained by a principal component analysis, with normalized zero mean and unit standard deviation.[11]

Finally, our empirical analysis exploits the information obtained through the Invalsi dataset merged with other variables that controls for different area characteristics: we control for macro-area dummies, plus we include a proxy for the wealth level of the school catchment area (per capita value added), a measure of the level of criminality, and a social capital indicator.[12] In fact, previous studies show that geographical location is an important determinant of Italian students test scores, with students in the Northern area usually outperforming those living in the South and differences in both economic and cultural factors may play a role. The complete list of variables is reported in Appendix A and Table 1 sums up the major characteristics of the variables used in regressions for our overall sample.

In our descriptives analysis we check if we observe a misalignment between the standardized scores and the teachers grades. Since the Invalsi and teachers' votes are expressed in different scales, in order to compare the two set of students results (blind/Invalsi tests vs non-blind/teachers) we

[9]Information is collected through a "Family Questionnaire" sent to each family before the test, a "Student Questionnaire" filled by each student the first day of the test and, finally a student general information part compiled from school administrative staff.

[10]First generation are students born abroad of foreign-born parents, while second generation students are native-born children of foreign-born parents. In using the variable "number of books at home" we follow Hanushek and Woessmann (2011) who argue that this is the best single predictor of students performance.

[11]They are the scores for the first principal component. The index is calculated considering the whole sample of sixth grade lower secondary school Italian students. See also Invalsi (2011) for details.

[12]We identify the following dummy variables: North-East, North-West, Centre, South, South-Islands.

firstly convert the scores to the same scale and, secondly, we calculate the z-scores: that is, we standardize them to a distribution with zero mean and unit standard deviation. Figures 1 to 4 show the kernel-density distribution of the two types of scores by gender and subject. For language (Fig.1 and Fig. 2) we observe a rightward shift of the Invalsi-score distribution relative to the teacher-score distribution for both boys and girls. The opposite is true for girls in math (Fig. 3), while the distributions of scores in math for boys almost overlaps (Fig. 4). In sum, assuming that the standardized test scores represent fair assessments, these figures seem to suggest that language teachers punish students with respect to the blind test results, while they tend to inflate girls scores in math.

Table 2 includes the non-standardized average scores by gender achieved in both the blind and non-blind test. This table compares the non-standardized scores, that is, both blind and non-blind test scores are only transformed to a 0-100 scale in order to make it easy to interpret. First, comparing the two types of tests, it seems that standardized tests are easier than the teachers assessment as scores are higher on average for both boys and girls. The opposite is true for math. Second, numbers show that, on average, in math boys outperform girls in the blind test scores. However, when assessed by teachers, girls obtain on average a higher score. For language the picture is different: girls are always better than boys in both types of tests.

Table 3 uses the z-scores and it includes a first measure of the teachers discrimination, calculated as a simple difference-in-differences: that is, teachers gender bias is defined as the average gap between non-blind and blind scores for boys, minus this same gap for girls. Overall, comparing the results in the two subjects, these numbers suggest that teachers discriminate boys in language by almost one-tenth of a standard deviation, and that the gender discrimination gap is higher (more than double) in math.

We finally report some descriptives on students noncognitive skills in Table 4. The Invalsi questionnaire includes several indicators related to students drive and motivation in studying a specific subject. In particular, it asks different questions designed to measure the self-assessment of boys and girls about their ability in Math (Q3) and Language (Q5) studies. In details, during the survey, Italian students are asked to indicate how much they agree with five different statements

about mathematics and language studies.[13] The specific questions asked and the results by gender are in Table 4. As expected, boys are more confident and enjoy more studying math, while girls are more confident in language studies. Overall, numbers show that the subject specific propensity for learning and achieving is very different between boys and girls and suggest that gender specific attitude may play a role in our analysis.

3 Results

Following Lavy (2008), we use the data pooled over the two types of scores, one blind and the other non-blind, in the two subjects (Math and Language) and use a difference-in-differences regression setting of the form:

$$y_{ijb} = \alpha + \beta Male_i + \gamma NB_{ijb} + \delta(Male_i \times NB_{ijb}) + \upsilon_{ijb} \tag{3.1}$$

where y_{ijb} is an indicator of performance of student i attending school j for both blind and non-blind scores b, $Male_j$ is the gender dummy (equal to one if male), and NB is the dummy identifying the teachers (non-blind) scoring procedure. Thus, the intercept is the average score obtained by female students on blind tests, β captures the score difference of male students in both types of tests, and γ measures the teachers effect, that is, the average differences in scores due to the type of tests. The parameter of interest is on the interaction term, δ, that measures the difference in scores obtained by male students due to teachers. As said, above, given the Invalsi testing protocol, we may assume that the standardized test score is free of bias that might be caused by stereotyped discrimination. Conversely, the non-blind score may possibly reflect biases teachers' gender stereotypes.

Table 5 and 6 show the results for the Language and Math scores respectively. In both Tables model 1 include the results of equation 1, our most parsimonious specification, while in the following models, we exploit a rich set of variables that control for student characteristics, including self-assessed ability and propensity for studying both Math and Language, and for school and area characteristics. Standard errors are clustered at class level. For Language (Table 5), we find that

[13] Invalsi uses the following scale: 1-moderately disagree, 2-moderately disagree, 3-somewhat agree, 4-strongly agree.

all coefficients in model 1 are significant and that, on average, female students perform better than boys: girls have advantages of 0.209 of a standard deviation of the blind score distribution in language. The mean difference between the teachers scores and the Invalsi scores is positive and significant, while our parameter of interest, the coefficient of the interaction term, is negative, suggesting that teachers' discrimination acts against male students. In sum, for language studies, results suggest that teachers widen an already existing female-male achievement difference.

A different picture emerges for math. In this case the advantage is for male students: the coefficient on the gender dummy, male, is positive (0.125) and it is statistically significant. And, as already seen in Table 3, the teachers bias is still against the boys. Therefore, we find no evidence that teachers gender stereotypes cause lower grades for girls in math, that represents one possible explanations of the bias against women existing in scientific, or STEM, fields. Conversely, the coefficient on the interaction term implies that the estimated bias in math represents 0.2 points of the standard deviation, and it almost doubles the teachers' bias coefficient found for language. This result is consistent with other evidence in the literature.[14]

For both math and language scores, the introduction of additional controls does not change these results. Model 2 introduces different variables that control for students demographics, while in Model 3 we increase the specification with more family characteristics in order to take into account for the student's socioeconomic background. Model 4 includes the school average socio-economic background (calculated by the ESCS index), the school size and the proportion of girls in each class. The latter variable should control for gender peer effects and it has been found to be an important determinant in these analysis (Lavy et al., 2011).[15] In particular, peers may directly influence gender differentiation by providing boys and girls with different learning opportunities and feedback. Unlike most studies on teachers discrimination, in model 5 we also introduce two different dummy variables that should capture students noncognitive abilities: the dummies "good in math/language" are equal to one for students that show a strong propensity for studying the specific subject.[16] Finally in model 6 we also control for area characteristics, including total value

[14]See for example Bjorn et al. (2011), Hanna and Linden (2012), and Breda and Ly (2015).

[15]Lavy et al. (2011) find that an increase in the proportion of girls improves boys and girls cognitive outcomes.

[16]Their answer in Q3A, Q3C, Q3D, Q3E, Q5A, Q5C, Q5D and Q5E is strongly agree, while they strongly disagree in Q3B and Q5B. See Table 4.

added per capita in 2001 that represents a standard proxy of an area economic performance, the rate of extortions over 1000 inhabitants, and a measure of social capital.[17] All these additional indicators should capture cultural features that may differently affect boys and girls students outcomes.

Overall, the estimated coefficients on our additional controls all show the expected signs and still confirm that boys who perform equally well as girls on language and math blind tests are graded less favorably by their teachers. They also confirm that, contrary to expectations from gender-stereotyping, discrimination goes more in favor of females in more scientific (or male) subjects.

4 Robustness checks

In this section we perform a set of robustness checks of the results discussed above. For these, we only report in our Tables the coefficients of the three main variables.

One problem of the analysis performed above is that it cannot rule out the hypothesis that the two types of test do not measure exactly the same skills. As found in Cornwell et al. (2013), even noncognitive skills my play an important role and they may be the main driver of our results. In general, characteristics such as oral expression, self-confidence, anxiety or shyness are likely to affect the candidates scores in different ways at the non-blind test and at the Invalsi blind tests: if there are systematic differences between males and females regarding these characteristics, we cannot interpret any gender difference between the two scores as reflecting discrimination, since we cannot disentangle the role of the teacher separate from that of the assessment process. For instance, the standardized test may be perceived by the students as a more pressured environment and, if girls are more anxious than boys, they may obtain lower results in standardized tests due to this. In our sample this should not represent a significant problem since, for sixth graders, the Invalsi test is not high stake, while the non-blind score may contribute their end of school year results. Overall, the two types of assessments should be equally stressful for the students.[18]

As seen above, our data includes different measures of self-assessed ability and propensity for studying Math and Language: boys indicate a more positive attitude than girls in studying math, while the opposite is true for language studies. Instead of including these variables among the list

[17] To this aim we use a synthetic social capital index at regional NUTS3 level, provided by Cartocci (2007), which merges data on 1) blood donations, 2) sport participation, 3) dissemination of newspaper and 4) voter turnout.

[18] If there is any stress difference, maybe the teachers' score should be more stressful.

of additional controls, we replicate our analysis for two subsamples of students that share the same level of attitude for learning a specific subject: the first only includes the group of students that are very confident in studying and being proficient in a certain subject, while the second includes only those that, conversely, seems to have a low attitude for studying.[19] Results are reported in Table 7 (for language) and 9 (for math), with Panel A showing the results for the students with strong propensity to learn and Panel B including those for the group with a low attitude for studying.

Further, we can also identify a representative and random sample of monitored classrooms where external inspectors invigilate students during the test and also help to both compute results and prepare the documentation relative to the test. This is an important feature of our dataset since there is evidence showing that Italian students in the non-monitored classrooms receive a more benevolent supervision, allowing student cheating behavior more easily (Lucifora and Tonello, 2015). Indeed, it is possible that the attitude towards cheating is different by gender. For this subsample we are also confident that the Invalsi test protocol has been thoroughly implemented and teachers, rather than students, did not manipulate the scores and, eventually, discriminate by gender. Evidence of teachers' manipulation has been found in Pereda-Fernndez (2016). This paper suggests that the cheating is concentrated in the South of Italy and, more important for us, it tends to favor female students.[20] In Panel C of Tables 7 (for language) and 9 (for math) we show the results when we replicate the analysis for the sub-sample of classes with the presence of an inspector.

We also replicate our analysis including fixed effects at class level, in order to capture all unobserved elements affecting scores in a given class, including also teachers' characteristics such as, for instance, teachers severity. Results are in Panel D (Table 7 for language and Table 9 for math). Unfortunately, our dataset does not include variables that control for teachers' characteristics. In particular, teachers' gender has been found to be an important variable in other studies in this literature, as it may influence students results through the presence of both a role model effect and/or a teacher bias effects (Lavy, 2008; Paredes 2013). In general, teachers may endorse

[19] In details, Panel A only includes students who strongly agree with the statement "I am proficient in Math/Language". In panel B results are obtained using the subsample of students that strongly disagree with the same statement. For more on this, see Table 4.

[20] Pereda-Fernndez (2016) uses the Invalsi data for the academic year 2012/13 and for different grades.

prejudices and show, for instance, preferences for same-gender individuals. Together with cultural stereotypes, prejudices influence teachers classroom behaviors and their assessment activities. The percentage of female teachers in Italy is among the highest across OECD countries (Education at glance, 2014): in our sample of lower secondary schools, the percentage of female teachers is almost 80%.[21]

Overall, evidence from Tables 7 and 9 shows that, even using subsamples that allow us to get rid of some important differences in noncognitive abilities, or controlling for class fixed effects, our main results are fully confirmed.

Finally, we replicate the analysis separately for the subsamples of northern and southern Italian regions. There is a vast literature showing that there exists a deep, persistent duality in Italy between the developed North-Center and the less developed South. The gap between the two areas is also in terms of culture and gender roles, and geographical location has been also found as an important determinant of Italian student test scores.[22] Thus, in principle it is possible that more educators in the southern regions endorse cultural gender stereotypes (e.g., math is easier for boys than girls) than in the northern ones. In this case, girls could be more discriminated in math when attending schools located in the southern rather than in northern regions. Tables 8 (language) and 10 (math) report the results, with Panel A showing the coefficients for the subsamples of northern regions, and Panel B for the South. We also replicate the same analysis using the subsample of inspected schools in Panel C and D. Results reveal no significant differences between the two areas of the country.

5 Conclusions

This study investigates if teachers have a grading bias against a specific students gender. To this aim, we exploit a unique dataset that, unlike other studies in this literature, enables us to use a rich set of variables for all Italian students attending the sixth grade. We apply a difference-in-differences approach using the information on both the teacher (non-blind) grades and the standardized test (blind) scores in two different subjects, math and language. We assume that teachers gender

[21] It is almost 100% (98%) in primary school, and 66% in Italian upper secondary schools. The OECD average is 82% in primary, 67% in lower secondary, and 57% in upper secondary schools.

[22] Cipollone et al. (2010), Di Liberto et al. (2015).

stereotypes are manifested through their evaluation of students, while the standardized test scores are an unbiased evaluation process.

Our results strongly suggest that Italian teachers tend to discriminate against boys, and that they do not discriminate more against girls in more scientific subjects. The teachers' bias is estimated in both math and language studies, but the coefficient of the former represents 0.2 points of the standard deviation and it almost doubles the latter. This result impinges the idea that school teachers directly contribute to the significant gender selection observed in STEM tertiary studies by discriminating more against girls in more scientific subjects. All robustness checks confirm these results. Our analysis takes into account for noncognitive skills and the possibility that the blind and the non-blind scores might not measure the same abilities, for the presence of different social norms and gender stereotyping in different areas, and it controls for fixed effects at class level.

In sum, this evidence may contribute to explain an important phenomenon such as the observed high dropout rate at school among boys. A potential explanation is that boys are systematically discouraged by teachers during their school career. Our findings corroborate the idea that teachers tend to favor some "girly" attitude in class, for instance, they punish boys for (bad) discipline. On this, our results are more suggestive rather than conclusive and these mechanisms need to be further investigated in future research.

References

[1] Bjorn, T.H., Hoglin, E., Johannesson, E. (2011), Are boys discriminated in Swedish high schools? *Economics of Education Review*, Vol. 30(4), 682-690.

[2] Blank, R. M. (1991), The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review *American Economic Review*, Vol. 81(5), 1041-1067.

[3] Breda, T., Ly, S. T. (2012), Do professors really perpetuate the gender gap in science? Evidence from a natural experiment in a French higher education institution. PSE Working Papers n.2012-13.

[4] Card D., Giuliano, L. (2015), Can Universal Screening Increase the Representation of Low Income and Minority Students in Gifted Education? NBER Working Paper No. 21519.

[5] Cartocci, R. (2007), Mappe del tesoro. Atlante del capitale sociale in Italia. Il Mulino, Bologna.

[6] Cipollone, P., Montanaro, P. and Sestito, P. (2010), Value-Added Measures in Italian High Schools: Problems and Findings, *Giornale degli Economisti* 69 (2):81114.

[7] Cornwell, C., Mustard D., Van Parys J., (2013), Non-cognitive Skills and Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School, *Journal of Human Resources*, Vol. 48(1), 236-264.

[8] Del Boca D. (2005), Editorial Foreword, *Labour*, vol. 19(s1), pp. 1-4.

[9] Di Liberto, A., (2008), Education and Italian Regional Development, Economics of Education Review, vol. 27, No.1, pp.94-107.

[10] Di Liberto, A. and M. Sideri (2015), Past dominations, current institutions and the Italian regional economic performance, *European Journal of Political Economy* 38, 12-41.

[11] Di Liberto, A., Schivardi, F., Sulis, G. (2015), Managerial Practices and Students' Performance, *Economic Policy*, Vol. 30 (84).

[12] Eurydice (2010), Gender Differences in Educational Outcomes: Study on the Measures Taken and the Current Situation in Europe, Brussels.

[13] Eurostat, Labour force survey 2014.

[14] Goldin, C., Rouse C. (2000), Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review*, Vol. 90(4), 715-741.

[15] Guiso, L., Monte F., Sapienza, P. and Zingales L. (2008), Culture, Gender, and Math, *Science* 320: 11641165.

[16] Hanna, R.N., Linden L. L., (2012), Discrimination in Grading, *American Economic Journal: Economic Policy*, Vol. 4(4), 146-68.

[17] Hanushek, E.A., Woessmann L. (2011), The Economics of International Differences in Educational Achievement, in: Hanushek E.A., Machin S., Woessmann L. (eds.) *Handbook of the Economics of Education*, Vol. 3, Amsterdam: North Holland.

[18] Hinnerich, B. T., Hoglin, E., Johannesson, M., (2011), Are boys discriminated in Swedish high schools? *Economics of Education Review*, Vol. 30, 682-690.

[19] Invalsi (2011), Rapporto tecnico sulle caratteristiche delle prove Invalsi 2011, *Technical Report, Invalsi.*

[20] Lavy, V. (2008), Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment, *Journal of Public Economics*, Vol. 92, 2083-2105.

[21] Lavy, V., Schlosser A. (2011), Mechanisms and Impacts of Gender Peer Effects at School, *American Economic Journal: Applied Economics*, 3: 133.

[22] Lavy, V., Sand E. (2015), On The Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers Stereotypical Biases, NBER Working Paper No. 20909.

[23] Lyche, C. (2010), Taking on the Completion Challenge: A Literature Review on Policies to Prevent Dropout and Early School Leaving, OECD Education Working Papers, No.53, OECD, Paris.

[24] Lucifora, C. and Tonello M. (2015), Cheating and social interactions. Evidence from a randomizedexperiment in a national evaluation program, *Journal of Economic Behavior and Organization*, 115, 4566.

[25] Machin S., Pekkarinen T. (2008), Global Sex Differences in Test Score Variability, *Science*, Volume 322(5906): 1331-2.

[26] OECD (2012), Equity and Quality in Education: Supporting Disadvantaged Students and Schools, OECD Publishing.

[27] OECD (2014),PISA 2012 Results: What Students Know and Can Do, OECD Publishing.

[28] Paredes, V., (2012), A teacher like me or a student like me? Role model versus teacher bias effect *Economics of Education Review*, Vol. 39(2014), 38-49.

[29] Pereda-Fernndez, S. (2016), Teachers and Cheaters. Just an Anagram?, Banca d'Italia, *mimeo*.

A Data sources

Description of Variables:

Dependent Variables:

- **Language_test**: Invalsi (blind) language test scores
- **Math_test**: Invalsi (blind) Math test scores
- **Language_Teacher**: Teachers' (non-blind) language scores
- **Math_Teacher**: Teachers' (non-blind) Math scores

Student and family characteristics:

- **Males**: dummy=1 if male
- **good at math**: see Table 4.
- **good at language**: see Table 4.
- **n_brothers**: number of siblings (4 indicates 4 or more)
- **manybooks**: dummy=1 if more than 100 books at home
- **degree_m**: dummy=1 if mother with a degree
- **degree_f**: dummy=1 if father with a degree
- **high_m**: dummy=1 if mother with a high school diploma
- **high_f**: dummy=1 if father with a high school diploma
- **housewife**: dummy=1 if mother housewife, dummy=0 otherwise
- **Dialect**: dummy=1 if language spoken at home is a dialect
- **Foreign language**: dummy=1 if language spoken at home is not Italian
- **Foreign1**: dummy=1 if students are 1st generation immigrants

- **Foreign2**: dummy=1 if students are 2st generation immigrants

School and Class characteristics:

- **stud_class**: number of students per class
- **f_m_ratio_class**: females_males ratio in class
- **school_size**: number of students per school
- **escs_school**: Average School Level ESCS Index. The Invalsi ESCS Index refers to the PISA index of economic, social and cultural status
- **Campione**: dummy=1 if class selected for external monitoring by Invalsi

All these variables are from Invalsi.

Area characteristics:

- **lnvapop09**: Total value added per capita, constant prices (base year 2000), 2001 data. Source: Fondazione Istituto Tagliacarne (2006). http://www.tagliacarne.it.
- **mean_est_99_02**: Extortions (1999-2001): average rate of extortions over 10,000 inhabitants. Source: Fiaschi, D., Gianmoena, L. and Parenti, A. (2011)
- **putnam**: Social capital indicator. Source: Cartocci (2007).

B Figures and Tables

B.1 Figures

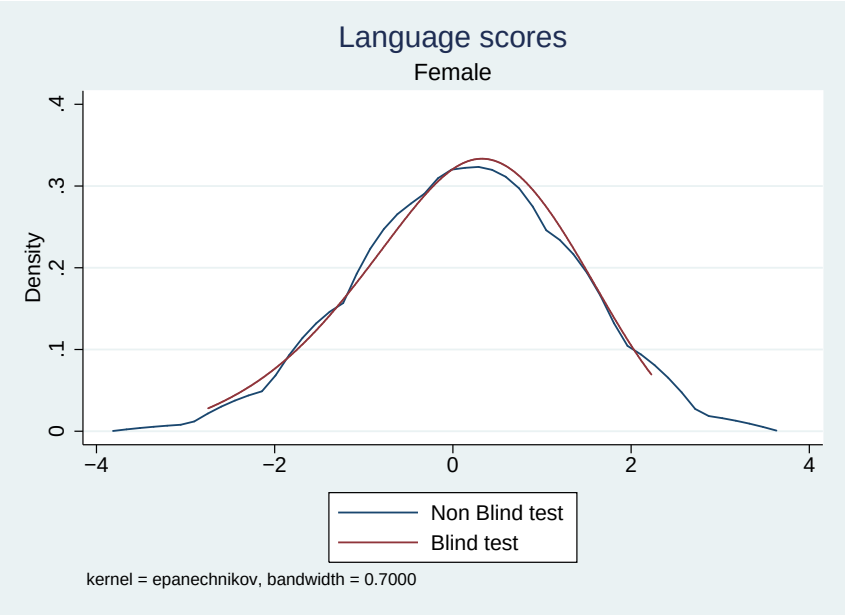

Figure 1: Language scores - Girls

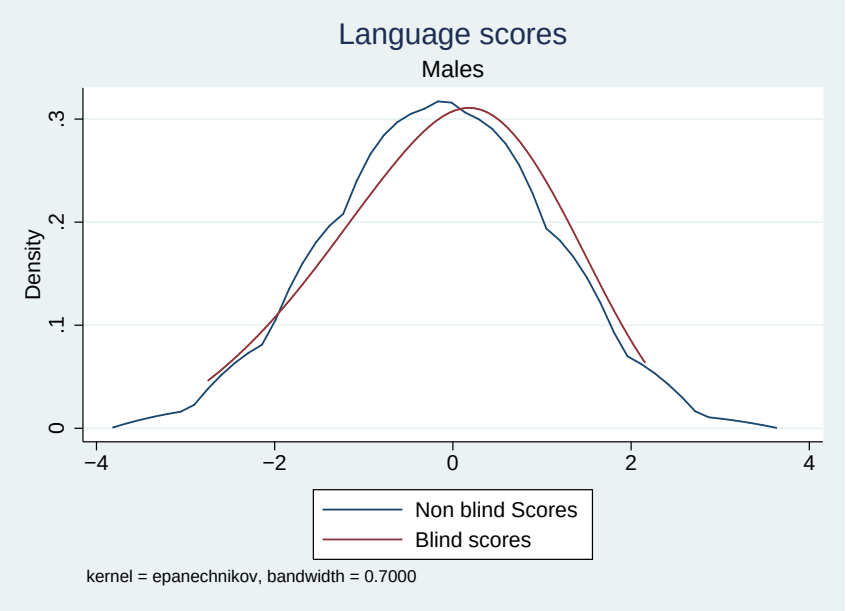

Figure 2: Language scores - Boys

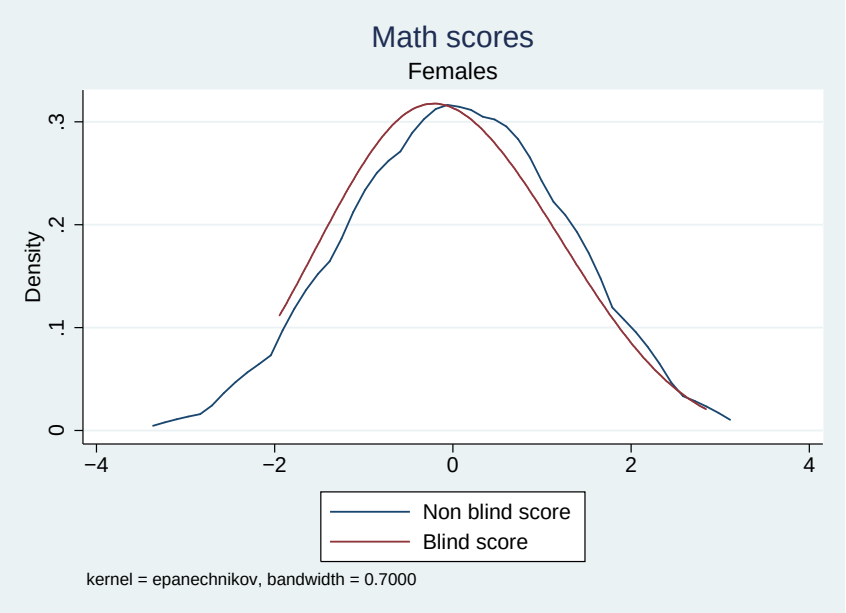

Figure 3: Math scores - Girls

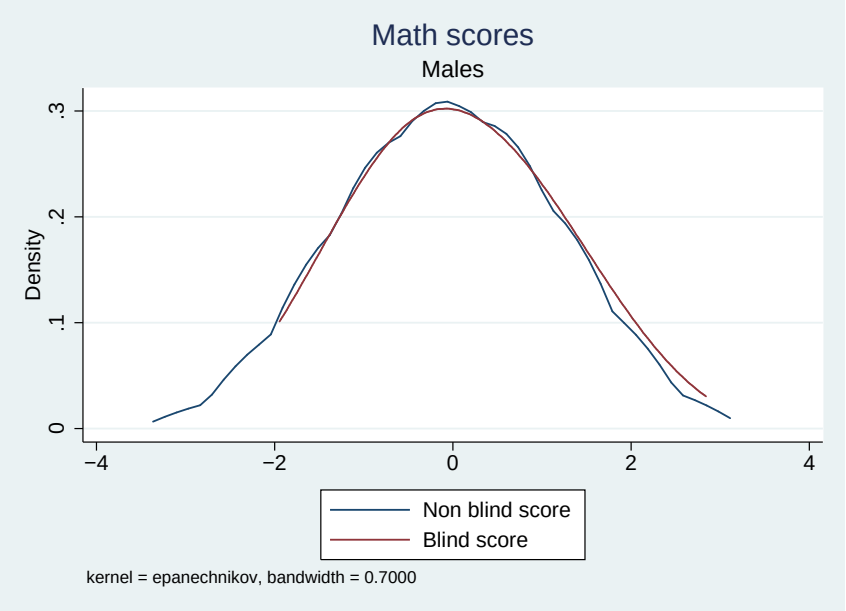

Figure 4: Math scores - Boys

B.2 Tables

Table 1: Descriptives statistics: overall sample

Variable	Obs	Mean	Std. Dev.	Min	Max
Dependent variables					
Language Test	498824	55.21	20.10	0.00	100.00
Math Test	498824	40.65	20.89	0.00	100.00
Language Teacher	498824	55.21	20.10	0.00	100.00
Math Teacher	498824	40.65	20.89	0.00	100.00
Student and family characteristics					
males	498824	0.51	0.50	0.00	1.00
good at math	492657	0.51	0.50	0.00	1.00
good at language	492172	0.48	0.50	0.00	1.00
n_brothers	462457	1.24	0.91	0.00	4.00
manybooks	498707	0.32	0.47	0.00	1.00
degree_m	418947	0.12	0.33	0.00	1.00
degree_f	412435	0.12	0.32	0.00	1.00
high_m	418947	0.38	0.49	0.00	1.00
high_f	412435	0.33	0.47	0.00	1.00
housewife_m	424056	0.40	0.49	0.00	1.00
dialect	467149	0.16	0.37	0.00	1.00
Foreign language	467149	0.07	0.26	0.00	1.00
Foreign 1st generation	498824	0.06	0.23	0.00	1.00
Foreign 2nd generation	498824	0.04	0.18	0.00	1.00
School and Class characteristics					
no stud class	498824	21.74	3.86	1.00	34.00
f_m ratio (class)	498824	0.46	0.11	0.00	1.00
no stud school	498824	147.14	77.54	1.00	417.00
escs_school	486597	-0.01	0.47	-2.39	1.78
campione	498824	0.08	0.27	0.00	1.00
Area characteristics					
lnvapop09	498824	10.04	0.29	9.50	10.47
mean_est_ 02	498824	6.50	3.74	1.71	19.45
putnam	498824	-0.69	3.16	-6.43	5.47
North_West	498824	0.25	0.43	0.00	1.00
Centre_North	498824	0.18	0.38	0.00	1.00
Centre_South	498824	0.23	0.42	0.00	1.00
Islands_South	498824	0.16	0.37	0.00	1.00

Table 2: Blind vs non-blind test: average results by gender

	Gender	Obs	Mean	Std. Dev.	Min	Max
Language - non-blind	Male	255032	49.12	16.56	0.00	100.00
	Female	243792	54.02	16.14	0.00	100.00
Mathematic - non-blind	Male	255032	51.58	20.01	0.00	100.00
	Female	243792	53.45	19.32	0.00	100.00
Language - blind	Male	255032	53.16	20.71	0.00	98.53
	Female	243792	57.36	19.21	0.00	100.00
Mathematic - blind	Male	255032	41.92	21.39	0.00	100.00
	Female	243792	39.31	20.28	0.00	100.00

Table 3: Means and Standard Deviations of Blind (B) and non-blind (NB) test and Teachers' Biases Measure at the Student Level by Gender

	Males			Females			
	non-blind Test	Blind Test	Difference Between N_B and B test	non-blind Test	Blind Test	Difference Between N_B and B Test	Teachers' Biases Measure (Student Level)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Math	-0.046	0.107	-0.107	0.049	-0.064	0.112	-0.219
	(-1.016)	0.956	(0.870)	(0.981)	(0.971)	(0.865)	
Language	-0.145	-0.102	-0.043	0.151	0.061	0.045	-0.088
	(1.001)	(1.030)	(0.891)	(0.976)	(1.024)	(0.856)	
Number of Students	255032	255032	255032	243792	243792	243792	498824

Notes: The Blind and non-blind scores are rescaled and standardized scores. The teachers' biases measured at the student level (column 7) are equal to the difference between boys' blind and non-blind scores (column 3) less the difference between girls' blind and non-blind scores (column 6). Standard errors are reported in parentheses.

Table 4: Ability in Math and Language studies: boys vs girls self-assessment

Please indicate how much you agree with the following statements or how true it is about you (mathematics) using the following scale: 1-moderately disagree, 2-moderately disagree, 3-somewhat agree, 4 strongly agree	**MALE**	**FEMALE**	**M vs F**
Q3_A - I am good at maths - I am proficient in maths	3.07	2.90	0.17
Q3_B - Studying math is more difficult for me than for most of my classmates	1.93	2.05	-0.12
Q3_C - It is easy for me to learn maths	3.05	2.87	0.18
Q3_D - Studying mathematics is fun	2.81	2.67	0.14
Q3_E - I would like to study more math at school	2.41	2.24	0.17
Please indicate how much you agree with the following statements or how true it is about you (Language) using the following scale: 1-strongly disagree, 2-moderately disagree, 3-moderately agree, 4 strongly agree	**MALE**	**FEMALE**	**M vs F**
Q5_A - I am good at language/Italian - I am proficient in Language/Italian	2.90	3.08	-0.18
Q5_B - Studying Language is more difficult for me than for most of my classmates	2.03	1.79	0.24
Q5_C - It is easy for me to learn Italian/Language	2.96	3.20	-0.24
Q5_D - Studying Italian/Language is fun	2.57	2.92	-0.35
Q5_E - I would like to study more Italian at school	2.15	2.47	-0.33

Table 5: Teachers gender bias in Language

Dependent Variable: Test results in Language (blind and non-blind)						
	(1)	(2)	(3)	(4)	(5)	(6)
Male	-0.209***	-0.166***	-0.171***	-0.168***	-0.176***	-0.176***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Non-blind score	0.045***	0.040***	0.046***	0.047***	0.046***	0.046***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Interaction	-0.088***	-0.090***	-0.091***	-0.092***	-0.092***	-0.092***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
dialect		-0.348***	-0.170***	-0.169***	-0.152***	-0.148***
		(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
for_language		-0.296***	-0.256***	-0.257***	-0.253***	-0.254***
		(0.007)	(0.008)	(0.008)	(0.008)	(0.008)
foreign1		-0.633***	-0.443***	-0.443***	-0.445***	-0.449***
		(0.008)	(0.009)	(0.009)	(0.009)	(0.009)
foreign2b		-0.431***	-0.267***	-0.267***	-0.274***	-0.279***
		(0.008)	(0.009)	(0.009)	(0.009)	(0.009)
n_brothers			-0.101***	-0.100***	-0.093***	-0.092***
			(0.002)	(0.002)	(0.002)	(0.002)
manybooks			0.213***	0.210***	0.173***	0.171***
			(0.003)	(0.003)	(0.003)	(0.003)
degree_m			0.419***	0.411***	0.387***	0.389***
			(0.005)	(0.005)	(0.005)	(0.005)
degree_f			0.332***	0.322***	0.302***	0.306***
			(0.005)	(0.005)	(0.005)	(0.005)
high_m			0.309***	0.306***	0.289***	0.290***
			(0.003)	(0.003)	(0.003)	(0.003)
high_f			0.225***	0.221***	0.208***	0.210***
			(0.003)	(0.003)	(0.003)	(0.003)
housewife_m			-0.025***	-0.022***	-0.028***	-0.024***
			(0.003)	(0.003)	(0.003)	(0.003)
no_stud_class				0.001**	0.001**	0.001**
				(0.001)	(0.001)	(0.001)
f_m ratio (class)				0.069***	0.079***	0.085***
				(0.021)	(0.021)	(0.021)
no_stud_school				-0.000***	-0.000***	-0.000***
				(0.000)	(0.000)	(0.000)
escs_school				0.039***	0.059***	0.055***
				(0.006)	(0.005)	(0.006)
good at math					0.275***	0.276***
					(0.003)	(0.003)
good at language					0.176***	0.178***
					(0.003)	(0.003)
lnvapop09						-0.033*
						(0.019)
mean_est_99_02						-0.006***
						(0.001)
social capital (putnam)						0.018***
						(0.001)
Constant	0.107***	0.408***	0.143***	0.106***	-0.080***	0.227
	(0.003)	(0.005)	(0.005)	(0.016)	(0.016)	(0.200)
Regional controls		YES	YES	YES	YES	YES
Observations	997,648	934,298	706,764	689,110	686,406	686,406
R-squared	0.016	0.093	0.193	0.194	0.222	0.223
No. classes	25819	25661	22928	22354	22350	22350

Table 6: Teachers gender bias in Mathematics

Dependent Variable: Test results in Math (blind and non-blind)						
	(1)	(2)	(3)	(4)	(5)	(6)
Male	0.125***	0.163***	0.161***	0.166***	0.088***	0.088***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Non-blind score	0.112***	0.113***	0.130***	0.131***	0.130***	0.130***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Interaction	-0.219***	-0.221***	-0.223***	-0.224***	-0.224***	-0.224***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
dialect		-0.314***	-0.148***	-0.150***	-0.131***	-0.127***
		(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
for_language		-0.222***	-0.183***	-0.186***	-0.192***	-0.194***
		(0.007)	(0.008)	(0.008)	(0.008)	(0.008)
foreign1		-0.494***	-0.314***	-0.314***	-0.317***	-0.321***
		(0.008)	(0.009)	(0.010)	(0.009)	(0.009)
foreign2b		-0.361***	-0.209***	-0.209***	-0.216***	-0.221***
		(0.008)	(0.010)	(0.010)	(0.009)	(0.009)
n_brothers			-0.072***	-0.071***	-0.066***	-0.064***
			(0.002)	(0.002)	(0.002)	(0.002)
manybooks			0.209***	0.207***	0.171***	0.168***
			(0.003)	(0.003)	(0.003)	(0.003)
degree_m			0.398***	0.393***	0.361***	0.362***
			(0.006)	(0.006)	(0.005)	(0.005)
degree_f			0.317***	0.313***	0.284***	0.288***
			(0.006)	(0.006)	(0.005)	(0.005)
high_m			0.286***	0.284***	0.261***	0.262***
			(0.003)	(0.004)	(0.003)	(0.003)
high_f			0.210***	0.207***	0.188***	0.191***
			(0.003)	(0.004)	(0.003)	(0.003)
housewife_m			-0.040***	-0.038***	-0.038***	-0.033***
			(0.003)	(0.003)	(0.003)	(0.003)
no_stud_class				0.001**	0.002***	0.002***
				(0.001)	(0.001)	(0.001)
f_m ratio (class)				0.086***	0.072***	0.078***
				(0.021)	(0.021)	(0.020)
no_stud_school				-0.000***	-0.000***	-0.000***
				(0.000)	(0.000)	(0.000)
escs_school				0.019***	0.041***	0.036***
				(0.006)	(0.006)	(0.006)
good at math					0.556***	0.558***
					(0.003)	(0.003)
good at language					-0.097***	-0.095***
					(0.003)	(0.003)
lnvapop09						-0.018
						(0.019)
mean_est_99_02						-0.007***
						(0.001)
social capital (putnam)						0.019***
						(0.001)
Constant	-0.064***	0.253***	-0.025***	-0.068***	-0.228***	-0.079
	(0.003)	(0.005)	(0.006)	(0.016)	(0.015)	(0.199)
Regional controls		YES	YES	YES	YES	YES
Observations	997,648	934,298	706,764	689,110	686,406	686,406
R-squared	0.003	0.070	0.160	0.161	0.238	0.240
No. classes	25819	25661	22928	22354	22350	22350

Table 7: Robustness checks 1: Language

Dependent Variable: Test results in Language (blind and non-blind)						
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: High achieving students (self-assessed)						
Male	-0.213***	-0.165***	-0.169***	-0.166***	-0.196***	-0.196***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.006)	(0.006)
Non-blind score	0.213***	0.208***	0.215***	0.216***	0.216***	0.216***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Interaction	-0.069***	-0.070***	-0.070***	-0.070***	-0.070***	-0.070***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Observations	*219,690*	*209,422*	*162,272*	*157,922*	*157,768*	*157,768*
Panel B: Low achieving students (self-assessed)						
Male	-0.273***	-0.245***	-0.243***	-0.233***	-0.210***	-0.210***
	(0.017)	(0.017)	(0.018)	(0.018)	(0.018)	(0.018)
Non-blind score	-0.047***	-0.064***	-0.067***	-0.067***	-0.066***	-0.066***
	(0.012)	(0.013)	(0.015)	(0.015)	(0.015)	(0.015)
Interaction	-0.052***	-0.043***	-0.047***	-0.050***	-0.051***	-0.051***
	(0.015)	(0.015)	(0.018)	(0.018)	(0.018)	(0.018)
Observations	*36,600*	*33,858*	*23,944*	*23,376*	*23,344*	*23,344*
Panel C: sub-sample of inspected schools						
Male	-0.208***	-0.170***	-0.175***	-0.172***	-0.181***	-0.182***
	(0.011)	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)
Non-blind score	0.015	0.013	0.019	0.019	0.019	0.019
	(0.012)	(0.012)	(0.013)	(0.013)	(0.013)	(0.013)
Interaction	-0.108***	-0.112***	-0.112***	-0.112***	-0.112***	-0.112***
	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)
Observations	*77,708*	*75,234*	*59,558*	*59,558*	*59,464*	*59,464*
Panel D: Class Fixed Effects						
Male	-0.20289***	-0.168***	-0.168***	-0.168***	-0.182***	-0.182***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Non-blind score	0.04481***	0.040***	0.044***	0.044***	0.043***	0.043***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Interaction	-0.08765***	-0.090***	-0.091***	-0.091***	-0.091***	-0.091***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Observations	997,648	934,298	627,874	627,874	625,418	625,418

Table 8: Robustness checks 2: Language

Dependent Variable: Test results in Language (blind and non-blind)						
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Northern regions						
Male	-0.214***	-0.177***	-0.170***	-0.169***	-0.186***	-0.186***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Non-blind score	0.141***	0.136***	0.134***	0.135***	0.134***	0.134***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Interaction	-0.071***	-0.076***	-0.081***	-0.080***	-0.080***	-0.080***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	*428,176*	*405,330*	*312,018*	*305,218*	*304,234*	*304,234*
Panel B: Southern regions						
Male	-0.216***	-0.155***	-0.179***	-0.171***	-0.171***	-0.171***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Non-blind score	-0.026***	-0.031***	-0.021***	-0.021***	-0.022***	-0.022***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Interaction	-0.105***	-0.107***	-0.104***	-0.104***	-0.103***	-0.103***
	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)
Observations	*392,322*	*363,858*	*275,138*	*267,764*	*266,522*	*266,522*
Panel C: North and sub-sample of inspected schools						
Male	-0.229***	-0.194***	-0.173***	-0.166***	-0.182***	-0.182***
	(0.017)	(0.016)	(0.016)	(0.017)	(0.017)	(0.017)
Non-blind score	0.080***	0.077***	0.073***	0.073***	0.073***	0.073***
	(0.017)	(0.017)	(0.018)	(0.018)	(0.018)	(0.018)
Interaction	-0.090***	-0.094***	-0.098***	-0.098***	-0.098***	-0.098***
	(0.014)	(0.014)	(0.015)	(0.015)	(0.015)	(0.015)
Observations	*29,512*	*28,708*	*22,856*	*22,856*	*22,824*	*22,824*
Panel D: South and sub-sample of inspected schools						
Male	-0.214***	-0.152***	-0.194***	-0.192***	-0.193***	-0.194***
	(0.019)	(0.018)	(0.019)	(0.018)	(0.019)	(0.019)
Non-blind score	-0.011	-0.013	-0.007	-0.007	-0.007	-0.007
	(0.021)	(0.022)	(0.023)	(0.023)	(0.023)	(0.023)
Interaction	-0.126***	-0.129***	-0.117***	-0.117***	-0.118***	-0.118***
	(0.016)	(0.016)	(0.017)	(0.017)	(0.017)	(0.017)
Observations	*31,064*	*30,020*	*24,178*	*24,178*	*24,128*	*24,128*

Table 9: Robustness checks 1: Mathematics

Dependent Variable: Test results in Math (blind and non-blind)						
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: High achieving students (self-assessed)						
Male	0.097***	0.119***	0.128***	0.132***	0.132***	0.133***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)
Non-blind score	0.261***	0.258***	0.267***	0.267***	0.267***	0.267***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Interaction	-0.263***	-0.262***	-0.265***	-0.265***	-0.265***	-0.265***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Observations	*244,350*	*233,552*	*181,242*	*176,658*	*176,416*	*176,416*
Panel B: Low achieving students (self-assessed)						
Male	0.037***	0.071***	0.067***	0.072***	0.042***	0.042***
	(0.011)	(0.012)	(0.013)	(0.014)	(0.013)	(0.013)
Non-blind score	-0.160***	-0.156***	-0.129***	-0.127***	-0.127***	-0.127***
	(0.008)	(0.008)	(0.009)	(0.010)	(0.010)	(0.010)
Interaction	-0.125***	-0.128***	-0.134***	-0.136***	-0.135***	-0.135***
	(0.011)	(0.011)	(0.014)	(0.014)	(0.014)	(0.014)
Observations	*45,728*	*42,554*	*29,998*	*29,200*	*29,130*	*29,130*
Panel C: sub-sample of inspected schools						
Male	0.142***	0.175***	0.169***	0.173***	0.086***	0.085***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Non-blind score	0.122***	0.125***	0.143***	0.143***	0.143***	0.143***
	(0.011)	(0.011)	(0.012)	(0.012)	(0.012)	(0.012)
Interaction	-0.244***	-0.245***	-0.249***	-0.249***	-0.249***	-0.249***
	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)
Observations	77,708	75,234	59,558	59,558	59,464	59,464
Panel D: Class Fixed Effects						
Male	0.131***	0.163***	0.167***	0.167***	0.088***	0.088***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
Non-blind score	0.112***	0.113***	0.130***	0.130***	0.130***	0.130***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Interaction	-0.220***	-0.221***	-0.224***	-0.224***	-0.224***	-0.224***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Observations	997,648	934,298	627,874	627,874	625,418	625,418

Table 10: Robustness checks 2: Mathematics

Dependent Variable: Test results in Math (blind and Non-blind)						
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Northern regions						
Male	0.132***	0.165***	0.171***	0.176***	0.080***	0.080***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Non-blind score	0.171***	0.171***	0.180***	0.180***	0.180***	0.180***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Interaction	-0.214***	-0.216***	-0.217***	-0.217***	-0.217***	-0.217***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	*428,176*	*405,330*	*312,018*	*305,218*	*304,234*	*304,234*
Panel B: Southern regions						
Male	0.105***	0.157***	0.143***	0.149***	0.095***	0.094***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.005)	(0.005)
Non-blind score	0.075***	0.077***	0.095***	0.096***	0.096***	0.096***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Interaction	-0.229***	-0.231***	-0.235***	-0.236***	-0.236***	-0.236***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Observations	*392,322*	*363,858*	*275,138*	*267,764*	*266,522*	*266,522*
Panel C: North - sub-sample of inspected schools						
Male	0.121***	0.152***	0.163***	0.171***	0.065***	0.064***
	(0.017)	(0.017)	(0.017)	(0.018)	(0.017)	(0.017)
Non-blind score	0.127***	0.127***	0.141***	0.141***	0.141***	0.141***
	(0.017)	(0.017)	(0.018)	(0.018)	(0.018)	(0.018)
Interaction	-0.223***	-0.225***	-0.227***	-0.227***	-0.227***	-0.227***
	(0.014)	(0.014)	(0.016)	(0.016)	(0.016)	(0.016)
Observations	*29,512*	*28,708*	*22,856*	*22,856*	*22,824*	*22,824*
Panel D: South - sub-sample of inspected schools						
Male	0.145***	0.196***	0.164***	0.170***	0.104***	0.103***
	(0.017)	(0.017)	(0.018)	(0.018)	(0.017)	(0.017)
Non-blind score	0.163***	0.166***	0.180***	0.180***	0.180***	0.180***
	(0.020)	(0.020)	(0.022)	(0.022)	(0.022)	(0.022)
Interaction	-0.280***	-0.280***	-0.280***	-0.280***	-0.279***	-0.279***
	(0.015)	(0.015)	(0.017)	(0.017)	(0.017)	(0.017)
Observations	*31,064*	*30,020*	*24,178*	*24,178*	*24,128*	*24,128*