Discriminating Behavior: Evidence from teachers' grading bias

Ferman, Bruno and Fontes, Luiz Felipe

Sao Paulo School of Economics - FGV, Sao Paulo School of Economics - FGV

14 May 2020

Discriminating Behavior: Evidence from teachers' grading bias

Bruno Ferman Luiz Felipe Fontes

(FGV-EESP)

May 14, 2020

Abstract

Recent evidence has established that non-cognitive skills are key determinants of education and labor outcomes, and are malleable throughout adolescence. However, little is known about the mechanisms producing these results. This paper tests a channel that could explain part of the association between non-cognitive skills and important outcomes: teacher grading discrimination toward student behaviors. Evidence is drawn from a unique data pertaining to students from middle and high-school in Brazilian private schools. Our empirical strategy is based on the contrasting of school-level tests graded by teachers and school-level tests that cover the same content but are graded blindly. Using detailed data on student classroom behaviors and holding constant performance in exams graded blindly, evidence indicates that teachers inflate the grades of better-behaved students while deducting points from worse-behaved ones. These biases are driven by grading discrimination in exams with open questions. Additionally, teachers' behavior does not appear to be consistent with statistical discrimination.

1 Introduction

Researchers have emphasized that socially productive skills include not only traditionally studied cognitive abilities, but also behavioral and socio-emotional factors such as perseverance, self-control, academic behaviors, and prosociality. In recent years, numerous studies have documented the central role played by these noncognitive skills in shaping educational attainment and adult outcomes (Segal, 2013; Heckman et al., 2006; Papageorge et al., 2019; Deming, 2017; Kautz and Zanoni, 2014).[1] Importantly from a policy standpoint, there is also ample evidence suggesting that these skills are malleable, and can be influenced by school and teacher quality, home environment, and educational interventions (Jackson, 2018; Bertrand and Pan, 2013; Heckman et al., 2013; Jackson et al., 2020; Alan et al., 2019). However, despite the importance of non-cognitive skills, and significant advances in understanding its causes and consequences, there is still limited empirical evidence on how they affect important outcomes.

In the present study, we propose grading discrimination as a potential mediator for gaps in attainment between students with different non-cognitive skills.[2] We examine its prevalence by testing whether classroom behaviors affect teacher grading.[3] The paper employs a unique administrative data from an educational company that manages more than one hundred private schools in Brazil. We use teachers' reports on their students' behavior to construct measures of good and bad in-class behaviors. Our empirical strategy is based on the contrasting of teacher-assigned and blindly-assigned scores from school-level tests that cover the same content. To deal with the incidence of measurement error on the blind test scores used as regressors, we use lagged scores as an instrument for the current ones. We show that if the exogeneity condition of the instrument does not hold, our discrimination parameters of interest are bounded by OLS and IV estimators under a few additional assumptions. We find evidence that teachers discriminate students with good behavior positively and discriminate students with bad behavior negatively. Between 20 and 30 per cent of the correlation between behaviors and teacher-assigned math scores seem to be explained by grading biases. We also find similar results for Portuguese and essay.

Our results are largely driven by grading discrimination in written exams, where teachers may exert discretion by assign partial credits for each question. Using blindly-assigned essay scores, we find no evidence supporting that our results are explained by potential biases from

[1] For surveys, see Almlund et al. (2011), Farrington et al. (2012), Heckman and Kautz (2012), and Heckman et al. (2019).

[2] By discrimination, we mean the unequal assessment of students on the basis of their in-class behaviors rather than their performance in examinations. On the one hand, teachers may statistically discriminate pupil grades by using in-class behaviors to evaluate the unobserved student scholastic aptitude, especially if signals of scholastic aptitude are hard to measure (Arrow, 1972; Phelps, 1972). On the other hand, teachers may have particular tastes for pupils with specific classroom behaviors, and this could also lie behind any grading bias towards student behavior (Becker, 2010).

[3] As previous papers, we proxy for students' non-cognitive skill using academic behaviors. Some studies have used classroom behaviors based on teacher reports like we do (Segal, 2013; Papageorge et al., 2019; Heckman et al., 2013). Other studies have used grades, disciplinary infractions, absences, and grade progression (Jackson, 2018; Kautz and Zanoni, 2014). Academic behaviors have also been associated with traditional non-cognitive skills such as patience (Alan and Ertac, 2018; Castillo et al., 2011; Sutter et al., 2013), self-control (Duckworth and Seligman, 2005), and conscientiousness (Segal, 2008). See Heckman et al. (2019) for further discussion.

math teachers toward student writing skills. Additionally, teachers' grading behavior does not appear to be consistent with statistical discrimination. First, grading biases are constant throughout the year, which is not compatible with a learning model of statistical discrimination. Second, limiting the sample to classrooms where badly-behaved students are as skilled as the well-behaved ones leaves the results unchanged. Overall, this paper shows that grading discrimination toward classroom behaviors does exist and mediates a significant share of the association between these non-cognitive skills and test scores. As suggested by Farrington et al. (2012), other non-cognitive factors affect performance through academic behaviors. Therefore, our results may be capturing discrimination toward a larger set of non-cognitive skills. What is more, as grading manipulation can have several consequences,[4] our results may also indicate that inaccurate grades can explain part of the relation between non-cognitive skills and other important life outcomes.

Our findings contribute to the literature highlighting the importance of non-cognitive skills. As already mentioned, little is known about the mechanisms behind the association between non-cognitive skills and educational and labor outcomes. Researches often speculate that non-cognitive skills produce good study habits which result in better life outcomes.[5] Evidence on that channel is mixed, depending on the skill being evaluated. While Lavecchia et al. (2016) show that impatient students report spending less time doing homework, other papers find that impatient students do not have lower study effort, even though they have much worse test scores. (De Paola and Gioia, 2017; Non and Tempelaar, 2016). Outside the economics literature, a few papers show that personality traits are correlated with study habits (Lubbers et al., 2010; Credé and Kuncel, 2008). A related explanation is that non-cognitive skills may affect the effort put during tasks to obtain good results. Evidence by Borghans et al. (2008) support this explanation. The authors show that individuals with high non-cognitive skills operate a low-stakes cognitive test at a high level, even without rewards (see Segal (2012) for a related result). Similarly, Cubel et al. (2016) show that personality traits predict performance in an experimental task that requires real effort. The authors argue that this finding suggests that at least part of the effect of personality on labor market outcomes operates through productivity. Overall, the few papers proposing mediators for the association between non-cognitive skills and schooling and labor market outcomes use small samples, analyze experimental outcomes, and are based on partial correlations. In our study, we analyze a significant number of students and make use of a quasi-experimental research design to estimate a different mechanism – discrimination in

[4]Recent papers show that grading biases have far-reaching consequences for the students, affecting their future performance in test scores, high-school graduation rates, college initiation rates, chosen field of education, and earnings (Lavy and Sand, 2018; Terrier, 2016; Dee et al., 2019; Nordin et al., 2019; Diamond and Persson, 2016).

[5]Segal (2013) theorizes a similar channel. Her empirical results provide evidence that childhood misbehavior is negatively correlated with educational attainment and labor market outcomes. Based on the results of Castillo et al. (2011) – which find that pupils with higher discount rates have more behavioral problems in school – she develops a model to interpret the mechanisms driving her results. In her model, individuals are endowed with both cognitive and non-cognitive human capital. They can enhance their cognitive human capital at each level of schooling by exerting costly effort. Those who value the future less (i.e., those with low non-cognitive skills) invest less effort in school and hence accumulate less cognitive human capital; as a result, they earn less.

grading – behind the relation between non-cognitive skills and an educational outcome. In particular, we evaluate performance in school-level exams, which is a non-experimental outcome that can have several consequences for the future of students.

This paper is closely related to the recent literature on teacher discrimination in grading. Some previous papers compare non-blindly graded exams and blindly graded exams across minority and non-minority students (Botelho et al., 2015; Burgess and Greaves, 2013; Hanna and Linden, 2012) and genders (Lavy, 2008; Hinnerich et al., 2011; Falch and Naper, 2013; Cornwell et al., 2013; Breda and Ly, 2015), and establish that grading discrimination exists in those dimensions. Besides ethnic and gender indicators, classroom behavior is another relevant student characteristic available to teachers during in-class interactions that may impact their judgment when grading.[6] We contribute to the literature by testing this hypothesis in detail, using a unique dataset containing information on objective indicators of student in-class behaviors. Previous papers recognize that classroom behavior may be one of the most critical cofounders in grading discrimination estimates. Hence, a few of them try to adjust for proxies of behavior.[7] Similar to those papers, we also find grading discrimination toward ethnicity and gender. However, in our setting, there is a small correlation between these characteristics and more objective measures of in-class behaviors. In this scenario, we were able to estimate discrimination effects toward behaviors, without ethnicity and gender being relevant confounders.

Despite being one of the first studies to examine grading biases toward pupil behaviors in detail, we are well aware that the question of whether teachers factor the behavior of students into grades is not new to the education and psychology literature. Researchers working with classroom assessments have been warning about the unreliability of grades and its consequences for a long time (e.g., Starch and Elliott (1912)). There is widespread agreement among measurement specialists that grades should be based exclusively on measures of current achievement. Still, grading practice studies show that a significant share of teachers report that their grades also reflect non-achievement factors such as behavior (McMillan, 2001; McMillan et al., 2002; McMillan, 2003; Frary et al., 1993; Cizek et al., 1995).[8] As far as we know, quantitative researchers within the education literature have not studied grading biases toward student behaviors. Similar to the economics literature, most of them focus on issues related to gender and race (Wen, 1979; Piché et al., 1977; Roen, 1992). We contribute by presenting such a quantitative analysis. Additionally, as we explore mechanisms behind the teachers' grading

[6] Mechtenberg (2009) refers to the behaviors as attitudes, which include habits, styles, and personality traits of the students that teachers may like/dislike.

[7] Lavy (2008) adjusts for past grades under the hypothesis that those should be correlated with students' past behavior in the classroom, which should be correlated with the students' current behavior. Botelho et al. (2015) also use previous grades as well as several other variables. Among them, physical education grades, attendance records, and the perception of parents regarding their children's engagement at school. Cornwell et al. (2013) controls for "attitudes toward learning", which are based on teacher reports. Alesina et al. (2018) use a subjective behavioral grade decided jointly by all the teachers. Terrier (2016) uses a variable of disruptive behavior that equals one if the student received a disciplinary warning from the class council or if he/she was temporarily excluded from the school by the school head because of violent behavior.

[8] Looking for remedies for the mixed signal sent by grades, specialists proposed a reform known as standards-based grading, which has been gaining momentum in U.S. schools for the past 20 years (see McMillan (2013) and references therein).

behavior, our results also contribute to researchers designing and evaluating technologies to reduce grading biases (e.g., Jae and Cowling (2009)).

Research in psychology has studied the effects of non-cognitive skills on test scores extensively. Several studies from the seventies and eighties show that student temperament in the classroom strongly predicts teacher-assigned grades. In a survey of these papers, Keogh (1986) concludes that teacher perceptions of student temperament in the classroom may influence their evaluations of pupil performance.[9] Similarly, new researches in psychology have evaluated the predictive power of student personality skills on grades and standardized achievement test (SAT) scores (for surveys, see Almlund et al. (2011) and Duckworth and Allred (2012)). An important finding from this literature is that, among the Big Five, conscientiousness – traditionally associated with student classroom behaviors – is the most predictive skill of both course grades and SAT scores. However, a few papers show that some facets of conscientiousness seem to be better predictors of course grades than of achievement scores. Following our previous discussion, researchers argue that this may be the result of those abilities inducing more positive study habits, which translate into higher course grades. As achievement tests require the students to solve relatively novel problems, its scores may not reflect study habits as much as test scores.[10] It is also speculated that some skills may help students to behave positively in the classroom, which could be directly factored into report card grades by teachers. Our study adds to the literature by providing evidence of this channel using an appropriate design. But, instead of evaluating SAT scores, we use blindly-assigned scores from school-level examinations that are high-stakes, and contrast them with teacher-assigned grades from examinations that cover the same content of the blind ones. In our setting, studying habits should impact both types of tests similarly.

This article is organized as follows. Section 2 describes our data and presents our behavior measures. Section 3 presents our empirical strategy. The main results are presented at Section 4. Section 5 presents robustness analyses. Section 6 studies statistical discrimination as a potential mechanism behind our results. Section 7 concludes.

[9]She interprets the results through the "Goodness of Fit" model, proposed by Thomas and Chess (1977). Their model describes how children whose attributes meet or exceed contextual presses (e.g. the demands or expectations of teachers) evoke stimulation from the context which more readily promotes adaptive functioning than is the case with the stimulation evoked by children whose characteristics did not fit the demands of their contexts. This concept can be applied in many concepts, but concerning student classroom behaviors it has been developed by Lerner and his colleagues (see, e.g., Lerner et al. (1985)). The general idea is that teachers expect from their students behavioral traits that they believe are consistent with the school environment. Students who come closest to the "optimum" demand evoke the most positive responses from teachers. As a result, classroom behaviors (as well as other types of socio-emotional skills) have consequences for achievement and perhaps especially for teachers' evaluations.

[10]Related to this channel, Borghans et al. (2016) find that IQ is a better predictor of SAT scores than of course grades.

2 Data

2.1 Background and Data

We employ administrative data from a Brazilian private education company. The company manages more than one hundred private schools located in the South, Southeast, and Center-West of Brazil. Enrollment corresponded to more than eighty thousand primary, middle, and high-school pupils in 2018. To examine teacher assessment biases, we take advantage of administrative dataset that contains students' scores on tests graded blindly and non-blindly.

Schools operate a two-term school year (first and second semester). Each term is divided into three cycles. Students perform, per each cycle, a test graded blindly and another graded by their teachers. Both types of tests are high-stakes and do factor into the pupils' end-of-term average score. Teacher-assessed grades are worth 50%, while the blind scores are worth 40%. Students also receive a subjective behavioral grade from each teacher, which factors 10% into the end-of-term average score.

Each subject has a specific teacher-graded examination. Most of the exams rely heavily on questions that require written answers. The exceptions are the third exams from the first and second semesters, which are only multiple-choice. When correcting open items, teachers have considerable arbitrariness to assign grades, mainly because of partial scores. Since teachers have permanent contact with the pupils they teach, these grades could potentially be biased by teacher stereotypes. Students from middle and high-school obtain the blind scores when they complete exams that test knowledge on four topics: mathematics, language (English and Portuguese), science (physics, biology, and chemistry), and humanities (geography, history, sociology and philosophy). All the questions from the blind exams are multiple-choice and corrected by a machine. Hence, the blind scores can be assumed to be free of any bias caused by stereotypes from examiners. Additionally, in most of the schools, students perform essays that are graded blindly by an external team. However, these essays are high-stakes for only a small sample of schools.

Both the blind and the non-blind examinations are created by the schools' pedagogical team, based on a bank of questions. For each cycle, they are designed to measure the same content. Both tests are also taken under the same conditions: they take place in the students' classroom and are supervised by inspectors, which are also responsible for giving general instructions. The non-blind exam is scheduled usually 5 to 30 days after the blind exams. However, neither teachers nor students know the current blind score before the non-blind test. The major difference between the tests is that the blind exams usually cover different subjects, while the non-blind exam is specific to each subject. The exceptions are math and essay. We focus mostly on math scores as most of the blindly-graded essays are not high-stakes.[11]

Besides evaluating student behaviors using subjective grades, teachers can report their

[11] We also test for teacher biases in the Portuguese and essay non-blind scores. We do not test for teacher biases in the other subjects since the blind examinations of humanities and science cover very different subjects.

pupils' classroom behavior at any specific class using a platform developed for this purpose, available in the schools' online system. Teachers must mark at least one of the following options when assessing their students' behavior: "dedication", "good interaction with classmates", "participation during the class", "excessive talking", "cellphone use", "disinterest during the class", or "did not complete the required tasks". All teachers are informed that they should use the platform regularly, although no sanctions are imposed on those who do not.

The dataset used in this study pertains to the school year 2018 and contains all the blind and non-blind scores of the students from middle school (grades 6-9) and the first two years of high-school (grades 10-11).[12] We select these students as before grade 6, pupils do not perform blindly-grade exams. Also, in the last year of high-school (grade 12), students do not perform teacher-assigned scores. Our working dataset is obtained after imposing restriction on the availability for each student of at least 4 (out of 6) blind and non-blind test scores. We also have access to the dataset coming from the schools' online system, which contains all the behavior assessments teachers made in 2018. We discuss next how we use these reports to construct behavior measures. Finally, our data also contain two major student characteristics: ethnicity and gender.

2.2 Behavior Measures

In order to estimate both a potential discrimination against badly behaved students, as well as a potential favoritism toward the best behaved ones, we propose and compute two behavior measures. To do so, we start classifying the behavior reports into good and bad assessments. In particular, "dedication", "good interaction with classmates", and "participation during the class" are classified as an assessment of good behavior. Now, "excessive talking", "cellphone use", "disinterest during the class", and "did not complete the required tasks" are classified as an assessment of bad behavior.

Based on this classification, we construct measures of good and bad classroom behavior defined on the interval $[0, 1]$ for each student and each subject. The measure of good (bad) behavior weights the number of good (bad) assessments a student received from his/her teacher of a specific subject by the maximum number of good (bad) evaluations received by a classmate from that same teacher. These measures are formally defined as follows. Let $\mathcal{I}$ denote the set of all students. For any $i \in \mathcal{I}$, define $\mathcal{C}(i) \subset \mathcal{I}$ as the set of students in the same classroom of pupil i, including himself/herself. Let b_{is} and g_{is} denote the number of bad and good behavior reports received by i from a subject s teacher. The good and the bad behavior measures are defined, respectively, as:

$$GB_{is} := \frac{g_{is}}{\max\{g_{js} : j \in \mathcal{C}(i)\}},$$

[12] In Brazil, children enter the first grade the year they turn six. So, our sample is composed mainly of students between 11 and 16 years old.

and

$$BB_{is} := \frac{b_{is}}{\max\{b_{js} : j \in \mathcal{C}(i)\}}.$$

Notice that the good (bad) behavior measure from subject s is not defined for pupils in classrooms where $g_{is} = 0$ ($b_{is} = 0$) for all $i \in \mathcal{C}(i)$. These classrooms are discarded from our final sample. However, the GB (BB) measure is well defined for pupils with no good (bad) behavior assessments in a specific subject provided they belong to classrooms where at least one of their classmates received such an assessment. In that case, $GB_{is} = 0$ ($BB_{is} = 0$). These students can be understood as "neutral" with respect to the respective behavior measure. To reduce the number of missings and neutral students we also use measures that are not a function of teacher s. These are based on the behavior assessments made by all the teachers. In robustness checks, we will also test alternative ways of using the behavior reports.

2.3 Descriptive Statistics

Table 1 reports the summary statistics for our sample. The data cover 14,777 students from grades 6-11 in 513 classrooms and 57 schools. At least 63% of the sample is white, 17% is *Pardo*, 3% is black, and 0.6% is yellow or indigenous. A large share of parents (16%) did not provide their children's ethnicity. The gender split is roughly even. In 94% of the classrooms, there are students with at least one behavior assessment. This share is lower if we consider only reports made by math teachers: 72%. Figure 1 plots the empirical CDF of the behavior measures for both cases. The measures computed using only the assessments from math teachers – panels (a) and (b) – assume value zero for a large share of students ($35\% - 40\%$). The share of neutral students reduces to $8\% - 20\%$ if we consider the behavior assessments from all the teachers – panels (c) and (d). In both cases students receive more good than bad behavior assessments. As reported in Table 1 students received, on average, 16 assessments of good behavior and 9 of bad behavior. Of these, 4 and 3 comes from math teachers, respectively.

Figure 2 displays the performance gap in teacher-assigned scores – converted into z-scores – between students with different behavior skills. In order to compare different grade distributions according to the student behaviors, we created binary variables that indicate whether students are in the top quartile of the bad and good behavior measures' distribution – $BB(Pct.75)$ and $GB(Pct.75)$, respectively. Students with $BB(Pct.75) = 0$ and $GB(Pct.75) = 1$ strongly outperform those with $BB(Pct.75) = 1$ and $GB(Pct.75) = 0$, respectively. Our goal from here consists in estimating whether part of this association can be explained by grading discrimination. As will be furthered discussed, we will use the blind scores as the counterfactual grades that students would receive if there were no grading discrimination. Figure 3 plots the blind and non-blind math scores after we took their average across the six examinations to reduce the effect of test scores measurement error. We can see visually that both grades are extremely correlated. This is confirmed after we fit the data points with a linear regression model and obtain an estimated slope of 0.81 and an R-squared of 59%.[13] This descriptive evidence supports

[13]When we pool the six examinations and correct test score measurement error using past scores as instrument

our research design that relies on the similarity between the blind and non-blind exams.

3 Empirical Strategy

We are interested in estimating a parameter of grading discrimination toward classroom behaviors, defined as the effect of in-class behaviors on test scores, conditional on the student proficiency in the subject and other characteristics that teachers may be biased at. To motivate our estimable equation, we assume a simple and intuitive statistical model for how test scores are defined. The non-blind scores of student i in exam j and subject s are determined by the following function:

$$S_{ijs}^{NB} = P_{ijs} + v_{ijs} + \Delta(W_{ijs}),$$

where P_{ijs} is student's proficiency. This component reflects factors such as i's knowledge in the subject s required by the exam j and his/her test-taking ability. The term v_{ijs} represents idiosyncratic factors, such as luck or how the student was feeling on a particular day, that are equal to zero in expectation. The term $\Delta(W_{ijs})$ represents potential bias by exam graders, who manipulate test scores based on student i's characteristics contained in W_{ijs}. In particular, we let

$$\Delta(W_{ijs}) = \beta' B_{is} + \phi' X_{ij} + \xi_{ijs},$$

where $B_{is} \coloneqq (GB_{is}, BB_{is})$ is a vector of student i's classroom behavior in subject s–; X_{ij} includes ethnicity indicators (Black, Indigenous, *Pardo*, Yellow, and White), gender, and the past performance of student i in blind examinations;[14] and ξ_{ijs} include i's characteristics unobserved by the econometrician that teachers observe and may discriminate against.

We refer to P_{ijs} as the test score that i would receive in expectation if there was no grading bias. However, we only observe a noisy signal from it, coming from scores in examinations that cover the same content of S_{ijs}^{NB}, take place under very similar conditions, but are graded blindly (S_{ijs}^{B}), and, hence, are free of any kind of bias from the graders. We assume that

$$S_{ijs}^{B} = P_{ijs} + e_{ijs},$$

where the error term e_{ijs} may not be necessarily idiosyncratic. As S_{ijs}^{B} could potentially measure different skills, we can decompose

$$e_{ijs} = \tilde{P}_{ijs} - P_{ijs} + u_{ijs},$$

where $\tilde{P}_{ijs}$ is the i's proficiency required in the blindly-graded examinations and u_{ijs} is an idiosyncratic noisy. To make explicit that potential biases could arise if both types of examination

for the current scores, we obtain an even higher slope.

[14] In our main specifications we adjust for the cumulative average performance in the blind examinations from science, humanities, and languages. When not using IV, we also adjust for past performance in math.

were to measure different abilities, we assume a simple relation between P_{ijs} and $\tilde{P}_{ijs}$:

$$P_{ijs} = \delta \tilde{P}_{ijs} + r_{ijs},$$

where r_{ijs} include factors that are required only by the non-blind examinations.

In a final step, as we pool the test scores from different examinations and classrooms, we add to our main specification classroom fixed effects (α_c) and exam fixed effects (π_j). Using previous definitions, we get that:

$$S^{NB}_{ijs} = \delta S^{B}_{ijs} + \beta' B_{is} + \phi' X_{ij} + \alpha_c + \pi_j + \varepsilon_{ijs}, \tag{1}$$

where $\varepsilon_{ijs} \coloneqq \xi_{ijs} + r_{ijs} - \delta u_{ijs} + v_{ijs}$. Our parameter of interest is the vector β. It measures the effects of classroom behaviors on teacher-assigned test-scores, conditional on student's proficiency (proxied by S^{B}_{ijs}), other characteristics that teachers may be biased at (X_{ij}), and exploring only within classroom variation (α_c). Its identification requires that we deal with unobserved heterogeneity ($\xi_{ijs} + r_{ijs}$) and measurement error in the blind scores (u_{ijs}).

Regarding unobserved heterogeneity, we claim that, conditional on ethnicity, gender, and past blind scores, if there are competencies not captured by the blind scores, varying systematically within the classroom, they are balanced between students with different in-class behaviors. In particular, we are adjusting for the characteristics the literature has shown teachers may be biased against. As will be shown main point estimates are stable in specifications with and without controls, which may suggest that omitted variable bias is not a major concern if selection on observables is informative about selection on unobservables (e.g., Altonji et al. (2005); Oster (2019)). Moreover, we advocate that $\mathbb{V}(r_{ijs}) \approx 0$. We already discussed some particularities of our design that may provide grounds for the plausibility of the assumption. Overall, we believe there are no apparent systematic differences in the exam-taking environment that could interact with i's characteristics. Both the blindly and the non-blindly graded exams are school-level tests that take place in the regular classes and are supervised by inspectors. Furthermore, both types of exams are high-stakes and designed by the pedagogical sector to cover the same content, based on a bank of questions. Our main concern is that, while the blind exams are only multiple-choice, most of the non-blind exams require written answers. If these questions require abilities not covered by the blind exams, correlated with in-class behaviors, our identification strategy could not be valid. To investigate whether this seems to be a potential concern we make use of a reduced sample of schools where students perform blindly-graded essays. First, we use the scores from these essays as proxies for student writing skills and show that controlling for it leaves the results unchanged. Second, we show that, in this subsample, essay teachers practice grading discrimination. In this case, we are able to present evidence of biases toward behaviors in a setting where both the blind and the non-blind exams have the same format.

To tackle the measurement error problem, we use the lagged blind math score (LS^{B}_{ijs}) as an instrument for the current one, under the commonly made assumption that measurement

errors are not serially correlated (e.g., Bond and Lang (2018)).[15] Within the discrimination in grading literature, this is the same strategy of Botelho et al. (2015).[16] In the context of Value-Added Models, simulations in Lockwood and McCaffrey (2014) suggest that using lagged scores as instruments for the current ones can eliminate the bias in treatment effects estimations originated from measurement error in test scores used as regressors.[17] Another important finding by the authors is that controlling for multiple prior test scores can mitigate the influence of test measurement error. For that reason, especially when we estimate (1) by OLS, X_{ij} include past cumulative performance in all other subjects: language, science, and humanities. One might still be worried about the validity of the exclusion restriction. It might be, for example, that teachers practice statistical discrimination by using students' past performance in blind math exams to reduce noisy about their proficiency. The exclusion restriction may be valid provided we adjust especially for past test scores in other subjects, but also for pupil's ethnicity, gender, classroom behavior, and classroom fixed effects. Otherwise, it is likely that lagged blind scores would be correlated positively with the unobserved skills that determine S_{ijs}^{NB}. Under this scenario, we show in Appendix A that OLS and IV produce upper and lower bounds for β. Due to test scores measurement error, the bias of the OLS estimator of β is bounded away from zero. The intuition is that behaviors measure part of δ through the correlation between behaviors and S_{ijs}^{B} once δ is estimated with attenuation bias. Contrary, if $\mathbb{C}(LS_{ijs}^{B}, \varepsilon_{ijs}) > 0$, δ is overestimated by IV, and hence, β is estimated with attenuation bias.[18] Anyway, we believe the non-validity of the exclusion restriction is not a major concern. We obtain nearly identical results when using more distant lags as instrument. What is more, when we use more than one lag as instrument, we perform over-identification tests, which provide evidence that all instruments are valid. Also, when we estimate β by OLS and consider additional proxies for P_{ijs}, we obtain upper bounds that are very close to the IV estimates. Finally, if we impose the restriction $\delta = 1$ (so that we do not need to deal with the measurement error problem) we obtain very similar results.

Regarding inference, standard errors are robust to heteroskedasticity and calculated with student-level clusters. We also tested for school-level clusters and the standard errors remained

[15] In a study of racial gaps in test scores over time, the authors present evidence suggesting that serial correlation of measurement errors is unlikely to be a significant problem in their data.

[16] In a study of gender discrimination in grading, Terrier (2016) estimates an equation similar to (1), in an additional analysis of her paper. She corrects the test score measurement error using quarter of birth as instrument for the blind scores. A few papers from the literature also use a similar specification to (1) but do correct for the measurement error problem (Cornwell et al., 2013; Alesina et al., 2018). Other studies – most studying gender discrimination – use a difference-in-differences (DID) specification (e.g., Falch and Naper (2013) and Lavy (2008)). Our results are very similar when we follow this strategy. The DID specification can be motivated in our setting when we set $\delta = 1$. In fact, the IV estimation of δ is not far from one.

[17] A drawback is that the IV estimation can lead to imprecise estimates. The authors propose several alternative methods based on the standard error of the test score – returned by Item Response Theory (IRT) – to correct for the measurement error bias. As the school examinations we study are not constructed using IRT, we can not use these methods. However, other papers that also use lagged scores as instrument to correct for the measurement error report that doing so using the standard errors returned by IRT leads to very similar results (Khwaja et al., 2011; Botelho et al., 2015).

[18] $\mathbb{C}(LS_{ijs}^{B}, \varepsilon_{ijs}) > 0$ if the lagged blind scores are correlated positively with the unobserved skills that determines S_{ijs}^{B} – $\mathbb{C}(LS_{ijs}^{NB}, \xi_{ijs} + r_{ijs}) > 0$ – and the measurement errors are not auto-correlated or the serial correlation is lower in comparison to $\mathbb{C}(LS_{ijs}^{NB}, \xi_{ijs} + r_{ijs})$.

nearly identical.[19] Additionally, in all our specifications, the test scores are standardized to a distribution with zero mean and a unit standard deviation. This procedure is applied within subjects to each test separately. To facilitate reading of results, in our main specifications B_{is} stands for binary variables that indicate whether students are in the top quartile of the behavior measures' distribution. We also present the results using continuous behavior measures. To deal with a possible simultaneity concern, we compute behavior measures that uses only reports that precede the teacher-graded examinations. Alternatively, we estimate the biases in the non-blind scores from the second semester, using only the behavior reports from the first semester.

4 Main Results

We begin by examining the association between behaviors and test scores. Table 2, column (1), presents the unconditional OLS estimates. We can see that the average math grades of students with bad in-class behaviors ($BB(Pct.75) = 1$) are 0.26 standard deviation (SD) below those with $BB(Pct.75) = 0$. The unconditional grade gap between students with $GB(Pct.75) = 1$ and $GB(Pct.75) = 0$ is even greater: 0.51 SD in favor of the better-behaved pupils. As several other non-cognitive skills studied by the literature, our behavior measures strongly predict pupil test scores. Of course, this does not imply directly that teachers are practicing grading discrimination. If, for example, students with better in-class behavior are those who prepare more for the tests, we should expect them to obtain higher grades.

Therefore, in column (2), we follow the strategy outlined in Section 3. We control for the blind math scores, our proxy for the grades that students would obtain if there were no grading bias. To tackle the measurement error problem we instrument current blind scores with its lagged values.[20] Under this specification, the behavior effects are significantly reduced, indicating that a share of the competence differences seen by teachers is captured by performance in the blindly-scored tests. This may be explained by a greater acquisition of cognitive skills – captured by blind scores – among the students with higher non-cognitive skills – captured by classroom behaviors. Still, the behavior effects are significant and high in magnitude, indicating that teachers discriminate student behaviors in grading. Our results suggest that the better-behaved students have their grades inflated by 0.11 SD. This amounts to 22% percent of the unconditional gap. Additionally, teachers seem to deduct, on average, 0.08 SD from worse behaved students, which represents 30% percent of the unconditional gap. As grading manipulation may have several consequences for the future of students, these results suggest an important channel whereby non-cognitive skills affect important life outcomes.

Our results remain virtually the same if we control for ethnicity and gender (column (3)). We are relieved that despite the vast literature showing grading biases toward these characteristics, they are not relevant cofounders in our setting. In the Appendix Section C.1

[19] Appendix Table B1 presents this result.

[20] Reflecting the cumulative nature of student performance, past scores are strongly correlated with current ones as it is suggested by the high first-stage F statistic. Additional first-stage summary statistics are available upon request.

we present evidence suggesting discrimination against boys and black students, in line with evidence from the literature. Results are, though, much weaker than the results we found for behavior. In column (4), we control for past scores in other subjects, which may proxy for unobservable competences required only by the non-blind exams. The point estimates remain very similar. Finally, in column (5), we follow the same specification of Botelho et al. (2015), which also correct for the measurement error in language test scores, and consider higher-order polynomials for the blind scores.[21] In particular, a cubic polynomial for the blind math scores, a linear function of the blind language scores, and the interaction between these. Our results remain nearly identical under this more flexible specification. We also analyze the different behaviors separately. Disinterest during class and dedication seem to be the most discriminated behaviors (Appendix Section C.2 discusses these results). Finally, in Appendix Section C.3, we repeat the same exercise made here, but looking at the non-blind Portuguese scores. Results found for this subject are similar to those found for math.

Table 3 presents the results when we tackle the measurement error problem using additional lags of the blind math scores as instrument. The results remain very similar when, instead of using the first lag, we use the second or the third lags (columns (2) and (3)). This gives some support that auto-correlation of measurement errors is not an issue. Otherwise, we should expect some variation in our estimates due to differential correlations between more distant measurement errors and the current ones. In column (4), we use both the first and the second lag as instrument. In column (5), we use all the first three lags. Point estimates do not change under these specifications. What is more, when using additional lags as instrument, we perform overidentification tests. The results from the tests suggest that we have no obvious reason to distrust the validity of the set of instruments employed.

As previously discussed, if serial correlation of the measurement errors is not a major concern, the non-validity of the exclusion restriction would probably leads us to underestimate the behavior effects. We then estimate equation (1) by OLS to obtain upper bounds for the true discrimination parameter of interest β. Table 4 presents the results. In column (1), we replicate the same specification from Table 2, column (2). The first thing to notice is that the relation between blind and non-blind scores estimated by OLS (0.42) is much lower than the estimated by IV (0.97), reflecting the attenuation bias due to test score measurement error. As a consequence, the behavior effects are higher when estimated by OLS. Another major difference is that the inclusion of past scores in column (3) changes significantly the magnitude of our estimates. The reason is that the past blind scores from other subjects serve as proxies for part of the student proficiency signal, which mitigates the biases from the estimated behavior effects. This allows us to estimate finer upper bounds for our parameter of interest. We also tried to reduce the biases even more by using past math scores as control, instead of instrument, but the results remained virtually the same (column (4)). Overall, we highlight that the upper bounds estimated by OLS using additional proxies for student proficiency (-0.10 and 0.17 for the BB

[21] Instrumenting language scores may be especially important if one believes that language skills are required only by the non-blind math scores.

and GB measures, respectively) are close to the IV estimates (−0.07 and 0.11, respectively), and hence underestimation of β does not seem to be a concern. Appendix Table B2 presents the results when we also estimate equation (1) by OLS, but restrict δ to 1. In this case, our dependent variable becomes the difference between blind and non-blind scores. Results are nearly identical those we obtain using IV.

Taken together, our results confirm that grading discrimination toward behaviors does exist in our sample and explain a significant share of the association between these non-cognitive skills and test scores. As already discussed, this may explain part of the effects of non-cognitive skills on education attainment and labor market outcomes. We also provided compelling evidence on the validity of our IV strategy to deal with test measurement errors. In the next sections, we present further robustness checks and study potential channels behind teachers' grading behavior.

5 Robustness

In this section, we conduct additional specification checks and explore data heterogeneity to assess empirically the validity of potential threats to the internal validity of our study.

5.1 Written vs. multiple-choice exams

We start testing whether the difference in exam format – where most of the non-blind tests require written answers and the blind ones are based on multiple-choice questions – poses a threat to the internal validity of our study. We start showing that our results are driven by grading discrimination in written exams, where teachers can exert discretion by assigning partial scores. Appendix Table B3 presents the results. We find evidence that students with bad in-class behaviors do not receive grade deductions in multiple-choice examinations. Indeed, the negative correlation of 0.18 points between bad behavior and math grades (column (1)) vanishes when we adjust for student proficiency.[22] Our findings also suggest that students with better classroom behavior still receive grade credits in objective tests (0.06, s.e 0.02), though in a much lower magnitude compared to exams that require written answers. In these exams, we find that teachers deduct, on average, 0.13 SD of the worse-behaved students and inflate the average grades of the well-behaved ones by 0.14 SD. These findings highlight an important mechanism on how teachers practice grading discrimination. The results are in line with evidence by Hanna and Linden (2012), which find that, in their experiment, graders made an effort to assign students partial credit. Differently from our study, though, they did not observe teachers grading exams that are only multiple-choice.

Biases are expected to be lower in exams that are only multiple-choice as teachers can only assign right or wrong to each item, having little leeway to discriminate grades. Still, one

[22] Appendix Table B4 presents the OLS estimations. In this case, we also find no grading discrimination in multiple-choice exams.

could interpret these results as a threat to the internal validity of our study. It can be that math teachers praise good handwriting and organization when grading questions that require written answers. If these skills correlate with behaviors, our estimates would be biased, and this could explain why the grading biases in the objective teacher-graded tests are lower. To test this hypothesis, we re-estimate our main results using a subsample of schools where students perform blindly-graded essays, and then adjust for the blind essay scores to check whether this seems to be an important confounder. This kind of examination, more than any other, should capture abilities like those mentioned before. Appendix Table B5 presents the results. In summary, we find nearly identical grading biases when we control for the essay scores, even if we correct its measurement error by using lagged scores. Overall, we are confident that our main results are not biased due to writing competencies praised by math teachers in grades.

We also use the blind essay scores to test whether teachers from this subject practice grading discrimination toward student behaviors. In this particular case, we were able to compare exams with exactly the same format. Appendix Table B6 presents the estimated biases in the non-blind essay scores. Unconditional results (column (1)) indicate that students with the worst classroom behavior (as reported by their essay teachers) have disadvantages of 0.34 SD. The well-behaved pupils have advantages of 0.36 SD. In column (2), we adjust for performance in blindly-graded essays. The students with good (bad) in-class behavior receive grade credits (deductions) of 0.12 SD (approximately 35% of the unconditional effects). Columns (3) and (4) present the results when we analyze only the essays that are high-stakes. In this reduced sample, though, we are underpowered to detect smaller effects. Hence, the BB estimate of -0.05 SD is not precisely estimated. Still, we can not reject under traditional levels of significance that this effect is statistically different from the effect we estimate using only non-high-stakes essays (column (5)). The GB estimate remains very similar (0.11, s.e 0.06). Overall, the estimates presented here are quantitatively similar to those obtained when analyzing math scores, even though here, we are comparing exams with exactly the same format, while in the mathematics' case, non-blind exams are mostly subjective and the blind ones are objective. This is further evidence suggesting that our results are not biased due to the potential skills required only by one type of exam.

5.2 Do the estimated biases reflect effect of different timing of exams?

We now examine the possibility that the results mainly reflect the effect of the specific pattern in the timing of the exams where, in most of the cases, the non-blind exam follows the blind exam. This different timing could, in theory, account for some of the gaps in performance between students with different behaviors in the blind and non-blind exams. For example, one could argue that while students with worse classroom behavior tend to rest between the blind and non-blind exams, those better behaved study. In this case, performance in the blind tests may not reflect precisely the students' level of proficiency at the time they take the non-blind

tests, especially if the time gap is large.[23] In our setting, the time difference between the two exams varies across schools, typically being between 5 to 30 days (see Appendix Figure B.1). However, the time difference varies less than ten days only for the first exams, which we can not explore in the IV estimation. In Appendix Table B7, we explore timing heterogeneity using two subsamples where the time difference between the blind and the non-blind exams differ from 10 to 20 days and more than 20 days. The estimated biases are very similar across these two subsamples. Appendix Table B8 presents the OLS estimates. We then explore an additional subsample where the time difference varies up to 9 days. Across these three subsamples, the estimated biases have the same sign and are high in magnitude. These results suggest that the specific pattern in the timing of the blind and non-blind exams is not the cause of the pattern in our main results.

5.3 Behavior Measures

We now test alternative ways of using the behavior reports. Besides considering only the behavior assessments by math teachers, we also test measures that use the assessments made by all teachers. One could argue that these measures depend less on the subjectivity of a teacher's type, and hence capture better the student behaviors.[24] Either way, Appendix Table B9 shows that the point estimates obtained in this case are identical to those obtained in our primary specification. Results also remain similar if we discard reports made by math teachers (Appendix Table B10). The results presented so far are based on a discretization of the behavior measures. One might be worried whether our evidence depends strongly on this transformation. Appendix Table B11 shows this is not the case. The share of the association between behaviors and test scores explained by grading discrimination remains the same when we use the continuous behavior measures (31% and 21% for the BB and GB measures, respectively). Additionally, we obtain similar results when using the overall number of good and bad assessments as regressors (see Appendix Table B12).[25] Finally, we use alternative data to test two other indicators of student behaviors. One of them is the behavioral grades that students receive from each teacher at the end of each semester. Although this is a subjective measure, it is based on mandatory assessments, unlike the behavior reports made by teachers on the school's system. The other measure we use is restricted to a small sample of schools where students take regular courses to improve their socio-emotional abilities. In these courses, students are assessed by teachers based on in-class examinations. We standardize the grades from these exams, and use it as a measure of non-cognitive skills. These last results are discussed in Appendix Section C.4. In summary, they are qualitatively similar to those estimated using the behavior reports.

[23] One could imagine several other mechanisms. Some, however, would not explain our results. For example, if students with worse classroom behavior tend to study for the exam later than well-behaved pupils, perhaps because they have a higher discount rate as suggested by the literature, than they might be better prepared for the second exam than for the first.

[24] In addition, as there are less missing data on behaviors when using the reports made by all teachers, we can show that our results are valid in a larger sample of students.

[25] We actually use the inverse hyperbolic sine of the behavior assessments.

5.4 Simultaneity

One might be worried about potential feedback effects between biased grades and classroom behavior. It might be, for example, that after receiving downward biased grades, students get mad at their teachers and start misbehaving in the classroom. To deal with that concern, we adopt two strategies. First, we estimate the biases in the non-blind math scores from the second semester, using only the behavior reports from the first (Appendix Table B13). The estimates are remarkably similar to our primary results. Second, we create measures that use only behavior reports that preceded the teacher-graded examinations.[26] Appendix Table B14 presents the results. In comparison to the grading discrimination estimated using the behavior measures that use the assessments from the whole year – column (2) – the BB estimate remains the same, while the GB estimate falls slightly – column (4). The conclusions, however, remain the same. Approximately 25% of the unconditional correlation between behaviors and teacher-assigned scores (column (3)) seem to be explained by teacher biases.

6 Potential Mechanisms

In this section, we propose and test empirically potential explanations for the teachers' grading behavior based on possible interpretations of statistical discrimination models.

6.1 Learning across the school-year

We start exploring predictions of models of statistical discrimination and employer learning (Altonji and Pierret, 2001). In the employer-learning model, employers observe workers' performance on the job and thus learn about workers' unobserved productivity. As they learn, they rely less on easily observed workers' characteristics to predict their productivity. The faster employers learn, the shorter the period during which firms statistically discriminate.[27] In our setting, under statistical discrimination, the more teachers observe their students' performance at school, the less they should use classroom behaviors to predict students' proficiency.[28]

Exploring the fact that we observe teachers grading six exams, we test whether estimated biases decay throughout the year. In practice, we compare the estimated biases in the first, second, and third exams from the second semester with its respective exams from the first semester. If statistical discrimination is at play, and there is some learning, we should expect lower effects in the second semester. When lagged blind scores are used as instrument for the current scores, we lose information from the first exam from the first semester. Then, Appendix

[26] The drawback of using these measures is that we lost many observations, mainly for the first exams, since among the teachers evaluating their pupils' behavior, not all have done so since the beginning of the school-year.

[27] For a test of learning in this context, see Lange (2007). He shows that employers learn quickly. Initial expectation errors decline by 50% within 3 years.

[28] In the context of racial grading discrimination, Botelho et al. (2015) find that while there is a bias toward black students attending classes with a teacher for the first time, no significant disparities are found among those that have already had classroom with that teacher before.

Figure 4 presents the results from the second and third exams only. The point estimates are remarkably stable across the semesters. Figure B.2 presents the results from OLS estimations, where we can also analyze the first exams. The estimates are also very similar across the semesters.[29] These evidence suggest that a model of statistical discrimination with learning does not seem to be explaining our results, perhaps because teachers do not statistically discriminate, or because learning is slow in our setting.

6.2 Biases when badly-behaved students are skilled in math

Statistical discrimination arises if teachers use classroom behavior to predict students' unobserved ability based on beliefs they have on the proficiency of students with certain characteristics. The question of whether such beliefs are based on real evidence or are unfounded is irrelevant to the outcome of statistical discrimination. However, it is plausible to imagine that these beliefs may be influenced by the superior average performance of students with better classroom behavior.

We explore heterogeneity in the performance of students in the blindly-graded examinations to test whether the estimated biases are lower in a subsample of students where those with worse classroom behaviors are as skilled as their classmates with better behaviors. To do so, we select classrooms where in the first semester students with $BB(pct.75) = 1$ or $GB(pct.75) = 0$ performed, on average, better in the blind math exams than their classmates with $BB(pct.75) = 0$ or $GB(pct.75) = 1$, respectively. We call this subsample of sample A. The subsample where the previous conditions are not satisfied is called of sample B. Appendix Figure B.3 shows that in sample A, students with good and bad behaviors are homogeneous in terms of math proficiency captured in the blind math scores from the first semester. In sample B, there is a striking gap between these groups of students.

Figure 5 presents the estimated biases for the full sample, sample A, and sample B. The estimates are similar across the samples, both when we use the behavior reports from the whole year (panel (a)) or when we use only the behavior assessments from the first semester (panel (b)). If anything, both the positive and the negative grading discrimination are higher within sample A, which is the opposite of what one should expect under statistical discrimination. These results may suggest that this form of discrimination is not explaining our results. Otherwise, teachers are not updating their beliefs based on the average performance of the math class they are grading.[30]

[29] Additionally, do observe that there is not much heterogeneity from the first to the second exams. The estimates differ a lot in the third exams as in most of the schools, these exams are only multiple-choice, which refers to our previous discussion.

[30] It could be, for example, that teachers think the grades from the worse-behaved students are less reliable. Maybe, because they believe that badly-behaved pupils are more prone to cheating.

7 Conclusion

Using a quasi-experimental design, we contrast blindly- and non-blindly- graded examinations that cover the same content and find that teachers inflate the grades of the better-behaved students while deducting points from the worse-behaved ones. Between 20 and 30 per cent of the correlation between behaviors and teacher-assigned grades seem to be explained by grading discrimination. As evidence from the literature point out that grading manipulation may have several consequences for the students, these findings suggest that grading discrimination is one mechanism whereby non-cognitive skills affect important life outcomes. Our results are robust to the incidence of test score measurement error, to the way student behaviors are measured, to potential feedback effects between behaviors and biased grades, and to the differences between blind and non-blind examinations. We also find that theses biases are driven by grading discrimination in written exams, where teachers can potentially be more discretionary. Finally, we show that teachers' grading behavior does not seem to be explained by statistical discrimination models.

In terms of policy, our results may be important for those interested in the effects of educational policies mediated by non-cognitive skills. As such interventions can also affect teacher grading practices – directly affecting the importance of non-cognitive skills to test scores – ignoring our results may lead to misleading interpretations of the results from these policies. Our results also shed light on policies aimed at reducing grading biases. We find that biasing grades based on student behaviors may be privately optimal from the perspective of teachers. The rationale of this practice as a policy would be to try to induce more positive school habits, which is essential not only for the students themselves but also for their peers. However, this grading behavior may have pervasive consequences for the students as inaccurate grades affect schooling and labor market outcomes. The reform known as standards-based grading emphasizes, among other things, the explicitly use of separate reports describing student performance and academic behaviors. The use of similar policies may reduce grading discrimination while not encouraging misbehavior. Additionally, our findings on the grading biases in written questions highlight the standardization in the use of partial credits as a potential mechanism to reduce discrimination. Alternatively, new technologies – such as the use of bar codes to assure student anonymity – can improve policy results. Evaluating the causal effects of such policies are left for future work.

References

Alan, S., Boneva, T., and Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics*, 134(3):1121–1162.

Alan, S. and Ertac, S. (2018). Fostering patience in the classroom: Results from randomized educational intervention. *Journal of Political Economy*, 126(5):1865–1911.

Alesina, A., Carlana, M., Ferrara, E. L., and Pinotti, P. (2018). Revealing stereotypes: Evidence from immigrants in schools. Technical report, National Bureau of Economic Research.

Almlund, M., Duckworth, A. L., Heckman, J., and Kautz, T. (2011). Personality psychology and economics. In *Handbook of the Economics of Education*, volume 4, pages 1–181. Elsevier.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political economy*, 113(1):151–184.

Altonji, J. G. and Pierret, C. R. (2001). Employer learning and statistical discrimination. *The Quarterly Journal of Economics*, 116(1):313–350.

Arrow, K. J. (1972). Models of job discrimination. *Racial discrimination in economic life*, 83.

Becker, G. S. (2010). *The economics of discrimination*. University of Chicago press.

Bertrand, M. and Pan, J. (2013). The trouble with boys: Social influences and the gender gap in disruptive behavior. *American Economic Journal: Applied Economics*, 5(1):32–64.

Bond, T. N. and Lang, K. (2018). The black–white education scaled test-score gap in grades k-7. *Journal of Human Resources*, 53(4):891–917.

Borghans, L., Golsteyn, B. H., Heckman, J. J., and Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences*, 113(47):13354–13359.

Borghans, L., Meijers, H., and Ter Weel, B. (2008). The role of noncognitive skills in explaining cognitive test scores. *Economic inquiry*, 46(1):2–12.

Botelho, F., Madeira, R. A., and Rangel, M. A. (2015). Racial discrimination in grading: Evidence from brazil. *American Economic Journal: Applied Economics*, 7(4):37–52.

Breda, T. and Ly, S. T. (2015). Professors in core science fields are not always biased against women: Evidence from france. *American Economic Journal: Applied Economics*, 7(4):53–75.

Burgess, S. and Greaves, E. (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics*, 31(3):535–576.

Castillo, M., Ferraro, P. J., Jordan, J. L., and Petrie, R. (2011). The today and tomorrow of kids: Time preferences and educational outcomes of children. *Journal of Public Economics*, 95(11-12):1377–1385.

Cizek, G. J., Fitzgerald, S. M., and Rachor, R. A. (1995). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational assessment*, 3(2):159–179.

Cornwell, C., Mustard, D. B., and Van Parys, J. (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human resources*, 48(1):236–264.

Credé, M. and Kuncel, N. R. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. *Perspectives on psychological science*, 3(6):425–453.

Cubel, M., Nuevo-Chiquero, A., Sanchez-Pages, S., and Vidal-Fernandez, M. (2016). Do personality traits affect productivity? evidence from the laboratory. *The Economic Journal*, 126(592):654–681.

De Paola, M. and Gioia, F. (2017). Impatience and academic performance. less effort and less ambitious goals. *Journal of Policy Modeling*, 39(3):443–460.

Dee, T. S., Dobbie, W., Jacob, B. A., and Rockoff, J. (2019). The causes and consequences of test score manipulation: Evidence from the new york regents examinations. *American Economic Journal: Applied Economics*, 11(3):382–423.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, 132(4):1593–1640.

Diamond, R. and Persson, P. (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. Technical report, National Bureau of Economic Research.

Duckworth, A. L. and Allred, K. M. (2012). Temperament in the classroom.

Duckworth, A. L. and Seligman, M. E. (2005). Self-discipline outdoes iq in predicting academic performance of adolescents. *Psychological science*, 16(12):939–944.

Falch, T. and Naper, L. R. (2013). Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review*, 36:12–25.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., and Beechum, N. O. (2012). *Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance–A Critical Literature Review.* ERIC.

Frary, R. B., Cross, L. H., and Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice*, 12(3):23–30.

Hanna, R. N. and Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*, 4(4):146–68.

Heckman, J. J., Jagelka, T., and Kautz, T. D. (2019). Some contributions of economics to the study of personality. Technical report, National Bureau of Economic Research.

Heckman, J. J. and Kautz, T. (2012). Hard evidence on soft skills. *Labour economics*, 19(4):451–464.

Heckman, J. J., Pinto, R., and Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, 103(6):2052–86.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics*, 24(3):411–482.

Hinnerich, B. T., Höglin, E., and Johannesson, M. (2011). Are boys discriminated in swedish high schools? *Economics of Education review*, 30(4):682–690.

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, 126(5):2072–2107.

Jackson, C. K., Porter, S. C., Easton, J. Q., Blanchard, A., and Kiguel, S. (2020). School effects on socio-emotional development, school-based arrests, and educational attainment. Technical report, National Bureau of Economic Research.

Jae, H. and Cowling, J. (2009). Objectivity in grading: The promise of bar codes. *College Teaching*, 57(1):51–55.

Kautz, T. and Zanoni, W. (2014). *Measuring and fostering non-cognitive skills in adolescence: Evidence from Chicago Public Schools and the OneGoal Program*. University of Chicago Chicago, IL.

Keogh, B. K. (1986). Temperament and schooling: meaning of" goodness of fit"? *New directions for child development*.

Khwaja, A. I., Andrabi, T., Das, J., Zajonc, T., et al. (2011). Do value-added estimates add value? accounting for learning dynamics. *American Economic Journal: Applied Economics*.

Lange, F. (2007). The speed of employer learning. *Journal of Labor Economics*, 25(1):1–35.

Lavecchia, A. M., Liu, H., and Oreopoulos, P. (2016). Behavioral economics of education: Progress and possibilities. In *Handbook of the Economics of Education*, volume 5, pages 1–74. Elsevier.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of Political Economy*, 92(10-11):2083–2105.

Lavy, V. and Sand, E. (2018). On the origins of gender gaps in human capital: Short-and long-term consequences of teachers' biases. *Journal of Public Economics*, 167:263–279.

Lerner, J. V., Lerner, R. M., and Zabski, S. (1985). Temperament and elementary school children's actual and rated academic performance: A test of a 'goodness-of-fit'model. *Journal of Child Psychology and Psychiatry*, 26(1):125–136.

Lockwood, J. and McCaffrey, D. F. (2014). Correcting for test score measurement error in ancova models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, 39(1):22–52.

Lubbers, M. J., Van Der Werf, M. P., Kuyper, H., and Hendriks, A. J. (2010). Does homework behavior mediate the relation between personality and academic performance? *Learning and Individual differences*, 20(3):203–208.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, 20(1):20–32.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational measurement: Issues and practice*, 22(4):34–43.

McMillan, J. H. (2013). *SAGE handbook of research on classroom assessment*. Sage.

McMillan, J. H., Myran, S., and Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The journal of educational research*, 95(4):203–213.

Mechtenberg, L. (2009). Cheap talk in the classroom: How biased grading at school explains gender differences in achievements, career choices and wages. *The review of economic studies*, 76(4):1431–1459.

Non, A. and Tempelaar, D. (2016). Time preferences, study effort, and academic performance. *Economics of Education Review*, 54:36–61.

Nordin, M., Heckley, G., and Gerdtham, U. (2019). The impact of grade inflation on higher education enrolment and earnings. *Economics of Education Review*, 73:101936.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Papageorge, N. W., Ronda, V., and Zheng, Y. (2019). The economic value of breaking bad: Misbehavior, schooling and the labor market. Technical report, National Bureau of Economic Research.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review*, 62(4):659–661.

Piché, G. L., Michlin, M., Rubin, D., and Sullivan, A. (1977). Effects of dialect-ethnicity, social class and quality of written compositions on teachers' subjective evaluations of children. *Communications Monographs*, 44(1):60–72.

Roen, D. (1992). Gender and teacher response to student writing. In *Gender issues in the teaching of English*, pages 126–141. Heinemann.

Segal, C. (2008). Classroom behavior. *Journal of Human Resources*, 43(4):783–814.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, 58(8):1438–1457.

Segal, C. (2013). Misbehavior, education, and labor market outcomes. *Journal of the European Economic Association*, 11(4):743–779.

Starch, D. and Elliott, E. (1912). Reliability of grading high-school work in english. *The School Review*, 20(7):442–457.

Sutter, M., Kocher, M. G., Glätzle-Rützler, D., and Trautmann, S. T. (2013). Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *American Economic Review*, 103(1):510–31.

Terrier, C. (2016). Boys lag behind: How teachers' gender biases affect student achievement. *IZA Discussion Paper*.

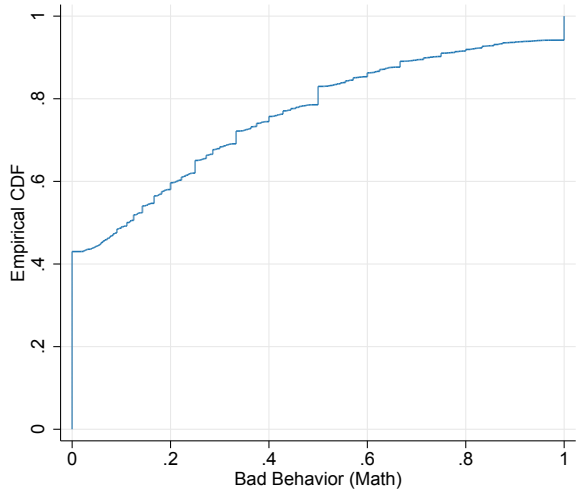
Thomas, A. and Chess, S. (1977). *Temperament and development.* Brunner/Mazel.

Wen, S.-s. (1979). Racial halo on evaluative rating: General or differential? *Contemporary Educational Psychology*, 4(1):15–19.

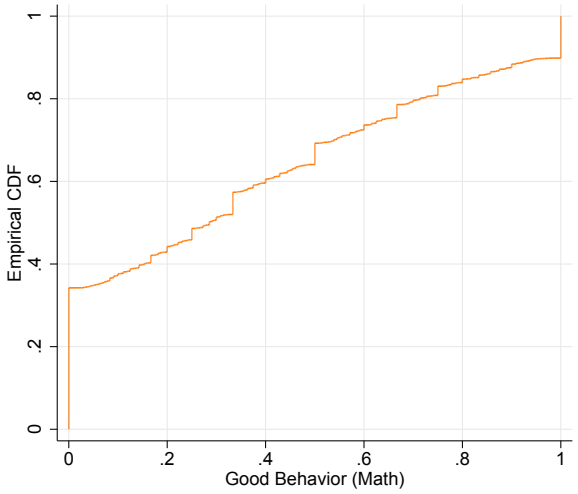
Tables and Figures

Figure (1) Empirical CDF of the Good Behavior (GB) and Bad Behavior (BB) Measures

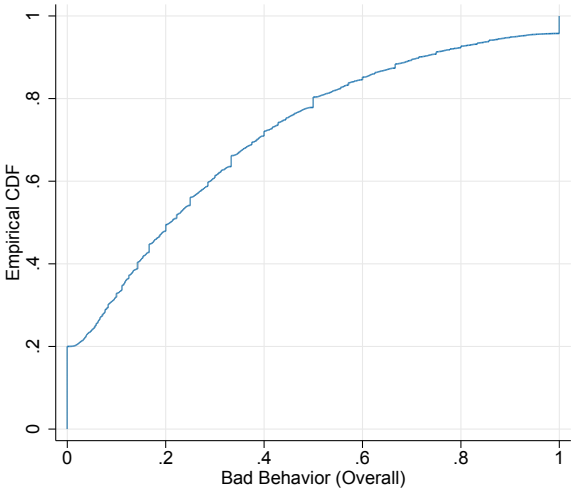
(a) ECDF of the BB measure – math



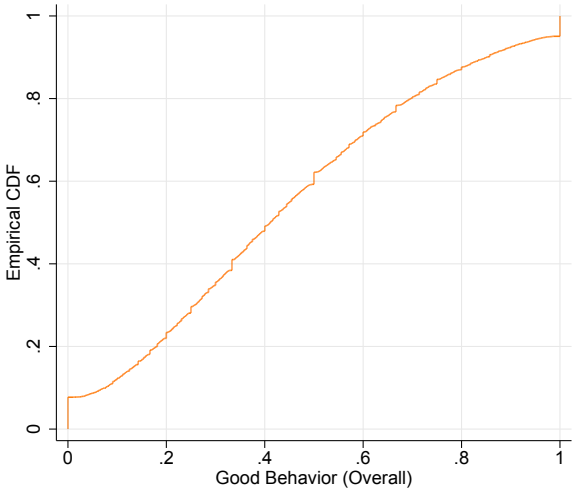
(b) ECDF of the GB measure – math



(c) ECDF of the BB measure – overall



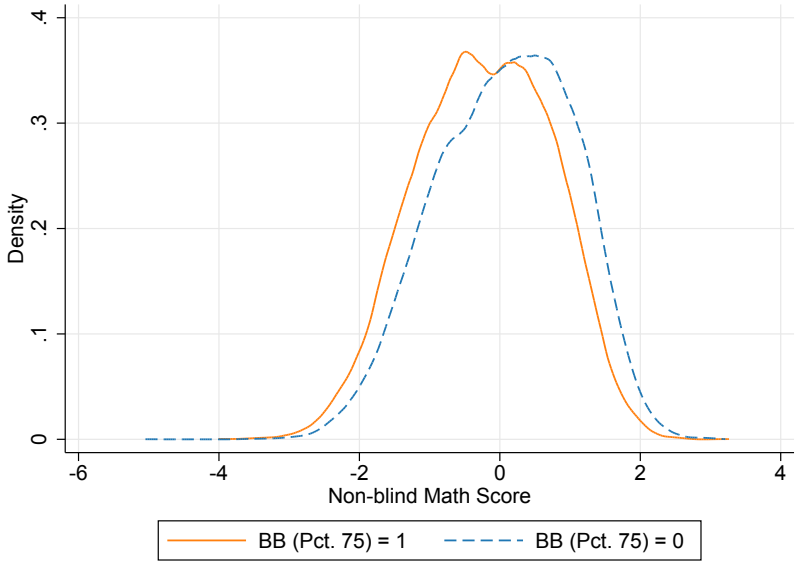
(d) ECDF of the GB measure – overall



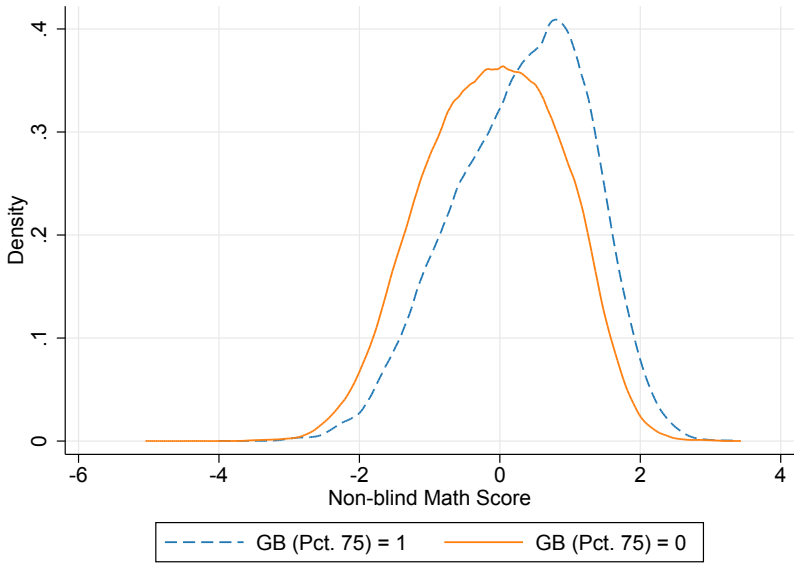
Note: This figure estimates the empirical cumulative distribution functions (ECDF) of the behavior measures. Panels a and b plot the ECDF of the bad and good behavior measures, respectively, computed using only the assessments by math teachers. Panels c and d plot the ECDF of the bad and good behavior measures, respectively, computed using the assessments made by all teachers.

Figure (2) Distribution of Blind and Non-blind Math Scores

(a) Across students with different BB measures

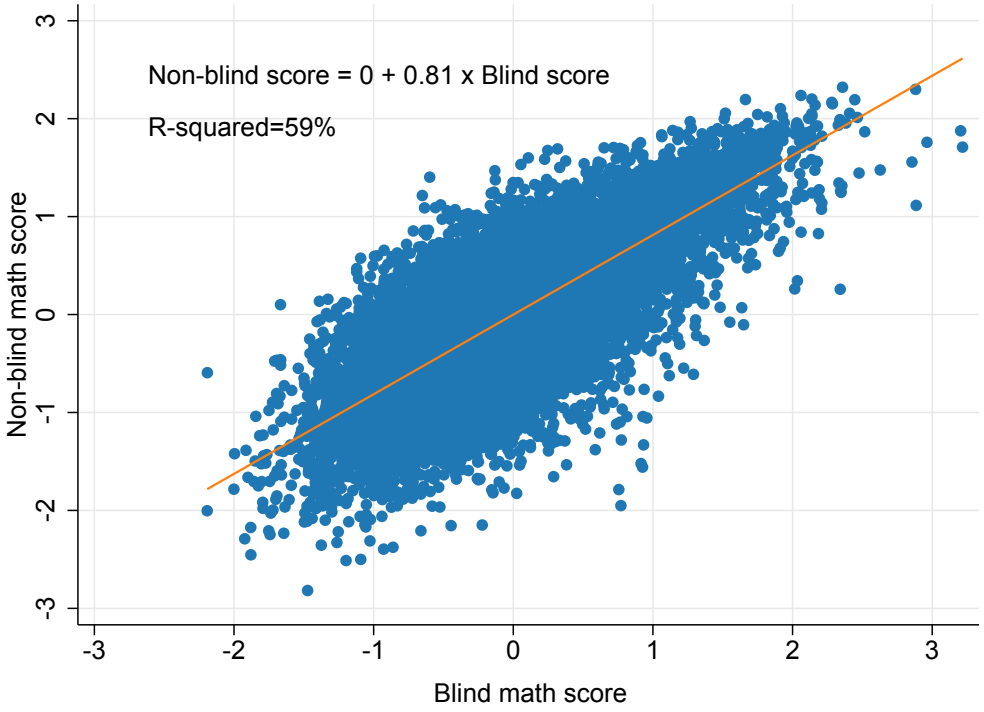


(b) Across students with different GB measures



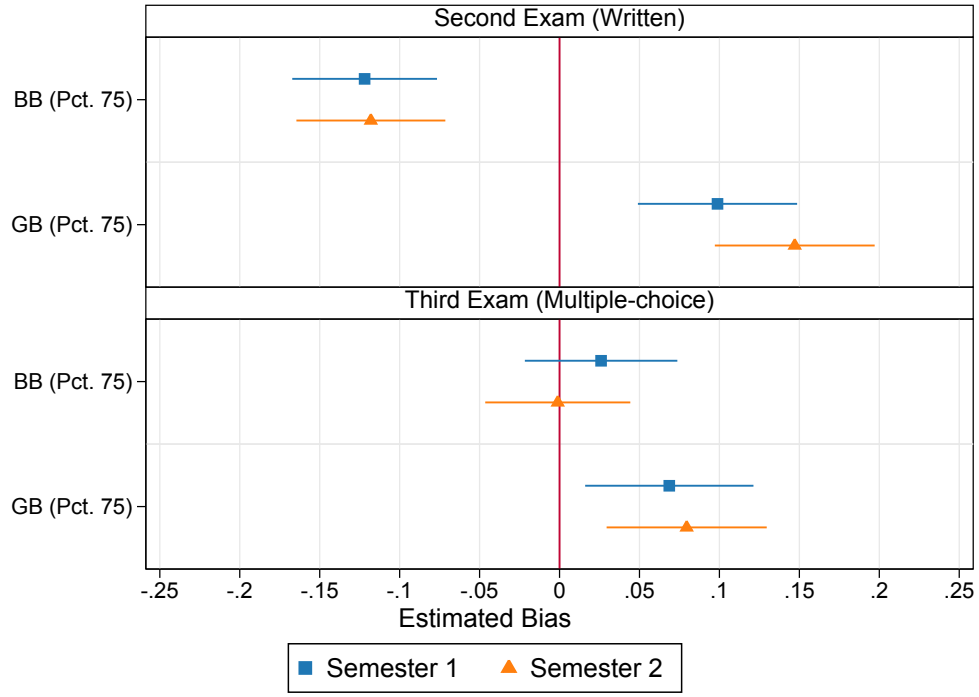
Note: These figures plot the distribution of blind math scores and non-blind math scores. Observations are at the student×exam level. $BB(75pct.)$ and $GB(75pct.)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. In panel (a), solid line represents students with $BB(Pct.75) = 1$ and the dotted line represents those with $BB(Pct.75) = 0$. In panel (b), solid line represents students with $GB(Pct.75) = 0$ and the dotted line represents those with $GB(Pct.75) = 1$. All test scores are standardized (the mean equals zero and the variance equals one).

Figure (3) Association Between the Blind and Non-blind Math Scores



Note: This figure plots the average of the blind and non-blind math scores across six examinations. The line fits the data points by OLS. All test scores are standardized (the mean equals zero and the variance equals one).

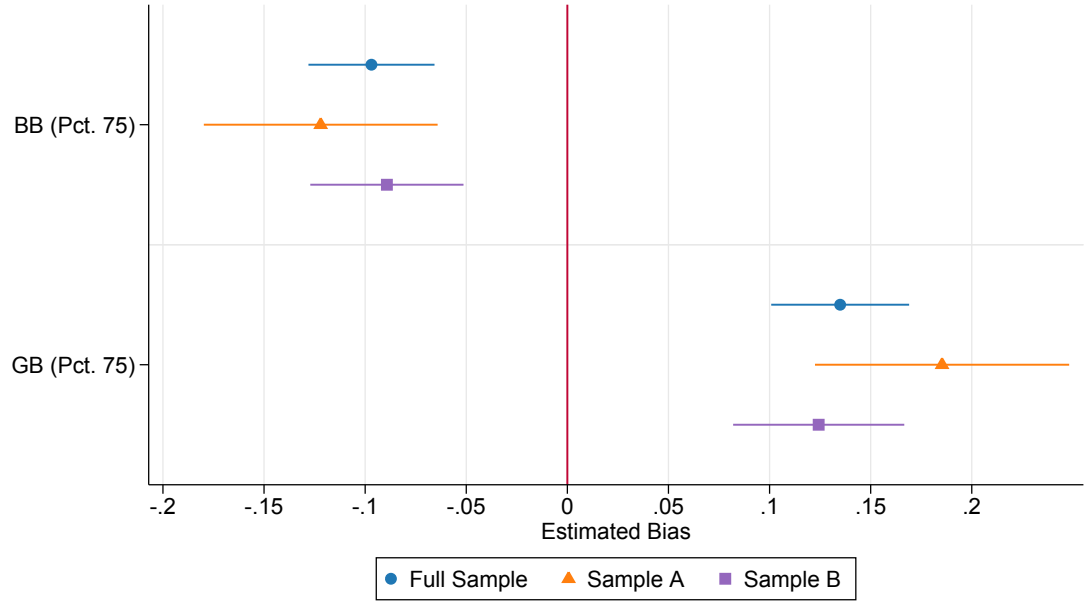
Figure (4) Heterogeneity Across Semesters and Exams - IV Estimation



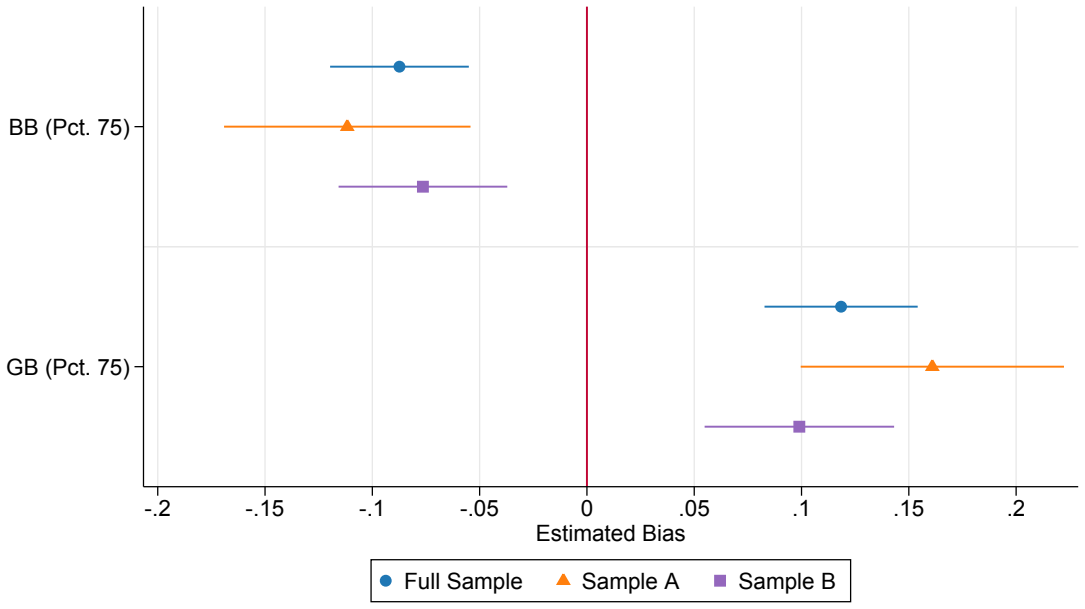
Note: This figure plots 90% confidence intervals and point estimates from student×exam-level OLS regressions of teacher-assigned math scores on classroom behavior, using subsamples that are specific for each exam. $BB(75pct.)$ and $GB(75pct.)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. All specifications follow Table 2, column 3.

Figure (5) Estimated biases in classrooms where students with different in-class behaviors differ (Sample B) and do not differ (Sample A) in their math proficiency

(a) Using All Behavior Reports



(b) Using Only Past Behavior Reports



Note: This figure plots 90% confidence intervals and point estimates from student×exam-level IV regressions of teacher-assigned math scores on classroom behavior, for different samples. Sample A selects classrooms where in the first semester students with $BB(pct.75) = 1$ or $GB(pct.75) = 0$ performed, on average, better in the blind math exams than their classmates with $BB(pct.75) = 0$ or $GB(pct.75) = 1$, respectively. The subsample where the previous conditions are not satisfied is called of sample B. Full sample uses samples A and B. All specifications follow Table 2, column (3).

Table (1) Summary Statistics

Variable	Mean	SD	Observations
Students			14,766
Schools			57
Classes			513
Grades			
Grade 6	14.61%		2,157
Grade 7	15.24%		2,250
Grade 8	17.30%		2,554
Grade 9	10.40%		1,536
Grade 10	22.19%		3,276
Grade 11	20.27%		2,993
Ethnicity and Gender			
Female	53.58%		7,912
White	63.36%		9,356
Pardo	16.88%		2,493
Black	2.72%		402
Other (Yellow or Indigenous)	0.63%		95
Refuse to report ethnicity	16.4%		2400
Behavior Data			
Classes with at least one behavior assessment	94%		
Classes with at least one behavior assessment (Math)	72.18%		
Good behavior reports	15.95	22.36	13,879
Good behavior reports (Math)	4.41	7.18	10,009
Bad behavior reports	8.58	13.00	13,766
Bad behavior reports (Math)	2.87	5.19	10,184
Good behavior measure	.43	.27	13,879
Good behavior measure (Math)	.35	.35	10,009
Bad behavior measure	.28	.27	13,766
Bad behavior measure (Math)	.23	.30	10,184

Note: This table reports summary statistics for our data. Data on grades, ethnicity and gender, and classroom behaviors are at the student-level.

Table (2) Estimated biases in the non-blind math scores toward classroom behavior - IV estimates

	(1) OLS	(2) IV	(3) IV	(4) IV	(5) IV
BB (Pct. 75)	-0.266	-0.081	-0.076	-0.073	-0.069
	(0.018)***	(0.014)***	(0.014)***	(0.013)***	(0.013)***
GB (Pct. 75)	0.511	0.116	0.115	0.112	0.115
	(0.021)***	(0.016)***	(0.016)***	(0.015)***	(0.015)***
Blind Math Score		0.968	0.970	0.911	0.969
		(0.013)***	(0.013)***	(0.028)***	(0.067)***
Ethnicity and Gender	No	No	Yes	Yes	Yes
Other Scores	No	No	No	Yes	Yes
Instrumenting Language Scores	-	No	No	No	Yes
High-Order Polynomials for Scores	-	No	No	No	Yes
Number of Observations	44979	44979	44979	44979	44979
Number of Clusters	9462	9462	9462	9462	9462
First-stage F Statistic		5432	5303	1367	10.73

Note: This table reports student×exam-level OLS (column 1) and IV (columns 2-5) regressions of teacher-assigned math scores on classroom behavior. $BB(75pct.)$ and $GB(75pct.)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. In the IV estimates, lagged blind math scores are used as instrumental variable for the current math scores. All specifications include classroom fixed effects and exams fixed effects. Other scores include the cumulative average performance in science and humanities, and current performance in language. High-order polynomials for scores include a third order polynomial for blind math scores, and an interaction term between math and language scores. In Column 5, we also use lagged language scores as instrumental variable for the current language scores. Controls for ethnicity include 5 indicators: Black, Indigenous, *Pardo*, Yellow, and White. We also include a dummy for students with missing data on ethnicity. Standard errors in parenthesis are robust and clustered at the student level.
** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (3) Estimated biases in the non-blind math scores toward classroom behavior - Testing additional lags as instrument

	(1) IV	(2) IV	(3) IV	(4) IV	(5) IV
BB (Pct. 75)	-0.096	-0.093	-0.094	-0.095	-0.095
	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***
GB (Pct. 75)	0.133	0.126	0.127	0.130	0.130
	(0.018)***	(0.018)***	(0.018)***	(0.018)***	(0.018)***
Blind Math Score	0.801	0.868	0.855	0.827	0.833
	(0.033)***	(0.040)***	(0.044)***	(0.029)***	(0.027)***
Number of Lags	1	2	3	1-2	1-3
Number of Observations	27940	27940	27940	27940	27940
Number of Clusters	9459	9459	9459	9459	9459
First-stage F Statistic	1012	766.3	619.1	804.8	649.7
Over-ID Test (p-value)				0.185	0.355

Note: This table reports student×exam-level IV regressions of teacher-assigned math scores on classroom behavior. Specifications follow Table 2, column (3), except for the instrumental variable. This table tests as instrumental variable the first, second, and third lags of the math blind scores separately (columns 1-3) and jointly (columns 4-5). The sample is restricted to observations from the second semester as we need at least three past exams when using the third lag of the blind scores as instrumental variable. Columns 4-5 present p-values for over-identification tests.

** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (4) Estimated biases in the non-blind math scores toward classroom behavior - OLS estimation

	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) OLS
BB (Pct. 75)	-0.261	-0.180	-0.185	-0.104	-0.099
	(0.018)***	(0.014)***	(0.014)***	(0.013)***	(0.013)***
GB (Pct. 75)	0.506	0.332	0.333	0.197	0.171
	(0.021)***	(0.016)***	(0.016)***	(0.014)***	(0.014)***
Blind Math Score		0.426	0.424	0.234	0.186
		(0.005)***	(0.005)***	(0.005)***	(0.005)***
Ethnicity and Gender	No	No	Yes	Yes	Yes
Other Scores	No	No	No	Yes	Yes
Past Math Scores	No	No	No	No	Yes
Number of Observations	46787	46787	46787	46787	46787
Number of Clusters	9462	9462	9462	9462	9462
Adjusted R-squared	0.0534	0.226	0.228	0.330	0.353

Note: This table reports student×exam-level OLS regressions of teacher-assigned math scores on classroom behavior. $BB(75pct.)$ and $GB(75pct.)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. Other scores include the cumulative average performance in science and humanities, and current performance in language. Past math scores include the lagged blind math score. Controls for ethnicity include 5 indicators: Black, Indigenous, *Pardo*, Yellow, and White. We also include a dummy for students with missing data on ethnicity. Standard errors in parenthesis are robust and clustered at the student level.

** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

A OLS and IV Potential Biases

We consider a simple econometric model to analyze the bias of the OLS if the blind test score is measured with error, and of the IV estimator when the exogeneity assumption of the instrument does not hold. For simplicity, we assume that all variables have expected value equal to zero, we consider only the measure of good behavior, and we suppress the ijs sub-index. A simplified version of equation 1 is then given by

$$S^{NB} = \beta GB + \delta S^B + \varepsilon, \tag{2}$$

where $\varepsilon = \xi + r - \delta u + v$. We assume that $\mathbb{E}[GB\varepsilon] = 0$, but $\mathbb{E}[S^B\varepsilon]$ is potentially different from zero. Following the discussion from Section 3, we consider the case in which $r \approx 0$, so that $\mathbb{E}[S^B\varepsilon] \neq 0$ because of the measurement error u. We also assume that ξ is uncorrelated with GB and S^B.

Assuming that u is uncorrelated with all other variables in the model, we have that the OLS estimator is such that

$$\begin{bmatrix} \hat{\beta}^{ols} \\ \hat{\delta}^{ols} \end{bmatrix} \rightarrow_p \begin{bmatrix} \beta \\ \delta \end{bmatrix} + \frac{1}{\sigma_x^2\sigma_w^2 - (\sigma_{xw})^2} \begin{bmatrix} \sigma_{xw}\sigma_u^2\delta \\ -\sigma_x^2\sigma_u^2\delta \end{bmatrix},$$

where $\sigma_x^2 = var(GB)$, $\sigma_w^2 = var(S^B)$, $\sigma_u^2 = var(u)$, and $\sigma_{xw} = cov(GB, S^B)$. If we define the linear projection $GB = \gamma S^B + h$, then $\gamma = \frac{\sigma_{xw}}{\sigma_w^2}$. Therefore, $\sigma_x^2\sigma_w^2 - (\sigma_{xw})^2 = \sigma_x^2\sigma_w^2\left[1 - \gamma^2\frac{\sigma_w^2}{\sigma_x^2}\right] > 0$, so the sign of the bias of the OLS estimator for β is determined by the signs of σ_{xw} and δ. Given model 1, we have that $\delta > 0$. Moreover, we can estimate σ_{xw} using the data, where we find $\hat{\sigma}_{xw} > 0$. Therefore, we should expect that $\hat{\beta}^{ols}$ is upward biased. The intuition is that the measurement error u implies that the estimator for δ will suffer from attenuation bias, which implies that it will not completely control for students' skills. If we consider instead our measure of bad behavior, then the correlation between BB and S^B is negative, which implies that the estimator associated with BB would be downward biased.

We consider next estimation of equation 2 using lagged blind test score LS^B as instrumental variable for S^B. This instrument clearly satisfies the relevance condition. If $\mathbb{E}[LS^B\varepsilon] = 0$, then the IV estimator would be consistent for β. We are worried, however, that the exogeneity condition for the instrument may not be valid. We assume that $\mathbb{E}[LS^Bu] \approx 0$, which is a standard assumption in classical test theory and applied papers (e.g., Bond and Lang (2018)). In section 4, we present evidence that gives some support on the validity of the assumption. Still, it may be that $\mathbb{E}[LS^B\xi] > 0$. For example, teachers may statistically discriminate students based on their past performance in blind scores or other correlated unobservable signals of scholastic ability. In this case, we have that the IV estimator will converge to

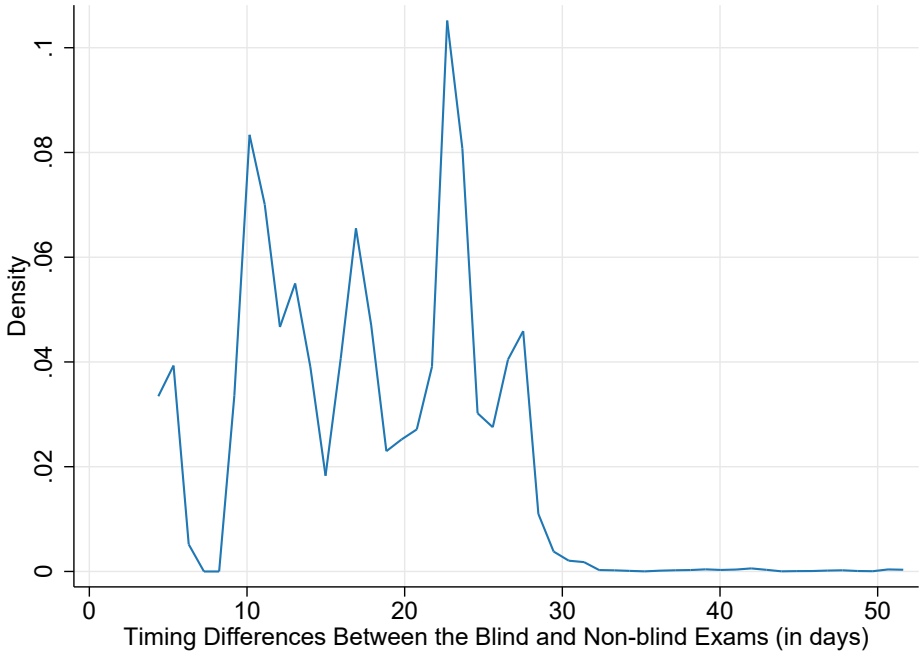
$$\begin{bmatrix} \hat{\beta}^{IV} \\ \hat{\delta}^{IV} \end{bmatrix} \rightarrow_p \begin{bmatrix} \beta \\ \delta \end{bmatrix} + \frac{1}{\sigma_x^2\sigma_{wz} - \sigma_{xw}\sigma_{xz}} \begin{bmatrix} -\sigma_{xw}\mathbb{E}[LS^B\varepsilon] \\ \sigma_x^2\mathbb{E}[LS^B\varepsilon], \end{bmatrix},$$

where $\sigma_{wz} = cov(LS^B, S^B)$ and $\sigma_{xz} = cov(LS^B, GB)$. Note that $\sigma_x^2\sigma_{wz} - \sigma_{xw}\sigma_{xz} = cov(e_1, e_2)$, where e_1 is the population error in the linear projection of GB on S^B, and e_2 is the population error in the linear projection of GB on LS^B. If we consider the residuals from a regression of GB on S^B and the residuals from a regression of GB on LS^B, then the correlation between these two residuals is positive, which provides evidence that $\sigma_x^2\sigma_{wz} - \sigma_{xw}\sigma_{xz}$ is positive. Given that $\sigma_{xw} > 0$ when we consider a measure of good behavior, if we have $\mathbb{E}[LS^B u] \approx 0$ and $\mathbb{E}[LS^B \xi] > 0$, then $\hat{\beta}^{IV}$ would be downward biased. Likewise, if we consider a measure of bad behavior, then the estimator associated with this variable would be upward biased.

Combining these results, we have that the discrimination parameters are bounded by the OLS and the IV estimators.

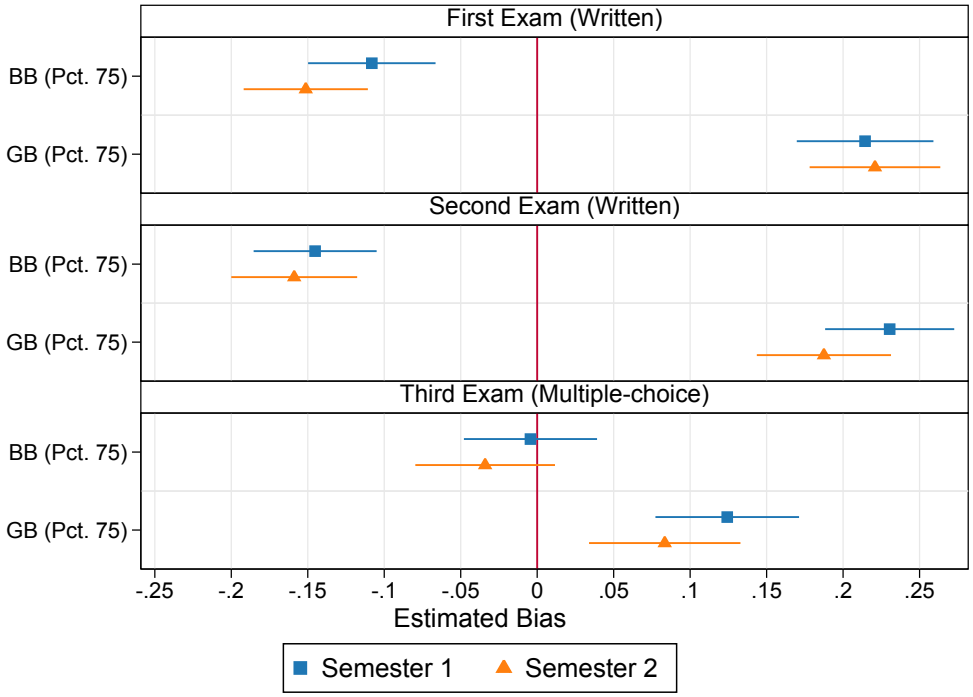
B Additional Tables and Figures

Figure (B.1) Density of the timing differences between the blind and the non-blind exams



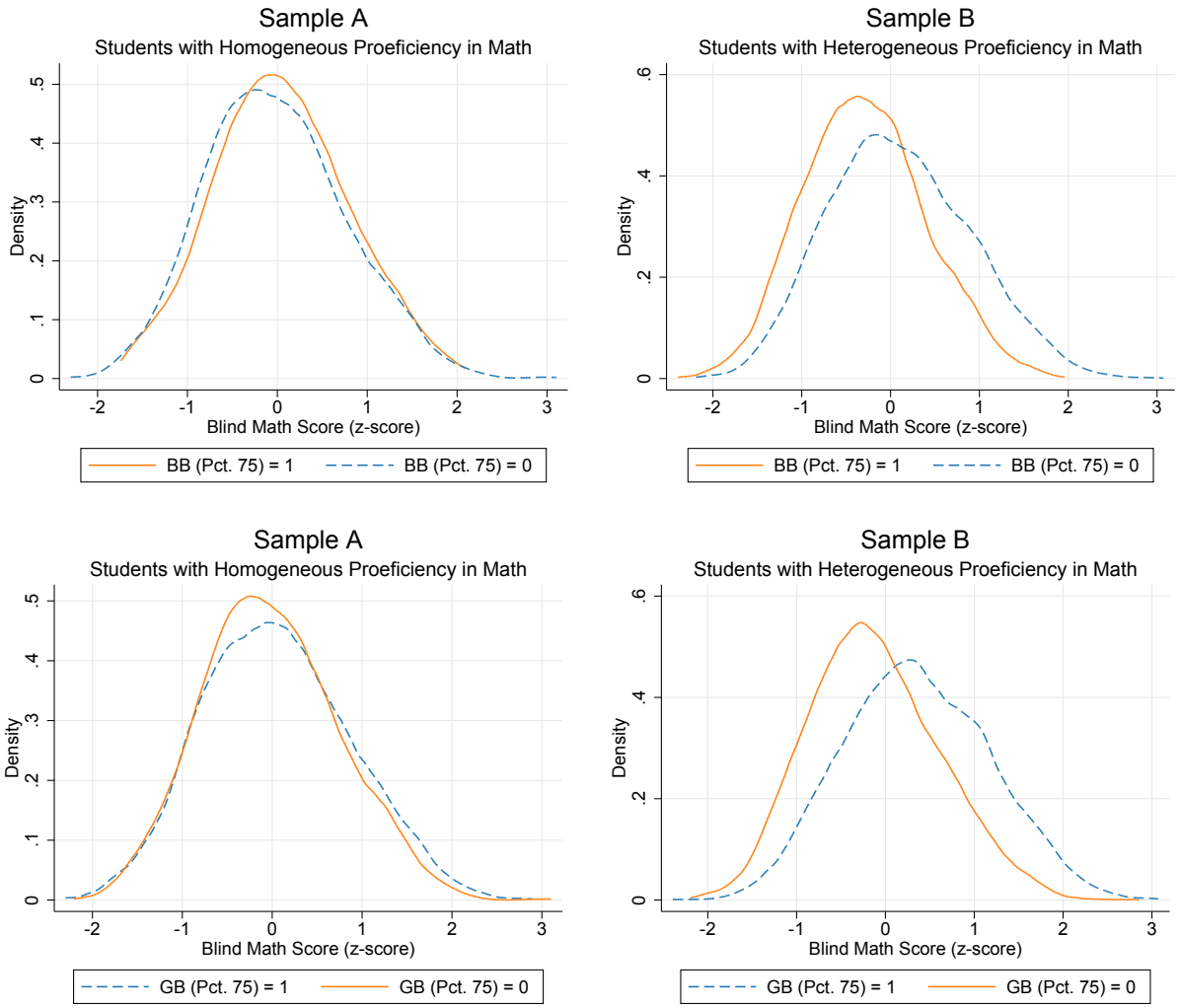
Note: This figure estimates the density of the timing differences between the blind and non-blind math exams, when we pool all the six exams.

Figure (B.2) Heterogeneity Across Semesters and Exams - OLS Estimation



Note: This figure plots 90% confidence intervals and point estimates from student×exam-level OLS regressions of teacher-assigned math scores on classroom behavior, using subsamples that are specific for each exam. $BB(75pct.)$ and $GB(75pct.)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. All specifications follow Table 4, column 5.

Figure (B.3) Distribution of Blind and Non-blind Math Scores from the 1st Semester



Note: These figures estimate the density of the blind math scores from the first-semester exams for students whose behavior indicators assume different values, using two different samples. $BB(75pct.)$ and $GB(75pct.)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. In the top two figures, solid line represents students with $BB(Pct.75) = 1$ and the dotted line represents those with $BB(Pct.75) = 0$. In the bottom two figures, solid line represents students with $GB(Pct.75) = 0$ and the dotted line represents those with $GB(Pct.75) = 1$. Sample A selects classrooms where in the first semester students with $BB(pct.75) = 1$ or $GB(pct.75) = 0$ performed, on average, better in the blind math exams than their classmates with $BB(pct.75) = 0$ or $GB(pct.75) = 1$, respectively. We call this subsample of sample A. The subsample where the previous conditions are not satisfied is called of sample B.

Table (B1) Estimated biases in the non-blind math scores toward classroom behavior – school-level cluster

	(1) OLS	(2) IV
BB (Pct. 75)	-0.266	-0.073
	(0.019)***	(0.012)***
GB (Pct. 75)	0.511	0.112
	(0.029)***	(0.017)***
Blind Math Score		0.911
		(0.035)***
Number of Observations	44979	44979
Number of Clusters	51	51
First-stage F Statistic		525.2

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior. Column 2 follows the same specification from Table 2, column 3, except for the standard errors that here are calculated with school-level clusters.
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B2) Estimated biases in the non-blind math scores toward classroom behavior – restricting $\delta = 1$

	(1) OLS	(2) OLS	(3) OLS
BB (Pct. 75)	-0.075	-0.070	-0.069
	(0.014)***	(0.014)***	(0.014)***
GB (Pct. 75)	0.103	0.103	0.101
	(0.015)***	(0.015)***	(0.015)***
Ethnicity and Gender	No	Yes	Yes
Other Scores	No	No	Yes
Number of Observations	44979	44979	44979
Number of Clusters	9462	9462	9462

Note: This table reports student×exam-level regressions of the difference between non-blind and blind math scores on classroom behavior. All test scores are standardized. All specifications include classroom fixed effects and exams fixed effects. Other scores include the cumulative average performance in science and humanities, and current performance in language. Controls for ethnicity include 5 indicators: Black, Indigenous, *Pardo*, Yellow, and White. We also include a dummy for students with missing data on ethnicity. Standard errors in parenthesis are robust and clustered at the student level.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B3) Estimated biases in the non-blind math scores toward classroom behavior by question type

	(1) OLS	(2) IV	(3) OLS	(4) IV
BB (Pct. 75)	-0.170	0.016	-0.330	-0.132
	(0.020)***	(0.021)	(0.021)***	(0.017)***
GB (Pct. 75)	0.405	0.058	0.578	0.145
	(0.024)***	(0.023)**	(0.023)***	(0.019)***
Blind Math Score		0.908		0.923
		(0.048)***		(0.034)***
Type of questions	Multiple-choice	Multiple-choice	Written	Written
Number of Observations	17701	17701	27278	27278
Number of Clusters	9282	9282	9457	9457
First-stage F Statistic		582.4		1081

Note: This table reports student×exam-level OLS (columns 1 and 3) and IV (columns 2 and 4) regressions of teacher-assigned math scores on classroom behavior, for two different subsamples. One of them is restricted to non-blind exams that are only multiple-choice, and the other to non-blind exams that require written answers. Columns 2 and 4 follow the same specification from Table 2, column 3.
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B4) Estimated biases in the non-blind math scores toward classroom behavior by question type – OLS estimation

	(1) OLS	(2) OLS	(3) OLS	(4) OLS
BB (Pct. 75)	-0.162	-0.020	-0.326	-0.152
	(0.020)***	(0.017)	(0.021)***	(0.015)***
GB (Pct. 75)	0.399	0.100	0.576	0.213
	(0.023)***	(0.019)***	(0.023)***	(0.016)***
Type of questions	Multiple-choice	Multiple-choice	Written	Written
Number of Observations	18463	18463	28324	28324
Number of Clusters	9335	9335	9462	9462

Note: This table reports student×exam-level OLS regressions of teacher-assigned math scores on classroom behavior, for two different subsamples. One of them is restricted to non-blind exams that are only multiple-choice, and the other to non-blind exams that require written answers. Columns 2 and 4 follow the same specification from Table 4, column 5.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B5) Estimated biases in the non-blind math scores toward classroom behavior – adjusting for blind essay scores

	(1) OLS	(2) IV	(3) IV	(4) IV
BB (Pct. 75)	-0.321	-0.133	-0.126	-0.109
	(0.022)***	(0.018)***	(0.018)***	(0.018)***
GB (Pct. 75)	0.568	0.129	0.126	0.121
	(0.024)***	(0.020)***	(0.019)***	(0.019)***
Blind Math Score		0.934	0.923	0.896
		(0.037)***	(0.037)***	(0.037)***
Essay Scores	No	No	Yes	Yes
Instrumenting Essay Scores	-	-	No	Yes
Number of Observations	23177	23177	23177	23177
Number of Clusters	8739	8739	8739	8739
First-stage F Statistic		912.4	903.4	320.6

Note: This table reports student×exam-level OLS (column 1) and IV (columns 2-3) regressions of teacher-assigned math scores on classroom behavior, in a subsample where essay scores are available. Column 2 follows the same specification from Table 2, column 3. Column 3 additionally controls for blind essay scores, and column 4 uses lagged essay scores as instrumental variable for the current ones.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B6) Estimated biases in the non-blind essay scores toward classroom behavior

	(1) OLS	(2) IV	(3) OLS	(4) IV	(5) Low−High
BB (Pct. 75)	-0.341	-0.125	-0.248	-0.0510	0.082
	(0.024)***	(0.021)***	(0.061)***	(0.053)	[0.157]
GB (Pct. 75)	0.356	0.117	0.375	0.112	-0.006
	(0.025)***	(0.022)***	(0.072)***	(0.058)*	[0.923]
Blind Essay Scores		0.565		0.759	
		(0.046)***		(0.104)***	
Stakes	High and Low	High and Low	High	High	
Number of Observations	17792	17792	3633	3634	
Number of Clusters	6257	6257	860	860	
First-stage F Statistic		423.4		96.48	

Note: This table reports student×exam-level OLS (columns 1 and 3) and IV (columns 2 and 4) regressions of teacher-assigned essay scores on classroom behavior, for two different subsamples. One of them uses all the essay exams, and the other is restricted to essay scores that are high-stakes. $BB(75pct.)$ and $GB(75pct.)$ stand for binary variables that indicate whether students are at the top quartile of the essay behavior measures' distribution. Columns 2 and 4 follow the same specification from Table 2, column 3. Column 5 reports the differences between the point estimates of $BB(Pct.75)$ and $GB(Pct.75)$ presented in column 4 and the point estimates associated with these behavior indicators obtained in a subsample of only non-high-stakes exams. In brackets, p-values for t-tests under the null that theses coefficients are equal.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B7) Estimated biases in the non-blind math scores toward classroom behavior – varying the timing between the blind and non-blind exams

	(1)	(2)
	IV	IV
BB (Pct. 75)	-0.118	-0.148
	(0.032)***	(0.033)***
GB (Pct. 75)	0.136	0.117
	(0.036)***	(0.038)***
Blind Math Score	1.093	0.924
	(0.065)***	(0.081)***
Timing differences	From 10 to 20 days	More than 20 days
Number of Observations	8610	6779
Number of Clusters	4350	5050
First-stage F Statistic	338.3	205.4

Note: This table reports student×exam-level OLS (column 1) and IV (columns 2-5) regressions of teacher-assigned math scores on classroom behavior, for two different subsamples according due to timing differences between the realization of the blind and non-blind exams. Column 1 uses exams where this difference varies from 10 to 20 days; Column 2, more than 20 days. All columns follow the same specification from Table 2, column 3.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B8) Estimated biases in the non-blind math scores toward classroom behavior – varying the timing between the blind and non-blind exams (OLS estimation)

	(1) OLS	(2) OLS	(3) OLS
BB (Pct. 75)	-0.112	-0.172	-0.129
	(0.039)***	(0.026)***	(0.027)***
GB (Pct. 75)	0.195	0.213	0.179
	(0.044)***	(0.028)***	(0.029)***
Timing differences	From 0 to 9 days	From 10 to 20 days	More than 20 days
Number of Observations	2436	8610	6779
Number of Clusters	2436	4350	5050

Note: This table reports student×exam-level OLS regressions of teacher-assigned math scores on classroom behavior, for three different subsamples according due to timing differences between the realization of the blind and non-blind exams. Column 1 uses exams in which this difference varies up to 9 days; Column 2, from 10 to 20 days; Column 3, more than 20 days. All columns follow the same specification from Table 4, column 5.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B9) Estimated biases in the non-blind math scores toward classroom behavior – using all the behavior assessments

	(1) OLS	(2) IV
BB (Pct. 75)	-0.240	-0.076
	(0.015)***	(0.011)***
GB (Pct. 75)	0.524	0.110
	(0.016)***	(0.012)***
Blind Math Score		0.870
		(0.022)***
Number of Observations	67495	67495
Number of Clusters	13654	13654
First-stage F Statistic		2194

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior. $BB(75pct.)$ and $GB(75pct.)$ stand for binary variables that indicate whether students are at the top quartile of the behavior measures' distribution, computed using the behavior assessments made by all teachers. Column 2 follows the same specification from Table 2, column 3.
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B10) Estimated biases in the non-blind math scores toward classroom behavior – using all the behavior assessments, expect those by math teachers

	(1) OLS	(2) IV
BB (Pct. 75)	-0.245	-0.073
	(0.015)***	(0.011)***
GB (Pct. 75)	0.483	0.102
	(0.017)***	(0.012)***
Blind Math Score		0.879
		(0.022)***
Number of Observations	65593	65593
Number of Clusters	13271	13271
First-stage F Statistic		2163

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior. $BB(75pct.)$ and $GB(75pct.)$ stand for binary variables that indicate whether students are at the top quartile of the behavior measures' distribution, computed using the behavior assessments made by all teachers, except those made by math teachers. Column 2 follows the same specification from Table 2, column 3.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B11) Estimated biases in the non-blind math scores toward classroom behavior – using the continuous behavior measures

	(1) OLS	(2) IV
BB	-0.357	-0.104
	(0.029)***	(0.021)***
GB	0.789	0.177
	(0.031)***	(0.022)***
Blind Math Score		0.906
		(0.028)***
Number of Observations	44979	44979
Number of Clusters	9462	9462
First-stage F Statistic		1356

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior. BB and GB stand for the math behavior measures. Column 2 follows the same specification from Table 2, column 3.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B12) Estimated biases in the non-blind math scores toward classroom behavior – using the number of behavior assessments

	(1) OLS	(2) IV
Ln BB Reports	-0.145	-0.052
	(0.012)***	(0.008)***
Ln GB Reports	0.320	0.074
	(0.014)***	(0.011)***
Blind Math Score		0.873
		(0.026)***
Number of Observations	46757	46787
Number of Clusters	9462	9462
First-stage F Statistic		1532

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior. Ln BB reports and Ln GB reports stand for the natural logarithm of the number of bad and good behavior assessments received by math teachers plus 1. Column 2 follows the same specification from Table 2, column 3. *** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B13) Estimated biases in the non-blind math scores from the 2nd semester towards classroom behavior – using behavior reports from the 1st semester

	(1) OLS	(2) IV
BB (Pct. 75) (1st Semester)	-0.253	-0.070
	(0.019)***	(0.014)***
GB (Pct. 75) (1st Semester)	0.488	0.103
	(0.022)***	(0.016)***
Blind Math Score		0.914
		(0.029)***
Number of Observations	40907	40907
Number of Clusters	8598	8598
First-stage F Statistic		1292

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior, for a subsample of exams from the second semester only. $BB(75pct.)$ $(1st\ Semester)$ and $GB(75pct.)$ $(1st\ Semester)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution, computed using only behavior assessments from the first semester. Column 2 follows the same specification from Table 2, column 3.
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (B14) Estimated biases in the non-blind math scores toward classroom behavior – using pre-exam behavior reports

	(1) OLS	(2) IV	(3) OLS	(4) IV
BB	-0.398	-0.118		
	(0.037)***	(0.027)***		
GB	0.811	0.213		
	(0.038)***	(0.027)***		
BB (Pre-exams)			-0.386	-0.108
			(0.033)***	(0.025)***
GB (Pre-exams)			0.728	0.182
			(0.033)***	(0.026)***
Blind Math Score		0.859		0.862
		(0.032)***		(0.033)***
Number of Observations	27303	27303	27303	27303
Number of Clusters	7296	7296	7296	7296
First-stage F Statistic		945.3		942.9

Note: This table reports student×exam-level OLS (columns 1 and 3) and IV (columns 2 and 4) regressions of teacher-assigned math scores on classroom behavior. $BB\,(Pre-exams)$ and $GB\,(Pre-exams)$ are the math behavior measures, computed using only assessments that preceded each examination. Columns 2 and 4 follow the same specification from Table 2, column 3. Columns 1 and 2 replicate Appendix Table B11, using a subsample of observations where the behavior measures that use pre-exam reports are not missing.
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

C Additional Heterogeneities and Robustness Checks

C.1 Racial and Gender Discrimination

Table C1 presents the estimated bias in the non-blind math scores toward black pupils. Blacks' average non-blind scores are 0.20 SD below than whites' (column (1)). This gap falls drastically (-0.03, s.e. 0.03) when we adjust for student proficiency captured by blind scores. Although not precisely estimated, this point estimate is similar to the obtained by Botelho et al. (2015) (-0.02, s.e. 0.005), which also analyze Brazilian students, though in a very different context. In their study, 18% of the students are black. In ours, only 3%. A major difference is that we are analyzing private schools, while they are studying public schools. This fact, associated with the income gap between black and white people in Brazil, probably explains the lower share of black students in our sample. They also evaluate a higher number of students, which allow them to precisely estimate discrimination effects of -0.02 SD. Assuming this is the true effect, considering our s.e. of 0.036, and fixing a 10% level test, our power is only 14%.

In Table C2 we present the estimated grading bias toward gender. Boys have advantages of 0.04 SD over girls in math test scores. This could indicate some favoritism toward boys. However, by controlling for non-blind math grades we find evidence that boys' math proficiency is under-assessed by teachers. The grading bias is equivalent to a taxation of 0.04 SD in non-blind scores. Our main estimate drops slightly when we control for student in-class behaviors, reflecting a small correlation between these non-cognitive skills and gender, but remains statistically different from zero. These findings are in line with several studies from the literature (e.g., Lavy (2008); Falch and Naper (2013)).

Taken together, our results indicate that despite suggestive evidence of biases toward boys and black students, results are much lower in comparison to discrimination toward behavior.

Table (C1) Estimated biases in the non-blind math scores against black pupils

	(1) OLS	(2) IV	(3) IV
Black	-0.205	-0.036	-0.040
	(0.051)***	(0.034)	(0.034)
Blind Math Score		0.974	0.946
		(0.012)***	(0.012)***
BB (Pct. 75)			-0.081
			(0.013)***
GB (Pct. 75)			0.121
			(0.015)***
Number of Observations	46974	46975	46975
Number of Clusters	9500	9500	9500
First-stage F Statistic		6351	5772

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on ethnicity. Black stands for a binary variable that indicates whether student is black. We also control for a binary variable that indicates whether student is Pardo, and for a binary variable that indicates whether student is yellow, indigenous or has missing data on ethnicity. The omitted category is white. $BB(75pct.)$ and $GB(75pct.)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. In the IV estimates, lagged blind math scores are used as instrumental variable for the current math scores. Columns 2-3 also controls for past blind scores of language, science, and humanities. All specifications include classroom fixed effects and exams fixed effects. Standard errors in parenthesis are robust and clustered at the student level. *** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Table (C2) Estimated biases in the non-blind math scores against boys

	(1) OLS	(2) IV	(3) IV
Boy	0.038	-0.040	-0.027
	(0.016)**	(0.011)***	(0.011)**
BB (Pct. 75)			-0.077
			(0.014)***
GB (Pct. 75)			0.119
			(0.015)***
Blind Math Score		0.977	0.949
		(0.012)***	(0.012)***
Number of Observations	46974	46975	46975
Number of Clusters	9500	9500	9500
First-stage F Statistic		6394	5788

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on gender. $BB(75pct.)$ and $GB(75pct.)$ stand for binary variables that indicate whether students are at the top quartile of the math behavior measures' distribution. In the IV estimates, lagged blind math scores are used as instrumental variable for the current math scores. Columns 2-3 also control for past blind scores of language, science, and humanities. All specifications include classroom fixed effects and exams fixed effects. Standard errors in parenthesis are robust and clustered at the student level.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

C.2 What are the behaviors driving the results?

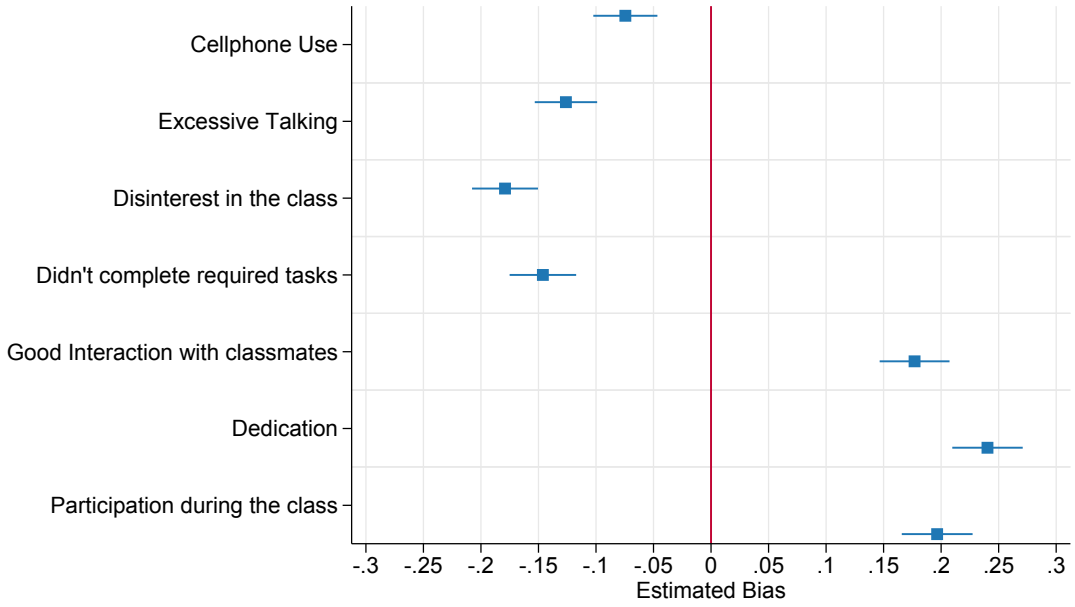
Here we estimate heterogeneous effects for each of the behaviors. To do so, we calculate disaggregated behavior measures. Take the behavior report "Dedication" as an example. Let d_{is} indicate the number of assessments i received under this category by a subject s teacher. The measure $Dedication_{is}$ is then defined as:

$$Dedication_{is} \coloneqq \frac{d_{is}}{\max\{d_{js} : j \in \mathcal{C}(i)\}}.$$

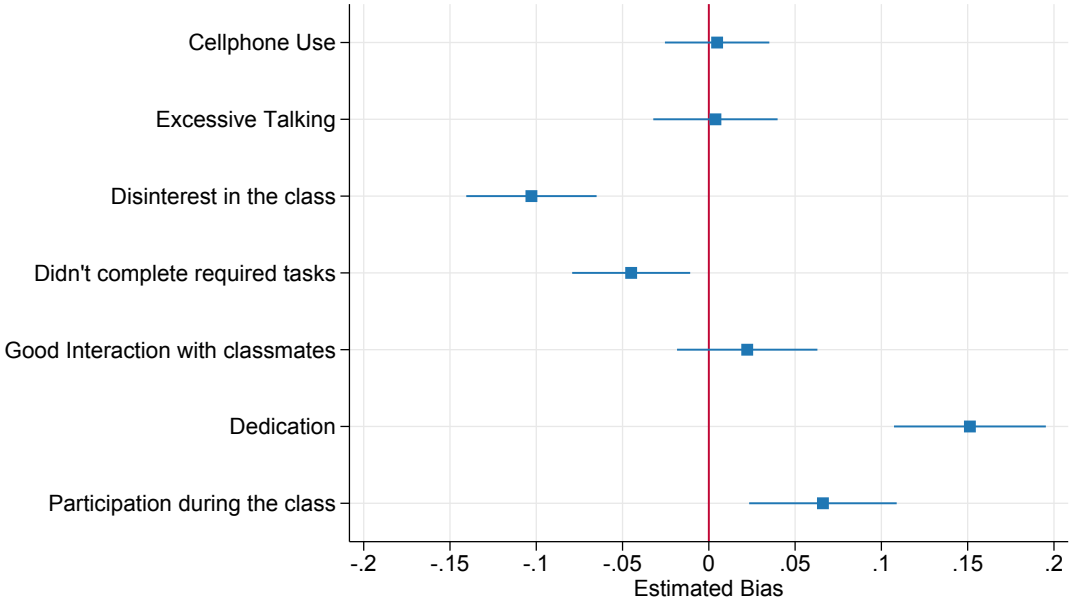
Figure C.1 presents our main estimates. In panel (a), we estimate the grading biases toward each of the behaviors separately. Overall, the point estimates are similar, indicating that they are all capturing correlated biases. Notice that in this case, the point estimates capture both positive and negative discrimination. In panel (b), we estimate the effects using the same regression model. The negative discrimination is driven by the disinterest of the students during the class (-0.10, s.e. 0.2) and is followed by 'Did not complete the required tasks' (-0.04, s.e. 0.2). Conditioning on all the disaggregated behavior measures, cellphone use and excessive talk do not seem to be factored by teachers in grades. The positive discrimination is driven by dedication (0.15, s.e. 0.2), and is followed by participation during the class (0.06, s.e. 0.2). Point estimate associated with good interaction with classmates is positive, though not statistically different from zero (0.2, s.e. 0.2).

Figure (C.1) Estimated biases toward each behavior

(a) Different regressions



(b) Same regression



Note: These figures plots student×exam-level IV regressions of teacher-assigned math scores on measures for each classroom behavior. Panel (a) plots point estimates from different IV regressions on each behavior. Panel (b) plots point estimates from an IV regression on all the behaviors. Both also plots 90% confidence intervals. All regressions follow the same specification from Table 2, column 3.

C.3 Biases in the Portuguese non-blind scores

Table (C3) Estimated biases in the non-blind Portuguese scores toward classroom behavior

	(1) OLS	(2) IV
BB (Pct. 75)	-0.259	-0.058
	(0.021)***	(0.016)***
GB (Pct. 75)	0.406	0.088
	(0.024)***	(0.017)***
Blind Portuguese Score		0.832
		(0.046)***
Number of Observations	31042	31042
Number of Clusters	6525	6525
First-stage F Statistic		585.6

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned Portuguese scores on classroom behavior. $BB(75pct.)$ and $GB(75pct.)$ stand for binary variables that indicates whether students are at the top quartile of the Portuguese behavior measures' distribution. Column 2 follows the same specification from Table 2, column 3.

** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

C.4 Alternative behavior measures

We start showing the results we obtain using the behavioral grades students receive at the end of each semester. Table C4 shows that students at the top of the math behavioral grades' distribution perform 0.54 points higher than others in non-blind math tests (column (1)). The point estimate drops significantly when we control for the blind scores, but remain statistically significant and high in magnitude (0.14, s.e 0.01). This amounts to 25% of the unconditional gaps. It is also equivalent to an increase of 0.18 of one SD in blind math scores. Figure C.2 presents the results when we use the standardized scores as regressors. In addition, we show that the results are similar if instead of using the math grades, we use the average grade from all subjects. For both cases, we also show that estimating the biases in the non-blind scores from the second semester, while using the behavioral grades from the first, leads to similar results.

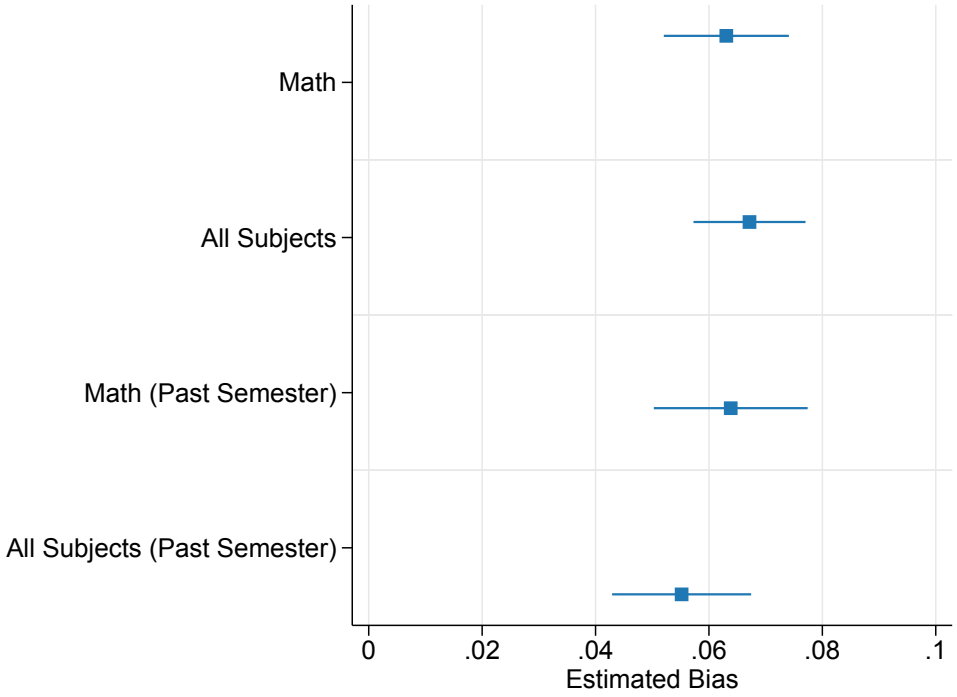
We now use the pupils' grades in the courses aimed at increasing their non-cognitive skills. Table C5 shows that the average math grade of students in the top quartile of the non-cognitive scores' distribution is 0.46 SD above the others (column (1)). We also find statistically significant results after controlling for student proficiency proxied by blind scores: 0.06 SD or 12% of the unconditional gap (column (2)). Table C6 shows that the results are similar when we use the continuous socio-emotional grades and slightly smaller we use only scores that precede the math examinations.

Table (C4) Estimated biases in the non-blind math scores toward classroom behavior (using behavioral grades)

	(1) OLS	(2) IV
Behavior Score (Pct.75)	0.547	0.142
	(0.019)***	(0.016)***
Blind Math Score		0.750
		(0.030)***
Number of Observations	35701	35701
Number of Clusters	11948	11948
First-stage F Statistic		1212

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior, measured by behavioral grades. *Behavior Score* (75*pct.*) stands for a binary variable that indicates whether students are at the top quartile of the math behavioral grade. Column 2 follows the same specification from Table 2, column 3.
*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$.

Figure (C.2) Estimated biases in the non-blind math scores toward in-class behaviors – using the behavioral grades



Note: This figure plots 90% confidence intervals and point estimates from student×exam-level IV regressions of teacher-assigned math scores on classroom behavior, measured by behavioral grades. Math indicates the average of the math grades through the year; All Subjects stands for the average of the grades from all subjects through the year; Math (Past Semester) indicates the first-semester math grades; and All Subjects (Past Semester) stands for the average of the first-semester grades from all subjects. In these last two cases, biases toward classroom behavior are estimated using only the teacher-assigned math scores from the second semester.

Table (C5) Estimated biases in the non-blind math scores toward classroom behavior – using the socio-emotional scores

	(1) OLS	(2) IV
Socio-emotional Score (Pct. 75)	0.466	0.057
	(0.042)***	(0.026)**
Blind Math Score		0.893
		(0.054)***
Number of Observations	9055	9055
Number of Clusters	1834	1834
First-stage F Statistic		341.5

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior, measured by grades from a non-cognitive skills course. Sotio-emotional Score (75pct.) stands for a binary variable that indicates whether students are at the top quartile of the grades from regular courses designed to improve sotio-emotional skills. Column 2 follows the same specification from Table 2, column 3.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$

Table (C6) Estimated biases in the non-blind math scores toward classroom behavior – using the continuous socio-emotional scores

	(1) OLS	(2) IV	(3) OLS	(4) OV
Socio-emotional Score	0.345	0.067		
	(0.023)***	(0.016)***		
Socio-emotional Score (Pre Exams)			0.296	0.051
			(0.021)***	(0.015)***
Blind Math Score		0.890		0.892
		(0.054)***		(0.054)***
Number of Observations	9055	9055	9055	9055
Number of Clusters	1834	1834	1834	1834
First-stage F Statistic		341.2		342

Note: This table reports student×exam-level OLS (column 1) and IV (column 2) regressions of teacher-assigned math scores on classroom behavior, measured by grades from a non-cognitive skills course. Socio-emotional Score stands for the average of the standardized scores that comes from grades in these courses. Socio-emotional Score (Pre Exams) stands for the cumulative average of these scores. Column 2 follows the same specification from Table 2, column 3.

*** $p < 0.01$; ** $p < 0.05$; *$p < 0.1$