Policy Research Working Paper 9593

Poor Expectations

Experimental Evidence on Teachers' Stereotypes and Student Assessment

Maria Gabriela Farfan Bertran
Alaka Holla
Renos Vakis

WORLD BANK GROUP
Poverty and Equity Global Practice
&
Education Global Practice
March 2021

Policy Research Working Paper 9593

Abstract

Do teachers' stereotypes of social class bias their assessment of students? This study uses a lab-in-the-field experiment among primary school teachers to test whether they are biased against poor students. Teachers assessed a student in a video of an oral exam after watching one of two versions of an introductory video that portrayed the child's home and playground. When the student in the exam video exhibited inconsistent performance, showing varying levels of scholastic aptitude and focus during the exam, teachers were far more likely to judge his scholastic aptitude as below grade-level if they had watched the introductory video portraying a poor background than if they had watched the introductory video portraying a middle-class background. The social class background portrayed in the introductory video did not affect teachers' behavioral assessments of the student. When the student in the exam video was consistently high achieving, showing high levels of scholastic aptitude and focus throughout the exam, teachers who watched the introductory video depicting a poor background were more likely to assess the student as above grade-level than teachers who watched the video conveying a middle-class background. In this case, however, they had a more negative assessment of the child's behavior when they thought he came from a poor background, deeming him to be less motivated and less emotionally mature than when the introductory video depicted a middle-class background. These findings suggest that stereotypes influence how teachers assess the scholastic aptitude and behavior of their students.

This paper is a product of the Poverty and Equity Global Practice and the Education Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at gfarfan@worldbank.org; aholla@worldbank.org; and rvakis@worldbank.org.

Poor Expectations: Experimental Evidence on Teachers' Stereotypes and Student Assessment*

Maria Gabriela Farfan Bertran, Alaka Holla, and Renos Vakis

Keywords: stereotypes, bias, education and inequality, lab-in-the-field
JEL codes: D91, I30, I240, O150

* We are grateful to colleagues at MineduLab within the Ministry of Education in Peru, especially Fabiola Caceres and Andrea Salazar for their support through design and implementation of the experiment. The paper and implementation benefited from comments from Karla Hoff, Ximena Del Carpio, Oscar Calvo-Gonzalez, Andress Yi Chang, Ines Kudo, Luciana Velarde, and participants at the 2019 BRIQ/IZA workshop on behavioral economics of education. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent. Contacts: gfarfan@worldbank.org; aholla@worldbank.org; rvakis@worldbank.org

1 Introduction

In all countries, there are disparities in student achievement across socio-economic classes (PISA, 2018). Across almost all countries, there are equally large socio-economic disparities in aspirations to complete tertiary education and the degree to which children believe that they can learn more if they work hard and persevere (a growth mindset). At the same time, there is a high degree of within-school variation in test scores, suggesting that what goes on within the classroom or within the home matters considerably for academic achievement. For example, in the latest test conducted by the Programme for International Student Assessment (PISA), within-school variation accounted for an average of 71 percent of the total variation observed in reading scores in OECD countries.

Poor children and their more affluent peers differ on an array of household and neighborhood inputs that might matter for their scholastic aptitude in school and ultimate educational attainment, such as cognitive stimulation in their early years (Fryer and Levitt, 2004; Bornstein and Putnick, 2012; PISA, 2018), respite from household chores and labor while in school (Angrist et al, 2002; Edmonds and Schady, 2012), information about the returns to school (Jensen, 2010), and neighborhoods with better schools and lower crime rates (Chetty and Hendren, 2017).

Some of the socio-economic disparities in achievement and within school variation may also stem from differences in children's classroom experiences. Evidence suggests that teachers do treat children from socially marginalized groups differently. They have lower expectations for them, and these expectations matter for children's educational trajectories. Black students in Brazil, for example, receive lower teacher assigned grades than their white peers with the same scores on a blindly-graded standardized assessment covering the same content (Botelho, Madiera, and Rangel, 2015), as do immigrant children in Italy when compared to their native peers (Alesina et al, 2018). Similarly, graders assign lower scores to low-caste children in India than high-caste children when they know students' background (Hanna and Linden, 2012).

This differential treatment in the classroom matters for a student's educational trajectory. Evidence from Israel, Italy, and Turkey suggests that teachers with higher measured gender bias can exacerbate gender gaps in math scores, and the female students of these teachers are less likely to enroll in more advanced classes or high schools (Lavy and Sand, 2018; Carlana, 2019; and Alan, Ertac, and Mumcu, 2018). Similarly, evidence from high school students in the United States suggests that their teachers' expectations matter for college completion and that black students suffer from systematically lower expectations held by their teachers (Papageorge, Gershenson, and Kang, 2020).

Identifying bias in the way teachers assess students, however, is empirically challenging. First, when a teacher rates one of her students from a marginalized group lower than a more affluent peer, she could be using a measure of competence that is correlated with social status and that she can observe but a researcher cannot. In an ideal experiment, a teacher would observe identical student performance from students of different social status, as Hanna and Linden (2012) designed at the group-level by randomly assigning children's background characteristics to exams.

A second challenge of detecting teacher bias stems from the form of assessment. While written examinations are important determinants of educational progression, teachers are more frequently assessing students' oral responses to questions and behavior in class to determine whether or not they understand the material that has been taught, and it is possible that this constant background assessment informs the way they grade written examinations, assign overall class grades, and communicate expectations to children.

In this study, we use a lab-in-the-field experiment modeled after the design of lab experiments in social psychology (Darley and Gross, 1983 and Baron, Albright, and Malloy, 1995) to address some of these methodological issues. In public schools in Metropolitan Lima, Peru, we show primary school teachers a video of a child responding to questions posed by a teacher and ask the teachers to assess the child's scholastic aptitude and behavior and to predict his future educational attainment. Before seeing this performance video, teachers see a brief segment that introduces the child and depicts him at play in his neighborhood and in his home. One version of the introductory segment depicts a poor background, while the other shows a neighborhood and home that are more associated with the middle-class.

We also experimentally varied the content of the performance video to identify what might underlie any observed differences in assessment across teachers who had seen the different introductory videos. In one variant, the child's performance is inconsistent. He answers 50 percent of the questions correctly. He sometimes gets an easy question wrong and a difficult question right. He sometimes fidgets and loses focus. In the other performance variant, the child is a high achiever. He answers most questions correctly. He demonstrates sustained focus on the questions.

Multiple studies show that people use social group membership to form judgments when information is absent or noisy. To use the language of Kahneman (2011), we take short-cuts (heuristics), where we use a fact like social group membership to fill in missing information. In multiple studies, observed bias declines in the presense of a strong signal or "individuating information" (Gneezy, List, and Price, 2012 and Rubinstein, Jussim, and Stevens, 2018). In economics, this reduction in bias is interpreted as evidence of stastistical discrimination, a rational form of Bayesian updating when traits of interest are unobservable. In models of statistical

discrimination, differential treatment results from a signal extraction problem (Becker, 1971; Phelps, 1972; Arrow, 1973; and Lundberg and Startz, 1983). In psychology, in contrast, how we choose to interpret data, extract signals, and update our priors are themselves choices, although possibly subconscious ones (Darley and Gross, 1983; Baron et al., 1995; Bertrand and Duflo, 2017; Kahneman, 2011). With incomplete or inconsistent information and with bounded attention, a heuristic that suggests that someone is of low quality leads us to selectively focus on negative information, which is used to justify the worse assessment for the marginalized group, a process akin to what behavioral economists have recognized as confirmation bias (Rabin and Schrag, 1999).

While our results do suggest that unambiguous signals of scholastic aptitude eliminate bias in the way teachers assess scholastic aptitude of children from different social classes, there is also evidence of selective attention induced by our introductory videos portraying social class. When teachers observe the performance video depicting inconsistent performance, they are 14 percentage points more likely to rate the poor child as performing below grade level than the middle-class child. They also predict a much lower likelihood of college attendance for the poor child than for the middle-class child. The introductory videos have no impact on teachers' relative assessments of the child's behavior.

On the other hand, when teachers observe the performance video where the child demonstrates high achievement and high focus, they assess the poor child as more able than the middle-class child in his academic performance. They estimate that the child is around 15 percentage points more likely to be performing *above* grade level when their introductory video depicted the poor background, compared to the middle-class background. In this case, however, teachers shift their negative assessment of the poorer child to the child's behavior. The teachers who viewed the poor introductory video deem the child to be less motivated and less emotionally mature than teachers who viewed the middle-class introductory video, and their predictions still imply a substantial gap in expected college attendance between the poor and middle-class child. Thus, irrespective of whether the videos showed a gifted student or a student of inconsistent performance, the teachers have higher expectations for the middle-class child than for the poor-child. These results hold both when we use the behavioral categories from the Darley and Gross (1983) and Baron et al. (1995) studies and when we use factor analysis to identify behavioral constructs from the data.

This study adds to the literature on discrimination in four important ways. First, we look for differential treatment in a school environment among real teachers who assess students' learning on a daily basis and who play a large role in determining the future education and earnings trajectory of children (Jackson, 2018 and Papageorge

et al. 2018). Other similar studies (Darley and Gross, 1983 and Baron et al., 1995) have been lab experiments with university students who do not affect any real-life students and who may be very unrepresentative of actual teachers, particularly in the way they assess students' aptitude and behavior.

Second, we examine a context where we do not expect discrimination to disappear with repeated interactions or where discriminators face a trade-off between discrimination and financial gain (Becker 1971). Unlike in activities such as job recruitment and housing search, teachers tend to have a lot of information about students, and market pressures to reduce bias in assessments are slight. In our experiment, the teachers in the sample would have had years of repeated interactions with the two student profiles we created.

A third contribution comes from the nature of the task that teachers had to do. All teachers were assessing the same child, not a pair of matched individuals as is done in many audit studies or correspondence studies (Betrand and Mullainathan, 2004 and Eriksson and Rooth, 2014). They were also making explicit assesssments of a student's observed aptitude and behavior, and not just predictions. In many experimental studies of discrimination in the labor and housing markets, we need to interpret differences in call-back rates as differences in the predictions that employers or landlords make about an individual's productivity or suitability as a worker or tenant, as the researchers do not get a chance to observe work or tenant conscientiousness in the experiments.

Finally, we extend the discrimination literature to social class, a variable that is rarely studied in the experimental literature in economics. Assumptions about social class, however, may partly underlie differential treatment across other social groups. For example, Bosch et al (2010) find that when emails from potential tenants to landlords reveal positive information about socio-economic status, the gap in landlord response rates narrows between potential tenants of Moroccan and native Spanish background. Similarly, gaps in landlord response rates to potential Black and White tenants disappear when email inquiries about rental units indicate high social class (Hanson and Hawley, 2011). Evidence from the United States suggests that teachers underestimate the potential of low-income students, not just of ethnic and racial minorities. Card and Giuliano (2016), for example, find that the introduction of universal screening with an IQ test substantially increased the fraction of economically disadvantaged students placed in gifted education programs compared to the previous system of allocating program slots, which relied on teacher and parent referral.

The next section describes the setting of the lab-in-the-field experiment – public primary schools in Metropolitan Lima, Peru. Section 3 details our experimental set up and measurement strategy, as well as the empirical specification we use to estimate

whether teachers assess differently a student when he is labeled as poor than when he is labeled as as middle class. Section 4 presents results, starting with a validation of the study design. The concluding section discusses the implications of our estimates for assessing whether poorer students face discrimination in schools.

2 Setting

We implemented our experiments with primary school teachers in public schools in Metropolitan Lima, Peru.[1] Within Peru, although it is not common among the population to refer to economic quintiles, the idea of five socio-economic strata (labeled A, B, C, D, and E from highest to lowest) is internalized and reflects a formal stratification of districts in Metropolitan Lima by the Statistics Office based on household consumption.

The public school system accounts for around 50 percent of the student population, though only 20 percent of schools (Minedu, 2014). While very few children from households in the top income quintile attend public schools (around 2 percent), 80 percent of children in the bottom quintile do (INEI, 2014).

There are no residence restrictions that determine which public schools children can attend, which means that in principle children from any district could enroll in any public school. In practice, most households choose schools in first grade (or kindergarten), leaving few slots available for transfer students in later grades. As a result, schools that are widely perceived as better quality schools are in high demand and often have to implement some selection criteria for prospective students. The Ministry of Education has established guidelines for these criteria, which include residence as one possible criterion, but it is ultimately the schools that determine their own selection process.[2] This set-up somewhat loosens the correlation between the socio-economic background of the neighborhood of the school and the background of the students, generating heterogeneity within schools.

[1] In Peru, primary education consists of six years of compulsory schooling for children 6 to 12 years of age.

[2] Proposed criteria include whether the child has a sibling in the institution and residence. Schools are forbidden to evaluate students and use the results of that evaluation as a condition for enrollment in kindergarten and first grade (when most of enrollment decisions are made). Schools have to publish their specific selection criteria before the start of the school year, but in practice it is unclear to what extent they actually do, and there is no enforcement mechanism to make sure rules are followed.

3 Experimental set-up and data

Because observational differences in how teachers assess scholastic aptitude and behavior among different socio-economic strata could reflect many things—true differences, statistical discrimination, or a stereotype that persists even in the presence of additional information, we set up a lab-in-the-field experiment among teachers in public primary schools to fix aptitude and behavior and identify the role that statistical discrimination or a persistent stereotype may play in teachers' assessment of students.

3.1 Sampling

To form our sample of teachers, we first sampled schools, starting with the universe of 1,016 public primary schools found in the Ministry of Education's school census in Metropolitan Lima and Callao.[3] We dropped 17 single-sex education schools and 3 schools that had closed between the school census and the start of our field work. From the remaining 996 schools, we excluded 293 schools that did not have at least 6 classes across grades 3, 4, and 5, as we aimed to sample 6 teachers per school in those grades. From the remaining 703 schools, we randomly selected 100 for the experimental sample. These schools are largely representative of public schools in Metropolitan Lima (Table 1, Panels A and B), however, they are slightly larger, with teachers that have more experience both in the profession and in the school that currently employs them, compared to the universe of public schools. The districts that sampled schools operate from are also largely representative of all districts in Metropolitan Lima (Table 2).

In each school, we randomly selected two teachers per grade from grades 3, 4, and 5 from a list of all teachers in the school and asked them to participate in the study during their lunch break or right after school. A pilot revealed that that the list of teachers provided by the Ministry was not up to date, and therefore selection of teachers for the study was done in two stages. First, prior to fieldwork, we randomly assigned treatment arms to grades within each school. Then, an enumerator visited each school twice to set up and monitor the experiment. During the first visit to the school, the enumerator would ask the school's director to provide a list of all teachers in grades 3 to 5. Typically there were two teachers per grade, but if one grade had more than two teachers, the enumerator would randomly select two for the study.

[3] The universe includes public schools run by the government. We did not include 48 privately managed public schools.

The enumerator would make an appointment for the same day with all six teachers selected for the study.

During the second visit to the school, enumerators presented the activities as a study of teacher evaluation methods, distributed tablets with preloaded videos, read the instructions, and signaled that they would be around to answer any questions when teachers had to fill out a self-administered questionnaire at the end of the experiment, which was also on the tablet (Appendix A presents the script in Spanish and English that teachers would have seen before encountering the videos and questionnaire).

3.2 Experimental groups

We stratified teachers by grade and assigned each to one of six experimental groups. Teachers in each group viewed a different set of videos on an individual tablet before they responded to questions about what they saw. We used the same child actor for the student and adult actor for the teacher in all videos.

Figure 1 illustrates our experimental design. Two groups of teachers viewed only a 22 minute video of a child, a nine-year old boy named Diego, showing him answering questions posed by a teacher (the performance video). The teacher in the video reads the questions from a flip chart, visible on the screen, and all questions are multiple choice. We worked with the Ministry of Education to design this oral exam following the 4th grade curriculum, which covered three subjects: Language, Math, and Science. In the video, the teacher is fully visible but only the back of Diego is visible, and the teacher does not react to Diego's responses (See Figure 2 for a still shot from the performance video).

There were two variants of this performance video, which were randomly assigned to teachers. Following Darley and Gross (1983), in one video, Diego's performance is inconsistent. He answers only 50 percent of the questions correctly. Sometimes he gets an easy question wrong; sometimes he gets a difficult question right. From time to time, he fidgets and loses focus. In a second variant that depicts the same teacher examining the same Diego, we follow Baron et al (1995) and made it obvious that Diego is performing well. He answers 75 percent of the questions correctly and more consistently focuses his attention on the questions.

Before watching this performance video, four groups of teachers first saw a four minute video that introduces Diego (the introductory video). He is shown playing in his neighborhood and inside his house, and a voice mentions Diego's age and his parents' education and occupation. Thus, it clearly identifies his social class. The videos show Diego from far away, and he does not face the camera.

Teachers were assigned to one of two introductory videos. In one video Diego comes from stratum D, which corresponds to the second quintile of the consumption distribution in Lima. In this variant, Diego's parents have completed primary school. His mother works as a domestic worker, while his father is a construction worker. The second variant of the introductory video portrays a child from stratum C, approximately the third quintile of of the consumption distribution or "the middle class." In this variant, Diego's parents have completed secondary education. His mother is a secretary, while his father is a local government employee. We chose these two adjacent quintiles to ensure that teachers would have seen children from these backgrounds in the same class in the same school. Contrasting more extreme cases (like the first and fourth or fifth quintiles) would have been less realistic, as it is less likely that those children would attend the same school. (See Figure 3 for stills from the introductory videos.)

3.3 Data and outcomes

After viewing the videos, teachers answered questions on their tablets. They were asked to rate the academic performance of the child and assess his behavior. Following this, they estimated the socio-economic background of the child and answered questions about their own family background, teaching experience, and exposure to students of different socio-economic backgrounds. The questionnaire took 58 minutes to complete on average, although this time includes the time teachers spent filling out an optional open-ended question where they were invited to share reflections on the video or on the questionnaire, which 70 percent of teachers filled out.

All six teachers sampled in a school had to do the entire exercise at the same time. There were seven schools where this did not happen, so we dropped 10 teachers that watched the video and answered the questionnaire after their peers did, as they might have discussed contents of the videos and questionnaire with their colleagues. We also dropped 6 teachers from a school that had technical issues with the tablets.[4]

Our main outcomes are the teachers' assessments of Diego's scholastic aptitude and behavior. We asked teachers to tell us what grade level they thought Diego was performing at, with response options running from first through sixth grade. For ease of interpretation, we have converted this variable into indicators corresponding to an assessment that Diego was performing at grade level (grade 4), above grade level

[4] Six teachers were unavailable at the time of appointment (mostly because they had to attend a last-minute request), and one teacher refused to participate. All 7 teachers were replaced before the experiment was introduced, and they are spread across all treatment arms.

(grades 5 or 6), or below grade level (grades 1-3). Since Diego's age was mentioned in the video, this might have anchored teachers' expectations about his grade and thus diluted variance in teachers' reports of Diego's grade level. Therefore, we also measure teachers' assessments of Diego's current scholastic aptitude by their responses to questions about whether they would recommend additional remedial support or placement in an advanced class for Diego.

To measure teachers' assessments of Diego's behavior, we used the same questions that appear in the Darley and Gross (1983) study.[5] Appendix Table 1 presents the categories used by Darley and Gross (1983) and Baron et al (1995) and their descriptions of the traits that respondents were supposed to rate, alongside the Spanish translations that we used with teachers in our sample. In our analysis below, we estimate effects both using the original categories of the earlier studies and using factor analysis to determine categories suggested by the data (Appendix Table 2 shows the overlap between the approaches).

3.4 Empirical specification

We make three main comparisons. First, we validate that the performance videos portrayed different levels of scholastic aptitude and behavior among the sample of teachers who did not watch an introductory video by comparing assessments of scholastic aptitude and behavior made by teachers who watched the video depicting inconsistent performance and teachers who watched the video where Diego is a high achiever. Second, among teachers who watched the video depicting inconsistent performance, we compare the assessments made by teachers who watched the introductory video depicting a poor background for Diego and teachers who watched the introductory video portraying a middle class background. Third, we make the same comparisons among teachers who watched the performance video in which Diego demonstrates high aptitude and high focus.

For each type of performance video and for each measure of a teacher i's assessment of scholastic aptitude or behavior, y_i, we estimate the following equation, where *Poor* takes a value of 1 when teachers first watched an introductory video depicting a home and neighborhood associated with a poor background:

$$y_i = \alpha + \beta Poor + \varepsilon_i \qquad (1)$$

[5] While the question wording was the same, we opted for a 5 point scale (instead of the 9-point scaled used by Darley and Gross (1983)) and gave teachers the option of not responding.

Since the randomization occurred at the individual level, we do not assume any clustering of the error term but rather estimate Huber-White robust standard errors. While with our experimental set-up, we cannot determine if a particular teacher exhibits any bias, the coefficient β indicates whether on average teachers assigned to a certain introductory video are using information in the video in their assessments of the performance video. A negative coefficient, for example, for the outcome grade-level would suggest that teachers assess aptitude to be worse when they think Diego comes from a poor background compared to when they think he comes from a middle class background.

Because there are so many items in the behavioral assessment, we first estimate β by computing the average effect size (AES) across all behavioral measures within a category (Kling et al., 2004 and Clingingsmith et al., 2009). We present estimates where categories are defined either by the original categories of Darley and Gross (1983) or by factor analysis. In the AES approach, each separate average treatment effect for each item in a category is scaled by the item's standard deviation, and the AES estimate is the average of these scaled treatment effects. In appendix tables, we also test robustness to an alternative strategy of simply first averaging items within a category and then using those averages as outcomes in our regressions.

4 Results

Tables 3 and 4 and Figures 4 and 5 present data from all samples to validate the study design. Tables 5-9 present estimates of the difference in the assessments of scholastic aptitude, behavior, and education potential made by teachers who first watched the introductory video depicting a poor background and teachers who first watched the video showing a middle-class background, for both the inconsistent and high performance videos. In Table 10, we check whether estimated effects vary by teacher, school, and neighborhood characteristics.

4.1 Validation of study design

Table 3 presents descriptive statistics of teachers by experimental group, which suggest that the randomization of teachers generated well-balanced groups. Teachers in the experimental sample are mostly middle-aged female civil-service teachers with a little over 20 years of teaching experience. This table presents the p-values for pair-wise t-tests for the experimental arms that will be compared to assess differences between the teachers that watched the poor introductory video and the teachers that watched the middle-class introductory video. Appendix Table 3 presents all possible pair-wise t-tests.

Responses from the teachers who only watched one of the two performance videos help establish whether teachers could discern a difference in the scholastic aptitude and behavior across the two performance variants (Table 4). First, both sets of teachers thought that the oral exam given to Diego was equally difficult at around a late third grade level, which suggests they were paying attention to content, as the same exam questions appeared in both videos and questions were taken from the official fourth grade syllabus. Teachers assigned to the high performance variant perceived that Diego answered more easy, moderate, and difficult questions correctly than the teachers assigned to the video depicting inconsistent performance; they also gave him higher behavioral assessment scores for the Darley and Gross (1983) categories. The majority of teachers who did not see an introductory video but only watched a performance video thought that Diego would eventually attend some tertiary education, although more teachers in the sample who watched the high performance video (92 percent) thought so relative to the inconsistent performance (74 percent). Given the average tertiary completion rate among 20-35 year olds in Lima in 2014 was 81 percent and 64 percent, respectively, for the top two consumption quintiles, these expectations suggest that without any explicit socio-economic labelling from an introductory video, teachers may be consciously or unconsciously assuming a relatively high socio-economic status for Diego.

To validate our portrayals of social class, we asked the sample that did see an introductory video which stratum they would place Diego in. Figure 4 demonstrates that teachers correctly inferred the socio-economic background of the child. Teachers who saw the introductory video depicting a poor background assessed Diego to be poorer than teachers who saw the video showing a middle-class background. Teachers also placed the respective Diegos in the correct stratum as well. The most common response for teachers assigned to the introductory video showing a poor background was Stratum D, while the most common response for teachers assigned to the video portraying a middle-class background was Stratum C. Although adjacent, distinctions between these strata are clearly meaningful for teachers and they can take clues from the child's physical surroundings depicted in the videos to correctly identify a student's background. This is not surprising as teachers report having the most exposure in their classrooms to children in the two strata shown in the videos (Figure 5).

4.2 Experimental results

4.2.1 Inconsistent performance

When Diego's performance is inconsistent, teachers assess his performance to be lower when the introductory video they watch depicts a home environment from the lower stratum (Table 5). On average, teachers watching the introductory video depicting a poor background assess a performance level that is 0.24 of a year (or just over 2 months) behind the level assessed by teachers who had seen the introductory video depicting a middle-class background (Column 1). When we define performing at grade level as a grade level assessment of 4, then teachers watching the poor introductory video were a significant 14.4 percentage points more likely to report that poor Diego performed below grade level (Column 2) compared to middle class Diego and 13.4 percentage points less likely to report he was performing at grade level (Column 3). Teachers in both experimental groups were equally unlikely to assess he was performing above grade level (Column 4). Consistent with these grade-level assessments, teachers who watched the introductory video portraying a poor background were 14.4 percentage points more likely to suggest that Diego needed additional support (Column 5) and equally unlikely to recommend placing him in an advanced class (Column 6), compared with teachers who had first watched the introductory video showing a middle-class background.

Also apparent from a comparison between Table 5 and Table 1 is that the assessments of teachers who received no information on Diego's socio-economic status appear more like the assessments of teachers who watched the introductory video depicting a middle class background.

Table 6 presents differences in teachers' behavioral ratings of Diego. Columns 1-5 report effects for the original Darley and Gross (1983) categories, while Columns 6-10 report corresponding effects for the categories suggested by factor analysis, where all estimates reflect average effect size estimates within each behavioral category. (Appendix Table 4 presents differences for each behavioral assessment item separately, and Appendix Table 5 presents results when items within a category are first averaged before they enter the regression.) While the differences suggest that teachers are more critical in their behavioral assessment of Diego when they first watch a introductory video depicting a poor background, these differences are not statistically significant, with the exception of cognitive ability category in the Darley and Gross (1983) groups, which is consistent with their lower grade level assessment, and the sociability and expression category in the factor analysis groupings.

Consistent with assessing Diego's current scholastic aptitude as lower when he is introduced as a child from a poor background, teachers also have lower expectations

for the highest level of education he will complete (Table 7). Teachers are 18.6 percentage points less likely to report that they expect him to complete college (Column 3) and correspondingly 17.5 percentage points more likely to report an expectation that he will complete only his secondary education, compared to teachers who watched the introductory video where Diego came from a middle-class background. At the bottom of the table, the average expectations of the teachers who watched no introductory background video and only saw a performance video were much higher than when they saw any introductory video. This could suggest that without an explicit prompt to think about the student's social class, teachers might have had as their reference a child from a much wealthier background, possibly more similar to their own.

4.2.2 High performance

When the performance video that teachers watch depicts a more consistently focused Diego who answers most of the questions correctly, they modify their assessments of scholastic aptitude in favor of the poorer Diego (Table 8). Teachers who saw the poor introductory video now provide a grade level assessment that is higher than teachers who saw the middle-class introductory video, although this difference is not statistically significant (Column 1). They are 11.8 percentage points significantly less likely to report that he is performing at grade level (Column 3) and correspondingly 14.5 percentage points more likely to report that he is performing above grade level (Column 4), compared with teachers who saw the same performance video but who first watched an introductory video depicting a middle-class background for Diego. That is, they consider Diego to be an exceptional student in terms of his scholastic aptitude when they think he is poor. While a quarter to a third of teachers think that Diego needs additional support and another quarter to a third recommend he be placed in an advanced class, the introductory videos portraying Diego's socio-economic background do not affect this assessment (Columns 5 and 6).

On the other hand, teachers watching the high-performance video are now harsher in their assessments of poorer Diego's behavior (Table 9) relative to their assessments of middle class Diego. When the behavioral assessment items are classified according to the original Darley and Gross (1983) groupings, teachers think Diego is much less motivated (Column 2) and emotionally mature (Column 4) when they watch the introductory video with a poor background compared to teachers who watched the introductory video with a middle-class background. The factor analysis-based groupings are consistent: poorer Diego is assessed as less motivated, less pleasant or assertive, and less likely to have a positive attitude. Consistent with this pattern of results, Appendix Table 4 presents differences for each behavioral item

separately, and Appendix Table 5 presents estimates items are averaged within each category before they enter the regression. A comparison of the bottom rows of Table 6 and Table 9 suggests that teachers who watch the high-performance video rate Diego's behavior as much better than teachers who watch the inconsistent performance video, both among the sample of teachers who saw only a performace video and among the sample that also saw an introductory video that portrayed Diego's social class. This positive difference in ratings across the performance variants, however, is much lower among teachers who first watched an introductory video depicting Diego as poor.

Columns 4-6 of Table 7 present teachers' expectations for future educational attainment. While the bottom rows suggest that expectations are higher among teachers watching the high-performance video than among teachers who saw the inconsistent performance, the gap in expectations observed in the inconsistent case remains in the high performance case. Teachers are 14.6 percentage points less likely to expect that Diego will complete college when they watch an introductory video depicting his background as poor than when the video suggests a middle-class background (Column 3).

4.2.3 Heterogeneity

In Table 10 we check whether the results on teachers' assessment of scholastic aptitude vary by teacher, school, and district characteristics. We interact the indicator for whether a teacher received an introductory video portraying a poor background with a teacher, school, or district characteristic and enter a main effect for the characteristic in our base regression. Some studies have found less bias when potential discriminators share more traits with the marginalized group (for example, when teachers are the same race as the students, as in Gershenson, Holt, and Papageorge, 2016) or have more experience making assessments (Hanna and Linden 2012) or regular exposure to the group (Finseraas and Kotsadam, 2017; Rao, 2019; Beaman et al, 2009). We find no evidence of this. We proxy a teachers' socio-economic background with the highest level of education she reports for her father. For both the inconsistent (Panel A Column 1) and high performance (Panel A Column 4) cases, we see no differences across teachers that came from different socio-economic backgrounds. Likewise, our estimates do not vary with teachers' experience teaching in primary school (Panel A Columns 2 and 5).

We use the poverty level of the school's district – specifically whether the district has a poverty rate that is above the median rate – to proxy for teachers' exposure to students from poor backgrounds. Increases in exposure do not change our results (Panel B Columns 1 and 4). Similarly, the experiment yielded statistically

indistinguishable results for teachers from high performing schools and teachers from low performing schools as measured by a school's average performance on standardized language and math tests (Panel B Columns 2,3,5, and 6).

5 Discussion and conclusion

We find that even when students display identical academic performance and behavior (as we constructed in our experiments), teachers exhibit bias when determining whether a student is performing at grade-level.

If we analyzed our results only on teachers' assessments of scholastic aptitude and their expectations for future educational attainment, we might conclude that our results are consistent with a model of statistical discrimination, as clearer information served to eliminate bias. When our fictitious student, Diego, exhibited inconsistent performance on the exam, teachers were far more likely to judge his performance as below grade level and to suggest additional support if they had watched the introductory video portraying a poor background than if they had watched the introductory video that portrayed a middle class background. Diego provided a noisy signal of his scholastic aptitude and teachers might have placed more weight on their prior expectation on the aptitude of poor children when making their assessment if they first viewed a video depicting Diego's background as poor. Once Diego provided a clearer signal of aptitude in the high performance video, teachers could put more weight on what they observed in the video than on their prior, and the poorer variant of Diego was not penalized as much in their assessments (in fact, teachers' observed bias seems to go in the other direction, in favor of poorer Diego, if they saw a high performance video). In both the inconsistent and the high performance cases, low expectations about future education attainment would be completely in line with observed data on socio-economic gradients in high school graduation and college attendance both in Peru (Sanchez and Singh, 2018; INEI 2014) and other low- and middle-income countries, even for students of identical scholastic aptitude (Das, Singh, and Yi Chang, 2020).

While the pattern of teachers' assessments of the child's scholastic aptitude are consistent with models of statistical discrimination in which teachers use a student's social background to help assess aptitude when signals are noisy but rely less on background when signals are clearer, such a model cannot explain why a gap emerges in teachers' assessment of behavior when they are given very clear signal of high scholastic aptitude and focused behavior. This pattern of results is consistent with models of confirmation bias and selective attention, where how we choose to focus our limited attention, interpret data, and update our priors are themselves choices, possibly subconscious ones.

When teachers saw the inconsistent performance video, where Diego's actions suggested a loss of focus from time to time, the introductory video they saw did not affect their behavioral assessment. The behaviour of poor Diego and middle class Diego was rated the same. When, however, teachers saw the high performance video where Diego fidgeted much less and exhibited greater focus, the behaviour of poor Diego and middle class Diego was not rated the same. They viewed poor Diego as performing significantly worse on multiple dimensions. This finding is consistent with the hypothesis from social psychology that our updating itself is a biased process: we pay attention to information selectively. In this case, knowledge of the child's poor socio-economic background led to a negative assessment of his behavior.

To put this another way, when the child's responses to exam questions easily permitted a negative assessment of his scholastic aptitude, teachers judged the child's academic ability more harshly when he was portrayed as poor than when he was portrayed as middle class. When the child's high rate of correct responses to exam questions made it more difficult to form a negative assessment of his academic ability, teachers shifted their harsher assessment of the poor child to his behavior.

This switching of a negative assessment (from aptitude to behavior) across the two performance conditions could still be consistent with statistical discrimination if teachers have a specific joint distribution of aptitude and behavior in mind for different social classes. As Heckman (1998) has argued in a critique of audit studies, if the joint distribution of two attributes meaningful for assessment differs across two groups and in an experimental context we equalize one attribute, then individuals making assessments will estimate different levels for the other attribute for the two groups. In this case, the switching of teachers' negative assessment from aptitude to behavior across the inconsistent and high performance variants would be consistent with statistical discrimination (without any bias in updating) if teachers believe that aptitude and behavior are positively correlated for middle-class children but less correlated for poorer children.

Regardless of whether this observed differential treatment comes from rational Bayesian updating or subconscious selective attention, it would be difficult to find a school system that would endorse using social class to assess students. Grades would be considered invalid, just as differential treatment in the workplace for workers exhibiting the same productivity would be considered illegal.

One limitation of our experiments is the one shot nature of the assessment. Because we implemented a lab-in-the-field experiment, we do not observe a long repeated interaction between a real teacher and real students. Teachers may learn about aptitude, behavior, and potential over time, consistent with Altonji & Pierret (2001), Botelho et al (2015) and Hanna and Linden (2012), and get more accurate estimates of ability.

On the other hand, even an initial inaccurate assessment by teachers can have lasting effects, particularly for stigmatized groups (Rosenthal and Jacobson, 1968; Jussim and Harber, 2005; and Papageorge et al., 2018). Spillover effects on children's own perception of self could have real consequences on their academic performance as suggested by the literature on stereotype threat (Steele and Aronson, 1995; Hoff and Pandey, 2006; Glover, Pallais, and Pariente, 2017; Lavy and Sand, 2015).

Because teachers' expectations have significant implications for students' later educational trajectories, policies that address these biases among teachers in the early grades of primary school may help decrease the socio-economic gradients observed in later test scores and educational attainment. Interventions that increase contact or exposure to groups facing discrimination have successfully changed mindsets (Finseraas and Kotsadam, 2017; Rao, 2019; Beaman et al, 2009), but teachers tend to have considerable exposure to their students already. Simply making teachers aware of their biases, however, or making certain pedagogical approaches more salient may help. In Italy, teachers increased grades assigned to immigrant students if they were made aware of their biases (through their scores on implicit association tests) (Alesina et al, 2018). In the United States, students whose teachers were trained to have a more empathetic approach to discipline were half as likely to be suspended during the school year compared to their peers in control classrooms (Okonofua et al, 2016).

It is also possible that social-psychological interventions targeting students are required to counteract any lower expectations they face from teachers. Self-affirmation exercises, for example, in which students write about something meaningful to them or something that makes them feel proud have been shown to reduce the achievement gap between minority and majority students in the United States, an effect that persists to the next school year (Cohen et al., 2006 and Cohen et al., 2009). Similarly, in a nationally representative experiment in the US, an online growth mindset module that taught students intellectual abilities are not fixed but can be developed improved the grades of lower-achieving students (Yeager et al, 2016), a finding that has been replicated in Norway (Bettinger et al, 2018), Peru (Outes-Leon et al, 2020), and Turkey (Alan et al, 2019).

While the current study used a lab-in-the-field experiment to demonstrate that teachers do indeed use a student's socio-economic background as a heuristic for assessing scholastic aptitude and behavior, future studies would ideally combine these kinds of metrics of bias, real assessments made by teachers and independent parties, and randomized interventions to quantify the impact of teacher bias on student learning and to identify successful strategies for either decreasing bias or mitigating its impact on students' educational trajectories.

6 References

Ahmed, A. M., L. Andersson, and M. Hammarstedt. 2010. "Can Discrimination in the Housing Market Be Reduced by Increasing the Information about the Aplicants?". *Land Economics*, 86(1):79-90

Alan, S., S. Ertac, and I. Mumcu. 2018. "Gender Stereotypes in the Classroom and Effects on Achievement". *The Review of Economics and Statistics*, 100(5): 876-890. DOI:10.1162/rest_a_00756

Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti. 2018. "Reveanling Stereotypes: Evidence from Immigrants in Schools". NBER Working Paper, DOI 10.3386/w25333

Alotonju, J. G., and C. R. Pierrer. 2001. "Employer Learning and Statistical Discrimination". *The Quarterly Journal of Econonmics*, 116(1):313-350, https://doi.org/10.1162/003355301556329

Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment". *American Economic Review*, 92(5):1535-1558, DOI: 10.1257/000282802762024629

Baron, R. M., L. Albright, and T. E. Malloy. 1995. "Effects of Behavioral and Social Class Information on Social Judgement". *Personality and Social Psychology Bulleting*, 21(4):308-315, https://doi.org/10.1177/0146167295214001

Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande, and P. Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias?" *The Quarterly Journal of Economics*, 124(4):1497-1540, https://doi.org/10.1162/qjec.2009.124.4.1497

Becker, G. S. 1957. The Economics of Discrimination. Economic Research Studies, 2nd Edition

Bertrand, M., and S. Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination". *American Economic Review*, 94(4):991-1013, DOI: 10.1257/0002828042002561

Bertrand, M. and E. Duflo. 2017. "Field Experiments on Discrimination". Handbook of Economic Field Experiments, 1:309-393, https://doi.org/10.1016/bs.hefe.2016.08.004

Bettinger, E., S. Ludvigsen, M. Rege, I. F. Solli, and D. Yeager. 2018. "Increasing Perseverance in Math: Evidence from a Field Experiment in Norway". *Journal of Economic Behavior and Organization*, 146:1-15, https://doi.org/10.1016/j.jebo.2017.11.032

Bornstein, M. H., and D. L. Putnick. 2012. "Cognitive and Socioemotional Caregiving in Developing Countries". *Child Development*, 83(1):46-61, https://doi.org/10.1111/j.1467-8624.2011.01673.x

Bosch, M., M. A. Carnero, and L. Farre. 2010. "Information and discrimination in the rental housing market: Evidence from a field experiment". 2010. *Regional Science and Urban Economics*, 40:11-19, doi:10.1016/j.regsciurbeco.2009.11.001

Botelho, F., R. A. Madeira, and M. A. Rangel. 2015. "Racial Discrimination in Grading: Evidence from Brazil". *American Economic Journal: Applied Economics*, 7(4): 37-52, http://dx.doi.org/10.1257/app.20140352

Card, D., and L. Giuliano. 2016. "Universal Screening Increases the Representation of Low-Income and Minority Students in Gifted Education". *PNAS*, 113(48):13678-13683, www.pnas.org/cgi/doi/10.1073/pnas.1605043113

Carlana, M. 2019. "Implicit Stereotypes: Evidence from Teachers' Gender Bias". *The Quarterly Journal of Economics*, 134(3):1163-1224, https://doi.org/10.1093/qje/qjz008

Chetty, R., and N. Hendren. 2018. "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects". *The Quarterly Journal of Economics*, 133(3):1107-1162, doi:10.1093/qje/qjy007

Clingingsmith, D., A. I. Khawaja, and M. Kremer. 2009. "Estimating the Impact of the Hajj: Religion and Tolerance in Islam's Global Gathering". *The Quarterly Journal of Economics,* 124(3):1133-1170, https://doi.org/10.1162/qjec.2009.124.3.1133

Cohen, G. L., J. Garcia, V. Purdie-Vaughns, N. Apfel, and P. Brzustoski. 2009. "Recursive Processes in Self-Affirmation: Intervening to Close the Minority Achievement Gap". *Science*, 324:400-403

Cohen, G. L., J. Garcia, N. Apfel, and A. Master. 2006. "Reducing the Racial Achievement Gap: A Social-Psychological Intervention". *Science*, 313:1307-1310

Darley, J. M., and P. H. Gross. 1983. "A Hypothesis-Confirming Bias in Labeling Effects". *Journal of Personality and Social Psychology*, 44(1):20-33

Das, J., A. Singh, and A. Yi Chang. 2020. "Test Scores and Educational Opportunities: Panel Evidence from Five Developing Countries." RISE Working paper 20/040

Edmonds, E. V., and N. Schady. 2012. "Poverty Alleviation and Child Labor". *American Economic Journal: Economic Policy*, 4(4):100-124, DOI: 10.1257/pol.4.4.100

Eriksson, S. and D-O Rooth. 2014. "Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment". *The American Economic Review*, 104(3):1014-1039, https://www.jstor.org/stable/42920727

Finseraas, H. and A. Kotsadam. 2017. "Does Personal Contact with Ethnic Minorities Affect Anti-immigrant Sentiments? Evidence from a Field Experiment". *European Journal of Political Research*, 56(3):703-722

Fryer Jr., R. G., and S. D. Levitt. 2004. "Understanding the Black-White Test Score Gap in the First Two Years of School". *The Review of Economics and Statistics*, 86(2):447-464

Glover, D., A. Pallais, and W. Pariente. 2017. "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores". *The Quarterly Journal of Economics*, 132(3):1219-1260, https://doi.org/10.1093/qje/qjx006

Gneezy, U., J. List, and M. K. Price. 2012. "Toward an Understanding of Why Peoople Discriminate: Evidence from a Series of Natural Field Experiments", NBER Working Paper 17855, http://www.nber.org/papers/w17855

Hanna, R. N., and L. L. Linden. 2012. "Discrimination in Grading". *American Economic Journal: Economic* Policy, 4(4):146-168, http://dx.doi.org/10.1257/pol.4.4.146.

Hanson, A. and Z. Hawley. 2011. "Do Landlords Discriminate in the Rental Housing Market? Evidence from an Internet Field Experiment in US Cities". *Journal of Urban Economics*, 70:99-114, doi:10.1016/j.jue.2011.02.003

Heckman, J. J. 1998. "Detecting Discrimination". *The Journal of Economic Perspectives*, 12(2):101-116, http://www.jstor.org/stable/2646964

Hoff, K. and P. Pandey. 2006. "Discrimination, Social Identity, and Durable Inequalities". *The American Economic Review*, 96(2):206-211, https://www.jstor.org/stable/30034643

INEI, Instituto Nacional de Estadística e Informática. 2014. Perú: Encuesta Nacional de Hogares 2014. Lima: INEI.

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

Jensen, R. 2010. "The (Perceived) Returns to Education and the Demand for Schooling". *The Quarterly Journal of Economics*, 125(2):515-548, https://doi.org/10.1162/qjec.2010.125.2.515

Jussim, L., and K. D. Harber. 2005. "Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies". *Personality and Social Psychology Review*, 9(2):131-155, https://doi.org/10.1207/s15327957pspr0902_3

Kahneman, D. 2011. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011

Kling, J., J. Liebman, L. Katz, and L. Sanbonmatsu. 2004. "Moving to Opportunity and Tranquility: Neighborhood Effects on Adult Economic Self-Sufficiency and Health from a Randomized Housing Voucher Experiment". Faculty Research Working Paper RWP04-035

Lavy, V. and E. Sand. 2015. "On the Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers' Stereotypical Biases". NBER Working Paper 20909, http://www.nber.org/papers/w20909

Lundberg, S. J., and R. Startz. 1983. "Private Discrimination and Social Intervention in Competitive Labor Market". *The American Economic Review*, 73(3):340-347, https://www.jstor.org/stable/1808117

Minedu, Ministerio de Educación del Perú. 2014. Censo Escolar 2014. Lima: MINEDU

Okonofua, J. A., D. Paunesku, and G. M. Walton. 2016. "Brief Intervention to Encourage Empathic Discipline Cuts Suspension Rates in Half among Adolescents". *Proceedings of the National Academy of Sciences*, 113(19):5221-5226, www.pnas.org/cgi/doi/10.1073/pnas.1523698113

Outes-Leon, I., A. Sanchez, and R. Vakis. 2020. "The Power of Believing You Can Get Smarter. The Impact of a Growth-Mindset Intervention on Academic Achievement in Peru". Policy Research Working Paper 9141

Papageorge, N. W., S. Gershenson, and K. M. Kang. 2020. "Teacher Expectations Matter". *Review of Economics and Statistics*, 102(2):234-251, https://doi.org/10.1162/rest_a_00838

Rabin, M., and J. L. Schrag. 1999. "First Impressions Matter: A Model of Confirmatory Bias". *The Quarterly Journal of Economics,* 114(1):37-82, https://doi.org/10.1162/003355399555945

Rao, Gautam. 2019. "Familiarity Does Not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools." *American Economic Review*. 109(3). 774-809.

Rosenthal, R. and L. Jacobson. 1968. "Pygmalion in the Classroom". *The Urban Review*. 3:16-20, https://doi.org/10.1007/BF02322211

Rubinstein, R. S., L. Jussim, and S. T. Stevens. 2018. "Reliance on Individuating Information and Stereotypes in Implicit and Explicit Person Perception". *Journal of Experimental Social Psychology,* 75:54-70, https://doi.org/10.1016/j.jesp.2017.11.009

Steele, C. M., and J. Aronson. 1995. "Stereotype Threat and the Intellectual Test Performance of African Americans". *Journal of Personality and Social Psychology,* 69(5):797-811, https://doi.org/10.1037/0022-3514.69.5.797

Yeager, D.S., P. Hanselman, G. M. Walton, et. Al. 2019. "A National Experiment Reveals Where a Growth Mindset Improves Achievement". *Nature,* 573:364-369, https://doi.org/10.1038/s41586-019-1466-y

7 Figures

Figure 1: Experimental design

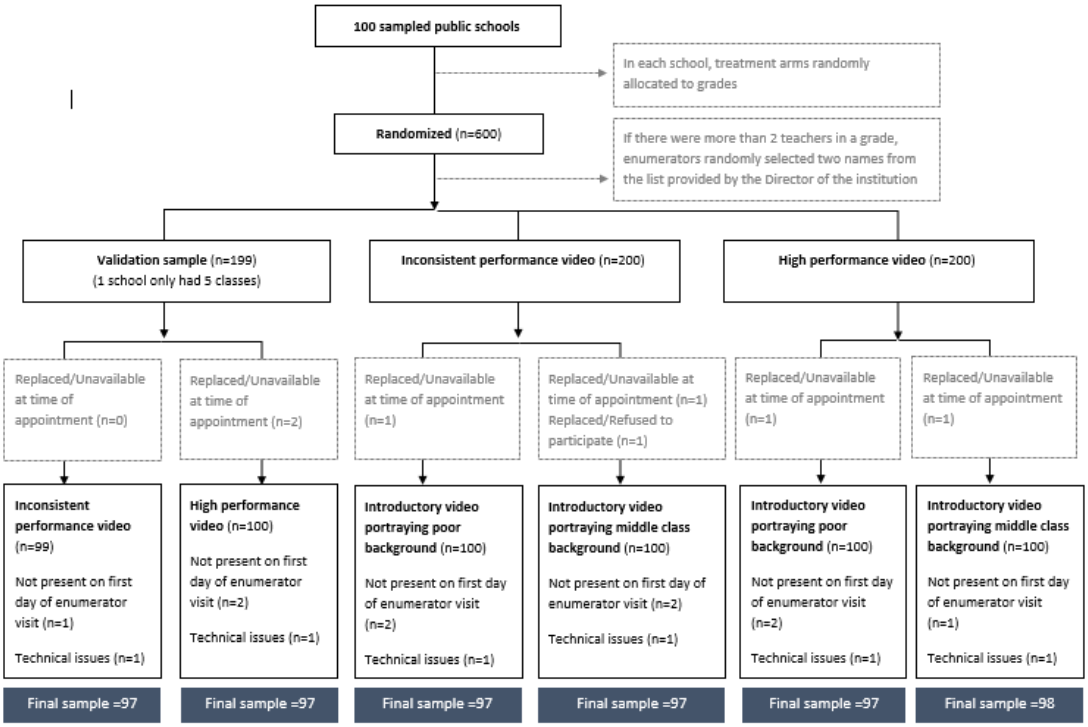


Figure 2: Still shot from performance video

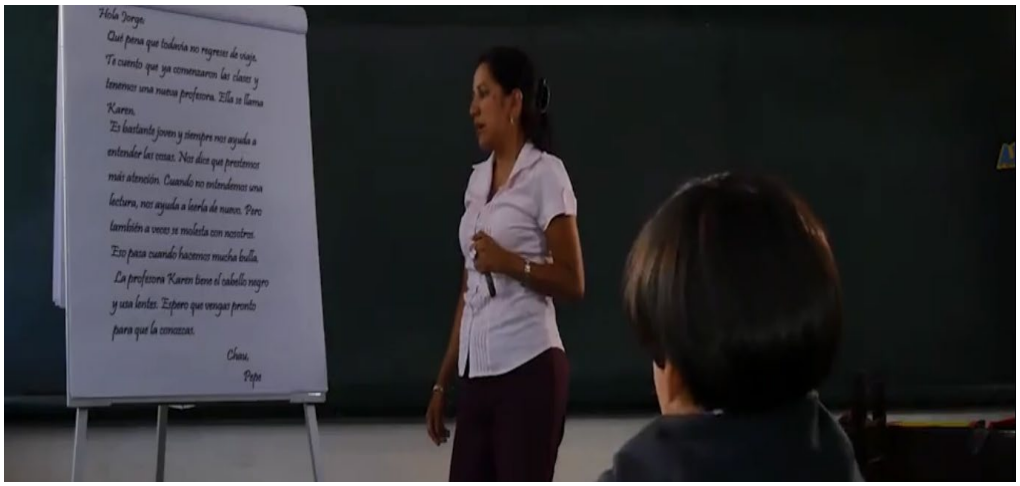
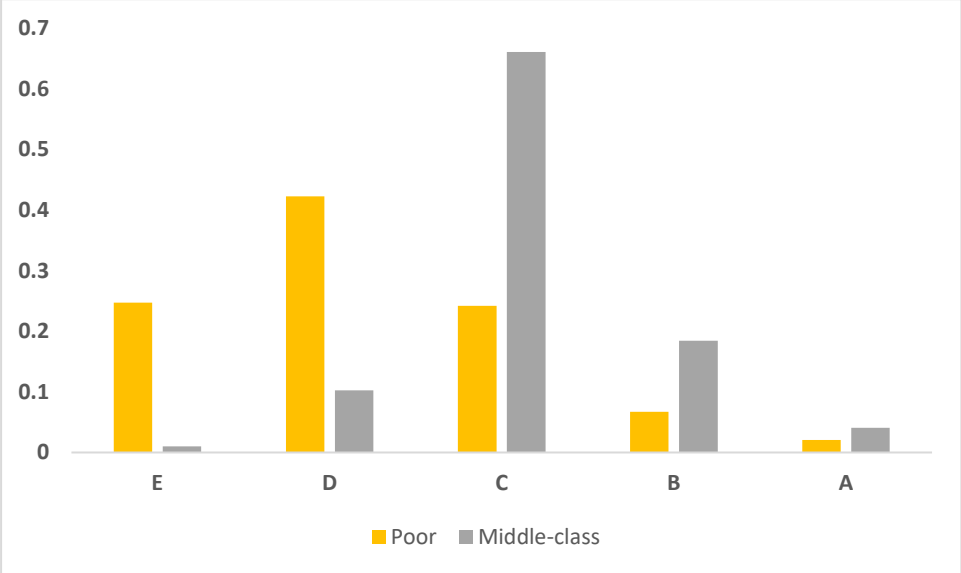


Figure 3: Still shots from introductory videos

Panel A: Poor background

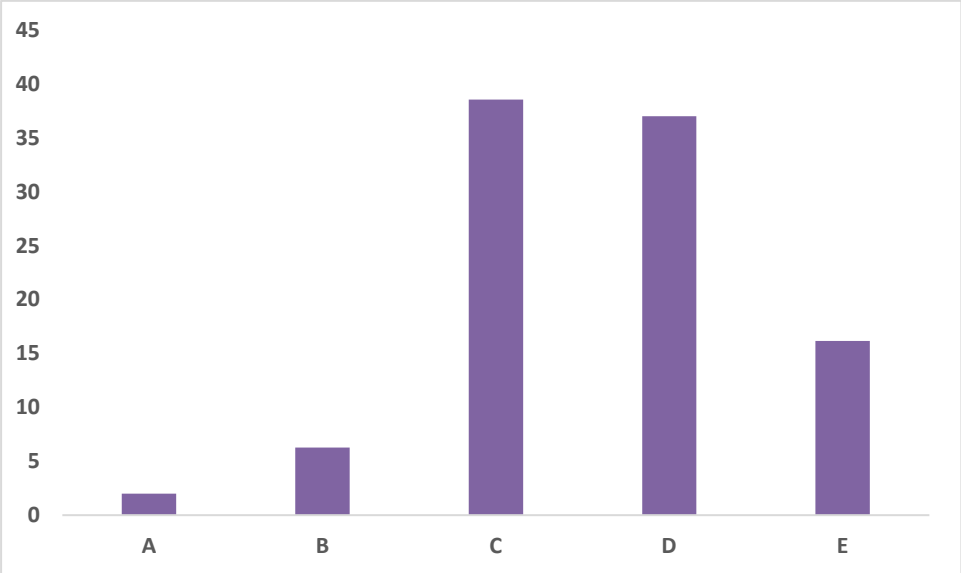
Panel B: Middle class background

Figure 4: Teachers' perception of child's socio-economic stratum (A-E), by introductory video



Note: The poor introductory video was meant to depict a child from stratum D, and middle-class introductory video was meant to depict a child from stratum C. This chart graphs teachers' responses to a question that asked them to slot the child in the video they watched into a stratum.

Figure 5: Teachers' perception of their own students' socio-economic stratum (A-E)



Note: This chart graphs teachers' responses to a question that asked them to identify the socio-economic stratum from which most of their students came from.

8 Tables

Table 1: A comparison of sampled schools and the universe of schools in Metropolitan Lima

	Metropolitan Lima	Sample	p-value of difference between (1) and (2)
	(1)	(2)	(3)
Panel A: Students			
Total in school	424.67	512.44	0.000
	(8.56)	(22.90)	
Percent female	0.80	0.79	0.348
	0.00	0.01	
Total in 3rd grade	69.42	85.04	0.000
	(1.40)	(3.86)	
Total in 4th grade	76.38	92.18	0.000
	(1.56)	(3.97)	
Total in 5th grade	74.86	90.49	0.001
	(1.57)	(4.34)	
Total grade retention	13.46	14.92	0.331
	(0.51)	(1.40)	
Panel B: Teachers			
Total in school	16.68	19.93	0.001
	(0.31)	(0.87)	
Percent female	0.80	0.79	0.348
	(0.00)	(0.01)	
Total in 3rd grade	2.66	3.16	0.001
	(0.05)	(0.13)	
Total in 4th grade	2.85	3.36	0.001
	(0.05)	(0.14)	
Total in 5th grade	2.77	3.34	0.000
	(0.05)	(0.14)	
Tenure in teaching (yrs)	14.26	16.86	0.001
	(0.27)	(0.75)	
Tenure in school (yrs)	15.25	18.13	0.002
	(0.29)	(0.86)	
Total number of schools	1,016	100	

Notes: Data source is the School Census 2014. Standard errors appear in parentheses.

Table 2: A comparison of districts of sampled schools and the universe of schools

	Metropolitan Lima	Sample	p-value of difference between (1) and (2)
	(1)	(2)	(3)
Panel A: Distribution of economic status			
Percentage poor households	13.55	13.88	0.871
	(1.36)	(1.51)	
Percentage households in strata A or B	34.60	31.53	0.695
	(5.72)	(5.72)	
Percentage households in stratum C	30.90	31.97	0.961
	(12.82)	(17.03)	
Percentage households in stratum D	9.90	7.13	0.738
	(5.67)	(5.70)	
Percentage households in stratum E	18.04	16.82	0.818
	(3.58)	(3.87)	
Household size	3.76	3.84	0.250
	(0.05)	(0.05)	
Years of education of household head	11.24	11.08	0.605
	(0.21)	(0.23)	
Panel B: Characteristics of homes			
Percentage of homes rented	19.66	20.11	0.839
	(1.44)	(1.64)	
Percentage with brick/concrete walls	80.81	81.74	0.761
	(2.07)	(2.27)	
Percentage with cement floor	43.81	46.06	0.648
	(3.30)	(3.65)	
Percentage with piped water in home	80.43	83.08	0.503
	(2.87)	(2.70)	
Percentage with refridgerator	74.54	73.76	0.786
	(1.93)	(2.11)	
Percentage with internet	39.40	38.27	0.806
	(3.08)	(3.40)	
Number of districts	50	37	

Notes: Data source is SISFOH, Peru. Standard errors appear in parentheses.

Table 3: Balance of teacher characteristics across experimental groups

	Inconsistent performance			High performance				
	Poor	Middle-class	No-Intro	Poor	Middle-class	No-Intro	p-value of (1)-(2)	p-value of (4)-(5)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Age	47.99	48.40	49.13	48.42	47.56	48.28	0.70	0.48
	(0.80)	(0.80)	(0.80)	(0.80)	(0.80)	(0.80)		
Male	0.22	0.16	0.23	0.18	0.12	0.23	0.36	0.30
	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)		
Third grade	0.35	0.29	0.34	0.18	0.32	0.32	0.36	0.92
	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)	(0.05)		
Fourth grade	0.33	0.34	0.31	0.31	0.40	0.33	0.88	0.60
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)		
Fifth grade	0.31	0.37	0.35	0.36	0.28	0.35	0.37	0.61
	(0.95)	(0.95)	(0.95)	(0.05)	(0.95)	(0.95)		
Years of experience	20.29	21.52	22.07	0.31	20.87	20.79	0.27	0.85
	(0.16)	(0.29)	(0.20)	(0.95)	(0.16)	(0.20)		
Graduate degree	0.34	0.34	0.32	20.64	0.28	0.39	1.00	0.50
	(0.05)	(0.05)	(0.05)	(0.15)	(0.05)	(0.05)		
Father with higher education	0.28	0.30	0.24	0.32	0.26	0.28	0.75	0.90
	(0.05)	(0.05)	(0.04)	(0.05)	(0.04)	(0.05)		
Mother with higher education	0.14	0.16	0.12	0.25	0.21	0.21	0.69	0.49
	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)		
Number of teachers	97	97	97	97	98	97		

Note: Standard errors in parentheses.

Table 4: The sample with no introductory video

	Inconsistent performance	High performance	p-value (1)-(2)
	(2)	(2)	(3)
Estimated grade level of exam	3.93	3.78	0.20
Estimated percentage of easy questions correct	56.76	72.00	0.00
Estimated percentage of moderate questions correct	36.57	51.99	0.00
Estimated percentage of difficult questions correct	24.61	36.95	0.00
Mean score working habits	3.46	4.10	0.00
Mean score motivation	3.08	3.74	0.00
Mean score sociability	3.76	4.06	0.02
Mean score emotional maturity	3.61	4.27	0.00
Means score cognitive ability	3.34	4.06	0.00
Expected educational attainment: primary	0.05	0.04	0.73
Expected educational attainment: secondary	0.21	0.04	0.00
Expected educational attainment: post-secondary	0.74	0.92	0.00
Number of observations	97	97	

Table 5: The impact of knowing social class on scholastic aptitude assessment | Inconsistent performance

		% reporting child is:			% suggesting child needs:	
	Predicted grade level (years)	Below grade level	At grade level	Above grade level	Additional support	Advanced placement
	(1)	(2)	(3)	(4)	(5)	(6)
Poor Diego	-0.24	0.14	-0.13	-0.01	0.14	-0.02
	(0.11)	(0.07)	(0.06)	(0.04)	(0.06)	(0.03)
Number of observations	194	194	194	194	194	194
Mean Poor Diego	3.22	0.71	0.22	0.07	0.78	0.03
	(0.08)	(0.05)	(0.04)	(0.03)	(0.04)	(0.02)
Mean Middle Class Diego	3.45	0.57	0.35	0.08	0.64	0.05
	(0.08)	(0.05)	(0.05)	(0.03)	(0.05)	(0.02)
Mean No-Intro	3.48	0.54	0.33	0.13	0.60	0.07
	(0.10)	(0.05)	(0.05)	(0.03)	(0.05)	(0.03)

Note: Robust standard errors are in parentheses. All regressions contain a constant term not reported here. *Poor Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as poor. *Middle Class Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as iddle class. The *no-intro* sample of teachers watched no introductory video and watched only a video portraying performance on an exam.

Table 6: The impact of knowing social class on behavior assessment (AES estimates) | Inconsistent performance

	Darley and Gross (1983) grouping					Factor analysis grouping				
	Work habits	Motivation	Sociability	Emotional maturity	Cognitive ability	Work habits	Motivation	Sociability and expression	Confidence and cooperation	Cognitive ability and learning attitude
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Poor Diego	-0.04	-0.10	-0.12	0.03	-0.28	-0.04	-0.10	-0.22	0.00	-0.16
	0.13	0.16	0.12	0.11	0.15	0.13	0.16	0.13	0.12	0.13
Number of observations	139	150	152	157	131	139	150	117	170	170
Mean Poor Diego	3.25	2.91	3.70	3.65	3.21	3.25	2.91	3.67	3.65	3.15
	0.09	0.12	0.07	0.06	0.08	0.09	0.12	0.07	0.07	0.08
Mean Middle Class Diego	3.35	2.98	3.71	3.65	3.39	3.35	2.98	3.70	3.61	3.27
	0.09	0.11	0.09	0.08	0.08	0.09	0.11	0.07	0.10	0.09
Mean No-Intro	3.46	3.08	3.76	3.61	3.34	3.46	3.08	3.72	3.63	3.22
	0.08	0.11	0.10	0.08	0.09	0.08	0.11	0.08	0.09	0.09

Note: Robust standard errors are in parentheses. All regressions contain a constant term not reported here. *Poor Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as poor. *Middle Class Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as iddle class. The *no-intro* sample of teachers watched no introductory video and watched only a video portraying performance on an exam.

Table 7: The impact of knowing social class on expected schooling attainment | Inconsistent and high performance variants

	Inconsistent			High performance		
	Primary	Secondary	Tertiary	Primary	Secondary	Tertiary
	(1)	(2)	(3)	(4)	(5)	(6)
Poor Diego	0.01	0.18	-0.19	0.04	0.10	-0.15
	(0.06)	(0.06)	(0.07)	(0.03)	(0.05)	(0.06)
Number of observations	194	194	194	195	195	195
Mean Poor Diego	0.23	0.38	0.39	0.08	0.23	0.69
	(0.04)	(0.05)	(0.05)	(0.03)	(0.04)	(0.05)
Mean Middle Class Diego	0.22	0.21	0.58	0.04	0.12	0.84
	(0.04)	(0.04)	(0.05)	(0.02)	(0.03)	(0.04)
Mean No-Intro	0.05	0.21	0.74	0.04	0.04	0.92
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)

Note: Robust standard errors are in parentheses. All regressions contain a constant term not reported here. *Poor Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as poor. *Middle Class Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as iddle class. The *no-intro* sample of teachers watched no introductory video and watched only a video portraying performance on an exam.

Table 8: The impact of knowing social class on scholastic aptitude assessment| High performance

		% reporting child is:			% suggesting child needs:	
	Predicted grade level (years)	Below grade level	At grade level	Above grade level	Additional support	Advanced placement
	(1)	(2)	(3)	(4)	(5)	(6)
Poor Diego	0.18	-0.03	-0.12	0.15	-0.03	0.03
	0.12	0.07	0.07	0.05	0.07	0.06
Number of observations	195	195	195	195	195	195
Mean Poor Diego	3.91	0.33	0.41	0.26	0.30	0.28
	0.09	0.05	0.05	0.04	0.05	0.05
Mean Middle Class Diego	3.72	0.36	0.53	0.11	0.33	0.24
	0.08	0.05	0.05	0.03	0.05	0.04
Mean No-Intro	4.05	0.23	0.51	0.27	0.21	0.31
	0.08	0.04	0.05	0.05	0.04	0.05

Note: Robust standard errors are in parentheses. All regressions contain a constant term not reported here. *Poor Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as poor. *Middle Class Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as iddle class. The *no-intro* sample of teachers watched no introductory video and watched only a video portraying performance on an exam.

Table 9: The impact of knowing social class on behavior assessment (AES estimates) | High performance

	Darley and Gross (1983) grouping					Factor analysis grouping				
	Work habits	Motivation	Sociability	Emotional maturity	Cognitive ability	Work habits	Motivation	Sociability and expression	Confidence and cooperation	Cognitive ability and learning attitude
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Poor Diego	0.03	-0.47	-0.10	-0.24	-0.26	0.03	-0.47	-0.23	-0.33	-0.06
	0.11	0.14	0.15	0.14	0.14	0.11	0.14	0.15	0.15	0.12
Number of observations	147	148	135	150	125	147	148	109	175	177
Mean Poor Diego	3.95	3.45	3.95	3.96	3.95	3.95	3.45	3.92	4.00	3.96
	0.06	0.09	0.08	0.07	0.07	0.06	0.09	0.07	0.08	0.07
Mean Middle Class Diego	4.00	3.76	4.03	4.22	4.04	4.00	3.76	4.08	4.28	4.03
	0.07	0.10	0.07	0.05	0.07	0.07	0.10	0.06	0.05	0.07
Mean No-Intro	4.10	3.74	4.06	4.27	4.06	4.10	3.74	4.08	4.34	4.13
	0.07	0.11	0.09	0.06	0.08	0.07	0.11	0.08	0.07	0.07

Note: Robust standard errors are in parentheses. All regressions contain a constant term not reported here. *Poor Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as poor. *Middle Class Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as iddle class. The *no-intro* sample of teachers watched no introductory video and watched only a video portraying performance on an exam.

Table 10: Heterogeneity treatment effects on assessing the child as below grade level

	Inconsistent performance			High performance		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Interactions with teacher characteristics						
Poor Diego	0.17	0.18		-0.04	0.01	
	(0.08)	(0.10)		(0.08)	(0.10)	
Father w/ higher ed	0.08			0.11		
	(0.11)			(0.11)		
Poor Diego x Father w/higher ed	-0.09			0.06		
	(0.15)			(0.16)		
Below median for primary school experience		0.17			0.07	
		(0.10)			(0.10)	
Poor Diego x Below median for primary school experience		-0.08			-0.07	
		(0.14)			(0.14)	
Number of observations	194	194		195	195	
Panel B: Interactions with district and school characteristics						
Poor Diego	0.111	0.125	0.196	0.0276	0.0208	-0.09
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Below median district poverty rate	-0.03			0.0612		
	(0.10)			(0.10)		
Poor Diego x Below median district poverty rate	0.0676			-0.11		
	(0.14)			(0.14)		
Above median school language performance		0.05			0.13	
		(0.10)			(0.10)	
Poor Diego x Above median school language performance		0.04			-0.09	
		(0.14)			(0.14)	
Above median school math performance			0.17			-0.06
			(0.10)			(0.10)
Poor Diego x Above median school math performance			-0.10			0.11
			(0.14)			(0.14)
Number of observations	194	194	194	195	195	195

Note: Robust standard errors are in parentheses. All regressions contain a constant term not reported here. *Poor Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as poor.

9 Appendix A: Instructions for teachers

Page 1	¡Bienvenido! *Welcome!* Como parte de una iniciativa para mejorar el sistema de educación pública, el Ministerio de Educación, en colaboración con el Banco Mundial, han diseñado este estudio exploratorio para mejorar métodos de evaluación de estudiantes. *As part of an initiative to improve the public education system, the Ministry of Education, in collaboration with the World Bank, have designed this exploratory study to improve student assessment methods.* No existen respuestas correctas o incorrectas, y los resultados del estudio no tendrán ninguna consecuencia o relación con las evaluaciones de desempeño docente. *There are no right or wrong answers, and the results of the study will have no consequence or relationship with the teacher performance evaluations.* A continuación le vamos a presentar un video de un niño tomando un examen, y luego le pediremos que responda un breve cuestionario. *Next we are going to show you a video of a child taking a test, and then we will ask you to answer a short questionnaire.* Todas sus respuestas serán confidenciales. *All your answers will be kept confidential.* ¡Muchas gracias por participar de este proyecto! *Thank you very much for participating in this project!* Pase a la siguiente página para acceder al video. *Go to the next page to access the video.*
Page 2	The video(s)
Page 3	Muchas gracias por ver el video. *Thank you very much for watching the video.* A continuación le pediremos que conteste a un breve cuestionario dividido en tres secciones: *Next we will ask you to answer a short questionnaire divided into three sections:* Sección 1: preguntas relacionadas al desempeño del niño durante el examen que vio en el video *Section 1: Questions related to the child's performance during the test you saw in the video*

	Sección 2: preguntas relacionadas a su experiencia como docente *Section 2: questions related to your teaching experience* Sección 3: información socio-demográfica básica *Section 3: basic socio-demographic information* Le recordamos que no hay respuestas correctas o incorrectas, y que la información es confidencial. *We remind you that there are no right or wrong answers, and that the information is confidential.* Le pedimos que intente ser tan preciso y objetivo como pueda. *We ask that you try to be as precise and objective as possible.* ¡Muchas gracias por su colaboración! *Thank you very much for your help!*

10 Appendix tables

Appendix Table 1: Original Darley and Gross (1983) behavioral domains and wording of behavioral measures

Domains	Attributes	Questionnaire wording for sampled teachers
Work habits	organization	El niño es organizado
	task orientation	El niño es autónomo
	dependability	El niño es responsable
	attention	El niño tiene capacidad de atención
	thoroughness	El niño es meticuloso
Motivation	involvement	El niño es involucrado
	motivation	El niño es motivado
	achievement orientation	El niño es proactivo
Sociability	popularity	El niño es popular
	verbal behavior	El niño usa lenguaje apropiado
	cooperation	El niño es cooperativo
Emotional maturity	confidence	El niño es seguro de sí mismo
	maturity	El niño es maduro
	disposition	El niño es predispuesto
	mood	El niño tiene buen carácter
Cognitive ability	articulation	El niño es bien articulado, se expresa bien
	creativity	El niño es creativo
	learning capability	El niño tiene buena capacidad de aprendizaje
	logical reasoning	El niño tiene buen razonamiento lógico

Appendix Table 2: Behavioral domains suggested by factor analysis

			Factor analysis grouping				
			Cognitively mature/ intelligent	Assertive/Pleasant	Work Habits	Motivation	Positive attitude
Original grouping	Work habits	organized			0.641		
		autonomous			0.494		
		responsible			0.674		
		capacity to pay attention	0.456		0.554		
		meticulous			0.518		
	Motivation	involved				0.679	
		motivated				0.628	
		proactive				0.554	
	Sociability	popular		0.502			
		appropriate language		0.651			
		cooperative		0.622			
	Emotional maturity	self-confident					0.515
		mature	0.441				
		predisposed	0.413				0.473
		good character		0.643			
	Cognitive ability	articulated		0.532		0.410	
		creative	0.408	0.445			
		good learning capacity	0.756				
		good logic reasoning	0.731				
	eigenvalue		2.926	2.871	2.549	2.323	1.066

Notes: only showing factor loadings >.4. Factor rotated with varimax solution

Appendix Table 3: Additional pair-wise t-tests for balance across experimental groups in Table 3

	p-value of (1)-(3)	p-value of (1)-(4)	p-value of (1)-(5)	p-value of (1)-(6)	p-value of (2)-(3)	p-value of (2)-(4)	p-value of (2)-(6)	p-value of (3)-(4)	p-value of (3)-(5)	p-value of (3)-(6)	p-value of (4)-(6)	p-value of (5)-(6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Age	0.31	0.72	0.71	0.80	0.49	0.99	0.91	0.56	0.16	0.44	0.90	0.52
Male	0.86	0.47	0.08	0.86	0.28	0.85	0.28	0.37	0.06	1.00	0.37	0.06
Third grade	0.88	0.54	0.61	0.65	0.44	0.76	0.64	0.65	0.72	0.76	0.88	0.96
Fourth grade	0.76	0.65	0.33	1.00	0.65	0.76	0.88	0.45	0.20	0.76	0.65	0.33
Fifth grade	0.54	1.00	0.61	0.54	0.77	0.37	0.77	0.54	0.26	1.00	0.54	0.26
Years of experience	0.13	0.77	0.63	0.66	0.60	0.43	0.50	0.22	0.30	0.26	0.89	0.95
Graduate degree	0.76	0.76	0.33	0.46	0.76	0.76	0.46	1.00	0.50	0.30	0.30	0.09
Father with higher education	0.51	0.63	0.72	1.00	0.33	0.42	0.75	0.87	0.77	0.51	0.63	0.72
Mother with higher education	0.68	0.56	0.20	0.26	0.42	0.85	0.46	0.32	0.09	0.12	0.59	0.89

Appendix Table 4: The impact of knowing social class on behavior assessment, by question

	Inconsistent performance			High performance		
	Poor Diego	Constant	N	Poor Diego	Constant	N
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Work habits						
Organized	-0.15	3.25	168	-0.16	3.84	176
	(0.17)	(0.11)		(0.15)	(0.11)	
Autonomous	-0.03	3.86	184	-0.04	4.24	187
	(0.14)	(0.10)		(0.10)	(0.08)	
Responsible	0.08	3.28	178	0.03	3.88	179
	(0.17)	(0.13)		(0.15)	(0.11)	
Capacity for attention	-0.09	3.33	188	-0.06	4.33	192
	(0.16)	(0.11)		(0.11)	(0.08)	
Meticulous	-0.24	2.99	170	0.05	3.63	172
	(0.19)	(0.13)		(0.15)	(0.11)	
Panel B: Motivation						
Involved	-0.06	3.08	172	-0.37	3.83	169
	(0.18)	(0.12)		(0.15)	(0.11)	
Motivated	-0.02	2.76	180	-0.41	3.77	172
	(0.19)	(0.13)		(0.18)	(0.13)	
Proactive	-0.20	3.16	166	-0.32	3.88	170
	(0.18)	(0.11)		(0.15)	(0.10)	
Panel C: Sociability						
Popular	-0.06	3.66	172	-0.15	3.99	165
	(0.16)	(0.12)		(0.14)	(0.09)	
Uses appropriate language	-0.15	3.79	175	-0.10	4.05	165
	(0.13)	(0.10)		(0.14)	(0.09)	
Cooperative	0.05	3.76	179	0.01	4.11	158
	(0.15)	(0.11)		(0.13)	(0.10)	
Panel D: Emotional maturity						
Self-confident	-0.02	3.70	185	-0.22	4.30	190
	(0.15)	(0.11)		(0.11)	(0.07)	
Mature	-0.05	3.42	183	-0.16	4.03	181
	(0.14)	(0.11)		(0.13)	(0.08)	
Good disposition	0.05	3.56	176	-0.28	4.23	176
	(0.14)	(0.10)		(0.11)	(0.07)	
Good character	0.03	3.94	177	-0.25	4.33	165
	(0.12)	(0.09)		(0.11)	(0.07)	
Panel E: Cognitive ability						
Articulate	-0.16	3.85	167	-0.32	4.21	151
	(0.14)	(0.10)		(0.13)	(0.08)	

Creative	-0.22	3.41	142	-0.16	3.95	138
	(0.18)	(0.12)		(0.15)	(0.11)	
Learning capacity	-0.15	3.30	182	-0.12	4.19	188
	(0.15)	(0.11)		(0.11)	(0.08)	
Logical	-0.22	3.15	188	0.10	3.89	189
	(0.14)	(0.10)		(0.12)	(0.09)	

Note: Robust standard errors appear in parentheses. Each row represents the dependent variable in a regression of the dependent variable on an indicator for being assigned to the group of teachers who saw an introductory video depicting Diego's background as poor. The behavioral categories in each panel are the groupings used in Darley and Gross (1983).

Appendix Table 5a: The impact of knowing social class on behavior assessment (average scores) | Inconsistent performance

	Darley and Gross (1983) grouping					Factor analysis grouping				
	Work habits	Motivation	Sociability	Emotional maturity	Cognitive ability	Work habits	Motivation	Sociability and expression	Confidence and cooperation	Cognitive ability and learning attitude
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Poor Diego	-0.09	-0.07	-0.01	0.00	-0.18	-0.09	-0.07	-0.04	0.04	-0.12
	0.13	0.16	0.12	0.11	0.12	0.13	0.16	0.10	0.12	0.12
Number of observations	194	189	191	194	192	194	189	192	191	193
Mean Poor Diego	3.25	2.91	3.70	3.65	3.21	3.25	2.91	3.67	3.65	3.15
	0.09	0.12	0.07	0.06	0.08	0.09	0.12	0.07	0.07	0.08
Mean Middle Class Diego	3.35	2.98	3.71	3.65	3.39	3.35	2.98	3.70	3.61	3.27
	0.09	0.11	0.09	0.08	0.08	0.09	0.11	0.07	0.10	0.09
Mean No-Intro	3.46	3.08	3.76	3.61	3.34	3.46	3.08	3.72	3.63	3.22
	0.08	0.11	0.10	0.08	0.09	0.08	0.11	0.08	0.09	0.09

Note: Robust standard errors are in parentheses. All regressions contain a constant term not reported here. *Poor Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as poor. *Middle Class Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as iddle class. The *no-intro* sample of teachers watched no introductory video and watched only a video portraying performance on an exam.

Appendix Table 5b: The impact of knowing social class on behavior assessment (average scores) | High performance

	Darley and Gross (1983) grouping					Factor analysis grouping				
	Work habits	Motivation	Sociability	Emotional maturity	Cognitive ability	Work habits	Motivation	Sociability and expression	Confidence and cooperation	Cognitive ability and learning attitude
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Poor Diego	-0.05	-0.31	-0.08	-0.26	-0.09	-0.05	-0.31	-0.16	-0.28	-0.07
	0.09	0.13	0.11	0.09	0.10	0.09	0.13	0.09	0.09	0.10
Number of observations	195	188	186	194	192	195	188	191	191	192
Mean Poor Diego	3.95	3.45	3.95	3.96	3.95	3.95	3.45	3.92	4.00	3.96
	0.06	0.09	0.08	0.07	0.07	0.06	0.09	0.07	0.08	0.07
Mean Middle Class Diego	4.00	3.76	4.03	4.22	4.04	4.00	3.76	4.08	4.28	4.03
	0.07	0.10	0.07	0.05	0.07	0.07	0.10	0.06	0.05	0.07
Mean No-Intro	4.10	3.74	4.06	4.27	4.06	4.10	3.74	4.08	4.34	4.13
	0.07	0.11	0.09	0.06	0.08	0.07	0.11	0.08	0.07	0.07

Note: Robust standard errors are in parentheses. All regressions contain a constant term not reported here. *Poor Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as poor. *Middle Class Diego* represents an indicator for assignment to the group of teachers that watched an introductory video depicting Diego's background as iddle class. The *no-intro* sample of teachers watched no introductory video and watched only a video portraying performance on an exam.