Discrimination Goes to School? Understanding Racial Differences in Non-Blind Grading*

Fernando B. Botelho†

Ricardo A. Madeira‡

Marcos A. Rangel§

Very Preliminary, Comments Welcome

PLEASE DO NOT CITE OR DISTRIBUTE
WITHOUT THE AUTHORS' PERMISSION

July, 2009

Abstract

Recent literature suggest that observed racial differentials in labor markets are result of lower investment in the accumulation of skills or of *pre-market factors* by individuals of African decent. If parents and children update investment decisions after extracting from school reports signals regarding scholastic abilities, differential errors in *perceived* ability could reinforce racial gaps in the accumulation of human capital. Evidence drawn from a unique data set from Brazilian elementary, middle and high-schools, suggests that teachers' grading (when compared to blindly scored tests of proficiency) suffers from cardinal and ordinal biases associated to a child's race .

*The views expressed in the paper, errors and omissions are sole responsibility of the authors. Data used in the analyses that follow were obtained under strict confidentiality agreement between the São Paulo Secretary of Education and the authors. We are thankful to the Secretary's staff for invaluable help with the data sets. An IRB approval has been obtained by the authors via Fundação Instituto de Pesquisas Econômicas (FIPE), in São Paulo.

†University of São Paulo and CenterIDEAS (fbotelho@fipe.org.br).

‡University of São Paulo and CenterIDEAS (rmadeira@fipe.org.br).

§University of Chicago and BREAD (rangelm@uchicago.edu).

1 Introduction

A number of studies detect significant association between individual characteristics used to infer African ancestry and various measures of socioeconomic success.[1] This historically-rooted stratification remains stark in a variety of contexts, even under sharp differences in patterns of economic development, institutional arrangements regarding racial segregation, and observed rates of miscegenation.[2] Consensus emerging from recent literature (mostly regarding the United States' case) indicates that those differences are the result of lower investment in the accumulation of skills or of *pre-market factors* by individuals of African decent, particularly in the case of labor markets.[3] Some of these contributions emphasize that differences in skills between blacks and whites emerge during infancy, affecting both cognitive and non-cognitive aspects of child development and tend to widen while children attend elementary schools.[4] The early appearance and the dynamics of such gaps suggest that, in order to uncover the determinants of racial differences in the rates of skill's acquisition, we need to better understand the interactions between parental investment decisions and the school environment's inner workings.

In the present article we tackle this issue by considering a simple conceptual framework in which parents and children themselves update investment decisions after extracting from school reports/transcripts signals regarding scholastic abilities. In this setting, any differential treatment received by children of different races (or complexions) within the classroom feeds back into the parental/individual decision process. If children's *perceived* ability increases the return to parental investments,[5] or if teacher's assessment influences some key non-cognitive aspects of a child's life (such as self-esteem, confidence and motivation), this mechanism could very well reinforce racial gaps in the accumulation of human capital. Therefore, we focus attention on indentifying racial differentials on the measurement of performance within classrooms, or, in other words, on racial

[1] Bertrand and Mullainathan (2004) provide experimental evidence that (randomly selected) firms in the United States are less likely to interview job-applicants with distinctively black names (see also Fryer and Levitt, 2004a). Using cross-sectional observational data, Goldsmith et al (2005), Gyimah-Brempong and Price (2006), Hersch (2006), Hunter et al. (2001) and Keith and Herring (1991) all find suggestive evidence of a complexion gap in terms of wages, legal punishments, education and unemployment among African Americans. Hammermesh and Biddle (1994) and Biddle and Hammermesh (1998) find evidence of appearance premia. Their reasoning could also be applied to hair curliness, nose width, lip thickness, steatopygia, and to any of the physical traits that can be linked to membership in the black or white ethnic groups. In fact, when the Apartheid regime was introduced in South Africa, skin color and physical traits were used in combination to establish the classification system imposed by government officials.

[2] See Alexander et al. (2001) for a perspective on the North-American, South African and Brazilian experiences. See also Herring et al. (2004) and Telles (2005) for comparatives studies between the United States and Brazil. Analysis of US historical data can be seen in Bodenhorn (2003), Bodenhorn and Ruebeck (2005), Dollard (1937), Freeman et al. (1966), Hill (2000), Ransford (1970), Reuter (1918) and Seeman (1946).

[3] See Carneiro et al. (2005), Heckman (1998) and Neal and Johnson (1996).

[4] See Carneiro et al. (2004) and Levitt and Fryer (2004b).

[5] See Behrman et al (1994), Rosenzweig and Schultz (1982) and Rosenzweig and Zhang (2006).

differences in the reliability of ability signals produced by school transcripts.

We examine data on teachers' grading of elementary, middle and high-school students based on non-blind evaluation methods and compare it with evaluations resulting from standardized tests taken by the same children, within the same classrooms, and scored without human interference.[6] Data from the largest public school system in Brazil (covering the entire state of São Paulo) are used. Differences between non-blind and blind scores/grades and rankings were associated to child characteristics, with robust and consistent underscoring and under-ranking of blacks relative to whites being identified both on Math and on Reading/Grammar evaluations. This suggests that teacher's assessment of dark skinned children may be blurred by stereotypes associated to the black population, constituting evidence compatible with perverse self-sustaining equilibrium of statistical discrimination models. We also uncover evidence that both mis-ranking and mis-scoring may influence the accumulation of human capital by affecting both the quantity and the quality of formal education provided to children.[7] Racial stereotypes (within the classroom) seem to be (at least partially) responsible for some forms of discrimination against blacks *before* they reach the labor market.

Any observer would quickly point-out that schools', neighborhoods', and marriage markets' levels of racial segregation in Brazil are not comparable to other contexts, in particular to the North American one.[8] We emphasize, however, that the relevance of the Brazilian experience to US policy makers resides on the fact that trends in background diversity within classrooms (due to government policies that lead to substantial increases in school enrollment rates) mirrors the one generated by the absorption of Hispanics into the US public schools' system.[9] The same is true for a number of developed countries facing large immigration flows (North-Africans in France, Turkish in Germany, among others). Moreover, our results also shed light on the experience of developing countries where stagnation in the quality of education seem to be associated to the democratization of access to schools (e.g.: South Africa, Kenya, India).[10] Therefore, this article offers important insights regarding teacher behavior within ethinicity/background-diverse classrooms that should

[6] There is a very recent contribution by Lam et al. (2007) examining a related question but with very different strategy for high-schools in South Africa. Lavy (2008) employs a strategy very similar to ours in order to examine the impact of gender stereotypes on grading college admission exams in Israel. Gender differentials in our data are examined on a companion paper (Botelho et al., 2008).

[7] The analysis here complements results presented in Rangel (2008), which showed that after controlling for observed and unobserved parental characteristics, light-skinned (European appearance) children are more likely to be enrolled in school than their dark-skinned (African appearance) siblings. Conditional on attendance, the latter are also less likely to attend higher-quality private schools.

[8] See Vigdor and Ludwig (2007) on school/neigborhood segregation and racial gaps in test scores. See Fryer (2007) on interracial marriage.

[9] This is also the case under school desegregation processes (see Guryan, 2006). For a perspective on impacts of racial compositional changes and its effects over learning, see Hoxby (2000).

[10] See Glewe and Kremer (2006).

not be uderestimated.

Besides the literature cited above, the discussions presented here draw on insights from (and aim at contributing to) different branches of the literature in education. This includes the measurement of grading standards and the estimation of its impact over learning (Becker and Rosen, 1992; Betts, 1995; Betts and Grogger, 2003; and Figlio and Lucas, 2004), the discussion of black-white test score gaps in Jencks and Phillips (1998), the analyses of statistical discrimination and stereotypes in Aigner and Cain (1977), Jessen and Fleming (1996), Lundberg and Starz (1998), and Zebrowitz (1996); and the discussion on grade inflation, signaling, and informative content of grades (Chan, Hao and Suen, 2007; Millman et al., 1983; and Ostrovsky and Schwarz, 2005).

The reminder of the article is organized as follows. Section 2 discusses in general lines the conceptual framework that guide the empirical strategy and informs the analyses of results. Section 3 presents the institutional background and previews the data. Section 4 overviews the econometric identification strategy and presents the results. Section 5 concludes.

2 Conceptual framework

We conceive a model in which parents and schools interact in order to "produce" children's proficiency in different school subjects. While traditional models in the economics of education emphasize the combination of school and family inputs, and attempt to identify their relative contribution to the accumulation of knowledge, we choose to take a different route. The focus here is on the asymmetric information context that characterizes parents-school interactions. This is particularly the case when parental formal education is inferior to the one already accumulated by their children.[11] Such parents tend to be over-reliant on the signal issued by transcripts when attempting to asses the scholastic abilities of their children.

Teachers translate observations in the school environment (private signals) into transcripts that (publicly) signal to children and their families (absolute or relative) levels of proficiency. Most importantly, by doing so, teachers inform the acting parties about the potential returns to additional investments in human capital accumulation. Parents find themselves, therefore, on the position of investors relying on the asset-return evaluations of more informed experts. Yet, depending on the use experts make of the private information available to them, they may (intentionaly or not) misrepresent the information and steer investors' decisions in one way or the other. In this setting, any differential treatment received by children of different races (or complexions) within classrooms feeds back into the parental/individual decision process.

[11] Or even in a different language, as the case of first-generation Mexican Americans.

Under this reasoning, it gains importance a better understanding of "signal translation technologies" available to teachers. In particular, we ought to discuss the use (and misuse) of alternative evaluation methods, including assessment schemes such as norm-referenced (relative ranking) or criterion-referenced (competency) grading. As further discussed below, most teachers are instructed by the central public-schools authority to use the latter, but that does not mean they refrain from using the former as a way of motivating students at the top of the proficiency distribution. For this reason, our empirical analysis contemplates both cardinal and ordinal dimensions of racial differences in evaluation outcomes. After all, parents may understand returns to investments in education as a the chance of their child winning a rank-order tournament (as in Lazear and Rosen,1981) - investment is attractive if a child can perform better than her peers; or they may simply understand the returns in the form of chances to achieve certification (promotion from one grade/level to the next).

3 Institutional background and data

3.1 Institutions

The Brazilian pre-college educational system is organized into four levels: preschool (attended by 6 year-olds), primary school (attended by 7 to 10 year-olds), middle/secondary school (attended by 11 to 14 year-olds) and high school (attended by 15 to 17 year-olds).[12] The primary school comprises four school years, the first four grades. The basic disciplines offered at the primary educational level are Language (Portuguese), Mathematics, Social Studies and Science. All the basic subjects are taught by the same teacher, but curricular activities also include activities such as Physical Education and the Arts, which are taught by specialized teachers. For middle and high-schools, teachers' subject-specialization is complete.

As in the case of most developing countries, in the 1990's Brazil witnessed significant progress in standard educational policy targets. There was, for example, a significant increase in the rates of enrollment of school-aged children. However, the surge in enrollment was not accompanied by an increase in the quality of education being provided, particularly in the public schools system.[13] To a large extent, however, the current situation of the Brazilian public school system can be attributed

[12] According to federal guidelines, local governments should run on elementary and middle schools (grades 1 to 8), with the States being responsible for high schools.

[13] In mathematics, Brazil occupies one of the last positions in the PISA 2006, with an average score of 370. This performance puts Brazil ahead only of Tunisia, Qatar, and Kyrgyzstan. Several developing countries such as Uruguay (score of 427), Turkey (424), Thailand (417) and Chile (411) outperform Brazil. The general picture for reading skills is very similar.

to the increased access to school that took place in the last fifteen years or so. This 'democratization' process has had a major impact on the composition of the student body, increasing heterogeneity within classes, and may very well have contributed to the sluggish improvement in the quality of instruction offered by public schools.

One of the chief policies adopted to boost enrollment rates was an automatic promotion scheme. Sao Paulo's public school system adopted such strategy in 1996. This policy groups contiguous grades into cycles, and retention can only occur at the end of each cycle. In terms of pre-college education there are three cycles. Cycle 1 encompasses grades 1 to 4, and cycle 2 covers grades 5 to 8. In the case of Sao Paulo, high-schools were not included in the automatic promotion scheme. Under such scheme, a student is promoted to the next grade if she attends more than 75% of the classes, irrespective of her mastery of the material that was covered during the academic year. Insufficient proficiency can result in grade retention only at the end of each cycle. In this case, the pupil must repeat the last grade within that cycle.[14] Several international organizations, including the World Bank, support this policy as an effective way to curb low grade completion and to decrease drop-out rates. The general lines of the argument are that grade retention could adversely affect some of the students' non-cognitive skills (like confidence and self-esteem), increasing anxiety levels and harming their learning capacity. In this scenario, a better alternative would be the promotion to the next grade despite the insufficient performance. One should keep in mind, however, that promotion of unfit students could cause the very same effect, since the underperforming pupils would be systematically behind their classmates, increasing the dropout rates later on.[15]

To some extent, one of the first impacts of automatic promotion schemes is to make grading by teachers somewhat irrelevant, at least in what concerns the bottom tail of the latent proficiency distribution. In fact, as a result of decentralized rules regarding grading scales under automatic promotion, at least 13 different methods of conveying grades were being adopted by schools within Sao Paulo's public education system as of August 2007. Different schools adopted different coding and cutoff grades, ranging from 0-100 numeric scales to A-F letter scales and including simple binary sufficient/insufficient classifications. Interestingly, this multiplicity of scales has called attention of education authorities in Sao Paulo. Internal communications provided to the authors indicate, for example, that the Secretary's staff had an explicit concern for parents having difficulties to assess leaning by children that attended schools with different grading scales.[16]

[14] There is anecdotal evidence, however, that after automatic promotions was implemented, retention rates also fell in end-of-cycle grades (4th grade in particular). It is not clear if this is the result of improved learning or if in fact end-of-cycle teachers feel they should not be responsible for failing students that the system allowed to get that far. To the extent failure rates are also interpreted by some as the result of poor teaching, individual incentives not necessarily align with the planner's.

[15] See King et al. (2008).

[16] The memorandum we had access to was dated July 2006.

In response to the issues raised by the coexistence of automatic promotion scheme and the signaling role of grades, the Sao Paulo's school authority adopted a uniform criterion-referenced rule in September, 2007. According to these new guidelines, all teachers had to attribute numeric integer grades ranging from 0 to 10. That is, students must now be allocated to a discrete grid with 11 categories, and passing grade is set at 5 points. Some schools had already been following this grading procedures, while others had only a couple of months to adapt.[17] The rule was also applied retroactively, since schools employ bimonthly evaluations and the school year starts in February. Teachers not only had to start evaluating students according to the new scale from October 2007, but they also had to review early grades in order to re-scale them according to the new guidelines.

Teachers and other school administrators were not given instructions on how to attribute grades as a function of a student's observed proficiency level beyond the guidelines imposed by a uniform school curriculum. The state administration provided pedagogical material aligned to such curriculum and teachers were supposed to evaluate students according to proficiency in such material. Nonetheless, no explicit guidance regarding the design of evaluations (except for questions included at the back of the teacher's booklet) was given. Nevertheless, schools and teachers still have complete autonomy to allocate the students in the 11 grading categories.

In our view, the new guidelines introduced noise on teachers' grading technology. Therefore, some of the empirical results presented below can be understood as a function of the adaptation to the new rules. Before proceeding with our econometric analysis we describe our data and introduce the computation of indicators we consider relevant.

3.2 Data

The Sao Paulo's Secretary of Education has agreed to share with the authors, under cooperation and confidentiality agreements, detailed information of the universe of students in the state's education system. We merged data sets from three distinct sections of their data bank: standardized tests of proficiency, matriculation information, and transcript records.[18] In what follows we refer to them as blind scores, flow measures and non-blind grades.

The blind scores: Blind scores are collect in the context of SARESP[19]. It consists of a statewide exam taken by students enrolled in grades 1, 2, 4 (elementary school), 6 and 8 (middle

[17] Unfortunately, the Secretary has no record of grading scales within the system (except for a small sample subject to an informal survey in 2006). The authors are working on a universal retrospective collection of such information.

[18] The Secretary itself had not attempted to combine these data. Each section of the data-set is responsability of a different department, so that this is the first time this data is used in the final integrated format.

[19] *Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo*, or State of São Paulo System for Evaluation of Scholastic Performance.

school), and 11 (high school) in the state government-run schools . The test has been applied in slightly different formats since 1996. This paper uses data from its 10th edition, which took place in 2007 and had over 1.8 million test-takers in approximately 5,200 schools. To gain a better understanding of the conditions, students, teachers, and school staff also answer a survey that asks questions on socioeconomic status, study habits, teaching and pedagogical practices, and perceptions about the school environment, among others.[20]

Its chief purpose of such exam is to measure the students' proficiency on the subjects (topics) assigned to each specific grade. The exam in 2007 had two sets of questions covering Math and Portuguese language. The Math set contained 30 multiple choice questions. The Portuguese language component depends on the grade being tested. For grades 1 and 2, the students had to write short answers to objective questions. For the other grades, students were asked to answer 30 multiple choice questions and write a short essay.[21] The exams were taken in late November, close to the end of the academic year, during class time and in the same place the students take regular classes. Grading is electronic for the multiple-choice questions: students mark their choices in a test sheet, that is scanned and graded automatically, without human interference. This grading procedure assures that a completely blind score is obtained.[22]

Students took the exam in two consecutive days, one for each subject. Since 4th, 6th, 8th and 11th graders, which are the focus of this paper, can attend classes either in the morning, in the afternoon or at night, a different exam (yet similar in difficulty) had to be prepared for each group. All students that usually attend classes during the same school shift took the same test. The State Secretary of Education hires an independent institution to prepare the exam, according to predetermined guidelines. To oversee the students during the test, teachers from other schools are mobilized, such that students are supervised by a teacher different from their regular ones. External observers are also assigned to each schools to guarantee the strict fulfillment of all rules.

Microdata on these tests' results are provided in two formats. First, *proficiency scores* in each subject are provided after application of Item Response Theory (IRT) methods. These scores are also converted into a (grade-subject-specific) four step classification system that reflects educators' consensus regarding levels of proficiency (below basic, basic, sufficient, more than sufficient). Proficiency in the essay portion of the Language exam is reported in a separate scale. Second, number of correct answers in the multiple-choice portion of the exam is reported. From these we construct

[20]Unfortunately, data on teachers were not collected with identifiers. We are, therefore, prevented from merging these sections of the data with the outcomes for students.

[21]The scores employed in our analysis do not take into consideration the written sample section of the exam.

[22]The exams taken by 1st and 2nd graders in 2007 were not graded in a strictly blind way. 1st graders are actually graded by their own teachers, while 2nd graders are graded by a team of teachers from the same school they attend. We, therefore, do not explore the SARESP data regarding these children in our analysis.

the *percentage of correct answers* and convert it into a 0-10 scale using standard nearest-integer rounding rules.

Individual-level results from SARESP are not made publicly available to children, parents, or schools. Until 2008 school-level results were not used in any explicit accountability system either, and have been serving the sole purpose of "diagnosing" the entire educational system.

The flow measures: Matriculations in the entire state of Sao Paulo is centralized by the Secretary of Education. Besides being a provider, the Secretary is responsible for regulation of the private schools system. It also maintains straight cooperation agreements with all Municipality-run schools across the state. The centralized system exists, therefore, as a way to avoid that parents matriculate their children in more than one school in order to guarantee a slot. In the past this practice has itself led to a number of children that could not be absorbed by the system (since some had taken two or three slots). This system offers interesting ways of tracking student mobility within the system, specially in the case of dropout and migration between public and private systems, or even between schools.[23]

The non-blind grades: we take advantage of a unique administrative data set containing detailed information regarding scores and attendance records for each student in the State of Sao Paulo's school system. The complete set of report cards available to us include information on bimonthly evaluations and attendance tracking for every school subject. We employ most of this information below. A grade is assigned as an integer between 0 and 10, while attendance rate is reported by teacher as an integer between 0 and 100. As mentioned above, the passing average grade per subject and for all schools is set at 5 points.

Table A1 in the appendix presents grade-specific averages of selected indicators from our final data set.

3.3 Relevant indicators of grading practices

Our premise is that despite the irrelevance of school grades to determine promotion in most of the grades, they still can serve as a signal to students and parents. In such context, we need to come up with measures of how much noise the new grading guidelines added to the transcripts signal sent to parents. In particular, we start by trying to describe heterogeneity of grading standards across classrooms within the system according to both cardinal and ordinal perspectives.

[23] The private schools system has the reputation of providing better quality education. Evaluations run by the federal government confirms that this is the case on average. There is a lot of heterogeneity, but parents seem to believe they could offer better chances to their children by enrolling them in private schools.

Consider first that teachers who cannot (or feel they should not) fail students due to insufficient proficiency, will inflate grades so that passing standards are achieved. In this static context grade inflation refers to the phenomenon of giving grades that overstate the underlying proficiency of students on a specific subject. We, therefore, define as grade inflation in a classroom the average increment in grades given by teachers relative to the proficiency levels indicated by blind scores. We convert teacher's 11 points scale into a 4-point scale by dividing the range into equally sized intervals (0 to 2.5, 2.5 to 5, 5 to 7.5 and 7.5 to 10). This new teacher scale is compared with blind-scores' proficiency scale. The classroom level average differences between non-blind and blind scales is then computed. Positive numbers indicate teachers are *inflating* proficiency levels. Figures 1A and 1B present histograms of this measures across schools by grade for Math and Language, respectively. In both cases inflation is widespread. Lower grades are more heterogeneous while on higher ones there is more inflation on average, particularly in the case of Mathematics.

Since grade inflation and grade deflation can be present within the classroom and may cancel each other out in the end, we also investigate an alternative aggregate indicator. We investigate the propensity of teachers to differentiate students within the same classroom. That is to say, we measure the probability of a teacher assigning the same level of proficiency to two different students and compare it with the probability obtained for that same classroom using the scale derived blind scores.[24] We refer to this measure as *bunching*. Teachers that produce more bunching are avoiding the use of their grades as an instrument of differentiation among students (what reduces the relative content of the information conveyed by transcripts). Histograms of this measures (where the unit of observation is the classroom) are presented in Figures 2A and 2B. Here the evidence indicates that, relative to the actual bunching of proficiency measured by standardized tests, the average teacher does not produce more bunching when grading their students on Math. The same is not true on Language evaluations, where the average teacher tends to relatively reduce differentiation among their students. In both cases, as in the case of grade inflation, a lot of heterogeneity across classrooms/teachers is observed.

These two indicators provides with insights regarding teachers behaviors, but they leave aside the possibility that students would receive transcripts that put them on a position relative to their classmates that is different from the one indicated by their blind proficiency scores. We therefore turn to measures of ranking within the classroom in order to turn the spotlight to relative positioning issues. We say that teachers produce *reshuffling* when their grades rearrange the ranking of students within the classroom.

[24] The mechanics is equivalent to the measurement of a fractionalization index, since we measure the chance of finding two students in the same classroom having achived the same proficiency level.

However, in the presence of bunching, ranking ties within class are not unusual.[25] We are, therefore, required to use ranking measures that are purely ordinal.[26] In order to tackle this issue we use two alternative ranking measures: fractional and classical ordinal ranking.[27] Since ordinal ranking attributes a distinct ordinal number for every student, including students that receive the same grade, an alternative way to construct such ranking under the presence of ties is to use a random criteria to brake ties. To verify the robustness of our results under the use of fractional ranking, we undertook 1,000 replications of our empirical exercise (using a random tie-braking criterion) and look for the dispersion of the distribution of estimates obtained.[28] The results reveal that the empirical distribution using ordinal rankings provide very tight bounds around the coefficient obtained using the fractional version of the rankings.

Figures 3A and 3B summarize findings regarding reshuffling across classrooms in our sample. The histograms plotted are measures of the average absolute change in percentile rankings within a classroom. They reveal that for both Math and Language evaluations, reshuffling tends to be more pertinent for students on higher grades.

3.4 Intra-classroom differentials

In the context described so far, a relevant question is to what extent different groups of students are affected by a noisier evaluation system. There are at least two reasons why teachers can systematically mis-evaluate the proficiency of students with certain attributes (like males, or blacks). First, the teacher may merely dislike people with those traits, imposing punishments that can take both cardinal and ordinal forms. Second, teachers may be more sophisticated, and want to evaluate proficiency according to some unobservable attribute which is correlated with the observed trait (race, gender). In this case, the characteristic itself conveys information, and can help the teacher to generate a better ranking according to this latent attribute. For example, suppose that teachers would like to rank students according to their overall scholastic aptitude, which is correlated with family background. One private signal of aptitude received by the teacher is the score in the regular

[25] Ties are particularly common for the non-blind grade ranks, since very often teachers attributes the same grade for different students of the same class.

[26] We say that a within class rank is purely ordinal if the mean of the rank variation between blind and non-blind ranks is always zero.

[27] According to the fractional ranking items that compare equal receive the same order indicator, which is the mean of what they would have under ordinal rankings. In ordinal ranking, all items receive distinct ordinal numbers, including items that compare equal. The assignment of distinct ordinal numbers to items that compare equal can be done at random, or arbitrarily, but it is generally preferable to use a system that is arbitrary but consistent, as this gives stable results if the ranking is done multiple times.

[28] For both ranks, fractional and ordinary, the actual within class ranking measure used in empirical exercises is given by $R_{ij} = \frac{r_{ij}-1}{n_j}$, where r_{ij} and n_j denotes the rank of the student i in the class j and the number of students in class j, respectively.

in-class exams, but attributes that are associated with family background – like personal hygiene, or how often parents visit the school and meet with teachers – may also reveal information relevant to this latent attribute – aptitude – the teacher wants to infer. In this circumstance, two students that have different traits (like personal appearance) but got the same score at the exam may receive a different grade. Of course, these two sources of discrimination are well know in the economics literature: the first is *taste discrimination*, whereas the second falls under the realm of *statistical discrimination*.

Considering the different nature of the exams applied within the school context by teachers and the standardized ones adopted for external monitoring of learning, any comparison of scores is not a trivial task. One can expect that, since teachers receive an uniform curriculum from the external examiner, their evaluation should reflect at least partially the same skills and cognitive abilities. Therefore, blind scores should approximate a counterfactual for non-blind ones. Nonetheless, since the blind and non-blind scores have different distributions for both subjects (Math and Language), some form of normalization is necessary for an appropriate comparisons. The panels in Figures 4A and 4B presents the histograms of both exams for both subjects for each race and each grade.[29] The panels indicate that the distributions of the non-blind scores (for both subjects and races) are skewed to the right in relation to the distributions of the blind scores with bunching around proficiency level 3 (or more precisely, grade 5). The existence of grade inflation for non-blind scores combined with the fact that the distributions of blacks' scores (for both subjects and tests) are skewed to the left when compared to their white counterparts makes cardinal comparisons particularly troublesome. Since blacks are overrepresented at the lower tail of the blind scores distributions, grade inflation *per se* might generate a discrimination *against* whites when simple cardinal comparisons are employed. However, the coexistence of cardinal discrimination *against* whites with ordinal discrimination against blacks is possible.[30]

We turn to testing for the presence of discrimination on grading by statistically testing if there is any difference between a child's assessment generated by the non-blind and the blind performance measures, and if such difference is strongly associated to his or her race. If this is the case, considering our discussion of grading criteria, either non-blind grading is "unfair" or standardized tests have a biased design that do not reflect characteristics considered important by teachers. Both yield important policy implications. In our empirical exercises we go a long way trying to eliminate potential biases from the omission of aspects that are both associated to race and may explain differences in proficiency measures. The identification strategy is detailed in the

[29] See also Table A2 for descriptive statistics.

[30] These two different forms of discrimination can have different consequences for students depending on how parents and children respond to teachers discrimination on the ordinal and cardinal dimensions.

next section.

4 Identification strategy

In this section, we outline our estimation strategy. We assume that differences in blind scores and non-blind grades are generated by error-components as in a signal extraction problem. We hypothesize that blind scores are subject to classical measurement error (mean zero, and zero covariance with latent proficiency and its components) while non-blind grading may be subject to non-classical components due to difficulties faced by teachers when translating exams results and classroom interactions into transcript-format grades.

4.1 A simple version

Consider two alternative measures of proficiency:

$$y_{ij}^{k} = y_{ij}^{*} + \epsilon_{ij}^{k} \tag{1}$$

where the evaluation system is indexed by k (NB=non-blind, B=blind), y_{ij}^{*} represents latent proficiency in a subject and ϵ_{ij}^{k} indicates a noise that characterizes a signal-extraction problem. We consider that proficiency is ultimately a function of individual/family characteristics, previous investments, student effort and teacher characteristics:

$$y_{ij}^{*} = X_{ij}^{'}\beta + \eta_{ij} \tag{2}$$

Both teachers and the econometrician face imperfect observability of proficiency's components, which we further decompose into unobservables at teacher/classroom and individual levels, respectively:

$$y_{ij}^{*} = X_{ij}^{'}\beta + \alpha_j + Z_i^{'}\gamma \tag{3}$$

We also assume that blind scores, as discussed above, cannot have errors that are non-classical. The same is not true for the non-blind grades. Now consider that we examine, in the context of the populational regression function, the influence of observable (by teachers and the econometrician) individual characteristics over available measures of proficiency within a classroom:

$$E\left[y_{ij}^{k}\,/X_{ij}, \alpha_j\right] = \alpha_j + X_{ij}^{'}\beta + E\left[Z_i^{'}\gamma + \epsilon_{ij}^{k}\,/X_{ij}, \alpha_j\right] \tag{4}$$

Ordinary least squares' estimators of the impact of x_r over observed proficiency would then be:

$$\begin{aligned}\frac{\partial E\left[y_{ij}^{k}/X_{ij},\alpha_j\right]}{\partial x_{ij}^{r}} &= \beta_r+\frac{\partial E\left[Z_i'\gamma+\epsilon_{ij}^{k}/X_{ij},\alpha_j\right]}{\partial x_{ij}^{r}} \\ &= \beta_r+\frac{\partial E\left[Z_i'/X_{ij},\alpha_j\right]}{\partial x_{ij}^{r}}\gamma+\frac{\partial E\left[\epsilon_{ij}^{k}/X_{ij},\alpha_j\right]}{\partial x_{ij}^{r}}\end{aligned} \tag{5}$$

a function of the actual impact over latent proficiency, the indirect impact filtered by the omission of other characteristics correlated with x_r, and the relation between such characteristic and the measurement error.

When exploring the availability of the two measures of proficiency, what we can uncover as the r-th slope parameter of the difference between non-blind grades and blind scores is:

$$\begin{aligned}\frac{\partial E\left[y_{ij}^{NB}-y_{ij}^{B}/X_{ij},\alpha_j\right]}{\partial x_{ij}^{r}} &= \frac{\partial E\left[y_{ij}^{NB}/X_{ij},\alpha_j\right]}{\partial x_{ij}^{r}}-\frac{\partial E\left[y_{ij}^{B}/X_{ij},\alpha_j\right]}{\partial x_{ij}^{r}} \\ &= \frac{\partial E\left[\epsilon_{ij}^{NB}/X_{ij},\alpha_j\right]}{\partial x_{ij}^{r}}-\frac{\partial E\left[\epsilon_{ij}^{B}/X_{ij},\alpha_j\right]}{\partial x_{ij}^{r}}\end{aligned} \tag{6}$$

and, since the blindly scored test is subject exclusively to the effects of classical measurement error, this is equivalent to:

$$\frac{\partial E\left[y_{ij}^{NB}-y_{ij}^{B}/X_{ij},\alpha_j\right]}{\partial x_{ij}^{r}}=\frac{\partial E\left[\epsilon_{ij}^{NB}/X_{ij},\alpha_j\right]}{\partial x_{ij}^{r}} \tag{7}$$

Therefore, when measuring the impact of race over the difference between blind and non-blind evaluations, we uncover the marginal effect of race over the transcripts' non-classical measurement error expected value.

4.2 A more general version

Consider alternatively that teachers proceed with their evaluation also directly weighting some child characteristics they consider relevant (say, non-cognitive dimensions observed by the teacher but not by the econometrician), so that:

$$y_{ij}^{B}=y_{ij}^{*}+\epsilon_{ij}^{B} \tag{8}$$

$$y_{ij}^{NB} = y_{ij}^{*} + \widetilde{\eta}_{ij} + \epsilon_{ij}^{NB} \tag{9}$$

where $\widetilde{\eta}_{ij} = \widetilde{\alpha}_j + Z_i'\widetilde{\gamma}$.

The replication of the exercise above yields:

$$E\left[y_{ij}^{NB}/X_{ij}, \alpha_j\right] = (\alpha_j + \widetilde{\alpha}_j) + X_{ij}'\beta + E\left[Z_i'/X_{ij}, \alpha_j\right](\gamma + \widetilde{\gamma}) + E\left[\epsilon_{ij}^{NB}/X_{ij}, \alpha_j\right] \tag{10}$$

so that:

$$\frac{\partial E\left[y_{ij}^{NB} - y_{ij}^{B}/X_{ij}, \alpha_j\right]}{\partial x_{ij}^{r}} = \frac{\partial E\left[Z_i'/X_{ij}, \alpha_j\right]}{\partial x_{ij}^{r}}\widetilde{\gamma} + \frac{\partial E\left[\epsilon_{ij}^{NB}/X_{ij}, \alpha_j\right]}{\partial x_{ij}^{r}} \tag{11}$$

In this case, ordinary least squares estimates identify the relevance of a particular individual characteristic via its impact over measurement errors in the non-blind test, but also any indirect influence due to correlation with non-observed components of what teachers call proficiency.

5 Results

We perform our empirical exercises in two steps. At first we try to identify the existence and robustness of the impact of race over the difference between non-blind and blind measures of proficiency. We then go about investigating if there is any detectable sign of such measures over observed behavior. We investigate five alternative measures of proficiency, and look at both cardinal and ordinal measurers. The cardinal measures used are: proficiency indicator (the student has achieved minimum requirement in the subject), and the direct difference between teachers' grades (in a 0 to 100 scale) and the percentage of correct answers in the SARESP exam. The ordinal measure encompass: an indicator for above the classroom's average proficiency; an indicator for best performance in classroom (achieving the maximum score within the classroom); and the percentile rank within the classroom.

Table 1 presents the estimations of models that focus on proficiency in both Math and Language. Blind grades unequivocally produce a more lax proficiency standard for both white and black students. An average black student is 7.5 times more likely to be recognized as proficient by her teacher than by the blindly scored test. The figures also reveal that blacks tend to benefit more from these lower standards. Raw differences after controlling for classroom fixed effects indicate, for example, that blacks are 1.1 percentage points more likely to be deemed proficient in Math (relative to the blind classification) than their white counterparts. This is equivalent to a 10% increase in the probability of being considered proficient. Results for Portuguese Language are even more striking, and reach a 4.9 percentage points difference. The inclusion of controls in the regression reduces some of these effects. We have attempted to measure socio-economic status (parental education and household environment) and intrinsic student ability on both subjects (scores on the 2005 version of SARESP). Most importantly we have seriously attempted to uncover measures that reflect student behavioral characteristics such as attendance records, attendance self-reports, grades on Physical Education activities, and history of failure and school transference. Under the automatic promotion policy, failure and transfers are almost entirely related to behavioral problems. The inclusion of these variables reduces the size of the detected effects, particularly in the case of Language, but the parameters are still significantly different from zero and positive. The bottom line is that blacks tend to be favored relative to whites in the context of non-blind grading. As discussed above, this is not surprising, considering that blacks' proficiency levels (blindly scored) sit to the left of the whites'. When inflating grades for all students, teachers necessarily favor blacks.

When we turn to ordinal measures the picture that emerges is different, however. Table 2 focuses on the probability of assigning grades above the classroom mean as a form of relativizing proficiency measures. The model reveals that in both Math and Language, teachers are less likely to rate black students as above the mean (relative to the blind scores) than their white counterparts. A black student is 0.6 percentage point less likely to be above the Math average (1.31% of the baseline average) and 0.4 percentage point less likely to be above the Language average (0.83% of the baseline). While the estimator seems insensitive to the inclusion of covariates in the case of Math, the same is not true for Language. This is the case because Language grading may be more intensive on classroom interactions (Language information is gathered when students speak, for example) than Math grading (based on problem solving abilities).

The ordinal discrimination patterns also emerge when we focus on the "best in class" indicator. Table 3 reveals that blacks are less likely to be at the top of their classes according to their teachers (relative to the blind measures) in comparison with their white classmates. For Math the difference in 0.5 percentage point and for Language it is 0.6 percentage point. In relative terms these correspond to 6.9 and 8.6%, respectively.

Table 4 summarizes both cardinal and ordinal differentials by looking at differences in grades in a 0 to 100 scale and on the percentile rank within the classroom, respectively. Both measures indicate that blacks are underscored and underanked by their teachers, relative to what blind scores would suggest and in comparison with white students. It is important to notice that the cardinal measure now reveals that, on average, blacks fare worse. Combined with the results on the proficiency indicator above, this suggests that most of the discrimination against blacks occur at the top of the distribution, where competition seem to be fiercer.

In order to shed more light into this issue, in Table 5 we present estimates of ordinal differences (percentile rank) at different sub-samples of students (defined according to proficiency levels). In practice, we stratify the sample and produce new estimates of ranking within classrooms in each subsample. Proficiency levels can be defined either in accordance to blind scores or according to grades assigned by teachers. Both strategies are employed. The first two columns, based on SARESP scoring indicate, even if not overwhelmingly, that ranking changes that favor whites are larger among proficient students in both Math and Language examinations. They are still significant among the non-proficient ones, however. When we turn to levels defined by teacher grades the results are more striking. All the ordinal discrimination against blacks seems to be concentrated in the group teachers allocate to the top fourth of the grading range. We confirm, therefore, that the teachers' ordinal discrimination practices are more important for students that perform better, while the ones in the back are basically being favored by the grade inflation imposed by automatic promotion.

These findings seem to confirm that cardinal measures of discrimination in grading may be deceiving. Blacks can be favored by teachers in order to achieve some minimum proficiency requirement, but the data also reveals that they are ordinally unfavored when disputing ranking positions at the top of their classes. We are not at a position to infer the source of such differential treatment, however. We conjecture, nonetheless, that most of the impact can be tied back to statistical discrimination when assigning and rounding grades within the classroom. Well intentioned teachers can very well try to explore all the information available to rank students as they should be ranked. This results in a self-sustaining equilibrium with blacks being unfavored because they are expected to be receiving less investments from parents, and parents investing less because of the signals sent by teachers about their children's scholastic abilities. To illustrate our reasoning, Table 6 presents the estimates of cardinal and ordinal measures of discrimination by school grade. What we consider revealing in these numbers is that discrimination against blacks is reduced after the elementary level is concluded. Our conjecture here is that statistical discrimination would imply reversion of the expectations, since teachers know black kids that get far in school are probably "survivors" of creaming implicit in the schooling process. In other words, teachers may start to consider that a black child that gets far in school is probably better than the average white classmate. Even though we do not find significant positive discrimination in later grades, there is clear evidence that the negative discrimination is dramatically reduced (if not eliminated).

The bottom line of the empirical exercises presented here is that depending on how parents respond to cardinal and ordinal differential treatment, or how labor markets are structured (say as tournament-like contracts or not), there may be incentives within the school that lead to differential pre-market factors' accumulation. We conclude our empirical analysis with a glance at the impacts of inflation and reshuffling measures on parental and child decisions related to the accumulation of skills. Table 7 presents results of regressions of dropout rates, migration to private schools (search for quality), students' satisfaction with leaning process, and students' motivation to learn Math and Language materials, against the cardinal and ordinal measures discussed above. Grade inflation has impacts that are implicit on the automatic promotion policy. Yet their impacts go beyond the promotion effect. A child that has her grade inflated above her actual proficiency level is less likely to drop out (even after conditioning on being promoted). The child is also more satisfied with the learning process, but seem to be less likely to migrate to a private school. Ordinal measures on the other hand, indicate that reshuffling tends to slightly reduce dropout among the favored children, it also increases the chance of investment in private education, and has positive impact over satisfaction and motivation. In that sense, the overall impact of discrimination of blacks in our study may go in both directions. There is little doubt, however, that ordinal discrimination has clear negative impacts over children's non-cognitive abilities and over the probability of parents investing in higher quality education.

6 Conclusions

The present article examined the hypothesis that racial differences in labor market outcomes are the result of lower investment in the accumulation of skills or of *pre-market factors* by individuals of African decent. The reasoning presented focused on the existence of differential treatment within schools. In particular, the article investigated teachers' role in informing parents of their children's scholastic abilities. If children's *perceived* ability increases the return to parental investments, or if teacher's assessment influences some key non-cognitive aspects of a child's life (such as self-esteem, confidence and motivation), any differential treatment by teachers could very well reinforce racial gaps in the accumulation of human capital.

We examined Brazilian data on teachers' grading of elementary, middle and high-school students based on non-blind evaluation methods and compared it with evaluations resulting from standardized tests taken by the same children within the same classrooms and scored without human interference. Robust and consistent underscoring and underranking of blacks relative to whites were identified both on Math and on Reading/Grammar evaluations. Ordinal discrimination was particularly true for children at the top of the distribution of latent proficiency, while positive discrimination was identified among the lower-proficiency kids. Differences between cardinal and ordinal measures of discrimination were discussed, and their impact over investments in the quantity and quality of education was investigated. We conclude that there are grounds to support the hypothesis that differential accumulation of skills between blacks and whites in our context can be associated to imprecise signals of scholastic abilities issued by teachers. This suggests that evaluation systems within schools are a potential target for public policies that attempt to curb racial differentials in schooling.

References

[to be added]

Table 1: Probability of Being Classified as Proficient (non-blind grades vs. blind scores)

	Levels (blind)	Differences	D-in-D's			
Mathematics						
White	16.154	73.253***				
		(0.120)				
Black	10.949	75.223***	1.969***	1.097***	1.543***	0.814***
		(0.105)	(0.103)	(0.086)	(0.086)	(0.080)
Language/Grammar						
White	39.021	52.400***				
		(0.121)				
Black	28.871	59.337***	6.937***	4.909***	4.796***	1.605***
		(0.109)	(0.112)	(0.101)	(0.101)	(0.093)
Controls						
Classroom fixed-effects			-	Yes	Yes	Yes
Demographics			-	-	Yes	Yes
SES			-	-	Yes	Yes
Ability: past blind scores			-	-	-	Yes
Additional ability: current writing sample			-	-	-	Yes
Behavior: teacher's attendance records			-	-	-	Yes
Behavior: self-reported attendance			-	-	-	Yes
Behavior: Physical Education grades and attendance			-	-	-	Yes
Behavior: Failure + Transfers History			-	-	-	Yes

Notes: Sample is 1,164,113 students. Standard errors are clustered at the classroom level.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 2: Probability of Being Classified as Above the Class Mean (non-blind grades vs. blind scores)

	Levels (blind)	Differences	D-in-D's			
Mathematics						
White	51.240	0.026				
		(0.090)				
Black	45.671	-0.683***	-0.709***	-0.697***	-0.07	-0.576***
		(0.092)	(0.113)	(0.118)	(0.117)	(0.115)
Language/Grammar						
White	55.687	-2.216***				
		(0.085)				
Black	48.066	-1.230***	0.986***	0.676***	1.163***	-0.372***
		(0.087)	(0.109)	(0.114)	(0.114)	(0.111)
Controls						
Classroom fixed-effects			-	Yes	Yes	Yes
Demographics			-	-	Yes	Yes
SES			-	-	Yes	Yes
Ability: past blind scores			-	-	-	Yes
Additional ability: current writing sample			-	-	-	Yes
Behavior: teacher's attendance records			-	-	-	Yes
Behavior: self-reported attendance			-	-	-	Yes
Behavior: Physical Education grades and attendance			-	-	-	Yes
Behavior: Failure + Transfers History			-	-	-	Yes

Notes: Sample is 1,164,113 students. Standard errors are clustered at the classroom level.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 3: Probability of Being Classified as Best in Class (non-blind grades vs. blind scores)

	Levels (blind)	Differences	D-in-D's			
Mathematics						
White	9.304	2.883***				
		(0.075)				
Black	7.209	2.132***	-0.751***	-1.002***	-0.746***	-0.482***
		(0.073)	(0.076)	(0.072)	(0.072)	(0.072)
Language/Grammar						
White	9.986	3.261***				
		(0.077)				
Black	7.005	2.964***	-0.297***	-0.789***	-0.552***	-0.594***
		(0.070)	(0.077)	(0.075)	(0.074)	(0.074)
Controls						
Classroom fixed-effects			-	Yes	Yes	Yes
Demographics			-	-	Yes	Yes
SES			-	-	Yes	Yes
Ability: past blind scores			-	-	-	Yes
Additional ability: current writing sample			-	-	-	Yes
Behavior: teacher's attendance records			-	-	-	Yes
Behavior: self-reported attendance			-	-	-	Yes
Behavior: Physical Education grades and attendance			-	-	-	Yes
Behavior: Failure + Transfers History			-	-	-	Yes

Notes: Sample is 1,164,113 students. Standard errors are clustered at the classroom level.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 4: Summary Cardinal and Ordinal Differences

	Grades (0-100 scale) vs. Percentage of Correct Answers in blindly scored test			Ranking of Grades vs. Ranking of IRT Proficiency Levels		
	Levels (blind)	Differences	D-in-D's	Levels (blind)	Differences	D-in-D's
Mathematics						
White	44.793	18.685***		50.444	0.139***	
		(0.075)			(0.029)	
Black	42.625	17.472***	-0.429***	46.723	-0.130***	-0.298***
		(0.068)	(0.031)		(0.028)	(0.059)
Language/Grammar						
White	59.209	5.981***		51.001	-0.258***	
		(0.059)			(0.027)	
Black	54.764	6.948***	-0.211***	46.109	0.181***	-0.304***
		(0.055)	(0.031)		(0.026)	(0.056)
Controls						
Classroom fixed-effects			Yes			Yes
Demographics			Yes			Yes
SES			Yes			Yes
Ability: past blind scores			Yes			Yes
Additional ability: current writing sample			Yes			Yes
Behavior: teacher's attendance records			Yes			Yes
Behavior: self-reported attendance			Yes			Yes
Behavior: Physical Education grades and attendance			Yes			Yes
Behavior: Failure + Transfers History			Yes			Yes

Notes: Sample is 1,164,113 students. Standard errors are clustered at the classroom level.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: Ranking of Grades vs. Ranking of IRT Proficiency Levels (Stratifications based on student's performance level)

	Proficient (according to blind scores)	NOT Proficient (according to blind scores)	Teacher Grade Level 1 (0 to 2.5)	Teacher Grade Level 2 (2.5 to 5.0)	Teacher Grade Level 3 (5.0 to 7.5)	Teacher Grade Level 4 (7.5 to 10)
Mathematics						
Black vs white	-1.076***	-0.674***	0.411	-0.017	-0.045	-0.590***
	(0.259)	(0.194)	(1.419)	(0.527)	(0.103)	(0.194)
Language/Grammar						
Black vs white	-1.137***	-0.955***	2.699**	0.003	-0.041	-0.655***
	(0.262)	(0.186)	(1.363)	(0.502)	(0.098)	(0.195)
Sample	*103212*	*208949*	*4423*	*31871*	*565822*	*160358*

Notes: Standard errors are clustered at the classroom level. All estimations include classroom fixed-effects.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 6: Proficiency, Above Average and Best-in-Class Classification Differentials by School Grade

	D-in-D's Proficiency	D-in-D's Above Average	D-in-D's Best	D-in-D's Scores	D-in-D's Ranking	
Mathematics						*sample*
4th grade	2.445***	-0.614***	-1.068***	-0.175**	-0.442***	*201590*
	(0.202)	(0.236)	(0.182)	(0.072)	(0.112)	
6th grade	1.537***	-0.839***	-0.500***	-0.409***	-0.409***	*381024*
	(0.166)	(0.193)	(0.116)	(0.053)	(0.097)	
8th grade	0.012	-0.610***	-0.431***	-0.458***	-0.294***	*353603*
	(0.118)	(0.218)	(0.124)	(0.055)	(0.112)	
11th grade	0.074	0.075	-0.302*	-0.273***	0.134	*226727*
	(0.125)	(0.290)	(0.181)	(0.067)	(0.158)	
Reading						
4th grade	2.133***	-0.474**	-0.937***	-0.190**	-0.381***	*201590*
	(0.217)	(0.238)	(0.190)	(0.075)	(0.113)	
6th grade	1.754***	-1.139***	-0.721***	-0.466***	-0.740***	*381024*
	(0.173)	(0.191)	(0.120)	(0.054)	(0.095)	
8th grade	1.839***	-0.043	-0.430***	-0.022	-0.056	*353603*
	(0.163)	(0.203)	(0.132)	(0.057)	(0.100)	
11th grade	0.891***	0.261	-0.668***	0.016	-0.025	*226727*
	(0.188)	(0.271)	(0.176)	(0.070)	(0.142)	
Controls	Yes	Yes	Yes	Yes	Yes	

Notes: Standard errors are clustered at the classroom level.

* significant at 10%; ** significant at 5%; *** significant at 1%

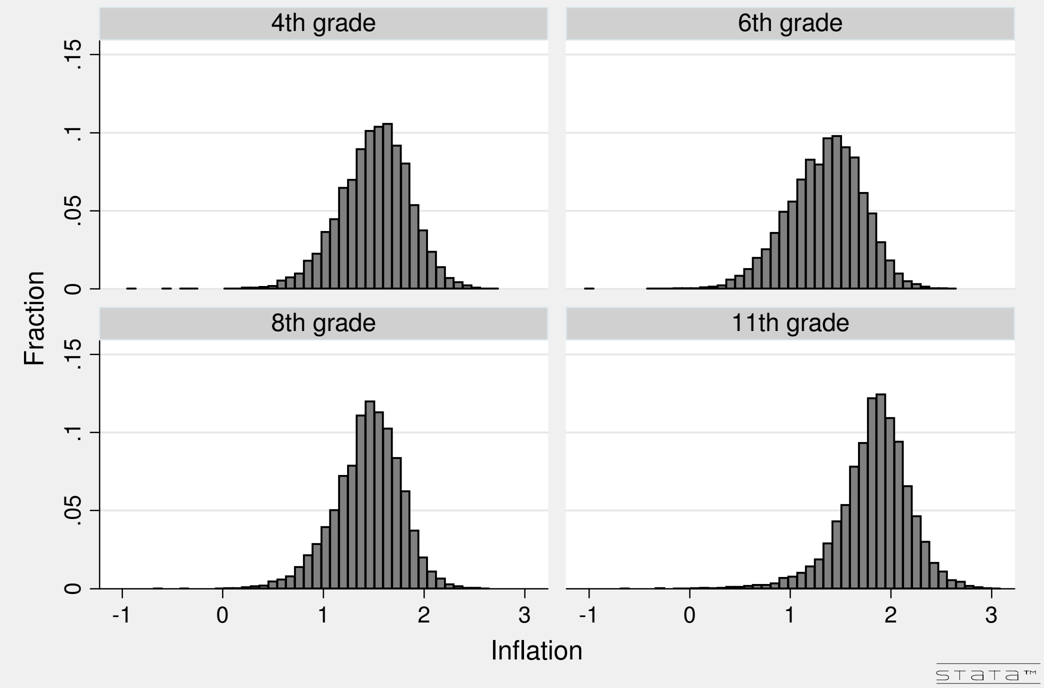
Table 7: Effects of mis-ranking and mis-scoring by teachers

	Dropout probability		Dropout probability		Private Education Switch Probability		Private Education Switch Probability		Satisfaction		Satisfaction		Motivation		Motivation	
Non-blind versus blind evaluations																
Mis-ranking in Math (+ means over-ranking)			0.001	-0.0011			0.0059***	0.0057***			0.1291***	0.1291***			0.2233***	0.2215***
			(0.0010)	(0.0010)			(0.0007)	(0.0007)			(0.0031)	(0.0030)			(0.0033)	(0.0033)
Mis-ranking in Language (+ means over-ranking)			-0.0049***	-0.0045***			0.0071***	0.0070***			0.0754***	0.0737***			0.0363***	0.0353***
			(0.0010)	(0.0010)			(0.0007)	(0.0007)			(0.0031)	(0.0030)			(0.0033)	(0.0032)
Proficiency indicator Math	-0.0009	-0.0111***	-0.0008	-0.0104***	-0.0014***	-0.0011***	-0.0032***	-0.0028***	0.0430***	0.0425***	0.0164***	0.0156***	0.0615***	0.0623***	0.0181***	0.0187***
	(0.0008)	(0.0009)	(0.0009)	(0.0009)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0026)	(0.0023)	(0.0027)	(0.0024)	(0.0027)	(0.0024)	(0.0028)	(0.0025)
Proficiency indicator Language	-0.0032***	-0.0201***	-0.0024**	-0.0193***	-0.0007	-0.0006*	-0.0026***	-0.0024***	0.0204***	0.0194***	0.004	0.0034	-0.0085***	-0.0087***	-0.0199***	-0.0194***
	(0.0010)	(0.0010)	(0.0010)	(0.0011)	(0.0004)	(0.0004)	(0.0005)	(0.0004)	(0.0030)	(0.0025)	(0.0030)	(0.0026)	(0.0029)	(0.0025)	(0.0030)	(0.0026)
Best in Math	0.0002	0.001			0.0041***	0.0042***			0.0432***	0.0427***			0.0599***	0.0597***		
	(0.0006)	(0.0007)			(0.0006)	(0.0006)			(0.0019)	(0.0019)			(0.0020)	(0.0020)		
Best in Language	-0.0010*	0.000			0.0047***	0.0046***			0.0343***	0.0339***			0.0365***	0.0366***		
	(0.0006)	(0.0006)			(0.0006)	(0.0006)			(0.0018)	(0.0018)			(0.0020)	(0.0020)		
Above average in Math	0.0008*	0.0008*			0.0013***	0.0013***			0.0437***	0.0433***			0.0746***	0.0733***		
	(0.0004)	(0.0005)			(0.0003)	(0.0003)			(0.0014)	(0.0014)			(0.0015)	(0.0014)		
Above average in Language	-0.0013***	-0.0007			0.0016***	0.0016***			0.0265***	0.0258***			0.0126***	0.0127***		
	(0.0004)	(0.0005)			(0.0003)	(0.0003)			(0.0014)	(0.0014)			(0.0015)	(0.0014)		
Average rate among blacks (%)	2.23	2.79	2.23	2.79	0.58	0.54	0.58	0.54	71.46	70.17	71.46	70.17	65.44	64.06	65.44	64.06
Sample	884635	972465	884635	972465	884635	972465	884635	972465	884635	972465	884635	972465	884635	972465	884635	972465
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Passing grade indicator	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Sample conditional on being promoted	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No

Notes: Standard errors are clustered at the classroom level. Sample is 972465 students in 4th, 6th and 8th grades. All estimations include classroom fixed-effects.

* significant at 10%; ** significant at 5%; *** significant at 1%

Figure 1A: Grade Inflation by grade: Math



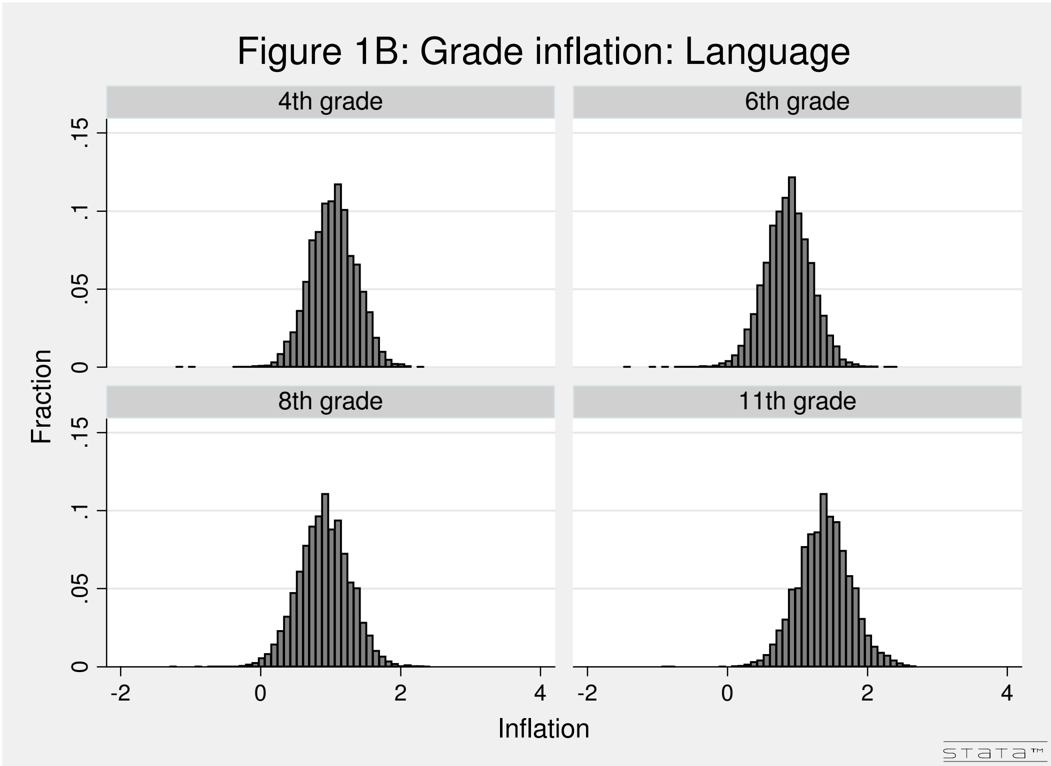


Figure 1B: Grade inflation: Language
4th grade
6th grade
8th grade
11th grade
Fraction
0
.05
.1
.15
-2
0
2
4
Inflation

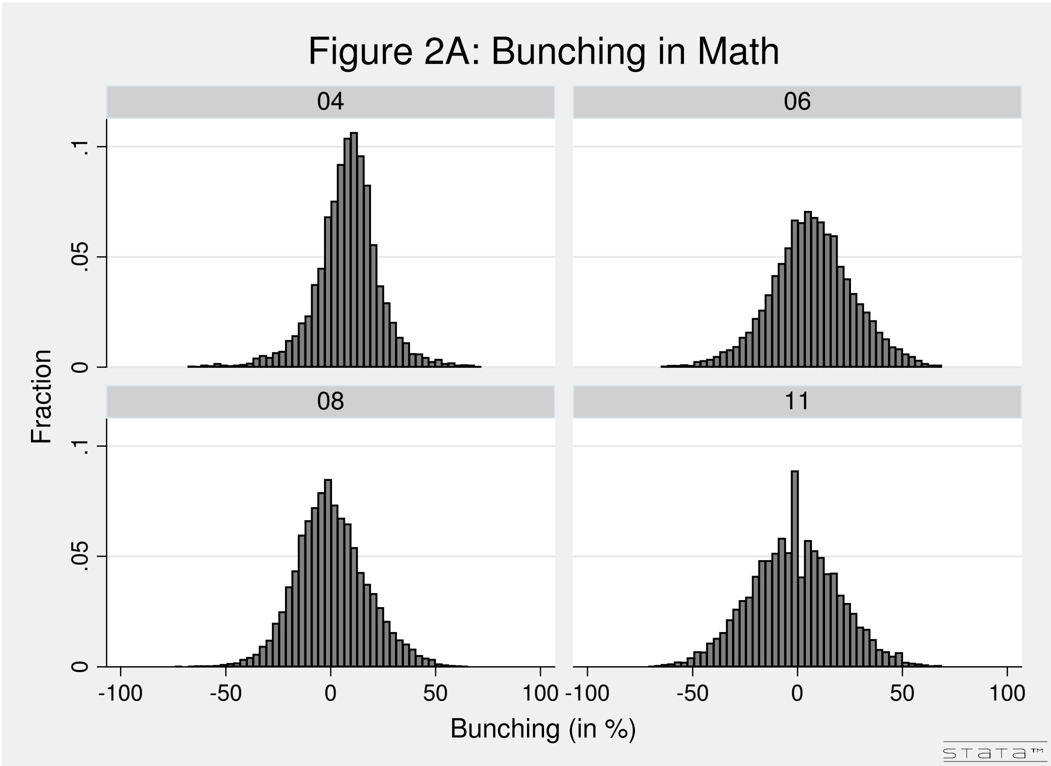
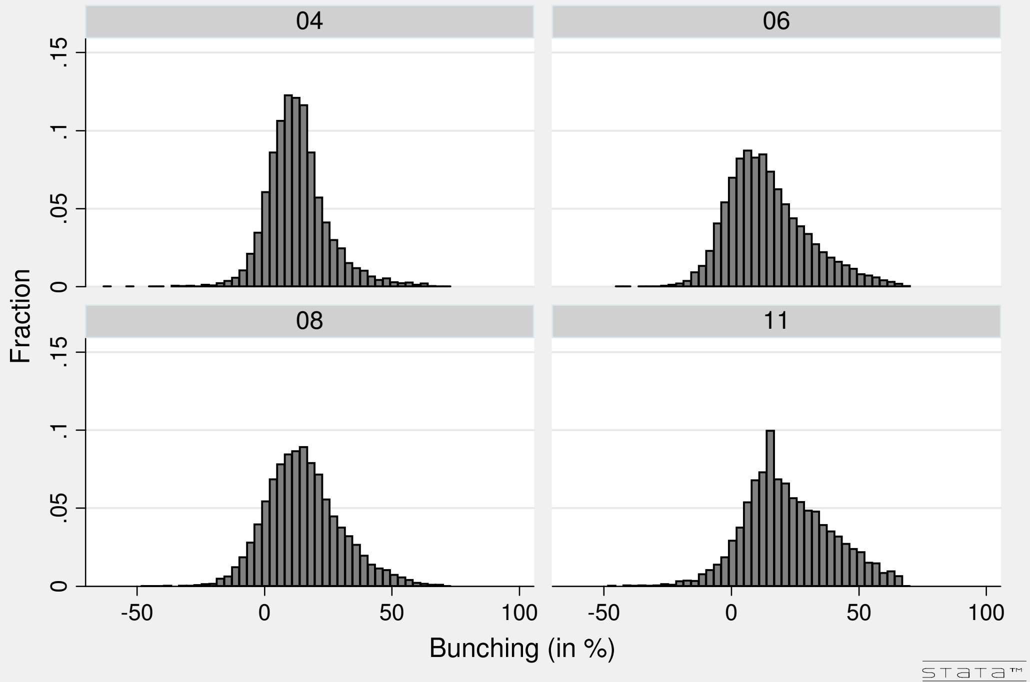


Figure 2A: Bunching in Math
04
06
08
11
Fraction
0
.05
.1
-100
-50
0
50
100
Bunching (in %)

Figure 2B: Bunching in Language



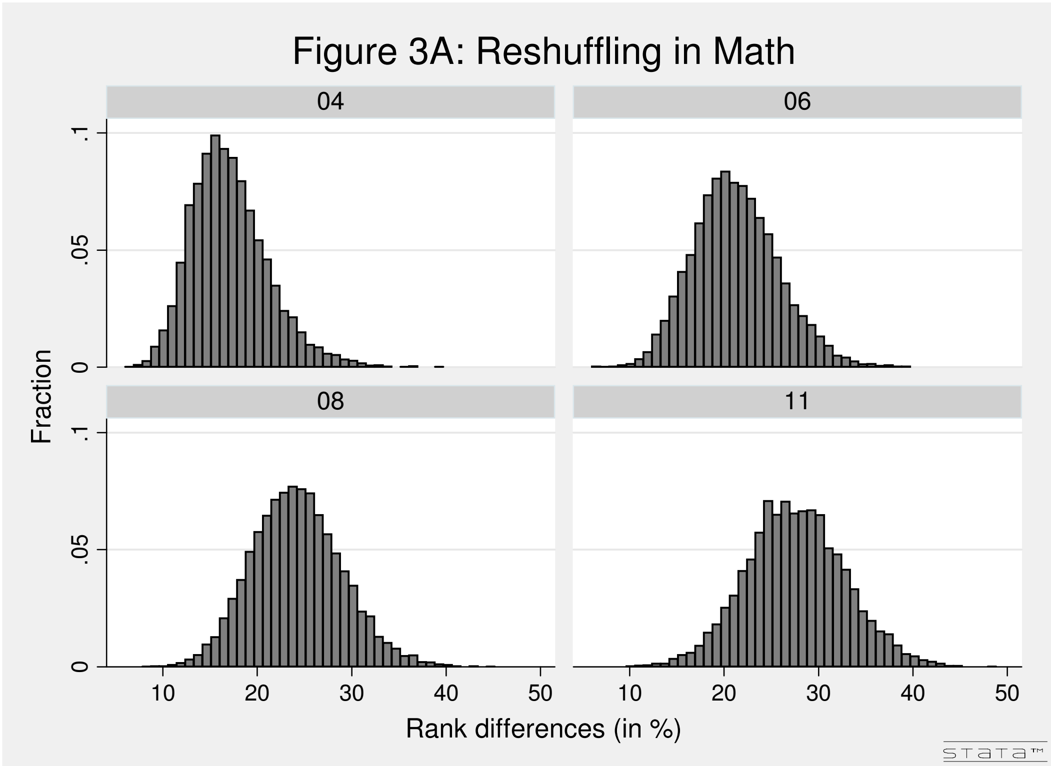


Figure 3A: Reshuffling in Math
04
06
08
11
Fraction
0
.05
.1
10
20
30
40
50
Rank differences (in %)

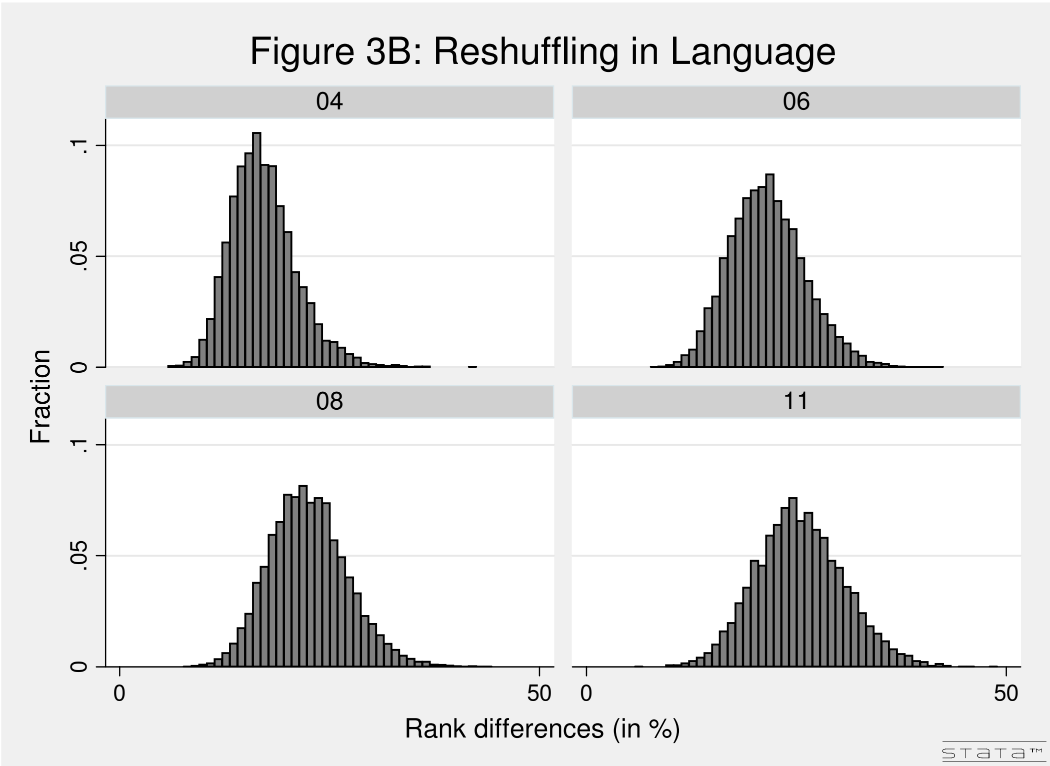
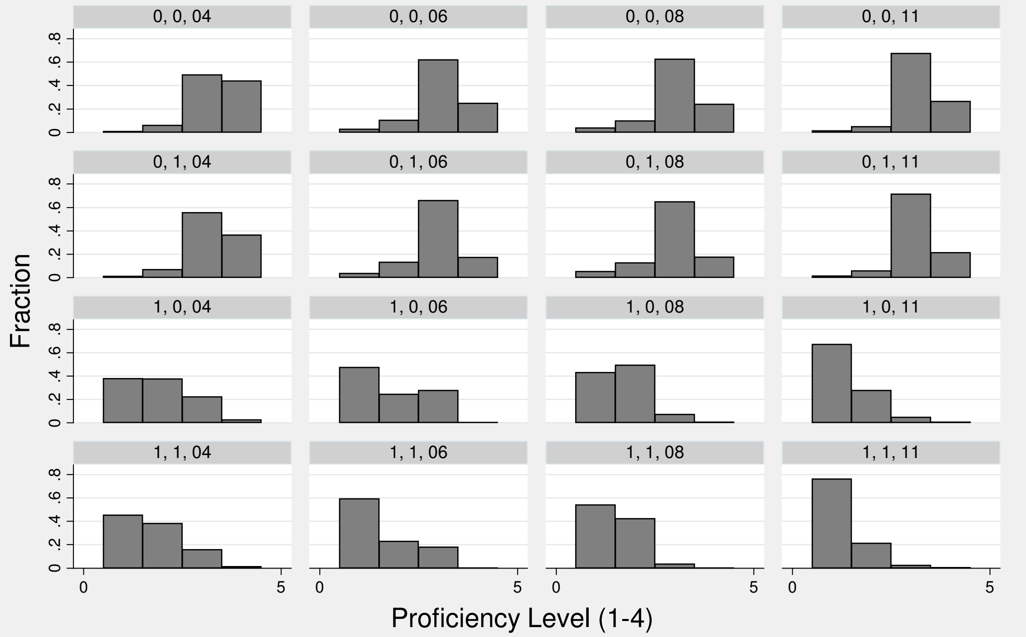


Figure 3B: Reshuffling in Language
04
06
08
11
Fraction
0
.05
.1
0
50
Rank differences (in %)

Figure 4A: Proficiency Lervel in Mathematics



Graphs are by triple (1{blind test}, 1{black}, grade)

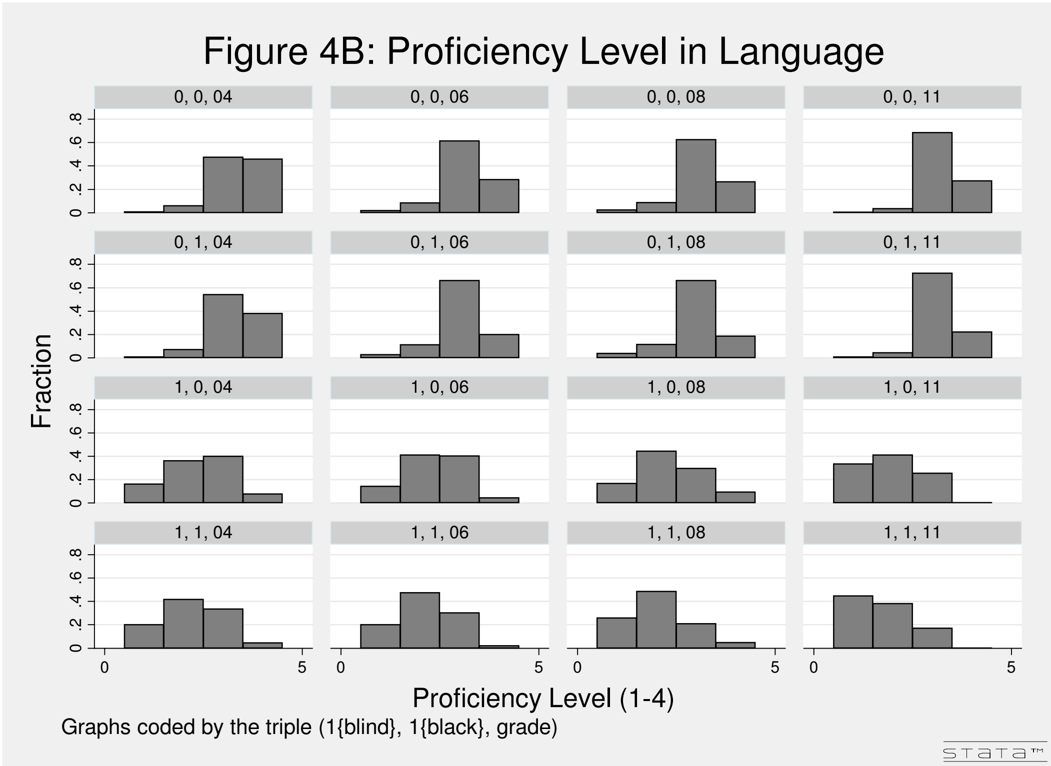


Figure 4B: Proficiency Level in Language

Table A1: Descriptive Statistics

	By School Grade			
	4th grade	6th grade	8th grade	11th grade
Demographics				
Share of black students	0.518 (0.0017)	0.507 (0.0013)	0.483 (0.0014)	0.418 (0.0017)
Share of male students	0.508 (0.0011)	0.502 (0.0009)	0.499 (0.0009)	0.451 (0.0012)
Year of birth	1996.612 (0.0026)	1994.498 (0.0022)	1992.422 (0.0024)	1989.187 (0.0060)
Month of birth	6.454 (0.0086)	6.466 (0.0064)	6.536 (0.0066)	6.257 (0.0081)
SES				
Household in paved road	0.830 (0.0017)	0.772 (0.0017)	0.804 (0.0018)	0.851 (0.0018)
Household with piped water	0.885 (0.0011)	0.911 (0.0006)	0.947 (0.0005)	0.968 (0.0005)
Household with electricity	0.950 (0.0006)	0.969 (0.0003)	0.981 (0.0003)	0.986 (0.0003)
Household has newspapers regularly	0.766 (0.0015)	0.510 (0.0012)	0.445 (0.0011)	0.362 (0.0013)
Household has dictionary	0.867 (0.0012)	0.878 (0.0008)	0.908 (0.0006)	0.900 (0.0008)
Household has Internet	0.383 (0.0017)	0.354 (0.0014)	0.375 (0.0016)	0.441 (0.0020)
Household has a place calm for study	0.789 (0.0012)	0.813 (0.0008)	0.828 (0.0008)	0.781 (0.0011)
Household has 20 books or more	0.459 (0.0015)	0.415 (0.0011)	0.515 (0.0011)	0.454 (0.0015)
Schooling history				
Failed school in 2005	0.011 (0.0003)	0.003 (0.0001)	0.004 (0.0001)	0.002 (0.0001)
Failed school in 2006	0.048 (0.0009)	0.005 (0.0001)	0.012 (0.0002)	0.002 (0.0001)
Changed schools 2005-2006	0.069 (0.0012)	0.009 (0.0002)	0.010 (0.0002)	0.003 (0.0001)
Changed schools 2006-2007	0.061 (0.0013)	0.002 (0.0001)	0.009 (0.0002)	0.002 (0.0001)
Attendance rate in Math	92.635 (0.0316)	90.662 (0.0294)	89.304 (0.0330)	87.332 (0.0498)
Attendance rate in Language	92.635 (0.0316)	90.931 (0.0285)	89.685 (0.0323)	87.771 (0.0481)
Outcomes				
Math Proficiency (blind IRT score)	184.181 (0.2571)	194.821 (0.1547)	232.334 (0.1703)	263.834 (0.2006)
Language Proficiency (blind IRT score)	188.482 (0.2134)	211.022 (0.1342)	244.412 (0.2157)	264.206 (0.1902)
Percentage of correct answers in Math (blind test)	55.148 (0.1202)	52.491 (0.0680)	36.026 (0.0601)	30.307 (0.0579)
Percentage of correct answers in Language (blind test)	62.671 (0.1195)	54.902 (0.0654)	60.668 (0.0706)	48.450 (0.0703)
Grade in Math (non-blind, 0-10 scale)	6.834 (0.0097)	5.951 (0.0070)	5.894 (0.0073)	6.327 (0.0089)
Grade in Language (non-blind, 0-10 scale)	6.900 (0.0097)	6.159 (0.0070)	6.090 (0.0073)	6.456 (0.0086)
Sample	201590	382173	353623	226727

Table A2: Descriptive Statistics

	By Race			
	Whites		Blacks	
Demographics				
Share of black students	0.000		100.000	
Share of male students	0.472	(0.001)	0.510	(0.001)
Year of birth	1993.028	(0.013)	1993.318	(0.012)
Month of birth	6.415	(0.005)	6.466	(0.005)
SES				
Household in paved road	0.834	(0.001)	0.786	(0.001)
Household with piped water	0.939	(0.000)	0.923	(0.000)
Household with electricity	0.979	(0.000)	0.969	(0.000)
Household has newspapers regularly	0.501	(0.001)	0.506	(0.001)
Household has dictionary	0.906	(0.000)	0.877	(0.001)
Household has Internet	0.443	(0.001)	0.325	(0.001)
Household has a place calm for study	0.830	(0.001)	0.790	(0.001)
Household has 20 books or more	0.472	(0.001)	0.449	(0.001)
Schooling history				
Failed school in 2005	0.003	(0.000)	0.005	(0.000)
Failed school in 2006	0.011	(0.000)	0.016	(0.000)
Changed schools 2005-2006	0.015	(0.000)	0.020	(0.000)
Changed schools 2006-2007	0.012	(0.000)	0.016	(0.000)
Attendance rate in Math	90.099	(0.022)	89.831	(0.021)
Attendance rate in Language	90.402	(0.021)	90.107	(0.021)
Outcomes				
Math Proficiency (blind IRT score)	226.099	(0.191)	211.321	(0.166)
Language Proficiency (blind IRT score)	237.049	(0.183)	220.243	(0.157)
Percentage of correct answers in Math (blind test)	44.793	(0.073)	42.625	(0.062)
Percentage of correct answers in Language (blind test)	59.209	(0.054)	54.764	(0.047)
Grade in Math (non-blind, 0-10 scale)	6.348	(0.005)	6.010	(0.005)
Grade in Language (non-blind, 0-10 scale)	6.519	(0.005)	6.171	(0.005)
Proficiency rate in Math (according to teacher)	0.894	(0.001)	0.862	(0.001)
Proficiency rate in Math (blind test)	0.162	(0.001)	0.109	(0.001)
Proficiency rate in Language (according to teacher)	0.914	(0.001)	0.882	(0.001)
Proficiency rate in Language (blind test)	0.390	(0.001)	0.289	(0.001)
Pctile Rank (blind Math proficiency)	0.504	(0.000)	0.465	(0.000)
Pctile Rank (blind Language proficiency)	0.510	(0.000)	0.461	(0.000)
Pctile Rank (non-blind Math grade)	0.506	(0.000)	0.464	(0.000)
Pctile Rank (non-blind Language grade)	0.507	(0.000)	0.463	(0.000)
Sample		1164113		

References

[1] Alexander, N.; A Guimarães; C. Hamilton; L. Huntley and W. James (2001); "Beyond Racism: Race and Inequality in Brazil, South Africa, and the United States," L. Rienner Publisher, Boulder, CO.

[2] Allen, W.; E. Telles and M. Hunter (2000); "Skin Color, Income and Education: A Comparison of African Americans and Mexican Americans" National Journal of Sociology, Vol (12), pp. 129-180.

[3] Antonovics, K. (2006); "Statistical Discrimination and Intergenerational Income Mobility," Unpublished manuscript, University of California at San Diego.

[4] Arias, O.; G Yamada and L.Tejerina (2004); "Education, family background and racial earnings inequality in Brazil", manuscript, Inter-American Development Bank, Washington, DC, USA.

[5] Barsh, G. S. (2003); "What Controls Variation in Human Skin Color?" PLoS Biol 1(1): e27.

[6] Becker, G. and N. Tomes (1976); "Child Endowments and the Quantity and Quality of Children" The Journal of Political Economy, Vol. 84(4), Part 2: Essays in Labor Economics in Honor of H. Gregg Lewis, pp. S143-S162.

[7] Behrman, J. (1988); "Intrahousehold Allocation of Nutrients in Rural India: Are Boys Favored? Do Parents Exhibit Inequality Aversion?," Oxford Economic Papers, Oxford University Press, Vol. 40(1), pp. 32-54.

[8] Behrman, J.; R. Pollak and P. Taubman (1986); "Do Parents Favor Boys?," International Economic Review, Vol. 27(1), pp. 33-54.

[9] Behrman, J.; R. Pollak and P. Taubman (1982). "Parental Preferences and Provision for Progeny," Journal of Political Economy, University of Chicago Press, Vol. 90(1), pp. 52-73.

[10] Beltrao, Kaizo, "Acesso a Educacao: Diferenciais entre os Sexos," IPEA Texto para Discussao No. 879, May, 2002.

[11] Bertrand, M. and S. Mullainathan (2004); "Are Emily and Brendan More Employable than Latoya and Tyrone? Evidence on Racial Discrimination in the Labor Market from a Large Randomized Experiment", The American Economic Review, Vol 94 (4), September, pp. 991-1013.

[12] Biddle, J. and D. Hamermesh (1998); "Beauty, Productivity, and Discrimination: Lawyers' Looks and Lucre," Journal of Labor Economics, University of Chicago Press, Vol. 16(1), January, pp. 172-201.

[13] Blume, L. (2006); "The Dynamics of Statistical Discrimination", The Economic Journal, Vol 116, November, pp. F480-F498.

[14] Bodenhorn, H. and C. Ruebeck (2005); "Colorism and African-American Wealth: Evidence from the Nineteenth- Century South," forthcoming in Journal of Population Economics.

[15] Bodenhorn, H. (2003); “The Complexion Gap: The Economic Consequences of Color among Free African Americans in the Rural Antebellum South,” in Kyle D. Kauffman, ed., Advances in agricultural economic history. Vol. 2. Amsterdam: Elsevier Science North-Holland, pp. 41–73.

[16] Bonilla-Silva, Eduardo (2002); “We are all Americans!: The Latin Americanization of racial stratification in the USA,”Race & Society, Vol 5, pp. 3–16.

[17] Brunsma, D. (2005); “Interracial Families and the Racial Identification ofMixed-Race Children: Evidence from the Early Childhood Longitudinal Study”, Social Forces, Vol. 84(2), pp.1131-1157.

[18] Campante, F.; A. Crespo and P. Leite (2004); “Wage Inequality across Races in Brazilian Urban Labor Markets: Regional Aspects.” Revista Brasileira de Economia, Vol. 58 (2), pp. 185-210.

[19] Campbell, Mary E. (2007). “A Rainbow Divide? Racial Stratification in a Multiracial Context,” unpublished manuscript, University of Wisconsin at Madison.

[20] Campbell, Mary E. and Jennifer Eggerling-Boeck (2006). “What about the Children? The Psychological and Social Well-Being of Multiracial Adolescents,” The Sociological Quarterly, Vol. 47(1), pp. 147–173.

[21] Carneiro, P.; F. Cunha and J. Heckman (2004); “The Technology of Skill Formation,” 2004 Meeting Papers 681, Society for Economic Dynamics.

[22] Carneiro, P.; J. Heckman and D. Masterov (2005); “Labor Market Discrimination and Racial Differences in Premarket Factors,” Journal of Law and Economics, Vol. 48(1), pp. 1-40.

[23] Carvalho-Silva, D.R.; F.R. Santos; J. Rocha and S. Pena (2001); “The phylogeography of Brazilian Y-chromosome lineages,” American Journal of Human Genetics, Vol. 68, pp. 281–286.

[24] Case, Anne, I-Fen Lin and Sara McLanahan, “Household Resource Allocation in Stepfamilies: Darwin Reflects on the Plight of Cinderella,” American Economic Review, Vol. LXXXIX (1999), No. 2, 234-248.

[25] Charles, K. and M. Luoh (2006); “Male Incarceration, the Marriage Market and Female Outcomes.” Unpublished Manuscript.

[26] Daly, Martin and Margo Wilson, The Truth About Cinderella: A Darwinian View of Parental Love (New Haven, CT, Yale University Press, 1998).

[27] Diamond, J. (1994); “Race Without Color”, Discover, Vol 15(11).

[28] Dollard, J. (1937); “Caste and Class in a Southern Town”, Yale University Press.

[29] Dunn, L. C. and T. Dobzhansky (1958); “Heredity, Race and Society: A Scientific Explanation of Human Differences,” Mentor New American Library Publisher.

[30] Freeman, H., D. Armor, J. M. Ross and T. J. Pettigrew (1966); “Color Gradation and Attitudes among Middle Income Negroes.” American Sociological Review, Vol.3 (1), pp. 365-374.

[31] Fryer, R.(2006); "Guess Who's Been Coming to Dinner? Trends in Interracial Marriage over the 20th Century." forthcoming in Journal of Economic Perspectives.

[32] Fryer, R. and S. Levitt (2004a); "Causes and Consequences of Distinctive Black Names," The Quarterly Journal of Economics, MIT Press, Vol. 119(3), pp. 767-805

[33] Fryer, R. and S. Levitt (2004b); "Understanding the Black-White Test Score Gap in the First Two Years of School," The Review of Economics and Statistics, MIT Press, Vol. 86(2), pp. 447-464.

[34] Gans, Herbert. (2006); "The Possibility of a New Racial Hierarchy in the 21st-Century United States," in M. Lamont (ed.) The Cultural territories of Race: University of Chicago Press, Chicago, pp. 371-390.

[35] Goldsmith, A.; D. Hamilton and W. Darity Jr. (2005); "From Dark to Light: Skin Color and Wages among African-Americans." Unpublished manuscript.

[36] Gyimah-Brempong, K. and G. Price (2006); "Crime and Punishment: And Skin Hue Too?" The American Economic Review, Vol 96(2), pp. 246-250.

[37] Hall, R. (1995); "The Bleaching Syndrome: African Americans' response to cultural domination vis-a-vis Skin Color ," Journal of Black Studies, Vol. 26, pp. 172-184.

[38] Hamermesh, D. and J. Biddle (1994); "Beauty and the Labor Market," The American Economic Review, Vol. 84(5), pp. 1174-1194.

[39] Heckman, J. (1998); "Detecting Discrimination" The Journal of Economic Perspectives, Vol. 12(2), Spring, pp. 101-116.

[40] Herring, C.; V. M. Keith and H. D. Horton (2004); "Skin Deep: How Race and Complexion Matter in the Color-Blind Era", University of Illinois Press, Chicago, IL.

[41] Hersch, J. (2006); "Skin Color and Wages Among New U.S. Immigrants" Unpublished manuscript.

[42] Hill, M. (2000); "Color Differences in the Socioeconomic Status of African American Men: Results of a Longitudinal Study." Social Forces, Vol. 78(4):1437-1460.

[43] Hunter, M.; W. Allen and E. Telles (2001); "The Significance of Skin Color among African Americans and Mexican Americans." African American Research Perspectives, Vol. 7(1), pp. 173–184.

[44] Keith, V. and C. Herring (1991); "Skin Tone and Stratification in the Black Community," American Journal of Sociology, Vol 97, pp. 760-778.

[45] King, J. (1971); "The Biology of Race," Harcourt Brace Jovanovich Publisher

[46] Lazear, E. (1980); "Family Background and Optimal Schooling Decisions," The Review of Economics and Statistics, Vol. 62 (1), February, pp. 42-51.

[47] Levhari, D. and Y. Weiss (1974); "The Effect of Risk on the Investment in Human Capital," The American Economic Review, Vol. 64 (6), december, pp. 950-963.

[48] Lopez, A. (2003); "Mixed-Race School-Age Children: A summary of Census 2000 data." Educational Researcher, Vol. 32 (6), 25-37.

[49] McLanahan, Sara and Gary Sandefur, Growing Up with a Single Parent: What Hurts, What Helps. (Cambridge, MA, Harvard University Press, 1994).

[50] Neal, D. and W. Johnson (1996); "The Role of Premarket Factors in Black-White Wage Differences," Journal of Political Economy, University of Chicago Press, Vol. 104(5), October, pp. 869-895.

[51] Park, J. and M. Lee (2005); "A Study of Skin Color by Melanin Index according to site, gestational age, birth-weight and season of birth in Korean neonates," Journal of KoreanMedical Science, Vol. 20, pp. 105-108.

[52] Parra, E. J.; R. A. Kittles and M. D. Shriver (2004); "Implications of Correlations Betweeen Skin Color and Genetic Ancestry for Biomedical Reserach" Nature Genetics Supplement, Vol. 36(11).

[53] Pena, S.; D. Carvalho-Silva; J. Alves-Silva; V. Prado and F. Santos (200); "Retrato Molecular do Brazil," Ciencia Hoje, Vol 27 (159).

[54] Perry, G.; O. Arias; J. H. López; W. Maloney and L. Servén (2006); "Poverty Reduction and Growth: Virtuous and Vicious Circles" The World Bank.

[55] Rangel, M. (2007); "Measurement and Mismeasurement of Skin Color, Racial and Ethnic Differentials: Socioeconomic and Health Outcomes" manuscript, University of Chicago.

[56] Ransford, E. (1970); "Skin Color, Life Chances and Anti-White Attitudes." Social Problems, Vol. 18, pp.164-178.

[57] Rees, J. (2003); "Genetics of Hair and Skin Color," Annual Review of Genetics, Vol 37, pp. 67-90.

[58] Relethford, J. (1997); "Hemispheric Difference in Human Skin Color," American Journal of Physical Anthropology, Vol. 104, pp. 449-457.

[59] Reuter, E. (1918); "The Mulatto in the United States: Including a Study of the Role of Mixed-Blood Races throughout the World". Richard G. Badger Publisher.

[60] Rockquemore, K. and T. Laszloffy (2005); "Raising Biracial Children." Altamira Press.

[61] Rosenzweig, M. and T. P. Schultz (1982); "Market Opportunities, Genetic Endowments, and Intrafamily Resource Distribution: Child Survival in Rural India," The American Economic Review, Vol. 72 (4), pp. 803-815.

[62] Seeman, M. (1946); "Skin Color Values in Three All-Negro School Classes." American Sociological Review, Vol 11, pp. 315-321.

[63] Sheshinski, E. and Y. Weiss (1982); "Inequality Within and Between Families" The Journal of Political Economy, Vol. 90 (1), pp. 105-127.

[64] Shih, M. and D. Sanchez (2005); "Perspectives and research on the positive and negative implications of having multiple racial identities." Psychological Bulletin, Vol. 131, pp. 569-591.

[65] Sturm, R.; N. Box and M. Ramsay (1998); "Human Pigmentation Genetics: the difference is only skin deep," BioEssay, Vol. 20, pp. 712-721.

[66] Telles, E. (2005); "Race in Another America: The Significance of Skin Color in Brazil." Princeton University Press.

[67] United Nations (2005); "Brazil: National Human Development Report", New York and Brasilia, Brazil.

[68] Yancey, George (2006); "Racial Justice in a Black/Nonblack Society," in D. Brunsma (ed.), Mixed messages: Multiracial identities in the Color Blind Era: Boulder, CO; Rienner, pp. 49-62.

[69] Zvoch, Keith, "Family Type and Investment in Education: A Comparison of Genetic and Stepparent Families," Evolution and Human Behavior, Vol. XX (1999), 435-464.