



Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev



Same work, lower grade? Student ethnicity and teachers' subjective assessments

Reyn van Ewijk*

VU University Amsterdam, De Boelelaan 1105, 1081 HV, Amsterdam, The Netherlands

ARTICLE INFO

Article history:

Received 26 August 2010
Received in revised form 10 May 2011
Accepted 14 May 2011

JEL classification:

I2
J15

Keywords:

Ethnicity
Discrimination
Grading
Experiment

ABSTRACT

Previous research shows that ethnic minority students perform poorer in school when they are taught by teachers belonging to the ethnic majority. Why this is the case was unclear. This paper focuses on one important potential explanation: I examine whether ethnic majority teachers grade minority and majority students differently for the same work. Using an experiment, I show that such a direct grading bias does not occur. I do find indirect evidence for alternative explanations: teachers report lower expectations and unfavorable attitudes that both likely affect their behavior towards minority students, potentially inducing them to perform below their ability level. Effects of having ethnic majority teachers on minority students' grades hence seem more likely to be indirect than direct.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Grades given by teachers are of crucial importance for students. On a very regular basis, students receive grades from their teachers for tests that, each individually, may be considered low-stakes, but that together determine decisions on track placements, ability grouping, grade retention, special education, and so on, to a much greater extent than national standardized tests.¹ Moreover, teachers' grading standards and factors such as the perceived fairness of grading are likely to affect students' motivation, self-confidence and longer term school outcomes (Betts & Grogger, 2003; Figlio & Lucas, 2004). Despite its importance, grading is usually a subjective evaluation procedure. Beside the criteria set consciously by

teachers, many other factors may influence grading, including interpersonal liking, group stereotypes, and physical attractiveness. Previous research suggests that biases in teachers' grading practices resulting from these factors may harm certain groups of students, depending on their sex, ethnicity, or socioeconomic status (Burgess & Graves, 2009; Dee, 2004, 2007; Figlio, 2005; Hanna & Linden, 2009; Lavy, 2008; Lindahl, 2007; Mechtenberg, 2009; Ouazad, 2008). Dee (2004), for example, shows that ethnic minority students obtain lower test scores if their teacher belongs to the ethnic majority than if their teacher belongs to their own ethnic group. His research, however, does not show whether this difference in test scores is indeed related to biased grading, or whether other factors play a role.

This paper focuses on the question whether and how ethnicity, independently of any of its correlates, affects students' grades; a question that is particularly relevant in light of the persisting achievement gaps in school between ethnic groups that exist in many countries.²

* Corresponding author. Tel.: +31 20 59 83616.

E-mail address: rewijk@feweb.vu.nl

¹ For example in many countries, including The Netherlands, where the current experiment was conducted, standardized tests are only taken at the end of primary or secondary school. Any decisions in between on track and ability group placements and retention are decided upon entirely based upon grades given by students' own teachers.

² Ethnicity definitions vary across countries. In American studies, ethnic (or racial) achievement gaps usually pertain to differences between

Although differences in background characteristics such as parental education and income inequality, and differences in school quality explain part of these gaps, a substantial part remains unexplained (e.g. Colding, Husted, & Hummelgaard, 2009; Fryer & Levitt, 2006). The research of Dee (2004) suggests that part of the achievement gaps can be explained by ethnic minority students obtaining lower test scores when they are taught by teachers belonging to the ethnic majority. How this ethnicity-pairing alters student achievement is not entirely clear. A partial explanation is given by Dee (2005), who finds that teachers rate students with a different ethnicity than their own as more disruptive, more inattentive, and more likely to rarely complete their homework. There are three ways in which these two findings of Thomas Dee can be reconciled: worse grades may be a result of worse student behavior; worse student behavior may be a reaction to behavior of the teacher, including unjust treatments such as biased grading practices; or worse behavior ratings may simply reflect a bias in teachers' perceptions and have to be regarded separately from findings concerning worse test performances. With respect to the second possibility, Ouazad (2008), Lindahl (2007), and Burgess and Greaves (2009) find evidence for a bias in grading practices: teachers give ethnic minority students assessments that deviate from what would be expected based on national standardized tests. The present paper aims to improve on previous work by disentangling why ethnic minority students perform differently when they are taught by teachers belonging to the ethnic majority: is this caused by teachers discriminating in their grading behavior, or by other aspects of student-teacher interactions?

Given the fact that most minority students are generally taught by teachers belonging to the ethnic majority, knowing how being taught by ethnic majority teachers affects their school performance, is essential for devising effective policy measures. For example, it is often proposed that more teachers who belong to an ethnic minority should be recruited in order to reduce the achievement gap (e.g. Carrington et al., 2000; Hope King, 1993). But except if extremely large numbers of such teachers are recruited, most minority students will continue to be taught by teachers belonging to the ethnic majority. It may therefore be more effective instead to try to alter biased grading practices or other behavior of the current teachers.

Using an experiment, I investigate how teachers' behavior differs with students' ethnicity. I focus on the question whether teachers belonging to the ethnic majority give

lower grades for similar work if a student belongs to an ethnic minority, but I additionally explore two alternative ways in which ethnic minority children may end up with lower grades than similar ethnic majority children. Both ways imply that interactions between teachers and students are influenced by ethnicity in such a way, that students indeed start to perform poorer. Teachers may either hold low expectations of individual minority students, or unfavorable attitudes toward ethnic minorities groups in general. Both the expectations and attitudes are likely to be noticed by the students, which may then lead them to adjust their efforts downwards, eventually leading to poorer performance.

Writing is one of the main competences children learn in school. Consequently, most schools assess students' skills at writing texts regularly and writing an essay is also an integral part of many standardized academic tests. Students initially start with short writing assessments and these get progressively longer as the students get older. In the experiment in this study, teachers graded a number of essays of the type and length that 11-year-old students would regularly write in school. Since writing texts is such an essential part of any school curriculum, all teachers in the experiment expectedly have ample experience grading such essays. By randomly manipulating the names on the essays, I make the teachers believe that students do or do not belong to an ethnic minority group. The underlying hypothesis is that the stereotypes and image of the student that this calls up in the teacher, affect perceived student performance and thus grades. Each of the essays is marked by all teachers in the sample and each essay alternately receives names that are typical for Turkish and Moroccan children (two major ethnic minority groups in The Netherlands, where this research is carried out), and names that are typical for native Dutch children. Teachers also state expectations for the secondary school track that they think the student will be able to attend. The teachers do not have any information about the students besides first names. Any effects of the ethnic origin of the names on the essays can therefore only be attributed to perceived group membership.

I find that student ethnicity does not directly affect the grades that teachers give, and that there are no subgroups of teachers that do exhibit such a grading bias in one or the other direction. Hence, it is not a direct grading bias that causes ethnic minority students to perform poorer when taught by teachers belonging to the ethnic majority. I do find that student ethnicity affects most teachers' expectations and that most teachers hold relatively unfavorable attitudes toward ethnic minorities. As explained in the next section, these may be alternative channels that might, in an indirect way, cause minority students to perform poorer when they are taught by teachers belonging to the ethnic majority.

This paper starts with an overview of the literature on how ethnic group membership may affect grades received by students. After that, I describe the experiment I carry out. The next section presents the results and the final section discusses my findings and their implications.

the ethnic majority (Whites) and Hispanics or Afro Americans. In the European context, the relevant ethnic minorities are immigrants from former colonies, refugees and guest workers from Mediterranean countries immigrating in the 1960s and early 1970s. In The Netherlands, the focus of this study, a person is classified in official statistics as ethnic Dutch only if both her parents were born in The Netherlands and as belonging to an ethnic minority group if either one or both of her parents were born in another country. In contrast, for example both the United States and United Kingdom censuses use self-identification for defining group membership.

2. Literature review

The ethnic origin of students' names may directly influence the grades given by teachers via their priors about ethnic minority students' performances and through ethnic stereotypes and attitudes. Based on general statistics that show that ethnic minority students on average perform poorer, teachers may expect individual ethnic minority students to have rather lower language skills on average. It is a priori unclear, however, how these priors will translate into grading behavior. In a classroom setting, teachers may give minority students higher than deserved grades if they want to provide encouragement. Teachers may also give minority students higher grades if the observed performance overcomes their expectations. On the other hand, they may give ethnic minority students lower grades if low expectations prevent them from recognizing performance. In that case, they 'do not believe what they see' and grade according to their priors. Also, stereotypes of ethnic minorities that they are "low performers", and negative attitudes toward such groups (disliking) may lead teachers to give biased grades to essays written by migrants. It is important to underline that such effects occur most probably in an unconscious way. Teachers are not likely to discriminate intentionally and if they are aware of such processes may even compensate and exhibit a bias in the opposite direction.

The direct effects of students' ethnicity on teacher evaluations of their work, I will henceforth refer to as "direct grading bias". Similar direct effects of ethnicity on outcomes have been described extensively in labor market studies, where simply changing the name on a resume alters the chances of being invited for a job interview (cf. Bertrand & Mullainathan, 2004; Carlsson & Rooth, 2007). There are, however, also other ways in which membership of an ethnic group, independent of any of its correlates, may influence students' grades. Dee (2005) divides the ways in which combinations of teacher and student demographic characteristics may affect school performance into two types. The first type refers to changes in teacher behavior as a result of student characteristics; the second type to changes in student behavior as a result of teacher characteristics. Some changes in teacher behavior (including direct grading bias) may also lead the student to change behavior, and vice versa, but the latter changes are a reaction to teachers' behavior, not to their demographic characteristics. Hence, effects are classified according to who originally instigates them.

Because the first type of effects implies the teacher changing behavior, Dee (2005) calls these "active" teacher effects. The described direct grading bias is the most obvious such effect. Indirect effects on grades arise when teachers change their behavior in class in reaction to a student's ethnicity, which subsequently provokes the student to perform worse. These changes in behavior may be unintentional and may go unnoticed to the teacher herself. They may be a result of stereotypes and attitudes they hold.³

³ Stereotypes are sets of beliefs about groups of people and their characteristics (Schneider, 2004). Attitudes are general evaluations of groups of

Stereotypes may lead teachers to treat students differentially, for example because they affect their interpretation of students' behavior. When an ethnic minority student asks a question, it may for instance be interpreted as a sign of ignorance, whereas the same question asked by a majority student might be interpreted as a sign of studiousness. Expectations, which can result from stereotypes and beliefs about traits such as effort and intelligence, can be a powerful determinant of teachers' behavior and students' performance. In psychological studies, teachers were told at the beginning of the year that their students had certain IQ's. These randomly allocated IQ-values affected the students' school performances at the end of the year. This self-fulfilling prophecy of teachers' expectations (the "Pygmalion effect") seems especially strong for students from stigmatized groups and low-achieving students (Jussim & Harber, 2005; Rosenthal & Jacobson, 1968). Similarly, teachers' negative attitudes toward an ethnic group may be communicated to the student through unintended changes in behavior. Examples of such behavior are that teachers call less on minority students to answer questions in class and help them less in finding the correct answer when they do ask them such questions (Casteel, 1998) and that teachers demand less from students about whom they have low expectations, give them less feedback, praise them less often for success and criticize them more frequently for failure (Good, 1987). Such differential treatment may affect students' motivation, self-confidence and, eventually, performance.

The second way in which ethnic group membership may affect students' school performance, Dee (2005) calls "passive teacher effects", because they imply students changing behavior in reaction to teacher demographic characteristics, while teacher behavior remains unchanged. Same-ethnicity teachers, for example, might improve performance by serving as a role model for minority students. Conversely, having a teacher of a different ethnicity may lead students to behave worse in class (Dee, 2005). Also, a different-ethnicity teacher may evoke "stereotype threat", meaning that, when confronted with the stereotype that people from their group supposedly perform poorly, people indeed start to perform poorly, because of a fear of confirming the stereotype in their performance, or a fear of being judged or treated according to the stereotype (Steele & Aronson, 1995). This may happen if ethnic minority students notice the teacher's different ethnicity and expect him/her to share this negative stereotype.

people, issues or objects (Ajzen, 2001). Both are cognitive strategies that people use to process information quickly and easily. In this way they help people cope with daily life without suffering from the information overload caused by having to evaluate each person or object as a complete blank slate (Fazio, 2000; MacCrae, Milne & Bodenhausen, 1994). When stereotypes and attitudes are inaccurate and refer to groups such as races or ethnic minorities, they may cause harm by leading to discrimination. It is to be stressed that stereotypes and attitudes should be viewed as naturally occurring, virtually ubiquitous and generally not ill-intended. The processes discussed here are subtle: I hypothesize that teachers, like most other people, hold stereotypes and negative attitudes about ethnic minorities and I will also investigate this, but this should not suggest that teachers hold racist attitudes. These are something of a very different nature and I have no indication about teachers holding such strong views.

Empirical evidence on whether and how student ethnicity, independent of any of its correlates, affects students' performance, is scarce. And none of the previous literature focused on the type of tests encountered by students in everyday school situations, that this paper focuses on. Fajardo (1985) manipulates author race on essays written by students applying for universities. He finds that Black students receive higher grades for the same essays. This result may not be readily generalizable to the everyday school situations I focus on, since the manipulation consisted of sometimes attaching a form stating Affirmative Action Status as "American Negro". This may have primed teachers towards reverse discrimination. Dee (2004) uses random assignment of students to teachers in Tennessee's project STAR to identify effects of having a different-ethnicity teacher and finds negative effects on test scores. Price (2010) shows that Black freshmen intending to major in science, technology, engineering, or math, are more likely to persist in this choice if they are taught in this field by Black instructors. Ouazad (2008) uses Early Childhood Longitudinal Study data and finds that White teachers give worse subjective assessments to Hispanic and Black children than would be expected based on formal tests. Burgess and Greaves (2009) find that English teachers give ethnic minority students assessments that are lower in comparison to national Key Stage 2 tests. This ethnic bias partially, and for some ethnic minority groups even completely, disappears when controlling for some roughly measured socioeconomic characteristics, so that it is not clear whether it is ethnicity that drives the results, or that a precisely measured socioeconomic background would be able to explain all of the effect. In a similar setting, Lindahl (2007), conversely, finds that Swedish teachers give non-natives assessments that are higher in comparison to national tests; an effect that could partially be explained by teachers in general being more generous to students who fail on the national test. Dee (2005) shows that students' behavior in class is rated as worse if their teacher is of a different ethnicity. This suggests that having a different-ethnicity teacher leads to changes in student behavior, which may consequently lead to lower test scores. However, since he uses teacher ratings of student behavior, the alternative explanation cannot be refuted that not student behavior, but teachers' *perceptions* of student behavior were changed. Also, worse student behavior may be either a cause or an effect of lower grades given by the teacher. Burgess and Greaves (2009) take the fact that teachers' biases differ between subjects as proof that it is not students' behavior (which is something that is observed by teachers, but not accounted for in formal tests) that drives the results, since they argue that students' behavior does not differ between subjects taught by the same teacher, whereas teachers' attitudes can be subject-specific. However, especially in reaction to existing subject-specific stereotypes, students' motivation and behavior can vary by subject. It is difficult to elucidate the causal pathways of how being taught by teachers belonging to the ethnic majority affects minority students' grades. Arguably, the best way to do so is by means of an experiment. This is what I do in the present study.

In this study, I deliberately remove any potential for effects caused by changes in student behavior in order to isolate effects caused by teacher behavior. My central focus is on direct grading bias. I find no evidence for such an effect: purported student ethnicity does not seem to affect the grade given to an essay. My experimental design also makes it possible to examine the prerequisites for two alternative, indirect, ways in which students' ethnic group membership may affect their grades. I show that teachers have lower expectations of ethnic Turkish or Moroccan students than of otherwise similar ethnic majority students. Such lower expectations, as argued, may become self-fulfilling prophecies through adjusted teacher behavior. Also, I show that ethnic Dutch teachers have relatively unfavorable attitudes toward the ethnic minority groups people in general, which may have similar effects.

3. The experiment and context

A sample of 113 Dutch teachers each graded the same set of ten essays written by 11-year-old students. By manipulating the names of the writers of the essays, following a procedure described below, the teachers were made to believe that some of these essays were written by ethnic Dutch students, and that others were written by students with a Turkish or Moroccan background. These latter groups form two of the major ethnic minority groups in The Netherlands. People from these two Mediterranean, predominantly Muslim countries originally came to The Netherlands in the 1960s and early 1970s to help alleviate a tight labor market. Now, they and their descendants together make up a little over 4% of the country's population. This share is higher among younger age-groups. These groups are particularly interesting to look at, because of three characteristics that they share with ethnic minorities in several countries. First, there is a sizeable achievement gap in school between the ethnic majority and these two groups. Data from the 2004 PRIMA study which contains test scores of a nationally representative sample of about 20,000 students between 10 and 12 years-of age, show that children with a Turkish or Moroccan background score about one full standard deviation lower on language test scores than native Dutch children. Second, the teachers these immigrants are taught by, are in overwhelming majority non-immigrant: in the same nationally representative sample, 84% of all primary schools do not have any non-ethnic Dutch teacher in any of the classes in their eighth grade levels. Third, like is the case in other countries, these minority groups may be vulnerable to grading bias, because of stereotypes and attitudes that many people from the Dutch ethnic majority arguably hold (e.g. Velasco Gonzalez, Verkuyten, Weesie & Poppe, 2008). This will be discussed in more detail later on.

3.1. The essays

The students received instructions in their own classroom from their teachers to write an essay with a length of about one page, which is typical for students at their age. The assignment for writing the text was also characteristic for students of this age: students had to choose a position

on a topic related to their everyday life, set up an argument and explain their point. The instruction emphasized that the text should be comprehensible for other readers and that the students should check for language mistakes. Thus, the essays used in this experiment are very typical for essays written in school by 11-year-old children: they are taught to write such texts and are regularly assessed at writing them. I removed those essays that gave cues about the true ethnicity of the writer and a few essays that were very short or of very low quality and picked a subset of ten from the remaining essays: five written by boys, and five written by girls. Before sending them to the participating teachers for grading, I alternately manipulated the names of the writers to be either typical for a Dutch child in this age-group (e.g. Sander and Charlotte), or to be typical for a Moroccan/Turkish child (e.g. Mohammed/Murat and Fatima/Beyza). The names were chosen using websites listing the popularities of names given to children of these three ethnic groups. Turkish and Moroccan names may be hard to distinguish from each other by non-experts, but both are easily distinguished from the typical Dutch names.⁴

In this experiment, it was of great importance that the participating teachers were aware of the (manipulated) ethnicity of the student. Earlier research on name manipulation on essays sometimes failed because teachers did not notice author names written above an essay, even if they were placed in a conspicuous place (Seraydarian & Busse, 1981). To avoid such problems, the assignment for the students writing the original essays was to write an essay about the topic “My best friend and I” in which students had to chose a very good friend and argue why this was their best friend. In this way, it was assured that teachers were confronted with student names throughout the text. Both the name of the writer and the name of the best friend were manipulated to be typical for the same ethnicity. The writer’s name appeared at the top of each essay and in six essays also at the bottom. The friend’s name appeared at least twice in each text and on average was mentioned 5.3 times.⁵ Effects from the manipulation hence may depend not only on the child’s own purported ethnicity, but also on that of the friend. This potential disadvantage is compensated for by an increased certainty that the manipulation had worked: although not asked about this, several participating teachers wrote comments to the questionnaire that related to the immigrant background of many students, e.g. that they were not used to grading ethnic minority students, or that they noticed that minority students made the same types of mistakes as

their own students. No teachers gave comments suggesting they had found out about the manipulation. The matching of students with same-ethnicity friends also reflects a reality in The Netherlands: the average ethnic minority child goes to a school in which 70% of all schoolmates are non-ethnic Dutch (Gijsberts, 2003). Because friendship networks within schools are often strongly segregated (Echenique & Fryer, 2007), the percentage of immigrant children whose best friend also has an immigration background most probably even exceeds this 70%.

3.2. The sample

The participants in this experiment came from an original sample of 128 Dutch primary schools. Two-third were randomly drawn from the population of all Dutch schools; the remainder was randomly drawn from the 16% of the schools where at least 25% of the students was non-ethnic Dutch. Teachers from the latter group have more experience with ethnic minorities, and may therefore be less influenced in their grading practices by the students’ presumed ethnicity (cf. Figlio, 2005). Also, in practice, ethnic minority children will more often be confronted with the latter type of teacher. Teachers teaching ten to 12-year-old students could participate and were promised a gift voucher for €25 – in exchange for their participation. Not all schools wanted to participate, and within schools often not all eligible teachers chose to participate. The final sample consisted of 115 teachers from 54 schools; two teachers were non-ethnically Dutch and were excluded from the data set.⁶ The purpose of the experiment described to the teachers was deliberately kept vague enough not to reveal the exact research questions described in the present paper. Deception of this type is common in discrimination experiments (cf. Bertrand and Mullainathan, 2004): telling the true goal of the study to participants may lead them to deliberately try to obscure discriminatory tendencies, which would bias results. Teachers were told that the university was examining the extent to which grades given by different teachers for the same essays corresponded; this in order to, e.g. improve teacher training on grading practices. In fact, the present experiment was indeed also intended to deliver data on other aspects of grading which will be described in separate papers. These other purposes did not necessitate any additional manipulations than those described here. After the experiment, more information about the purposes, including those described in the present paper, was given to the teachers.

Participating teachers received one out of four sets of essays. Each set contained the same ten essays, but the names on the essays were rotated in such a way, that each essay was in two sets purportedly written by an ethnic Turkish or Moroccan student and in the two other sets by an ethnic Dutch student. Two sets contained seven essays by ethnic minority children, and two sets contained three essays by minorities. Having high numbers of ethnic minority students in the set was credible, since teachers were

⁴ For practical reasons, the original essays were hand written. Because this made it difficult to manipulate the names, the essays were typed over exactly (including the original mistakes) and set in lay-outs copied from other essays which had been written by same-aged children on the computer. Teachers participating in the experiment were told that the essays they received had been typed as they were.

⁵ To mimic a maximally natural situation for grading essays written by primary school children, and in order not to raise suspicion on the true purpose of the experiment, I chose not to accompany each essay by a title page with name and/or demographic information of the child (cf. Hanna & Linden, 2009), but instead to let the names appear on the essay as the children themselves did.

⁶ The average school contributed 2.1 teacher. There was one school with seven participating teachers; the rest contributed fewer teachers.

Table 1

Characteristics of the 113 participating teachers and their assignment to the four sets of essays.

	All	Set 1	Set 2	Set 3	Set 4	p-Value
Male	0.36 (0.48)	0.35 (0.49)	0.23 (0.43)	0.41 (0.50)	0.48 (0.51)	0.21
Age	38.68 (11.67)	39.35 (10.17)	39.06 (12.38)	37.32 (12.42)	38.59 (12.32)	0.93
Years of teaching experience	14.52 (11.40)	13.94 (10.22)	14.47 (12.07)	13.82 (11.57)	15.72 (12.22)	0.77
Ethnicity: Dutch	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	
At least two years experience teaching classes with ≥ 5 ethnic minority children	0.56 (0.50)	0.61 (0.50)	0.53 (0.51)	0.50 (0.51)	0.59 (0.50)	0.92
Number of years experience teaching classes with ≥ 5 ethnic minority children	5.08 (7.00)	4.84 (6.24)	4.63 (7.36)	4.68 (7.18)	6.10 (7.51)	0.85
7 out of 10 essays attributed to Turkish/Moroccan child (vs. 3 out of 10)	0.53 (0.50)	0.00 (0.00)	1.00 (0.00)	0.00 (0.00)	1.00 (0.00)	
N	113	31	31	22	29	

Table shows means and, in parentheses, standard deviations for the entire sample of teachers and for the sub samples of teachers receiving each of the four sets of essays. The right column shows the p-value (where applicable) from an ANOVA test of equality of the characteristics of teachers receiving the four different sets of essays.

told that the essays were written by students from schools in Amsterdam; a city with a very high share of Turkish and Moroccan children among the school-aged population. The teachers received specific instructions on which aspects of the essays to pay attention to when grading. These criteria pertained to the content, style and use of language (cf. Follman & Anderson, 1967).⁷ Uniform criteria help reducing noise in grades resulting from variation in the aspects of essays teachers pay most attention to when grading, which may otherwise be considerable (Meadows & Billington, 2005). Teachers were asked to first grade the essays on a scale of one to ten,⁸ and after that, to state an expectation of the type of secondary school that the writer of the essay would be able to attend. Dutch children go to secondary school at age twelve. There are seven sub-types of secondary school in The Netherlands, ranging from practical education and basic vocational education to university preparatory education. The decision which of these secondary school tracks the child will be attending usually depends half on teacher advice, and half on standardized tests.

The participating teachers sent in their evaluations via the Internet, after which they completed an additional questionnaire. Teachers from the same school received the same set of essays/names, in order to avoid that it became known that the names had been manipulated. The Internet questionnaire contained questions on teacher background characteristics and on grading practices and strategies. It also contained questions about attitudes toward different topics, social groups, and toward ethnic minority groups. These attitudes will be discussed in more detail later. Importantly, because these questions appeared after teach-

ers had sent in the evaluations, there was no risk that they might affect the grades and expectations teachers gave, e.g. by potentially giving away the goal of the experiment. Table 1 shows the background characteristics of the participating teachers and presents tests on the randomness of the assignment of teachers to the four sets of essays and Table 2 shows descriptives for their evaluations.

4. Results

4.1. Direct grading bias

Table 3 shows the effects of purported ethnicity on the grade that a teacher gives for an essay. The left column shows an OLS estimate of grade on an “ethnic minority name”-dummy; the model shown in column 2 adds essay fixed effects to control for differences in quality between the ten essays; in column 3 teacher fixed effects are added to control for unobserved teacher characteristics. In column 4, both fixed effects are included simultaneously. In all models, standard errors were clustered at the school level. Note that teachers in this experiment had no incentive to grade more consistently than they would usually do. Their rewards were independent of grading performance and the instructions they received strongly emphasized

Table 2

Descriptive statistics for the essays.

	Essay attributed to	
	Dutch student	Turkish/Moroccan student
Grade (scales: 1–10)	6.83 (1.08)	6.80 (1.06)
Expectation (scales: 1–7)	4.67 (1.50)	4.53 (1.52)
Observations	551	579

Table shows means and, in parentheses, standard deviations. Each teacher accounts for ten observations.

⁷ I thank my former university's department of Educational Sciences for their advice on these criteria.

⁸ In line with the commonly used grading system in The Netherlands, teachers could also give “broken” grades: 7 1/2 (=7.5), 7+ (=7.25), 7- (=6.75), etc.

Table 3
Effect of purported student ethnicity on grade.

	OLS (1)	Essay f.e. (2)	Teacher f.e. (3)	Two-way f.e. (4)
Ethnic minority student	−0.028 (0.105)	−0.018 (0.082)	−0.027 (0.113)	−0.015 (0.056)
Observations	1130	1130	1130	1130

Table shows coefficients and, in parentheses, standard errors (clustered by school). f.e., fixed effects. Each teacher accounts for ten observations (evaluations of essays).

* $p < 0.10$.

** $p < 0.05$.

*** $p < 0.01$.

that all data would be processed anonymously. The effect turns out not to be significantly different from zero, and very small in an absolute sense: a bit over 1% of the grades' standard deviation in the two-way fixed effects model.

Although this suggests that teachers on average do not exhibit a direct grading bias, it is possible that subgroups of teachers do exhibit such a bias. It could even be that one group of teachers has a bias in one direction, while another group has a bias in the opposite direction so that both effects cancel each other out in the presented estimate of the average effect. I therefore estimate per teacher whether (s)he exhibits direct grading bias and look at the distribution of these biases. First, I take the difference between the given grade and the grade that is predicted from a regression of grade on essay and teacher dummies. Next, per teacher, I rank-sum test whether ethnic majority or minority students systematically end up with higher than predicted grades. Four teachers gave essays with ethnic Dutch names significantly higher grades at the 10% significance level; for three of them, this was also significant at the 5% level. Seven teachers gave essays with ethnic Turkish/Moroccan names significantly higher grades at the 10% level; for one of them, this was also significant at the 5% level. These numbers do not differ from what would roughly be expected by chance in a sample of 113 observations. Fig. 1 shows the distribution of z-values from the rank-sum tests per teacher. If certain groups of teachers exhibited direct grading bias in one or the other direction, the distribution should be non-normal, with fat tails on one or either side. This does not seem to be the case and the distribution does not differ signifi-

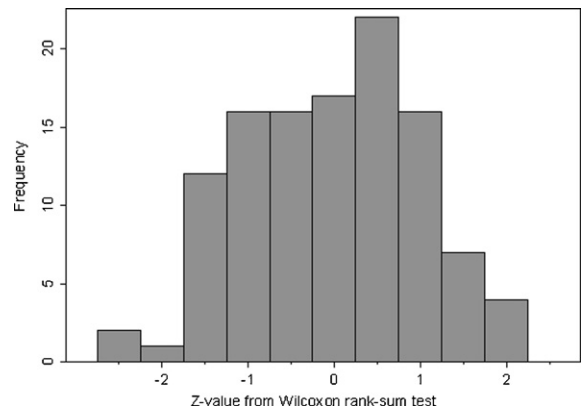


Fig. 1. Distribution of z-values from rank-sum tests per teacher on direct grading bias. Figure shows the z-values from rank-sum tests per teacher ($N = 113$) in which a comparison is made between observed minus predicted grade for essays that were purportedly written by ethnic Dutch students and observed minus predicted grade for essays purportedly written by ethnic minority students. Negative (positive) values indicate a higher ordering for essays written by ethnic Dutch (ethnic minority students).

cantly from a normal one (in a skewness–kurtosis test for normality: adj. $X^2(2) = 0.64$; $p = 0.72$; skewness = -0.08 ; $p(\text{skewness}) = 0.73$; kurtosis = 2.63 ; $p(\text{kurtosis}) = 0.47$).

To test whether specific, relevant subgroups of teachers exhibit direct grading bias, I next add interaction effects to the previously described two-way fixed effects models. As Table 4 shows, the near-zero average effect does not vary with the sex of the teachers. Teaching experience

Table 4
Interaction effects of purported student ethnicity on grade.

	Interaction of “ethnic minority student” variable is with:			
	Teacher is male (1)	Years of teaching experience (2)	Experience w. ethnic minority children (3)	High share of ethnic minorities in set of essays (4)
Ethnic minority student	−0.013 (0.063)	−0.015 (0.082)	−0.014 (0.074)	0.033 (0.075)
Interaction effect	−0.006 (0.108)	0.000 (0.005)	0.001 (0.099)	−0.090 (0.105)
Observations	1130	1120	1130	1130

Table shows coefficients and, in parentheses, standard errors (clustered by school). All models include essay and teacher fixed effects. A high (low) share of ethnic minorities in the set of essays means seven (three) out of ten essays. Experience with teaching ethnic minority children: counted if the teacher taught a class with ≥ 5 ethnic minority children for at least two years.

* $p < 0.10$.

** $p < 0.05$.

*** $p < 0.01$.

may enable teachers to grade more consistently and thus to be less influenced by students' ethnicity, but this interaction is virtually zero. A potential concern in this study is that teachers, even though not realizing that names had been manipulated, might become aware that one of the research aspects was ethnicity, because of the high number of typical foreign names that they were confronted with. If so, and if they reacted to this by adjusting their grading behavior, then, arguably, effects should differ between the condition in which seven out of ten essays were purportedly written by Turkish/Moroccan students and the condition in which this was three out of ten. However, as the table shows, how often teachers were confronted with foreign names, does not affect their tendency to be biased in their judgments.⁹ Finally, one could expect teachers with more experience teaching ethnic minority children to exhibit less grading bias than teachers with little or no such experience (cf. Figlio, 2005). However, I find no difference between these two groups: neither group directly discriminates ethnic minority children in grading. All-in-all, I conclude that no subgroups of teachers can be identified that give ethnic minority children different grades (either higher or lower) than ethnic majority children for the same essays.

Note that students were unaware of the goal of this experiment and had no incentive to alter their writing style so as to influence the experiment. Hence, there is no reason discrimination should systematically vary between the essays in my set and any other essays. But potentially, incidental characteristics of my essays might have led them to be prone to less discrimination. I cannot directly check this hypothesis, but if characteristics of essays influence the amount of direct grading bias, I should arguably observe substantial variations between my ten essays in the amount of grading bias that they evoke. I therefore once more take the difference between the given and the predicted grade and per essay test whether teachers exhibit direct grading bias. With 113 observations per essay, I test parametrically by means of a *t*-test whether there is a significant difference between the obtained grades and the predicted grade for essays "written" by ethnic majority vs. essays "written" by ethnic minority children. Table 5 shows the average direct grading bias per essay and the corresponding *t*-statistic. The average grading biases seem evenly distributed around zero; at a 10% significance level, there is one essay on which ethnic majority students receive higher grades and one essay on which ethnic minority students receive higher grades. These two essays do not represent extremes in their characteristics: they, respectively, rank fourth and eighth on the average grade they receive; sixth and eighth on essay length and second and joint fifth/sixth on the number of times a name appears on the essay. This strengthens the belief that the grading bias per essay fits with a random distribution and that the absence of direct grading bias found in the previous analyses cannot be explained by the specific characteristics of the essays.

⁹ I also found no difference in effect between purported Turkish and Moroccan ethnicity.

Table 5

Average direct grading bias and corresponding *t*-value per essay.

Average grading bias	<i>t</i> -Value
-0.27	-1.82
-0.16	-1.56
-0.07	-0.55
-0.06	-0.40
-0.05	-0.36
0.00	0.04
0.03	0.27
0.04	0.25
0.17	1.41
0.25	2.19

Table shows the average grading bias (calculated as the difference between observed and predicted grades) and the corresponding *t*-value per essay. Negative values indicate that an essay on average receives higher grades when it is attributed to an ethnic Dutch student; positive values indicate higher grades when it is attributed to an ethnic Turkish/Moroccan student.

4.2. Expectations

Having found no evidence that student ethnicity directly affects grades given by teachers, I now look at an alternative way in which ethnicity may indirectly affect achievement. I test whether teachers have lower expectations of the secondary school level that students will be able to attend from ethnic minority students than from otherwise similar students who belong to the ethnic majority. Lower expectations may be a sign of statistical discrimination: aggregate group information is used to substitute for lacking information on true abilities (Phelps, 1972). Lower expectations may unintentionally be communicated to the student, leading him/her to indeed perform poorer. Table 6 presents ordered probit estimates of the secondary school track, the teacher thinks will be feasible for the student in about a year's time, on ethnicity and, from columns 1 to 4, no fixed effects, essay fixed effects, teacher fixed effects and two-way fixed effects. There are seven secondary school tracks in The Netherlands: practical education, basic vocational education, advanced vocational track, combined track, theoretical track, senior general secondary education, and university preparatory education. Because of low frequencies for the lowest level, I pool it together with the second-lowest level. The upper panel shows ordered probit regressions that do not control for grade; the lower panel adds grade as a control. If ethnicity affects grades, grade is an endogenous control and the coefficients for ethnicity are biased. The previous paragraph showed no statistical proof for such an effect; nevertheless, in subsequent analyses, I will use specifications that do not control for grade. I find that the ethnic majority teachers have lower expectations from children who purportedly belong to an ethnic minority, than from similar children who purportedly belong to the ethnic majority. Fig. 2 visualizes this result, by showing the predicted probabilities (derived from the two-way fixed effects model) for ethnic minority and majority children to receive an expectation for each of the levels of secondary school. The probability that a teacher expects the feasible secondary school track to be levels 5, 6 or 7 (the general tracks) is lower for ethnic minority children, while the probability of the expected track to be level 4 or one of the levels below that (the vocational tracks), is higher.

Table 6
Effect of purported student ethnicity on expectations.

	OLS (1)	Essay f.e. (2)	Teacher f.e. (3)	Two-way f.e. (4)
<i>Not controlling for grade</i>				
Ethnic minority student	−0.108 (0.100)	−0.125 (0.122)	−0.119 (0.108)	−0.167 ** (0.085)
<i>Controlling for grade</i>				
Grade	1.059 *** (0.110)	0.909 *** (0.131)	1.791 *** (0.118)	1.608 *** (0.143)
Ethnic minority student	−0.141 (0.089)	−0.149 (0.106)	−0.198 *** (0.077)	−0.223 *** (0.083)
Observations	1130	1130	1130	1130

Table shows coefficients and, in parentheses, standard errors (clustered by school) from ordered probit regressions. f.e., fixed effects. Each teacher accounts for ten observations (evaluations of essays).

* $p < 0.10$.

** $p < 0.05$.

*** $p < 0.01$.

Table 7
Interaction effects of purported student ethnicity on expectations.

	Interaction of "ethnic minority student" variable is with			
	Teacher is male (1)	Years of teaching experience (2)	Experience w. ethnic minority children (3)	High share of ethnic minorities in set of essays (4)
Ethnic minority student	−0.213 ** (0.099)	−0.158 (0.099)	−0.137 (0.108)	−0.143 (0.121)
Interaction effect	0.127 (0.154)	0.000 (0.006)	−0.032 (0.153)	−0.047 (0.160)
Observations	1130	1120	1130	1130

Table shows coefficients and, in parentheses, standard errors (clustered by school) from ordered probit regressions. All models include essay and teacher dummies. A high (low) share of ethnic minorities in the set of essays means seven (three) out of ten essays. Experience with teaching ethnic minority children: counted if the teacher taught a class with ≥ 5 ethnic minority children for at least two years.

* $p < 0.10$.

** $p < 0.05$.

*** $p < 0.01$.

To examine whether the effect is driven by particular subgroups of teachers, in Table 7, I estimate models where the dummy for ethnic minority student is interacted with teacher background characteristics. Having experience teaching ethnic minority children is not related to a significant smaller or larger bias in expectations. Teaching experience and the number of essays that was purportedly written by ethnic minority children moderate the effect neither.¹⁰ The point estimate for the interaction with teacher sex suggests that effects may be smaller for male teachers, but this interaction is far from reaching significance. Since subgroups of teachers with a particularly strong bias in expectations may not be defined along the lines of the background characteristics I observe, I also look at the distribution of biases per teacher to determine whether most teachers are somewhat vulnerable to this effect, or whether it is a small number of teachers that is strongly biased. First, I estimate the predicted probabilities for each of the secondary school tracks from an ordered probit regression of expectations on essay and

teacher dummies. Next, I take the expectation given by the teacher and calculate the probability that the teacher would have given an expectation that is lower than this. Per teacher, I now rank-sum test whether essays purportedly written by ethnic majority, or essays written by ethnic minority students end up with expectations that fall higher in the distribution of predicted probabilities. Fig. 3 shows the distribution of z-values from these tests per teacher. If most teachers are vulnerable to the bias in expectations, I should see a distribution that is centered somewhat below zero (in line with the average effect described above) and that is normally distributed around this point. I cannot refute that this is the case: I find no significant deviations from a normal distribution (skewness–kurtosis test for normality: adj. $X^2(2) = 1.11$; $p = 0.57$; skewness = -0.03 ; $p(\text{skewness}) = 0.88$; kurtosis = 2.54 ; $p(\text{kurtosis}) = 0.30$) and the distribution is indeed centered just below zero (median = -0.114). In line with the proposed distribution, nine (three) teachers had significantly higher expectations of ethnic majority students at the 10% (5%) significance level and five (one) had significantly higher expectations from minority students at the 10% (5%) level. I conclude that the average bias I found is not caused by subgroups of teachers, but seems present to a certain extent in a large share of the teachers.

¹⁰ I also found no difference in effect between essays purportedly written by Turkish and Moroccan students.

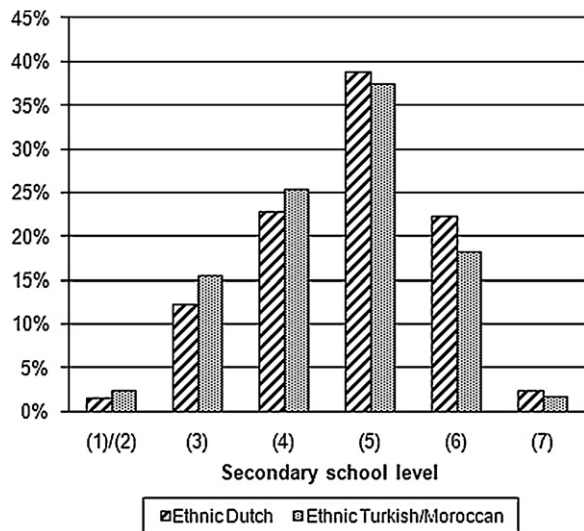


Fig. 2. Predicted probabilities for secondary school level expectations. The figure shows the predicted probability that a student is expected to be able to attend a certain secondary school level, for ethnic Dutch students vs. ethnic Turkish/Moroccan children. There are seven (sub-)levels of secondary school: 1, practical education; 2, basic vocational education; 3, advanced vocational track; 4, combined track; 5, theoretical track; 6, senior general secondary education; 7, university preparatory education. Because of low frequencies for level 1, it is pooled together with level 2. All predicted probabilities are derived from ordered probit models which include essay and teacher dummies.

4.3. Attitudes

If teachers hold negative attitudes toward ethnic minority groups, chances are high that students belonging to these groups will notice these, even if it is only through unintended changes in teacher behavior. Attitudes are general evaluations of groups of people, issues or objects that are automatically used by the brain (Ajzen, 2001). The

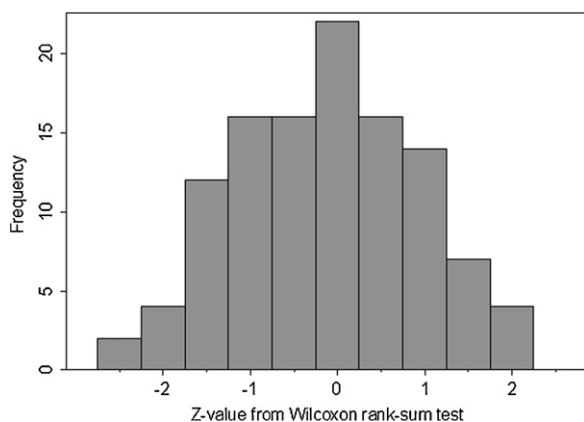


Fig. 3. Distribution of z-values from rank-sum tests per teacher on bias in expectations. Figure shows the z-values from rank-sum tests per teacher ($N = 113$) in which I compare the probability that the teacher would have given a lower expectation than the one (s)he has really given between essays purportedly written by ethnic Dutch students and essays purportedly written by ethnic minority students. Negative (positive) values indicate a higher ordering for essays written by ethnic Dutch (ethnic minority students).

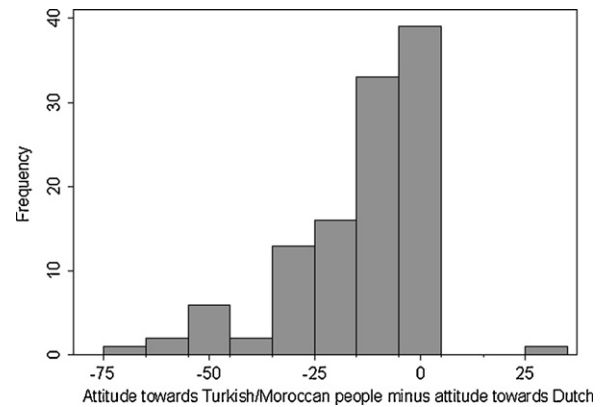


Fig. 4. Distribution of the attitude gap (attitude toward ethnic minorities minus attitude toward Dutch). Figure shows the distribution of the attitude gap over teachers ($N = 113$). Attitude gap is defined as the attitude toward ethnic minorities (being the average of the attitude toward Turkish people and the attitude toward Moroccan people) minus the attitude toward Dutch people, both measured on a scale of 0–100.

cognitive capacity of the human brain is limited. It therefore uses automatic strategies to help reduce the burden of information processing. Using general evaluations (attitudes) instead of evaluating each separate case anew is one such strategy. To give an example: it is less costly in terms of cognitive demand for the brain to rely on the general attitude that you do not like smoking than to make a complete evaluation of the pros and cons of smoking each time you are offered a cigarette. In a similar way, people unconsciously use attitudes to form quick evaluations of others who belong to different (ethnic) groups. Such a cognitive energy saving strategy may result in unduly negative evaluations of individuals who belong to groups that are generally perceived in a negative way. As argued, the resulting unintended changes in teacher behavior this may cause may in an indirect way negatively affect students' school performance.

Negative attitudes toward ethnic minority groups are generally seen as undesirable and are hard to measure, since people tend not to admit to holding them if asked about them explicitly (Greenwald, McGhee & Schwartz, 1998). I therefore take a less explicit approach. At the end of the Internet questionnaire, I added a section in which I told the teachers I wanted to ask some additional questions about their opinion toward several issues, as part of a broader, sociographic research project intended to map and compare the attitudes of several groups of Dutch people. The teachers then were presented with twelve “feeling thermometers”, which were sliding scales on which they could indicate their feelings toward a subject on a scale of 0 (very cold/unfavorable) to 100 (very warm/favorable). These subjects included Politicians, the European Union, Lawyers, and similar other subjects that will not be used in this paper, but that were included only to divert attention from the fact that I was mainly interested in the attitudes toward three other subjects: Dutch, Turkish and Moroccan people. As Table 8 and Fig. 4 show, the great majority of teachers hold less positive attitudes toward Turkish and Moroccan people than toward ethnic Dutch people.

Table 8
Attitude gap.

	Attitude ethnic minorities minus attitude Dutch	N	p difference (A)–(B)
All teachers	–14.2 (16.6)	113	–
(A) Female teachers	–13.7 (17.1)	72	0.677
(B) Male teachers	–15.0 (15.9)	41	
(A) Teachers aged <45	–14.4 (16.7)	68	0.867
(B) Teachers aged ≥45	–13.8 (16.6)	45	
(A) <2 year experience teaching minorities	–15.7 (14.8)	49	0.363
(B) 2+ year experience teaching minorities	–12.8 (18.0)	63	

Attitude gap is attitude toward ethnic minorities (being the average of the attitude toward Turkish and the attitude toward Moroccan people) minus attitude toward Dutch people. Standard deviations are reported in parentheses. Attitudes are measured on a scale of 0–100. For each of the seven presented groups of teachers, the attitude gap is <0 with $p < 0.0001$. Experience teaching ethnic minority children: counted if the teacher taught a class with ≥ 5 ethnic minority children.

* $p < 0.10$.

** $p < 0.05$.

*** $p < 0.01$.

The average attitude gap (difference between the attitude toward Dutch and the mean of the attitude toward Turkish and the attitude toward Moroccans) is about 14 points on the 100-point scale.^{11,12} This gap does not seem to differ between teachers with and without substantial experience teaching ethnic minority children, nor do I find differences between male and female, or younger and older teachers: each group reports similar attitude gaps. Teachers with larger attitude gaps do not exhibit more of a direct grading bias, nor is the size of teachers' attitude gaps related to the size of their bias in expectations: regressions similar to the ones presented in Tables 4 and 7 yield attitude gap*ethnic minority student interaction parameters of -0.001 (SE: 0.002) for a regression of grades, and -0.001 (SE: 0.003) for an ordered probit of expectations.

5. Discussion

The grades students receive in everyday class-situations are of crucial importance for their school careers. Previous research suggests that ethnicity, independent of any of its correlates, may affect performance of students in school (Burgess & Greaves, 2009; Dee, 2004; Lindahl, 2007; Ouazad, 2008). Several explanations for this have been proposed: teachers may give ethnic minority students different grades than majority students for the same work;

differential treatment by teachers may induce students to perform worse; and students may change their behavior in class in reaction to the teacher's ethnicity (Dee, 2005). Using an experimental approach, in which ethnic majority teachers grade essays on which names have been manipulated so that some appeared to be written by ethnic majority students and others by ethnic minority students, I study how teachers react to student ethnicity. My results show that one way in which ethnicity may affect grades can be ruled out: teachers do not give lower grades for essays that were purportedly written by ethnic minority students than for the same essays when purportedly written by ethnic majority students. For two alternative ways in which ethnicity may affect grades, I do find some indirect evidence. Teachers report lower expectations for ethnic minority students than for similar students belonging to the ethnic majority. Also, they report having relatively unfavorable attitudes toward ethnic minority groups in general. Both can affect teacher behavior toward minority students, and, even if this happens in unwitting or subtle ways, this may lead students to adjust their efforts downward and to perform poorer (Jussim & Harber, 2005).

The a priori hypothesis in this study was that teachers would hold stereotypes and consequent expectations of ethnic minorities, and attitudes toward them, that would affect their grading behavior. I find no effects on grading, although I do find evidence for the expectations and attitudes. Why then did these not affect grading?

Differential effects. A first point to note is that expectations and attitudes may both induce teachers to give lower grades and to give higher grades (e.g. Lindahl, 2007). Perhaps some teachers exhibit a direct grading bias in one direction and others in the other direction, leading to the average zero effect. However, I find no evidence for this: there were no identifiable subgroups of teachers who exhibited a bias in one or the other direction. That specific characteristics of my set of ten essays drive the absence of an effect, is also unlikely, since the effects per essay were about normally distributed around zero, with no suggestion that some essays evoked more direct grading bias than others.

Potential failure of the manipulation. Alternatively, one could argue that my manipulation might have failed: teachers might have seen through the manipulation and therefore deliberately avoided giving minority students lower grades. Or the realization that they were being observed may have led teachers to grade more consistently. These explanations, however, would be at odds with the effect on expectations that I do find, which contradicts that teachers tried to reduce biases. Comments given by several teachers that were related to them noticing the ethnicity they believed the students to have, indicate that the manipulation has worked as well, as do the reported attitudes toward ethnic minorities: if teachers would have tried – and managed – to suppress their tendency to give lower grades to ethnic minority students, they would arguably also have done more to hide their unfavorable attitudes toward ethnic minorities than was evidently the case now, especially since these would have been easy to hide for someone intending to do so. Also, if a large number of essays written by minority children

¹¹ The average attitude toward Dutch people was 68.3 (standard deviation: 14.8); toward Turkish people: 57.0 (15.8) and toward Moroccan people: 51.3 (19.0). The difference with the attitude toward Dutch people is significant at the 0.0001 level for both ethnic groups.

¹² As noted before, this should not be interpreted as teachers having racist attitudes. It rather indicates that teachers in their honest reports do not differ from other humans, although this potentially has unintended negative consequences for their students.

would have made teachers aware of an emphasis in this study on ethnicity, then this would probably have differentially affected the condition in which seven out of ten essays were purportedly written by an ethnic minority student and the condition in which this was three out of ten. However, I found no difference in effects between the two conditions. Finally, even if teachers did not consciously do their best to grade more consistently and to suppress discriminatory tendencies, the realization that their grading was being observed might still have unwittingly led them to grade more consistently and this greater consistency might have suppressed unconscious discriminatory tendencies.¹³ The experimental set-up aimed to minimize the possibilities for this happening in several ways. Teachers had no incentive to grade more consistently than they were used to. The instruction put great emphasis on the fact that all data from the experiment would be processed anonymously, while the reward they received was independent of how they graded. Consequently, there was substantial variation in the grades each essay received from different teachers. With an average standard deviation of the grades per essay of 0.86, there is an arguably sufficiently large subjective component in the grading process that would reveal unconscious discriminatory tendencies if they existed. Also, increased consistency due to the knowledge of being observed suppressing discriminatory tendencies would arguably be at odds with the fact that I do find lower expectations and more negative attitudes, even though these responses were being observed as well.

Teachers are not influenced by student ethnicity when grading. So, if the absence of a direct grading bias seems unlikely to be related to characteristics of this specific study, why then, do teachers not exhibit a direct grading bias if they do hold the preconditions for this? It may be that teachers are just able to not let themselves be influenced by student ethnicity in their grading behavior. Either they have no tendency to be biased in their grading, or they are aware of such a tendency and deliberately adjust for this. If this is the case, this study's results indicate that this ability to grade unbiasedly does not come with years of teaching experience, or with experience teaching minority children. In this study, I did find a bias in expectations, but this is something different from a bias in grading. Bias in expectations is a form of statistical discrimination: statistical information on ethnic groups is used as a substitute for lacking information on the individual's true ability (Phelps, 1972). The statistical information that is used for this, is generally contained in stereotypes and may hence be either correct or wrong. A bias in grading would have pointed to taste-based discrimination (Becker, 1957) and, as shown, seems absent, despite the clear differences in taste evidenced by the attitudes.

Attitudes as predictors of behavior. Importantly, previous research has shown that negative evaluations of minority groups in general do not always need to translate into prej-

udiced behaviors. Particularly, general attitudes seem poor predictors of specific behaviors (Ajzen & Fishbein, 1977). So in this case, the negative attitudes I measured toward ethnic minority groups in general, may not lead to the specific behavior of giving low grades to individual ethnic minority students. As argued, students may notice their teachers' negative attitudes toward their group at some point, but teachers then reveal those attitudes in much more subtle and unconscious ways than through their grading behavior and potential effects on grades are then only indirect.

School situations are different from labor market situations. If teachers in their assessments are not influenced by ethnicity, this sets them apart from employers in labor market studies, who did show to be sensitive to manipulations of names on resumes (cf. Bertrand & Mullainathan, 2004). There are a few reasons that make it plausible that teachers may indeed be less influenced by ethnicity information when grading than employers are when assessing resumes. First, employers have a clear incentive to hire the best person. They are interested in the candidate's future performance. Hence, using aggregate group information to substitute for lacking information on the individual's true abilities can be beneficial for them. Teachers, however, have no monetary incentive to grade in one way or the other and are thus less prone to statistical discrimination. Second, as Bertrand and Mullainathan (2004) note, employers receive so many resumes that they may resort to heuristics, such as rejecting as soon as they see an Afro American name, as a strategy to quickly filter these. Something similar is unlikely to happen in school situations, where teachers, arguably, want to give each essay due attention. Hanna and Linden (2009) show that some Indian teachers discriminate lower-caste students in grading. Their setting is a competition, which makes it more similar to the labor market situation where the task is to select one best individual. In line with this, they find that only students who were objectively performing below-average and would therefore not have a chance to win the prize are discriminated. Perhaps teachers graded more sloppily and were guided more by superficial clues such as caste membership once they realized they were grading a poor student. Third, as psychological research shows, the more someone knows about, and feels similar to, another person and the more the other is seen as an individual instead of as a member of a certain group, the smaller the tendency to discriminate the other person (Schneider, 2004). In school situations, teachers know the students, and in the present study, they will also have felt as if judging a real individual child, because the essays were about who the child's best friend was and what activities (s)he liked doing with him/her. This makes the student-teacher relation much less distant than the relation between an applicant and an employer, who is only interested in selecting one suitable candidate, being more or less indifferent toward the rest.

Nevertheless, although student ethnicity does not directly affect grades given by teachers, teachers do have lower expectations of ethnic minority children and do report relatively unfavorable attitudes toward ethnic minority groups. These may both indirectly affect grades. Teachers were asked to give expectations about the sec-

¹³ Note that in a similar way, teachers might grade more consistently in non experimental exam settings in which a second grader anonymously checks a first grader's evaluations.

ondary school track the student will be able to go to in about a year. Track placement should only be based on ability, but teachers expected the feasible starting track to be lower for minority students than for majority students who, according to the grades they gave them, had the same ability. Teachers had no full information on students' ability and let their expectations be influenced by demographic cues, which is a sign of statistical discrimination. Thus, in giving expectations about future performance, teachers resemble employers in labor-market more closely.

In real class settings, more discrimination may occur. Experimental studies virtually always deviate in some respects from real-world situation. This study is no exception. There are a few aspects where my experiment deviates from real-life classroom situations. First, teachers received guidelines on which aspects of the essays to pay attention to when grading. In non-experimental situations, teachers will sometimes use similar guidelines and sometimes not. In the latter case, the subjective component in the grading process might be larger and the potential for discrimination correspondingly so. (Do note that, as mentioned, even with the current grading guidelines, there was still a large subjective component in the grades each essay received from the various teachers.) Second, in the experiment, teachers did not know the students they were evaluating personally. This was a necessary condition to single out the pure, uncontaminated effect of ethnicity on grades. The choice of essay topic tried to minimize this difference, but some difference remains. Third, in real class situations, other forms of discrimination may occur than the type I studied here: the primary goal of this study was to examine whether and how students' ethnicity, independently of its correlates, affects their grades for written work. The focus was on direct grading bias, referring to teachers adjusting grades in reaction to students' ethnicity per se. In the experiment, I hence deliberately removed any potential for effects caused by systematic differences in behavior between minority and majority students. I found no evidence for a direct grading bias. This, however, does not mean that no other forms of discrimination occur. Particularly, in real class settings, student behavior can give rise to biases in grading that lead to lower grades for minority students than for equally competent majority students. Specifically, teachers may use grades to express their (dis-)approval of students' general attitudes and social behavior. If they note that minority students' attitudes and behaviors are worse, or if they dislike their personalities, they may give grades that are biased to reflect these judgements (Ouazzad, 2008). Importantly, this bias can have two causes with different implications.

First, if teachers perceive minority students' behaviors as worse, this may not reflect a reality, but may only be a bias in the teachers' own perceptions. My findings on teachers' expectations and attitudes indicate that biased perceptions are indeed plausible. If behavioral ratings affect grading in this way, the resulting bias would be an indirect grading bias. Although I demonstrated the absence of direct grading bias, there would then still be discrimination in teachers' grading practices, albeit in another, more indirect, form. Future research will have to show whether such indirect grading bias indeed occurs.

Second, the teachers' ratings of minority students' attitudes and social behavior may be correct, i.e. minority students indeed behave worse than majority students. If so, and if teachers express their evaluations of student behavior in their grades, this is a form of taste-based discrimination based on student behavior that, although it is not ethnic discrimination, still disproportionately affects minority students. Some people will argue that it is legitimate for teachers to express evaluations of student behavior in grades. But even from this viewpoint, if minority students behave worse only as a reaction to differential teacher behavior (Dee, 2005), lower grades because of worse behavior may still imply a form of indirect discrimination. Alternatively, minority students may behave worse as a reaction to the teacher's demographic characteristics. This would be an example of "passive teacher effects", e.g. caused by stereotype threat (Dee, 2005; Steele & Aronson, 1995).

One situation in which discrimination may occur that I did not study in this experiment is the judging of oral contributions in class. When teachers have to grade work from several students, as was the case in this experiment, they may use relative grading, which may make it easier to correct for their biases. The more direct interactions and the absence of the possibility of relative grading in combination with potentially biased judgments of students' behavior in class may entice teachers to exhibit more taste-based discrimination when grading oral contributions in class (Casteel, 1998; Good, 1987).

A limitation of this study is that it only looks at whether direct grading bias based on ethnicity affects minority students' grades for written work. It shows that this potentially important type of discrimination does not occur, but it does not give a full picture of all the potential other teacher biases that may affect minority students' grades and what role they play in explaining the achievement gap. This study does show that teachers hold the prerequisites for other types of biases, but more research is needed to establish the occurrence and importance of each of those other types.

To conclude, this paper shows that teachers belonging to the ethnic majority do not give lower – nor higher – grades to ethnic minority students, which rules out one potentially important direct cause for under performance of minority students in comparison to their ability level. Alternative ways in which being taught by ethnic majority teachers might negatively affect minority students' school performance are by behavioral changes induced by teachers' low expectations from ethnic minority students and negative attitudes toward minority groups. I do find evidence for the presence of both lower expectations and negative attitudes, although making the link with altered behavior is beyond the scope of the present research. If expectations and attitudes indeed exert such an influence, it would be important to change stereotypes and attitudes held by teachers. Changing stereotypes and attitudes is not easy: it requires extensive training programs and might fail if teachers are influenced by others who share the same stereotypes and attitudes. Making them aware of the presence and potential consequences of their stereotypes and attitudes might then be a first step in effectively dealing with it. Interventions aimed at reducing bias in ethnic

majority teachers' grading practices, given the current evidence, seem unlikely to reduce minority students' under performance.

Acknowledgements

I would like to thank Adam Booij, Monique de Haan, Charlotte Dignath-van Ewijk, Hessel Oosterbeek, Erik Plug, Peter Slegers and Maresa Sprietsma for discussion and suggestions and participants at the 2009 EEEPE meeting and the 2008 IACCP conference for their useful comments. I thank Harm Kluter and Mirjam van der Geijs for their help in collecting the essays and Amos van Gelderen from the University of Amsterdams Educational Department for his advice on writing/grading instructions for students and teachers.

References

- Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology*, 52, 27–58.
- Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, 84(5), 888–918.
- Becker, G. (1957). *The economics of discrimination*. Chicago: University of Chicago Press.
- Bertrand, M., & Mullainathan, S. (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.
- Betts, J. R., & Grogger, J. (2003). The impact of grading standards on student achievement educational attainment and entry-level earnings. *Economics of Education Review*, 22, 343–352.
- Burgess, S., & Greaves, E. (2009). Test scores, subjective assessment and stereotyping of ethnic minorities. CMPO working paper, no. 09/221. The Centre for Market and Public Organisation.
- Carlsson, M., & Roothe, D. O. (2007). Evidence of ethnic discrimination in the Swedish labor market using experimental data. *Labour Economics*, 14(4), 716–729.
- Casteel, C. A. (1998). Teacher–student interactions and race in integrated classrooms. *Journal of Educational Research*, 92, 115–120.
- Carrington, B., Bonnett, A., Nayak, A., Skelton, C., Smith, F., Tomlin, R., et al. (2000). The recruitment of new teachers from minority ethnic groups. *International Studies in Sociology of Education*, 10(1), 3–22.
- Colding, B., Husted, L., & Hummelgaard, H. (2009). Educational progression of second-generation immigrants and immigrant children. *Economics of Education Review*, 28, 434–443.
- Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, 86(1), 195–210.
- Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, 95(2), 158–165.
- Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, 42(3), 528–554.
- Echenique, F., & Fryer, R. G. (2007). A measure of segregation based on social interactions. *The Quarterly Journal of Economics*, 122(2), 441–485.
- Fajardo, D. M. (1985). Author race, essay quality, and reverse discrimination. *Journal of Applied Social Psychology*, 15(3), 255–268.
- Fazio, R. H. (2000). Accessible attitudes as tools for object appraisal: Their costs and benefits. In G. Maio, & J. Olson (Eds.), *Why we evaluate: Functions of attitudes* (pp. 1–36). Mahwah, NJ: Lawrence Erlbaum.
- Figlio, D. (2005). Names, expectations and the Black–White test score gap. NBER working papers, no. 11195. National Bureau of Economic Research.
- Figlio, D., & Lucas, M. (2004). Do high grading standards affect student performance? *Journal of Public Economics*, 88, 1815–1834.
- Follman, J. C., & Anderson, U. A. (1967). An investigation into the reliability of five procedures for grading English themes. *Research in the Teaching of English*, 1(2), 190–200.
- Fryer, R. G., & Levitt, S. D. (2006). The Black–White test score gap through third grade. *American Law and Economics Review*, 8(2), 249–281.
- Gijsberts, M. (2003). Minderheden in het Basisonderwijs (Minorities in Primary Education). In J. Dagevos, M. Gijsberts & C. van Praag (Eds.), *Rapportage minderheden 2003. Onderwijs, arbeid en sociaal-culturele integratie (Report minorities 2003. Education, labor and socio-cultural integration)*. Den Haag, The Netherlands: Sociaal en Cultureel Planbureau.
- Good, T. L. (1987). Two decades of research on teacher expectations: Findings and future directions. *Journal of Teacher Education*, 38(4), 32–47.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Hanna, R., & Linden, L. (2009). Measuring discrimination in education. NBER working papers, no. 15057. National Bureau of Economic Research.
- Hope King, S. (1993). The limited presence of African-American teachers. *Review of Educational Research*, 63(2), 115–149.
- Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, 9(2), 131–155.
- Lavy, V. (2008). Do gender stereotypes reduce girls' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, 92(10–11), 2083–2105.
- Lindahl, E. (2007). Comparing teachers' assessments and national test results—Evidence from Sweden. IFAU working paper 2007:24.
- MacCrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, 66(1), 37–47.
- Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability*. London: Qualifications and Curriculum Authority.
- Mechtenberg, L. (2009). Cheap talk in the classroom: How biased grading at school explains gender differences in achievements, career choices and wages. *Review of Economic Studies*, 76, 1431–1459.
- Ouazad, A. (2008). Assessed by a teacher like me: Race, gender, and subjective evaluations. INSEAD working paper.
- Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review*, 62(4), 559–661.
- Price, J. (2010). The effect of instructor race and gender on student persistence in STEM fields. *Economics of Education Review*, 29, 901–910.
- Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectations and student intellectual development*. New York: Holt, Rinehart & Winston.
- Schneider, D. J. (2004). *The psychology of stereotyping*. New York: Guilford.
- Seraydarian, L., & Busse, T. V. (1981). First-name stereotypes and essay grading. *Journal of Psychology*, 108(2), 253–257.
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5), 797–811.
- Velasco Gonzalez, K., Verkuyten, M., Weesie, J., & Poppe, E. (2008). Prejudice towards Muslims in The Netherlands: Testing integrated threat theory. *British Journal of Social Psychology*, 47, 667–685.