Empirical essays on education and health

van Ewijk, R.J.G.

Chapter 3

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

3.1 Introduction

Grades given by teachers generate crucial incentives for student learning in school: even in low-stakes tests, the grade received, and factors such as the types of skills rewarded and the perceived fairness of grading are likely to affect student effort, self-confidence and potentially also long-term school outcomes. Despite its importance, grading is often a subjective evaluation procedure. Beside the criterions set consciously by teachers, many other factors may influence grading, including interpersonal liking, group stereotypes, and physical attractiveness. Based on such factors, previous research suggests that biases in teachers' grading practices may harm certain groups of students, depending on their sex, ethnicity, or socio-economic status (Dee, 2004; Dee, 2007; Figlio, 2005; Lavy, 2004; Lindahl, 2007; 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 the ethnic minority group, too. His research, however, does not show whether this difference in test scores is indeed directly related to subjectivity in grading practices, or whether other factors play a role.

This paper focuses on the question whether and how ethnic group membership, independent 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. 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 seems to remain unexplained (e.g. Fryer & Levitt, 2006). The research of Dee (2004) suggests that part of the achievement gaps may be explained by ethnic minority students receiving lower grades from ethnic majority teachers than what would fit to their ability level. This may happen either because these teachers give them grades that are too low for the quality of their work, or because these students indeed perform poorer when their teacher belongs to the ethnic majority, probably as a consequence of changes in interactions between teachers and students.

Given the fact that most minority students are generally taught by majority teachers, 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 ethnic minority teachers should be recruited in order to reduce the achievement gap (e.g. Carrington et al., 2000; Hope King, 1993). But this may not be the easiest or best way to deal with effects arising from being taught by majority teachers, and except if extremely large numbers of minority teachers are recruited, most minority students will continue to be taught by teachers belonging to the majority. More effective measures may therefore be those that connect to how teachers' and students' behaviors are influenced by each other's ethnicities.

Using an experiment, I investigate how ethnic Dutch teachers' behavior differs with students' ethnicity. I focus on the question whether teachers 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 are likely to be noticed by the students at some point, which may then lead them to adjust their efforts and performance downwards.

The main part of the experiment consists of letting teachers grade a number of essays. By randomly manipulating names, 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), 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. But I do demonstrate that it affects the expectations they hold from students and that this bias seems present in a large share of teachers. These lower expectations and the relatively unfavorable attitudes toward ethnic minorities that teachers report, expectedly lead to differential treatment and thus, in an indirect way, may negatively affect ethnic minority students' performance in school.

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 following section presents the results and the final section discusses my findings and their implications.

3.2. Literature review

The ethnic origin of students' names may directly influence the grades given by teachers via expectations they hold from ethnic minority students and through ethnic stereotypes and attitudes. Based on statistics, teachers may expect ethnic minority students to have rather lower language skills on average. It is a priori unclear, however, how these lower expectations will translate into grading behavior. In a classroom setting, teachers may give minority students higher than deserved grades if they want to provide encouragement. In this experiment's setting, this is unlikely since the teachers do not know the students and were asked to grade the essays for a scientific purpose. The teachers therefore expectedly did their best to grade the essays objectively. Another possibility why teachers may give minority students higher grades is that the observed performance overcomes their expectations. On the other hand, teachers 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 expectations. Also, stereotypes of ethnic minorities that they are "low performers", and negative attitudes toward such groups (disliking) may lead teachers to judge essays written by migrants less objectively and to give lower grades. It is important to underline that such effects occur most probably in an unconscious way. Teachers are not likely to discriminate intentionally, especially in the context of the study. Teachers who 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 (cf. Bertrand & Mullainathan, 2004 and Carlsson & Roothe, 2007), where simply changing the name on a resume alters the chances of being invited for a job interview. 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 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. Changes in teachers' behavior often arise unintentionally, as a result of stereotypes and attitudes that they hold. Stereotypes are sets of beliefs about groups of people and their characteristics (Schneider, 2004). Attitudes are general evaluations of groups of

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). Unfortunately, stereotypes and attitudes may also be inaccurate. When they refer to groups such as races or ethnic minorities, they may cause great harm. Potentially, they may then lead to discrimination. I stress here 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.

Stereotypes may lead teachers to treat students differentially, for example because it affects how they interpret students' behavior. A question asked by an ethnic minority student may for instance be interpreted as a sign of ignorance, where it might have been interpreted as a sign of studiousness, had the same question been asked by a student belonging to the ethnic majority. Also, expectations, which proceed directly from stereotypes and their constituent beliefs about traits such as effort and intelligence , can be a powerful determinant of teacher behavior and of students' performance: psychological research shows that manipulating the expectations a teacher has of a student, leads the student to perform worse in school. Especially students from stigmatized groups and low-achieving students seem vulnerable to this self-fulfilling prophecy which is commonly known as the "Pygmalion effect" (Jussim & Harber, 2005; Rosenthal & Jacobson, 1965). In a similar way, if a teacher holds negative attitudes toward an ethnic group, these may be communicated to the student through unintended changes in behavior. An example of such unconscious processes is given by Casteel (1998), who shows that white teachers call less on Afro American students than on white students to answer questions in class and help them less in finding the correct answer when they do ask them such questions. Good (1987), in a review, adds that teachers demand less from students about whom they have low expectations, give them less feedback, pay less attention to them, praise them less often for success and criticize them more frequently for failure. 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 that students' behavior is changed in reaction to teacher demographic characteristics, while teacher behavior remains unchanged. Teacher ethnicity may positively affect student behavior if a same-ethnicity teacher serves as a role model for minority students. Conversely, a different-ethnicity teacher may evoke "stereotype threat". This is the phenomenon that if a stereotype exists that says that people from a certain group perform poorly, a member of this group will indeed start to perform poorly when this negative stereotype about the own group gets activated (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. Thinking about the stereotype leads to a fear of being judged or treated according to the stereotype, or simply of confirming the stereotype in their performance, which consequently causes them to perform poorer. Also, having a teacher of a different ethnicity may lead students to behave worse in class (Dee, 2005).

Empirical evidence on whether and how student ethnicity, independent of any of its correlates, affects assessments, is scarce. 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. 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, perhaps because teachers included more information than only ability in their judgments, such as behavior in class. 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 both researchers use 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. It is difficult to elucidate the causal pathways of how being taught by majority teachers 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 unfounded 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.

3.3. The experiment and context

A sample of 113 Dutch teachers each graded the same set of ten essays written by 11-year old students. Following a random assignment procedure, 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 Dutch 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. This will be discussed in more detail later on.

3.3.1. The essays

The ten essays used in this experiment were written by students from two primary schools that did not participate in the later study. Their teachers instructed them to write an essay for researchers from the university, "who were interested in how well primary school children are able to write". 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. These essays were each about one page in length – fitting with essays for writing assignments regularly given to children of this age. 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, Charlotte), or to be typical for a Moroccan/Turkish child (e.g. Mohammed/Murat, 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.[8]

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 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. On average, in each of the essays, the child's own name appeared 1.6 times and the friend's name appeared 5.3 times. 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%.

3.3.2. 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 Dutch 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 twelve-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

[8] For practical reasons, the original essays were hand written. Because this made it difficult to manipulate the names, the essays were typed over 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 by 11-year old children.

115 ethnically Dutch teachers from 54 schools; two of those were non-ethnically Dutch and were excluded from the dataset. 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. 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 sometimes purportedly written by an ethnic Turkish or Moroccan student and sometimes 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 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. Advice on these criteria was given by my university's department of Educational Sciences. Uniform criteria help reducing noise in grades resulting from variation in grading practices, which may otherwise be considerable. Teachers were asked to first grade the essays on a scale of one to ten[9], 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 teachers 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 3.1 shows the background characteristics of the participating teachers and table 3.2 shows descriptives for their evaluations.

[9] In line with the commonly used grading system in The Netherlands, teachers could also give "broken" grades: 7 ½ (=7.5), 7^{+} (=7.25), 7^{-} (=6.75), etc.

Table 3.1: Characteristics of the 113 participating teachers

	Mean	SD
Male	0.363	0.483
Age	38.7	11.7
Years of teaching experience	14.5	11.4
Ethnicity: Dutch	1.000	0.000
At least two years experience teaching classes with ≥ 5 ethnic minority children	0.563	0.498
Number of years experience teaching classes with ≥ 5 ethnic minority children	5.08	7.00
7 out of 10 essays attributed to Turkish/Moroccan child (vs. 3 out of 10)	0.531	0.501

Table 3.2: Descriptive statistics for the essays

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

Standard errors are reported in parentheses. Each teacher accounts for ten observations.

3.4. Results

3.4.1. Direct grading bias

Table 3.3 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 are clustered by teacher. 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.

Table 3.3: 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	-0.018	-0.027	-0.015
	(0.074)	(0.059)	(0.074)	(0.049)
N	1130	1130	1130	1130

* p<0.10, ** p<0.05, *** p<0.01

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

Table 3.4: Number of teachers giving ethnic Dutch or ethnic Turkish/Moroccan higher grades for the same essay

	Higher grade for ethnic Dutch	Higher grade for ethnic Turkish/Moroccan
at p = .10	4	7
at p = .05	3	1

Table shows, at two significance levels, the number of teachers for whom rank-sum tests indicate that they gave ethnic Dutch or ethnic Turkish/Moroccan students higher than expected grades.

Figure 3.1: Distribution of z-values from rank-sum tests per teacher on direct grading bias

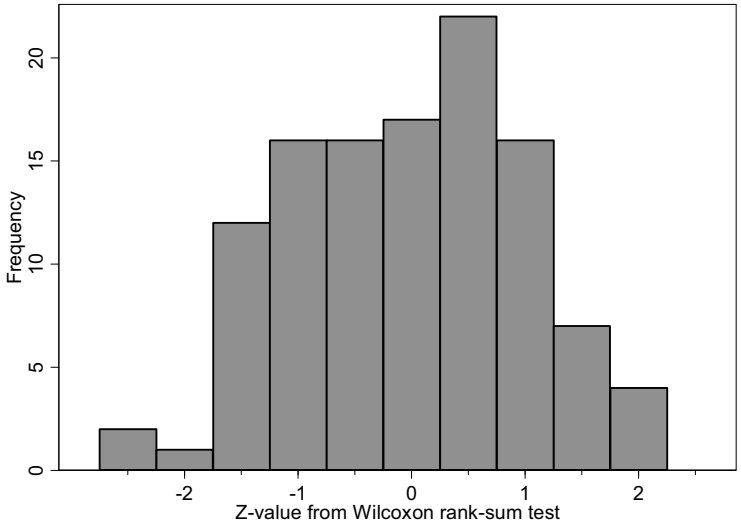


Figure shows the z-values from rank-sum tests per teacher (N = 113) in which a comparison is made between observed minus predicted grade for essays that were purportedly written by ethnic Dutch students and observed minus predicted grade for essays purportedly written by ethnic minority students. Negative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

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 Dutch or ethnic Turkish/Moroccan students systematically end up with higher than predicted grades. Table 3.4 shows the number of teachers for whom essays with ethnic Dutch or ethnic Turkish/Moroccan names end up with significantly higher grades at the 10% and 5% significance levels. These numbers do not differ from what would roughly be expected by chance in a sample of 113 observations. Figure 3.1 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 significantly from a normal one (in a skewness-kurtosis test for normality: adj. $\chi^2(2)$ = 0.64; p = 0.72; skewness = -0.08; p(skewness) = 0.73; kurtosis = 2.63; p(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 3.5 shows, the near-zero average effect does not vary with the sex of the teachers. Teaching experience may enable teachers to grade more objectively 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. 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.

Table 3.5: Interaction effects of purported student ethnicity on grade

	Interaction of "ethnic minority student" variable is with:			
	Teacher is male	Years of teaching experience	Experience w. ethnic minority children	High share of ethnic minorities in set of essays
Ethnic minority student	-0.013	-0.015	-0.014	0.033
	(0.061)	(0.089)	(0.074)	(0.066)
Interaction effect	-0.006	0.000	0.001	-0.090
	(0.100)	(0.006)	(0.099)	(0.092)
Number of observations	1130	1120	1130	1130

* p<0.10, ** p<0.05, *** p<0.01

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

Table 3.6: Average direct grading bias and corresponding t-value per essay

Average grading bias	t-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.

Another potential concern is that perhaps I do not find an effect, just because of specificities of my sample of ten essays: the direct grading bias may depend on characteristics of essays and perhaps I would have found an effect if I would have had a different set of essay. I cannot directly check this hypothesis, but I can compare the average grading biases that I obtain for each essay. Arguably, if the characteristics of the essays influence the amount of direct grading bias, I should 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 Dutch versus essays "written" by ethnic Turkish/Moroccan children. Table 3.6 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 Dutch receive higher grades and one essay on which Turkish/Moroccans 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.

3.4.2. Expectations

Although I find no evidence that student ethnicity directly affects grades given by teachers, there may be more indirect ways in which ethnicity affects achievement. I test whether teachers have lower expectations of the secondary school level that a student will be able

to attend, from ethnic Turkish/Moroccan students than from otherwise similar students who belong to the ethnic majority. Unfounded lower expectations may unintentionally be communicated to the student, leading him/her to indeed perform poorer. Table 3.7 presents ordered probit estimates of the secondary school track, measured on a seven-point scale, the teacher thinks will be feasible for the student in about a year's time, on ethnicity and, from column 1 to 4, no fixed effects, essay fixed effects, teacher fixed effects and two-way fixed effects. Because of low frequencies for secondary school level 1, I pool it together with level 2. I find that the ethnic Dutch teachers have lower expectations from children who purportedly belong to an ethnic minority, than from similar children who purportedly belong to the ethnic majority.[10] Figure 3.2 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 level 5, 6 or 7 is lower for ethnic minority children, while the probability of the expected track to be level 4 or one of the levels below that, is higher.

Table 3.7: Effect of purported student ethnicity on expectations

	No f.e. (1)	Essay f.e. (2)	Teacher f.e. (3)	Two-way f.e. (4)
Ethnic minority student	-0.108	-0.125*	-0.119	-0.167**
	(0.067)	(0.072)	(0.073)	(0.070)
Number of observations	1130	1130	1130	1130

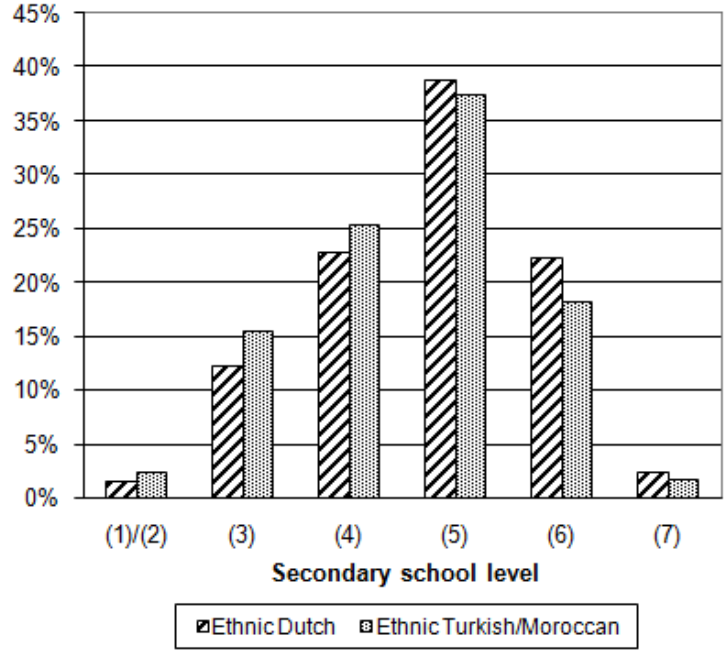
* p<0.10, ** p<0.05, *** p<0.01

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

To examine whether the effect is driven by particular subgroups of teachers for whom the bias in expectations is strong, I test per teacher whether (s)he exhibits a bias in expectations. 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 Dutch, or essays written by Turkish/Moroccan students end up with expectations that fall higher in the distribution of predicted probabilities. Figure 3.3 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. $\chi^2(2) = 1.11$; $p = 0.57$; skewness = -0.03; p(skewness) = 0.88;

[10] Appendix table 3.A-1 shows similar ordered probit regressions which also correct for the grade given to the student (potentially endogenous control). These effects conditional on grade are highly similar to the ones presented in table 3.7.

Figure 3.2: 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.

Table 3.8: Number of teachers having higher conditional expectations from ethnic Dutch or ethnic Turkish/Moroccan for the same essay

	Higher expectation for ethnic Dutch	Higher expectation for ethnic Turkish/Moroccan
at p = .10	9	5
at p = .05	3	1

Table shows, at two significance levels, the number of teachers for whom rank-sum tests indicate that they had higher expectations for ethnic Dutch or for ethnic Turkish/Moroccan students.

kurtosis = 2.54; p(kurtosis) = 0.30) and the distribution is indeed centered just below zero (median = -0.114). In line with the proposed distribution, table 3.8 shows that there are a few more teachers that have significantly higher expectations from ethnic Dutch students than teachers who have higher expectations from ethnic minority students.

In table 3.9, using interaction effects, I test whether the bias in expectations differs for specific subgroups of teachers. 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. 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. 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.

Figure 3.3: Distribution of z-values from rank-sum tests per teacher on bias in expectations

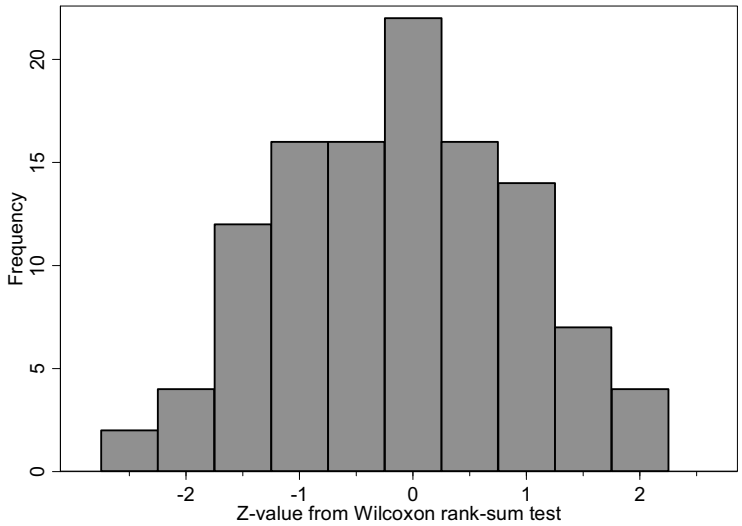


Figure shows the z-values from rank-sum tests per teacher (N = 113) in which I compare the probability that the teacher would have given a lower expectation than the one (s)he has really given between essays purportedly written by ethnic Dutch students and essays purportedly written by ethnic minority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

Table 3.9: Interaction effects of purported student ethnicity on expectations

	Interaction of "ethnic minority student" variable is with:			
	Teacher is male	Years of teaching experience	Experience w. ethnic minority children	High share of ethnic minorities in set of essays
Ethnic minority student	-0.213**	-0.158	-0.137	-0.143**
	(0.090)	(0.113)	(0.100)	(0.097)
Interaction effect	0.127	0.000	-0.032	-0.047
	(0.144)	(0.006)	(0.138)	(0.137)
Number of observations	1130	1120	1120	1130

* p<0.10, ** p<0.05, *** p<0.01

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

3.4.3. Attitudes

If teachers hold negative attitudes toward ethnic minority groups, chances are high that students belonging to these groups will notice this, even if it is only through unintended changes in teacher behavior. As argued, this may 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 it explicitly (Greenwald, McGhee & Schwartz, 1998). I therefore take a less explicit approach. At the end of the 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 topics such as Politicians, the European Union and Lawyers in order to divert attention from the fact that I was mainly interested in attitudes toward Dutch, Turkish and Moroccan people. As table 3.10 and 3.11, and figure 3.4 show, the great majority of teachers hold less positive attitudes toward Turkish and Moroccan people than toward ethnic Dutch people. 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. 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. As table 3.12 shows, teachers with larger attitude gaps do not exhibit a larger bias in expectations, nor is a larger attitude gap related to exhibiting a direct grading bias. However, the fact that so many teachers report attitudes toward ethnic minorities that are considerably less positive than their attitudes toward ethnic Dutch people, may potentially affect ethnic minority students' school achievement in a negative way. In school, many of these children will have teachers who have relatively unfavorable attitudes toward the group they belong to. This may influence their relations with the teacher, their motivation and, eventually, their school performance. As noted before, the discussed attitudes should not be interpreted as teachers having racist attitudes. It rather indicates that teachers in their honest reports do not differ from other humans, but that this may have unintended negative consequences.

Table 3.10: Attitude toward Dutch people and toward ethnic minorities

Attitude toward	Mean	SD	Sample size	p (difference with attitude toward Dutch people)
Dutch people	68.3	(14.8)	113	--
Turkish people	57.0	(15.8)	113	p < .0001
Moroccan people	51.3	(19.0)	113	p < .0001

Table shows the average attitudes of teachers toward three ethnic groups. Attitudes are measured on a scale of 0 (indicating a very unfavorable attitude) to 100 (indicating a very favorable attitude).

Table 3.11: Attitude gap

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

* p<0.10, ** p<0.05, *** p<0.01

Table shows 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. Standard deviations are reported in parentheses. Attitudes are measured on a scale of 0 (indicating a very unfavorable attitude) to 100 (indicating a very favorable attitude). For each of the presented seven groups of teachers, the difference (attitude gap) is < 0 with $p < 0.0001$. Experience with teaching ethnic minority children: counted if the teacher taught a class with ≥ 5 ethnic minority children for at least two years.

Figure 3.4: Distribution of the attitude gap (attitude toward ethnic minorities minus attitude toward Dutch)

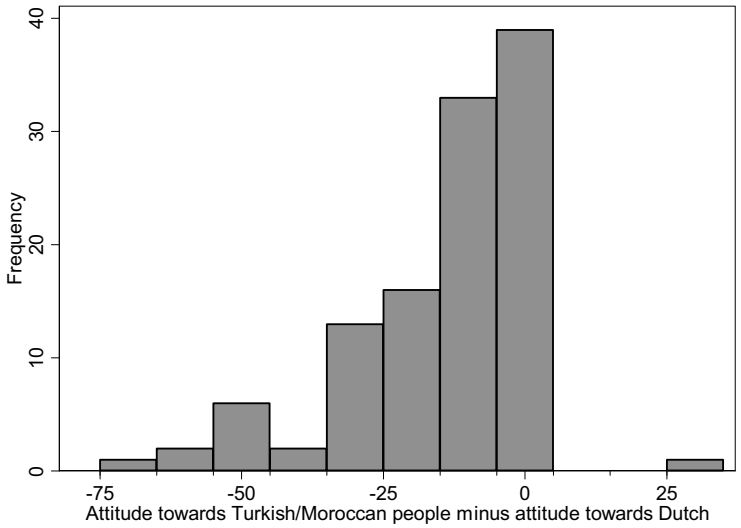


Figure shows the distribution of the attitude gap over teachers (N = 113). Attitude gap is defined as the attitude toward ethnic minorities (being the average of the attitude toward Turkish people and the attitude toward Moroccan people) minus the attitude toward Dutch people, both measured on a scale of 0 (indicating a very unfavorable attitude) to 100 (indicating a very favorable attitude).

Table 3.12: Interaction effect attitude gap * purported student ethnicity

	Dependent variable is:	
	Grades	Expectations
	(1)	(2)
Attitude gap * Ethnic minority student	-0.001	-0.001
	(0.002)	(0.004)
Number of observations	1130	1130

* p<0.10, ** p<0.05, *** p<0.01

Table shows coefficients and, in parentheses, the standard errors (clustered by teacher). All models include essay and teacher fixed effects. Each teacher accounts for ten observations (evaluations of essays). Attitude gap is 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 (indicating a very unfavorable attitude) to 100 (indicating a very favorable attitude).

3.5. Discussion

Previous research suggests that ethnicity, independent of any of its correlates, may affect the grades students receive in school (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 Dutch teachers grade essays on which names have been manipulated so that some appeared to be written by ethnic Dutch students and others by ethnic Turkish or Moroccan 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 more indirect ways in which ethnicity may affect grades, I do find some 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 are likely to 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, and attitudes toward ethnic minorities, 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 objectively. These explanations, however, would be at odds with the effect on expectations that I do found, which contradicts objectivity in teachers' judgments. 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 with an incentive to do so. Finally, if a large number of essays written by minority children 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.

Objectivity. 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 be objective does not come with years of teaching experience, or with experience teaching minority children.

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 prejudiced 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, it is likely that students will notice their teachers' negative attitudes toward their group at some point, but teachers may reveal those attitudes in much more subtle and unconscious ways than through their grading behavior. Effects on grades are then only indirect.

Teacher-student relations are relatively personal. 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, as Bertrand & 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. Second, psychological research has shown that 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 the present study, teachers may 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 towards the rest. Personal information is of course not disclosed by students in most tests they take in school; but then again, in a real school situation, teachers know much more about the child they are grading than was the case in this study. Therefore in general, grading in school does seem a different situation in this respect from applicant selection by employers.

Nevertheless, although student ethnicity does not directly affect grades given by teachers, teachers do have lower expectations for ethnic minority children and do report relatively unfavorable attitudes toward ethnic minority groups, which may both indirectly affect the grades that students get. One could argue that giving lower expectations to

ethnic minority students is reasonable and does not indicate a bias in these expectations, since in The Netherlands ethnic minority children do on average get less far in their school career than ethnic Dutch children with similar abilities. However, teachers were not asked to give expectations about a final level of schooling that will be achieved, but about the secondary school track the student will be able to go to in about a year. For students with similar abilities, the track they are able to start with in secondary school should be the same. If teachers do expect this track to be lower for minority students, it means that their expectations are indeed biased.

To conclude, this paper shows that ethnic Dutch teachers do not give lower – nor higher – grades to ethnic minority students, which rules out one potentially important direct cause for underperformance of minority students in comparison to their ability level. Nevertheless, the picture is not all bright: teachers have lower expectations from minority students than from similar children belonging to the ethnic majority, and they report relatively unfavorable attitudes toward ethnic minority groups in general. Both may in indirect ways induce minority students to perform below their ability level.

Stereotypes and attitudes held by teachers are not easy to change: this would require extensive training programs and might fail if teachers are influenced by others who share the same stereotypes and attitudes. However, awareness of the problem may be a first step in effectively dealing with it.

Appendix 3

Appendix table 3.A-1: Effect of purported student ethnicity on expectations, conditional on grade

	No f.e. (1)	Essay f.e. (2)	Teacher f.e. (3)	Two-way f.e. (4)
Ethnic minority student	-0.141**	-0.149**	-0.198***	-0.223***
	(0.070)	(0.074)	(0.072)	(0.075)
Number of observations	1130	1130	1130	1130

* p<0.10, ** p<0.05, *** p<0.01

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