Discussion Paper No. 09-074

Discrimination in Grading? Experimental Evidence from Primary School

Maresa Sprietsma

ZEW
Zentrum für Europäische Wirtschaftsforschung GmbH
Centre for European Economic Research

Discussion Paper No. 09-074

Discrimination in Grading? Experimental Evidence from Primary School

Maresa Sprietsma

Non-technical summary

Migrant children are lagging behind in terms of educational performance in many industrialised countries. The relatively low educational level of their parents as well as language difficulties are well-known reasons behind the test score gap between migrant and non-migrant children. However, little is known about the role of teacher expectations in explaining this gap.

According to the psychological literature, teacher expectations may affect pupil performance in several ways. For instance, there is evidence that teacher expectations may lead to (unconscious) changes in teacher behaviour that affect the actual performance of the pupils. But besides inducing changes in teacher behavior and pupil performance, expectations may also influence the way in which teachers grade pupils' work. Grades are the main signal of ability and performance that teachers give to their pupils and they can have long term consequences for student achievement. However, grading is a subjective process. Even if teachers generally use point schemes, many subjective impressions may play a role in estimating the quality of student work.

In this paper, we randomly assign typical German or Turkish names to identical sets of essays to test the effect of teacher expectations regarding pupil background on grades. We find that essays bearing Turkish names receive significantly worse grades. Moreover, teachers recommend the highest track of secondary school with a 10% lower probability when an essay bears a Turkish name.

The effects we measure are relatively small and cannot explain the full gap in the attendance of the highest secondary school track between migrant and German pupils. However, considering the numerous disadvantages faced by migrant pupils, removing the additional penalty resulting from lower teacher expectations would be a welcome step forward. It is therefore good news that the observed grading and recommendation bias originates from a small group of teachers only. Hence most teachers do not grade or recommend different tracks based on ostensible pupil origin. This implies that lower expectations do not necessarily affect teachers' judgment of a pupil's potential and there is scope for getting rid of these biases for the remaining group. For instance, increased awareness about the importance of teacher expectations through teacher training, could contribute to reducing the grading and expectation bias.

Zusammenfassung

Kinder mit Migrationshintergrund zeigen in vielen industrialisierten Ländern einen Rückstand im Bildungserfolg. Der relativ geringe Bildungsstand der Eltern sowie Sprachschwierigkeiten sind wichtige Erklärungsfaktoren für die Leistungslücke zwischen Kindern mit und ohne Migrationshintergrund. Allerdings weiß man bislang nur wenig darüber, welche Rolle die Erwartungen der Lehrer für die Entstehung dieser Lücke spielen.

Die psychologische Literatur weist auf mehrere Einflusskanäle hin, über die Lehrererwartungen die Leistung von Schülern beeinflussen können. Es gibt z.B. Hinweise darauf, dass die Erwartungen zu (unbewussten) Verhaltensänderungen der Lehrer führen können, wodurch die Leistung der Schüler tatsächlich beeinflusst wird. Neben Änderungen im Verhalten der Lehrer könnten Lehrererwartungen sich aber auch auf die Benotung auswirken. Noten sind das wichtigste Signal der Begabung und Leistung, das die Lehrer den Schülern geben, und sie können langfristige Konsequenzen für deren schulischen Erfolg haben. Jedoch ist Benotung ein subjektiver Prozess. Obwohl Lehrer bei der Beurteilung von Schülern überwiegend Notenschemata nutzen, können viele subjektive Eindrücke (unbewusst) bei der Bestimmung der Note eine Rolle spielen.

In diesem Papier wird der Effekt der über den Namen angedeuteten Herkunft eines Schülers (Migrationshintergrund oder nicht) auf die Benotung von Aufsätzen in einem Experiment ermittelt. Zu diesem Zweck verändern wir systematisch die Schülernamen bei einer Reihe von Aufsätzen von Schülern der 4. Klasse. Die Ergebnisse zeigen, dass Aufsätze mit einem türkischen Namen eine signifikant schlechtere Note erhalten. Zudem sprechen Lehrer mit einer 10% geringeren Wahrscheinlichkeit eine Gymnasialempfehlung aus, wenn ein Aufsatz einen türkischen Namen trägt.

Die Effekte sind relativ klein und deshalb nicht der wichtigste Erklärungsfaktor für die Leistungslücke zwischen Schülern mit und ohne Migrationshintergrund. Da Schüler mit Migrationshintergrund aber ohnehin schon mit vielen ungünstigen Umständen konfrontiert werden, scheint es erstrebenswert, diesen zusätzlichen Nachteil zu beseitigen. Die Notenverzerrung und die geringeren Lehrererwartungen treten nur bei einer kleinen Gruppe von Lehrern auf. Die meisten Lehrer lassen sich also nicht bei der Benotung und in ihren Erwartungen durch die Herkunft der Schüler beeinflussen. Es sollte daher möglich sein, die negativen Effekte der Lehrererwartungen auch in der betroffenen Gruppe von Lehrern zu beseitigen. Zum Beispiel könnte eine Weiterbildung, die das Bewusstsein für diese Effekte steigert, dazu beitragen, die negativen Folgen von unterschiedlichen Lehrererwartungen zu verringern.

Discrimination in Grading ? Experimental Evidence from Primary School

Maresa Sprietsma[*†]

November 2009

Abstract

This paper studies the effect of teacher expectations on essay grades in an experimental setting. To this purpose, we randomly assign Turkish or German first names to a set of essays so that some teachers believe a given essay was written by a German native pupil, whereas others believe it was written by a pupil of Turkish origin. We find that essays obtain significantly lower grades and lower secondary school recommendations when bearing a Turkish sounding name.

JEL code: I20

Keywords: Experiment, discrimination, grading, pupils with migration background

[*]Centre for European Economic Research (ZEW), L7,1. 68 161 Mannheim, Germany. *Email*: sprietsma@zew.de

[†]**Acknowledgements**: First of all, I am very grateful to the teachers that agreed to participate in the study for their time and cooperation. I would like to thank Professor Michael Lechner and Professor Bernd Fitzenberger for helpful comments and suggestions. Finally thanks are due to Katja Coneus and Julia Horstschräer for their comments as well as to the participants to the ZEW Werkstattseminar.

1 Introduction

Migrant children are lagging behind in terms of educational performance in many industrialised countries. The relatively low educational level of their parents as well as language difficulties are well-known reasons behind the test score gap between migrant and non-migrant children. However, little is known about the role of teacher expectations in explaining this gap. According to the psychological literature, teacher expectations may affect pupil performance in several ways. Firstly, expectations may lead to (unconscious) changes in teacher behaviour that affect the actual performance of the pupils (Rosenthal and Jacobson, 1968; Figlio, 2005). For the US, there is evidence that teachers have different expectations regarding the performance of pupils according to their ethnic background (Tenenbaum and Ruck, 2007) and it has been shown that African-American students tend to receive a less favorable treatment in the classroom if their teacher is of a different ethnic origin. For instance, in Casteel (1998), African-American students received significantly less praise and less direct questions from their teacher as the Caucasian students did. In Ferguson (2003), African-American students received less feedback after mistakes than their peers from different ethnic backgrounds. Such changes in teacher behaviour are thought to affect pupil performance as they interact with the development of students' self-perception and behaviour in a way that reinforces the teacher's expectations (Ferguson, 2003; Dee, 2005).

But besides inducing changes in teacher behaviour and pupil performance, expectations may also affect the way in which teachers evaluate pupils' performance. This is particularly the case when the skills to be evaluated leave room for a lot of subjectivity such as in (Dee, 2005), where African-American students were more likely to be considered inattentive and disruptive in class than their Caucasian peers when their teacher was from a different ethnic background. Similar bias in evaluation can arise with respect to grades (see

e.g. (Lavy, 2008)). Grades are the main signal of ability and performance that teachers give to their pupils and grades can have long term consequences for student achievement. However, grades are rarely an objective measure of the performance at a specific test. First of all, because teachers tend to use grades for other purposes besides valuing the performance at a particular test. They may for instance use grades to punish or reward behaviour in class, or to encourage pupils with low self-esteem. Moreover, grading is a subjective process. Even if teachers generally use point schemes, many subjective impressions may play a role in estimating the quality of student work.

If teachers hold low expectations as to the performance of migrant pupils, as can be expected given the existing performance gap, they may be surprised by the quality of an essay and react in various different ways. A first possibility is that teachers may give the unexpectedly good essay a better grade than they would otherwise have, as a reward to the pupil for overcoming supposed language or background difficulties. On the other hand, the essay may, on average, obtain a lower grade than if the teacher believed it to be from a German pupil, because the teacher looks harder for additional errors that confirm his/her expectations. In fact, psychological research has shown that persons are likely to search harder for evidence in favor of their expectations than the other way round. This is called the expectation-confirmation bias (Darley and Gross, 2005).

In this paper, we focus on the effect of pupil names on essay grades in an experimental setting. To this purpose, we manipulate the names appearing on a set of 4th grade essays. Whereas some teachers believe a given essay was written by a native German pupil, others believe it was written by a pupil of Turkish origin. We want to find out whether the presence of a Turkish name, i.e. teacher beliefs as to who wrote the essay, affects the essay grade. In addition, we ask the teachers to emit a secondary school recommendation based on the essays, and test whether the expected feasible secondary school

for a given essay varies according to the type of name appearing on the essay. Finally, we measure teachers attitudes towards German versus Turkish people using feelings thermometers, that allow teachers to express the warmth of their feelings concerning groups of persons and ideas. The goal of the latter two research questions is to investigate whether teachers hold different expectations or attitudes with regard to pupils with different backgrounds. If teacher did hold relatively lower expectations or attitudes concerning pupils from Turkish background, this would imply that there is a foundation for changes in teacher grading and classroom behaviour. Nevertheless, in our experiment, changes in pupil performance due to teacher behaviour are excluded since the essays were written by pupil unknown to the teachers. For the same reason, the teacher expectations in our experiment are based solely on the ostensible pupil origin as suggested by pupil names, rather than on additional personal knowledge of the pupil as would be the case in real life.

We would like to underline that teacher expectations and their impact on teacher behaviour are not to be confused with intentional pupil discrimination. Although we can speak of discrimination as soon as a person's behaviour is affected by his/her beliefs about another group, this need not be intentional. Actually, psychological research has shown that behaviour is often affected by beliefs and expectations involuntary, this is called 'implicit bias'. Moreover, implicit bias is only weakly correlated with explicit judgments (Hofmann et al., 2005). However, there is also evidence that developing the awareness of the existence of implicit bias may limit its effects (Rudman et al., 2001). This is one of the motivations for our work.

The contribution of this paper is twofold. Firstly, we provide experimental evidence on direct effects of teacher expectations on grades. Although the effect of names on the probability of being hired has been investigated in the recent literature (Bertrand and Mullainathan, 2004; Carlsson and Rooth, 2007), to our knowledge, causal evidence on the impact of names on grades is scarce. An exception is the paper by Lindahl (2007), who compares grades

given by the pupils own teachers with grades obtained at anonymous centralised examinations. Using Swedish data, she finds that non-native students are more generously rewarded when assessed by their own teachers. Secondly, we provide evidence as to differences in teacher attitudes and expectations by student background for a European country. The companion paper by (van Ewijk, 2009) presents the results for this same experiment in the Netherlands. We therefore refer to the Dutch results on several occasions in the interpretation part. The remainder of the paper is structured as follows. Section 2 presents the experiment design and the data collection. Section 3 contains the empirical results and interpretations. Section 4 concludes.

2 Methodology

The methodology of our experiment is similar to that used to assess discrimination in hiring. In these experiments, identical application letters and CVs are sent out, bearing either European names or foreign names (Bertrand and Mullainathan, 2004; Carlsson and Rooth, 2007). We sent identical sets of 10 essays, collected from two anonymous 4th grade classes in Germany (that do not participate in the experiment afterwards) to different teachers, randomly assigning pupil names to the essays. As a consequence, some teachers believed a given essay was written by a native German pupil, whereas others believed it to be written by a pupil of Turkish origin. The names we used were typical German or Turkish and we selected names that were frequently given 10 years ago (around the birthyear of the essay writers). German names for instance included Max, Stefan, Anja or Melanie whereas the Turkish nameset contained names like Sevda, Gönül, Hakan or Coskun. The full list of names is presented in Table 1. We chose to use Turkish names because this nationality represents one of the largest migrant communities in Germany: 42% of non-German primary school pupils have Turkish nationality[1]. The disparity

[1]Statistisches Bundesamt, 2007/08

in test scores between second generation migrant pupils and native pupils is very large in Germany. Second-generation students lag behind their native peers by 93 score points, which is equivalent to one and a half proficiency levels. In addition, the test score gap for pupils of Turkish background is the highest when compared to Asian or Eastern European migrants OECD (2006).

Four different name sets were allocated to the ten essays. This implies that approximately 25% of the teachers receive the same nameset. In two of the namesets, 30% of the names were of Turkish origin, in the other two 50%. We do not increase the share of migrant origin names more because the share of Turkish pupil is generally lower than 50% in German schools. In order for the relatively high shares of Turkish pupils to be credible to the teachers, they were told the essays come from pupils in a large city which is known to have many Turkish migrants. Half of the essays were written by girls and names were not manipulated with regard to gender. A girl's essay always bore a girl's name because the essays were quite gender specific in terms of mentioned activities and of the description of the friend.

We sent information letters to public primary schools in two German regions, asking the school directors to inform their 4th grade german teachers about the study. If teachers were interested, the essays and questionnaires were sent to the school as well. The names of the teachers were unknown to us or were kept confidential if we were contacted directly. Participation was thus anonymous and voluntary. In order for the assessment of the name effect to work, it was important that teachers did not know the exact purpose of the experiment. They were told that the study aimed to assess the determinants of grading in order to e.g. improve future teacher training. In order to limit variation in criteria of evaluation, we provided the teachers with three essay characteristics (content, style and language) to be taken into account when giving the grades.

In order to ensure that the teachers noticed the names on the essays

(which was problematic in Seraydaran and Busse, 1981) we chose the topic of the essay to be 'My best friend and I'. This implied that the name of the friend was mentioned in the essay several times (on average 3,7 times). We also manipulated the friends' names to have the same origin as the ostensible author, including when several friends were mentioned. Essays that were clearly identifiable as being from a Turkish or from a German background because of the choice of topics (for instance religious activities) were not included in our set of essays. We also excluded extremely bad or short essays that were unlikely to generate much variation in grades across teachers. The essays were about 2/3 of a page long and were copied by us (including mistakes and formatting) to be in typewritten form. Teachers in the same school always received the same namesets to prevent the name manipulation from being discovered. We have no reason to believe that teachers found out about the name manipulation. One teacher even refused to participate because of his lack of experience in grading migrant pupils.

In addition, we asked teachers to give a secondary school track recommendation to the pupils based on the essays. In Germany, pupils are generally separated into different secondary school tracks at the end of 4th grade. Based on teachers recommendations and parents opinions, pupils may attend the Hauptschule (lowest track), the Realschule (middle track) or the Gymnasium (which grants access to the university). Giving such a recommendation based on only one essay is a rather difficult task, but all participating teachers agreed to give a recommendation. Finally, we asked the teachers to fill out twelve so-called feeling thermometers. They were asked to indicate on a scale of 0 (very cold/uncomfortable) to 100 (very warm/comfortable) how they felt concerning 12 specific topics such as politicians, environmental policy, and German and Turkish people. It was clearly specified that these were subjective answers, and that there was no right answer. This indicator allowed us to proxy attitudes towards migrants.

Grades ranged from 1 (very good) to 6 (very insufficient), as is common

practice in German schools. Please note that this implies that higher grades mean worse performance. Teachers could give grades like 1- or 4+. This is also common practice in German schools and is translated in numbers accordingly (1- would be 1.25 whereas 4+ corresponds to 3.75).

3 Results

Eighty-eight teachers from two German regions participated in our study. Although the response rate to the information letters was low, 80% of the teachers that requested to participate sent back the completed questionnaire. All teachers are German natives and quite experienced. Eighty-three percent were female, they were on average 48 years old and most had at least 2 years of teaching experience with migrants (Table 2). Very few teachers have other degrees than that required to become a primary school teacher. These figures are in line with representative statistics as in the schoolyear 2008/2009, on average 77% of primary school teachers in the regions where our sample was gathered were female[2]. Moreover, a large share of the teacher population in Germany is older than 50 years of age, because of an increase in hiring during the 60ties, when the babyboomers went to school. The trend in the teachers age is decreasing in recent years, as these teachers start to retire. As a result, on average, primary school teachers were 50 in 2007-2008. The teachers took about 2,5 hours on average to correct the 10 essays.

Table 3 shows the descriptive statistics of grades by ostensible pupil origin. On average, the essays receive 0.12 points lower test scores (out of 6) if the essay bears a Turkish name. This difference is significant at thte 5% level of confidence. Regression estimates of essay grades on ostensible pupil origin are presented in Table 4. We start by regressing the essay grades on a dummy equal to one if the essay had a Turkish name by ordinary least squares (Column 1). We then include essay fixed effects to control for ef-

[2]Statistisches Bundesamt, 2007/08

fective differences in essay quality (Column 2), and teacher fixed effects in order to capture a teacher's tendency to grade severely or generously (Column 3). Because the residuals are correlated between teachers and between essays, we compute the standard errors clustered by teacher. Alternatively, we could cluster standard errors by essay or by school (or teacher and essay or school). The latter specifications yield smaller standard errors. Our level of significance, using clustering on teacher only, is therefore a lower bound. The explanatory power of the ostensible pupil origin is non significant in the OLS specification. However, the standard errors become much smaller once we include important determinants of grades such as teacher and essay dummies. In the latter specifications, we find that essays obtain significantly higher (worse) grades when bearing a Turkish sounding name. The size of the effect is relatively small with around 10% of a standard deviation in test scores but considering that the distribution of grades is centered around 3, the passing grade, such a small difference in grades may still play an important role.

In order to find out how the grading bias is distributed among teachers, we compute the grading bias between essays with Turkish names and those with native German names for each teacher and look at the distribution of these biases. To this purpose, we compute the difference between the grade given to each essay by the teacher and the average grade obtained by the essay in the sample. We then ranksum these differences by ostensible pupil origin for each teacher, and compute the percentage of teachers for whom the difference to the mean are significantly larger when the essay bare a Turkish name. Table 5 shows that the grading bias comes from a minority (14%) of teachers that gave essays with Turkish names lower grades at the 10% confidence level.

In a next step, we therefore try to identify whether the observed grading bias is related to teacher characteristics. We include crossed effects of Turkish name dummy times teacher experience with migrant pupils, the teacher's age, and the attitude gap between German and Turkish people. None of

these crossed effects are significantly different from zero as can be seen in Table 6. However, the bias seems to arise more strongly if teachers have at least some experience with migrants. The latter supports the hypothesis that teachers base their expectations on their own experience. A probit estimation of teacher characteristics on the probability of presenting a grading bias shows that the teachers that present grading bias do not differ from the other teachers in any of the observed characteristics (Table 7).

Nevertheless, the crossed effect of the name dummy with the share of Turkish names in the received set of essays is significantly different from zero. If 50% of the essays in a set bear Turkish names (as compared to only 30%) the grading bias is significantly smaller (Table 6). It is possible that the larger observed diversity in essays from ostensibly Turkish pupils reduces the teachers expectations bias. This interpretation is in line with recent psychological evidence on discrimination. For instance, in Lebrecht et al., 2009, Caucasians' implicit bias toward African-Americans diminished after they learned to individuate faces of that race. The realisation of diversity inside the unknown group reduced unconscious stereotypes.

Although the estimated grading bias is limited in size, our results stand in contrast to the results obtained in the Netherlands (van Ewijk, 2009), where no effect of names on grades was found. However, the set-up of the experiment is the same in both countries, and teacher characteristics are not related to the grading bias in either country. One possible explanation could be that there are differences in the degree of expectation bias awareness between Dutch and German teachers. Teacher training to evaluate pupils can be expected to affect such awareness but there are no significant differences in the amount of training the teachers of the two countries received to evaluate pupils' written work. Both in the Netherlands and in Germany, less than 40% of teachers participated in such training. In Germany, evaluation training tends to occur more often after, rather than during, the studies. Crossed effects with the name dummy are not significantly different from zero, mean-

ing that the existing courses do not affect grading bias. However, we do not know (and could not ask without endangering the experiment) what was taught in the evaluation courses and whether it was relevant in reducing the expectation bias. Moreover, differences in awareness of the expectation bias may exist for other reasons as well.

Our second research question relates to the teachers' secondary school recommendations for pupils of different origins. In this section, we investigate whether there is a foundation for the observed grading bias in terms of attitudes and expectations. As mentioned in the introduction, colder feelings and lower expectations with regard to the performance of migrant pupils could be the source of differences in behavior in teaching but also in grading. The feeling thermometers show that German teachers (just like Dutch teachers, van Ewijk (2009)) do have less positive attitudes towards Turkish people than towards German people. The average attitude gap amounts to 8,5 points on a 100 point scale (Table 8). For teachers with less than two years experience of teaching migrant pupils, the attitude gap is significantly larger but on the other hand their attitudes are more positive overall. Other teacher characteristics are not correlated with a larger attitude gap.

Descriptive statistics of the recommended secondary school tracks by ostensible pupil origin are presented in Table 9. It appears that essays bearing a Turkish name receive more recommendations for the Hauptschule (lowest track) and less for the Gymnasium (highest track). Table 10 displays the results of probit estimations[3] of the probability of receiving a recommendations to attend each of secondary school tracks. We include teacher and essay fixed effects in the estimation and standard errors are clustered by teacher.

[3] Estimates from linear regression using the same specification yields similar results. We chose not to use an ordered probit model because the repeated probit estimations impose less strong assumptions. Since results look different for the lower tracks in ordered probit, we preferred the more flexible specification presented in the paper.

Our results indicate that teacher expectations with regard to pupils' capacity for attending different secondary school tracks are significantly affected by the name appearing on the essay. Based on the same essay, teachers tend to recommend the Gymnasium (highest track) to essays bearing Turkish names with an 11% lower probability. This means that teacher expectations are lower for pupils with Turkish background. No significant effects are found for the lower tracks. These results are in line with the results found for the Netherlands by van Ewijk (2009). Estimates including crossed effects of the name with teacher characteristics are presented in Table 11. Each line represents the estimates from the probit estimation including the specified crossed effect for each of the three secondary school types. The effect of the name on the expected feasible secondary school track does not vary with teachers' attitude gap between German and Turkish people, nor is it related to the absence of experience with migrant pupils or the share of Turkish names in the received set of essays. However, the bias against higher tracks is stronger for teachers with more than 5 years experience teaching migrant pupils. This is in line with the intuition that teachers probably base their expectations on their own experience with migrant pupils. In order to find out how the expectation bias is distributed among teachers, we again rank order the expectation biases of the teachers. It appears that, similarly to the grading bias, only a small fraction of teachers (around 10%) presents such a bias. The majority of teachers do not have different expectations according to pupil background.

The different secondary school recommendations for pupils with an ostensibly Turkish background as well as the difference in attitudes show that there is a foundation for potentially different behavior towards migrant pupils. The fact that teachers hold lower teacher expectations regarding migrant pupils may also be one of the reasons behind the observed grading bias. Finally, the less favorable secondary school recommendations are a direct disadvantage for pupils with Turkish names because attending a lower track may have

long-term negative effects. For instance, the lowest track does not give access to tertiary education, and the Realschule only partly.

Conclusion

In this paper we randomly assigned typical German or Turkish names to identical sets of essays to test the effect of teacher expectations regarding pupil background on grades. We find that grades are significantly lower when an essay bears a Turkish name. Although the effect of names on grades is small, this result stands in contrast to that found for the Netherlands, where names have no effect on grades. Observed teacher characteristics are not correlated with presenting a grading bias but it is possible that an increased awareness of the expectation bias for cultural reasons limits this effect in the Netherlands. In effect, the observed grading bias decreases when a larger share of the essays bears a Turkish name, which points to the potential relevance of having a differentiated view of migrant pupils. Furthermore, in line with the Dutch results, we find that essays from ostensibly Turkish pupils receive recommendations for the highest secondary school track with an 11% lower probability and that teachers hold less positive attitudes towards Turkish than towards German people. These differences in expectations and attitudes constitute a foundation for potentially different behavior towards migrant pupils. The effects we find are relatively small and cannot explain the full gap in the educational attainment between migrant and German pupils. Well-known determinants of pupil performance such as parental education or language difficulties play an important role as well. However, considering the numerous disadvantages already faced by migrant pupils, removing the additional penalty resulting from lower teacher expectations would be a welcome step forward. It is therefore good news that the observed grading and recommendation bias originates from a small group of teachers only. Most teachers do not grade or recommend different tracks based on ostensible pupil

origin. This implies that lower expectations do not necessarily affect teachers' judgment of a pupil's potential and that there is scope for getting rid of these biases for the remaining group. For instance, increased awareness about the importance of teacher expectations through teacher training, could contribute to reducing the grading and expectation bias. Further research is needed to reveal whether training teachers to evaluate pupil verbal skills or training that aims to increase awareness of involuntary discrimination can decrease the observed grading bias.

Table 1: First names used on the essays

German Names		**Turkish Names**	
Essay writer	**Friend**	**Essay writer**	**Friend**
Boys			
Julian	Lars	Murat	Hamid
Stefan	Tobias	Yusuf	Onur
Timo	Paul	Mehmet	Burak
Max	Alexander	Engin	Osman
Florian	Tom	Hakan	Mustafa
Denis	Frederik	Aziz	Selim
Lukas	Philip	Coskun	Idris
Christian	Jan	Enis	Kemal
Niklas	Sebastian		
Daniel	Jonas		
Niels	Lennart		
Andreas	Mark		
Girls			
Nina	Svenja	Fatma	Leyla
Lisa	Katrin	Ayse	Zehra
Julia	Lara	Sevda	Hayat
Jennifer	Anja	Gönül	Sibel
Claudia	Jacqueline	Zeynep	Meryem
Marie	Hanna	Dilara	Sengül
Laura	Leonie	Burcu	Selin
Anna	Sarah	Sevim	Fazilet
Kristina	Natalie		
Melanie	Michelle		
Lena	Paula		
Sandra	Vanessa		

Table 2: Descriptive statistics

	Mean	Std dev	Nb of obs.
Female teacher	0.83	0.38	88
<2 years experience with migrant pupils	0.23	0,42	88
Years of teacher experience total	22.29	11.22	88
Age teacher	48	11	88
Only required teacher degree (Staatsexam Lehramt)	0.83	0.38	88
Time taken to correct the essays	2.4	3.32	88
Essay grade	3.02	0.9	880

Table 3: Average grade by ostensible pupil origin

	Mean	Std Dev	Nb of obs.
German name	2.98	0.84	532
Turkish name	3.11	0.98	348

{*Note:* Grades range from 1 (very good) to 6 (very insufficient)}

Table 4: The effect of ostensible pupil origin on grades

	OLS	Essay fixed effects	Essay and teacher fixed effects
Name of Turkish origin	0,13	0,09**	0,11**
Std Error	0,09	0,04	0,04
Observations	880	880	880
R squared	0,004	0,56	0,65

{ *Note:* Grades range from 1 (very good) to 6 (very insufficient) as in common practice in German schools,**, *** indicate significance at the 10, 5 and 1% level. Dependant variable: grades. Standard errors are clustered by teacher}

Table 5: Proportion of teachers giving German or Turkish names higher grades for the same essay

	Worse grade if Turkish name	No grading bias	Worse grade if German name
At the 10% confidence level	13,6%	86,4%	0%
At the 5% confidence level	3,4%	96,6%	0%
Observations	88	88	88

Table 6: Interaction effects of ostensible pupil origin on essay grades

Name is interacted with:	50% Turkish names in set of essays	Turkish-German attitude gap	$<$ 2 years experience with migrant pupils	Teacher's age below 45
Name of Turkish origin	0,22***	0,10**	-0.08	0,10**
	(0,07)	(0,05)	(0,05)	(0,05)
Interaction effect	-0,19*	-0,01	-0,15	0,02
	(0,10)	(0,02)	(0,10)	(0,09)
Nb of observations	880	880	880	880
R squared	0,65	0,65	0,57	0,65

{*Note:* *,**, *** indicate significance at the 10, 5 and 1% level. Own data. Dependant variable: grades. Specification including teacher and essay fixed effects. Standard errors are clustered by teacher

Table 7: Probit estimates of teacher characteristics on the probability of presenting a grading bias

	Marginal effect	Std. Error
Male teacher	-0,09	(0.07)
Teacher is less than 45 years old	-0,02	(0.07)
Teacher found it difficult to grade the essays	-0.05	(0.07)
Parttime	0.02	(0.10)
Only required teacher degree	-0.05	(0.10)
More than 5 year experience with migrant pupils	-0.03	(0.07)
Attitude gap	0.00	(0.02)
Read all essays before grading	0.02	(0.07)
Parttime	0.00	(0.10)
Followed course in evaluating pupils verbal skills	-0.04	(0.07)
More than 2 hours to correct the essays	0.12	(0.08)
Pseudo R-squared	0.05	

{ *Note:* *,**, *** indicate significance at the 10, 5 and 1% level. Dependant variable: probability of presenting a grading bias}

Table 8: Attitudes towards German versus Turkish people

Attitude towards	German people	Turkish people	p (difference in attitudes)
All teachers	58,5 (19,4)	49,9 (17,8)	p<.0001
< 2 years experience with migrant pupils	69,1 (15,9)	56,1 (13,2)	p<.0001
> 2 years experience with migrant pupils	55,4 (19,3)	48,1 (18,6)	p<.0001

{*Note:* Attitudes measured on a scale of 0 (cold feelings) -100 (warm feelings)

Table 9: Secondary school recommendation by ostensible pupil origin

	Hauptschule	Realschule	Gymnasium	Nb of observations
Name of German origin	34.4%	38.2%	27.4%	532
Name of Turkish origin	38.5%	39.4%	22.1%	348

Table 10: Probit estimates of the recommended secondary school track (average change in probability)

	Hauptschule	Realschule	Gymnasium
Name of Turkish origin	0,04	0,06	-0,11***
Std Error	0,05	0,04	0,03
Essay fixed effects	yes	yes	yes
Teacher fixed effects	yes	yes	yes
Observations	880	880	880
Pseudo R squared	0,33	0,33	0,33

{Note: *,**, *** indicate significance at the 10, 5 and 1% level. Dependant variable: expected feasible secondary school track. Standard errors are clustered by teacher.

Table 11: Recommended secondary school tracks: Interaction effects of ostensible pupil origin with teacher characteristics

	Hauptschule	Realschule	Gymnasium
Name of Turkish origin	0,03	0,07	-0,15*
	(0,10)	(0,07)	(0,08)
Tu Name * 50% Turkish names	0,01	-0,02	0,06
	(0,13)	(0,10)	(0,12)
Name of Turkish origin	0,03	0,04	-0,10**
	(0,06)	(0,04)	(0,04)
Tu Name * Attitude gap	-0,00	-0,02	0,03
	(0,01)	(0,02)	(0,02)
Name of Turkish origin	-0,06	0,03	-0,03
	(0,09)	(0,05)	(0,07)
Tu Name * >5 years experience	0,16	0,06	-0,12*
	(0,12)	(0,07)	(0,06)
Name of Turkish origin	0,03	0,03	-0,08**
	(0,05)	(0,05)	(0,04)
Tu Name * Teacher age <45	0,03	0,10	-0,10
	(0,11)	(0,07)	(0,06)
Name of Turkish origin	0,05	0,06	-0,12***
	(0.05)	(0,04)	(0,03)
Tu Name * No experience	-0,06	0,00	0,06
migrants	(0,13)	(0,08)	(0,15)
Teacher and essay fixed effects	yes	yes	yes
Nb of observations	880	880	880
Pseudo R squared	0,43	0,15	0,37

{Note: *,**, *** indicate significance at the 10, 5 and 1% level. Dependant variable: expected feasible secondary school track. Standard errors are clustered by teacher.

Table 12: Proportion of teachers giving German or Turkish names higher recommendations for the same essay

	Lower track if Turkish name	No tracking bias	Lower track if German name
At the 10% confidence level	13,6%	82,95%	3,4%
At the 5% confidence level	9,1%	87,5%	3,4%
Observations	88	88	88

References

Bertrand, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labour market discrimination. *American Economic Review*, 94(4):991–1013.

Carlsson, M. and Rooth, D.-O. (2007). Evidence of ethnic discrimination in the swedisch labor market using experimental data. *Labour Economics*, 14:716–729.

Casteel, C. A. (1998). Teacher-student interactions and race in integrated classrooms. *Journal of Educational Research*, 92(2):115–120.

Darley, J. and Gross, P. (2005). *Social cognition: key readings*, chapter A hypothesis-confirming bias in labelling effects. Taylor and Francis Ltd.

Dee, T. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, 95(2):158–165.

Ferguson, R. F. (2003). Teachers' perceptions and expectations and the black-white test score gap. *Urban Education*, 38(4):460–507.

Figlio, D. (2005). Names, expectations and the black-white test score gap. *NBER Working Paper*, 11195.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., and Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin*, 31:1369–1385.

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

Lebrecht, S., Pierce, L., Tarr, M., and Tanaka, J. (2009). Perceptual other-race training reduces implicit racial bias. *PLoS ONE*, 4(1). doi:10.1371/journal.pone.0004215.

Lindahl, E. (2007). Comparing teachers' assessments and national test results : evidence from sweden. *Institute for Labour Market Policy Evaluation Working Paper*, 24.

OECD (2006). *Where immigrant students succeed - A comparative review of performance and engagement in PISA 2003*. Organisation for Economic Cooperation and Development, Paris, programme for international student assessment edition.

Rosenthal, R. and Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectations and student intellectual development*. New York: Holt, Rinehart and Winston.

Rudman, L., Ashmore, R., and Gary, M. (2001). 'unlearning' automatic biases: The malleability of implicit prejudice and stereotypes. *Journal of Personality and Social Psychology*, 81(5):856–868.

Seraydaran, L. and Busse, T. (1981). First name stereotypes and essay grading. *Journal of Psychology*, 108(2):253–257.

Tenenbaum, H. R. and Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for european american students? a meta-analysis. *Journal of Educational Psychology*, 99(2):253–273.

van Ewijk, R. (2009). Same work, lower grade? student ethnicity and teachers' subjective assessments. *Universiteit van Amsterdam Discussion Paper*.