Empir Econ (2013) 45:523–538
DOI 10.1007/s00181-012-0609-x

Discrimination in grading: experimental evidence from primary school teachers

Maresa Sprietsma

Received: 31 January 2011 / Accepted: 24 April 2012 / Published online: 6 June 2012

Abstract This paper studies the effect of teacher expectations on essay grades in an experimental setting. For 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 the same essays obtain significantly worse grades and lower secondary school recommendations when bearing a Turkish sounding name.

Keywords Discrimination · Migrant background · Grading · Field experiment

JEL Classification I20 · I24 · C93

1 Introduction

Children with a migrant background are lagging behind in terms of educational performance in many industrialized countries. In Germany, the performance gap is particularly large as second-generation migrant pupils lag behind their native peers by 93 test score points (out of 500 average test score points) on PISA. This is equivalent to one and a half proficiency levels[1] or to around one standard deviation in test scores (OECD 2006). and has long-term consequences for their labor market perspectives. For instance, there is consistent evidence that better cognitive skills yield higher earnings (Murnane et al 1995; Harmon et al 2003; Card 1999).

The less favorable social and educational background level of their parents as well as language difficulties (Ammermüller 2007; Casey and Dustmann 2008) are important

[1] PISA uses 6 proficiency levels in Maths and Reading.

M. Sprietsma (✉)
Centre for European Economic Research (ZEW), L7,1, 68 161 Mannheim, Germany
e-mail: sprietsma@zew.de

reasons behind the lower educational performance of migrant children in Germany. However, little is known about the role of teacher expectations in explaining the lower educational performance of children with a migrant background.

Existing evidence shows that teachers have different expectations regarding the performance of pupils according to their ethnic background (Tenenbaum and Ruck 2007), and that teacher expectations affect pupil performance in several ways. First, teacher expectations with respect to pupil ability may lead to (unconscious) changes in teacher behavior that affect the actual performance of their pupils. In one of the first experiments on the self-fulfilling prophecy by Rosenthal and Jacobson (1968), teachers were led to believe that a random set of pupils had higher ability. At the end of the school year, these pupils actually performed significantly better than their classmates. Unconscious differences in teacher behavior depending on pupil background have indeed been observed in more recent studies. There is e.g., evidence that African–American pupils receive significantly less praise, less direct questions from their teachers and less feedback after mistakes than their peers from different ethnic backgrounds (Casteel 1998; Ferguson 2003). Such differences in teacher behavior related to expectations are thought to affect pupil performance all the more as they interact with the development of pupil self-perception and behavior in a way that reinforces teacher expectations (Ferguson 2003; Dee 2005).

Second, besides inducing changes in teacher behavior and actual pupil performance, teacher expectations may also affect the way in which teachers perceive and grade pupil performance. This is particularly the case when the skills to be evaluated leave much room for subjectivity such as behavior in class (Dee 2005). Grades are the main indicator of ability and performance that teachers give to their pupils and can have long-term consequences for pupil achievement and thereby on future employment perspectives (Papay et al. 2011). For instance, in Germany, teacher grades are to a great extent the basis for the recommended secondary school track.[2] 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 behavior in class, or to encourage pupils with low self-esteem. Moreover, even if teachers generally use a set of specific evaluation criteria to assess pupil work, many subjective impressions and expectations may play a role in estimating the quality of pupil work.

Given the existing performance gap between pupils with and without a migrant background, teachers rightly hold different expectations as to the average performance of this group of pupils, and the quality of an essay may not fit their expectations given the pupil's supposed origin. Teachers may react to this distortion in various different ways. A first possibility is that teachers may give the essay a better grade than they would have otherwise, 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 or her expectations. In fact, psychological research has shown that persons are likely to search harder for evidence in

[2] In the majority of the German regions, there is no standardized test, and parents do not have the final word in this decision.

favor of their expectations than the other way round (Darley and Gross 2005). Interestingly, names seem to be a particularly strong trigger for expectations, even more than explicit information on ethnic origin or socio-economic background (Anderson-Clark et al. 2008; Figlio 2005).

In this paper, we assess the effect of pupils' supposed origin on essay grades in Germany in an experimental setting. To this purpose, we randomly assign Turkish and German first names to a set of 4th grade essays. As a result, 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 the origin of the pupil that wrote the essay, affects the essay grade.

Hanna and Linden (2009) use a similar experimental design to measure grading discrimination in India. They randomly assigned age, gender, and caste of pupils to the cover sheets of exams. They find that teachers on average assigned scores to pupils from low castes that were 3 to 9 % lower than those of pupils who were described as being from a higher caste group. Van Ewijk (2011) performs a similar experiment with pupil names in the Netherlands and finds no effect of first names associated with a migrant background on grades. Since the present experiment uses the same design as the study by Van Ewijk (2011), we refer to the Dutch results on several occasions in the interpretation part.

In non-experimental studies that assess the effect of stereotypes on grading such as Lindahl (2007); Lavy (2008), or Kristen (2006), grades given by the pupils' own teachers are compared with grades obtained at centralized examinations for different groups of pupils. However, these studies do not allow distinguishing the effect of teacher stereotypes on a pupil's origin from the effect of the teacher's personal knowledge about pupils. The grades given by a teacher to his or her own pupils are indeed likely to be affected by the full set of information and impressions the teacher has about the pupil. By contrast, in our experimental setting, the teacher expectations can be triggered only by the pupil's name, because the children who wrote the essays are unknown to the teachers who correct the essays.

In order to measure teachers' expectations with respect to pupils' educational achievement, we ask the teachers to emit a secondary school recommendation based on the essays. We 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 conscious teacher attitudes toward German versus Turkish people using feeling thermometers, which allow teachers to express the warmth of their feelings toward groups of persons and ideas.

We find that teachers hold lower expectations with respect to pupils of supposedly Turkish origin and that essays receive significantly lower grades when bearing a Turkish name. However, these effects originate from a minority of teachers in the sample.

We would like to underline that teacher expectations and their impact on teacher behavior are not to be confused with intentional pupil discrimination. Although, we can speak of discrimination as soon as a person's behavior is affected by his/her beliefs about another group, this need not be intentional. In fact, psychological research has shown that behavior is often affected by beliefs and expectations involuntarily, and

that such implicit bias is only weakly correlated with explicit judgments (Hofmann et al. 2005). However, there is also evidence that creating awareness of the existence of implicit bias may limit its effects (Rudman et al. 2001). This is one of the motivations for our study.

This paper contributes to the literature by providing experimental evidence on the effect of pupils' names on grades based on a sample of German primary school teachers. We compare our findings for Germany with those from the same experiments in the Netherlands and India. Replicating experimental designs in different countries is important to assess the external validity of results. Moreover, differences in results across countries may point to specific problems in the given national context.

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 design of this experiment (as well as those by Hanna and Linden (2009); Van Ewijk (2011)) is similar to that used to assess discrimination in hiring. In these experiments, identical application letters and CVs are sent out, bearing either European or foreign names (Bertrand and Mullainathan 2004; Carlsson and Rooth 2007; Kaas and Manger 2010). Following the same principle, we collected essays from two anonymous 4th grade classes (4th graders are about ten years old) that do not participate in the experiment afterwards. We then sent the same set of ten essays 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. Essays that were clearly identifiable as not being from a Turkish background because of the choice of topics (for example religious activities such as related to Christmas) were not included in our set of essays. None of the included essays were actually written by pupils of Turkish background[3] We also excluded extremely bad or short essays that were unlikely to generate much variation in grades across teachers.

The names we used were popular German and Turkish sounding names that were frequently given 10 years ago (around the birth year of the essay writers). The names were taken from websites with first-name statistics. German names for instance included Max, Stefan, Anja or Melanie whereas the Turkish name set contained names like Sevda, Gönül, Hakan or Coskun.[4] We chose to use Turkish names because this nationality still represents one of the largest migrant communities in Germany: 37 % of non-German primary school pupils have Turkish nationality.[5] Moreover, the PISA

[3] As far as the first names and choice of topics in the original essays reveal.

[4] Note that the association of a first name to a social and cultural group is a local perception. The same first name, for instance "Xavier," may sound Afro-American to an American citizen and French native to a French citizen. The first names we use have been confirmed to sound "Turkish" by a German person of Turkish background.

[5] Statistisches Bundesamt (2008–2009).

Table 1 Name sets used on the essays

Set 1		Set 2		Set 3		Set 4	
Essay writer	Friend	Essay writer	Friend	Essay writer	Friend	Essay writer	Friend
Julian	Lars	Max	Alexander	Daniel	Jonas	Coskun	Idris
Lisa	Katrin	Sevda	Hayat	Anna	Sarah	Sevim	Fazilet
Fatma	Leyla	Marie	Hanna	Dilara	Sengül	Sandra	Vanessa
Murat	Hamid	Florian	Tom	Hakan	Mustafa	Niels	Lennart
Stefan	Tobias	Mehmet	Burak	Lukas	Philip	Aziz	Selim
Jennifer	Anja	Gönül	Sibel	Laura	Leonie	Lena	Paula
Nina	Svenja	Ayse	Zehra	Zeynep	Meryem	Melanie	Michelle
Timo	Paul	Denis	Frederik	Christian	Jan	Enis	Kemal
Julia	Lara	Claudia	Jacqueline	Kristina	Natalie	Burcu	Selin
Yusuf	Onur	Engin	Osman	Niklas	Sebastian	Andreas	Mark

performance of pupils with Turkish background is lower than that for pupils with Asian or Eastern European backgrounds (OECD 2006).

Four different name sets were allocated to the ten essays (see Table 1). This implies that approximately 25 % of the teachers received the same name set. In two of the name sets (sets 1 and 3), 30 % of the names were of Turkish origin, in the other two 50 %. We did not increase the share of migrant origin names above 50 % because the share of Turkish pupils is generally lower in German schools. In order for the relatively high shares of Turkish pupils not to appear unrealistic to the teachers, they were told that the essays came 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 the description of the friend.

We sent information letters to all public primary schools in two German regions[6] by mail, asking the school directors to inform their 4th grade German teachers about the study. If teachers were interested, the essays, the questionnaires and the instructions were sent to the school by mail 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. No financial incentives were given to participate (contrary to the teachers in the Dutch experiment by Van Ewijk (2011), who received 25 Euro).

In order for the experiment to work, it was important that teachers did not know the real purpose of the experiment. They were told that the study aimed to assess the determinants of grading 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.

[6] Data protection legislation for teachers is regional in Germany. The two regions we use for our experiment present the advantage that school directors are allowed to decide whether or not their school participates.

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 three times). We also manipulated the friends' names to have the same origin as the ostensible author, including when several friends were mentioned. The essays were about 2/3 of a page long in longhand. This corresponds to 222 words on average per essay. We copied the handwritten essays (including mistakes and formatting) into the computer to be in typewritten form of the same length. Teachers of the same school always received the same name sets 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 answer several questions in a short questionnaire. The grades and recommendations were to be filled out at the beginning of this questionnaire. Using so-called feeling thermometers, the teachers were asked to indicate on a scale of 0 (very cold/uncomfortable) to 100 (very warm/comfortable) how they felt concerning twelve specific topics such as politicians, environmental policy, and German and Turkish people. This indicator allowed us to proxy attitudes towards migrants. The feeling thermometers were included towards the end of the questionnaire, and the teachers were told these were meant for another survey on the opinions of different population groups including teachers. Included topics for the feelings thermometers were German politicians, federalism, the EU, the USA, financial investment, environmental policy, the police, the German legal system, the Islam, German people, Turkish people, and East-German people.

Teachers were asked to give grades in line with common practice in German schools, that is, grades ranging from 1 (very good) to 6 (very insufficient). Moreover, teachers could give grades like 1− or 4+, which is translated in numbers accordingly (1− would be 1.25 whereas 4+ corresponds to 3.75). In the rest of the paper, we reverse the common German grading scale so that increasing grades correspond to better performance as is the case in most other countries. Please recall that in what follows grade 1 is thus the lowest grade and 6 the best grade.

3 Results

Eighty-eight primary school teachers (from 58 different schools) from two German federal regions agreed to participate in our study. The information letters were sent to *all* public primary schools in these regions (these are around 3500 schools), and

Table 2 Descriptive statistics as compared with full teacher population

	Sample	Region 1	Region 2	Germany
Share female (%)	83	87	91	88
Average age	49	47	44	43
Share part-time (%)	42	55	40	51
Nb of observations	88	25991	10590	190299

Note: These figures are from the federal statistical office (Statistisches Bundesamt 2008–2009) and refer to primary school teachers for 2008–2009

Table 3 Descriptive statistics by assigned name origin

	German name on essay		Turkish name on essay	
	Mean	Std.dev	Mean	Std.dev
Female teacher	0.83	(0.38)	0.83	(0.38)
2 years experience with migrant pupils	0.24	(0.43)	0.20	(0.40)
Years of teacher experience total	22.27	(11.22)	22.32	(11.23)
Age teacher	48.52	(10.77)	48.64	(10.71)
Holds only the required[a] teacher degree	0.84	(0.37)	0.82	(0.39)
Time taken to correct the essays	2.43	(3.36)	2.37	(3.23)
Essay grade	4.02	(0.84)	3.89	(0.98)
Number of observations	532		348	

Note: Grades range from 1 (very insufficient) to 6 (very good)
[a] The required degree to become a primary school teacher is the Staatsexam Lehramt

we had only 88 participants. It is thus clear that the response rate was extremely low and that our sample is not representative of the primary school teacher population of these regions. The participating teachers are likely to be more engaged and motivated as their colleagues. 80 % of the teachers who agreed to participate did send back the completed questionnaire.

Table 2 presents compared teacher characteristics from our sample with administrative teacher statistics in the regions where our sample originates from in the school year 2008/2009. Unfortunately, these statistics are available for only three variables: gender, age, and share of part-time teachers. In the sample as well as in the full population most primary school teachers are female. Moreover, Table 2 shows that the primary school teacher population in Germany is relatively old, because of an increase in hiring during the sixties, when the baby boomers went to school. The trend in the teachers' age is decreasing in recent years as these teachers start to retire. As a result, primary school teachers were about 43 years old on average in the school year 2008–2009. Finally, around half of the teachers in our sample and in the full teacher population work part-time.

Table 3 shows descriptive statistics of the teachers in the sample by assigned name type. All teachers are German natives and on average have 22 years of experience. Eighty-three percent are female, they are on average 49 years old, and most had at

Table 4 Interrater reliability agreement on essay rank by quality

Rankings	K alpha	95 % Confidence interval	No. teachers	Pairs
All essays	0.6	(0.62; 0.67)	88	38280
Content as important criterion	0.62	(0.60; 0.65)	38	7030
Style as important criterion	0.52	(0.48; 0.56)	17	1360
Language as important criterion	0.71	(0.70; 0.73)	35	5950

Table 5 The effect of assigned pupil origin on grades (OLS estimates)

	(1)	(2)	(3)
Name of Turkish origin	−0.13	−0.09**	−0.11**
Std error	(0.09)	(0.04)	(0.04)
Essay fixed effects	No	Yes	Yes
Teacher fixed effects	No	No	Yes
Observations	880	880	880
R squared	0.004	0.56	0.65

Note: Grades range from 1 (very insufficient) to 6 (very good)
*, **, *** indicate statistical significance at the 10, 5 and 1 % level of confidence, dependent variable: grades. Standard errors in parentheses and clustered by teacher

least two years of teaching experience with migrants. Very few teachers have other degrees than that required to become a primary school teacher (i.e., the state examination Lehramt). It took the teachers about 2.5 hours on average to correct the ten essays. Table 3 also makes clear that teacher characteristics do not vary significantly according to the assigned names as we would expect as a result of the random assignment procedure. The last line of Table 3 shows the descriptive statistics of grades by ostensible pupil origin. On average, the essays receive 0.13 points lower test scores (out of 6) when they bear a Turkish name. This difference is significant at the 5 % confidence level.

In order to assess to what extent the teachers agree on the ranking of the essays by quality, we compute a measure of Interrater Reliability Agreement. To this purpose, we compute the Krippendorff's alpha (Hayes and Krippendorff 2007) based on the rankings of the essays by teacher from 1 (best essay) to 10 (worst essay). Table 4 shows that the agreement on the ranking of the essays is at 0.65. Moreover, the degree of agreement is higher when teachers use language or content rather than style as an important evaluation criterion.

Regression estimates of essay grades on ostensible pupil origin are presented in Table 5. 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 effective 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). The latter specifications

Table 6 Proportion of teachers giving German or Turkish names lower 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	0 %	96.6 %	3.4 %
Observations	88	88	88

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 not significant in the baseline specification. Standard errors become much smaller once we include important determinants of grades such as teacher and essay dummies. Moreover, the sample size is relatively small, and therefore standard errors are relatively large. However, the point estimates are similar in all three specifications. In the latter specifications, we find that essays obtain significantly lower 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.

A grading bias occurs if a teacher on average grades essays with assigned Turkish names differently than essays with assigned German names. In order to find out how the grading bias is distributed among teachers, 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 rank the sums of 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 bears a Turkish name. Table 6 shows that the grading bias comes from a minority (14 %) of teachers that gave essays with Turkish names worse grades at the 10 % confidence level.

In a next step, we therefore try to identify whether the observed grading bias is correlated with certain teacher characteristics. For this purpose, we included crossed effects of the Turkish name dummy times teacher experience with migrant pupils, the teacher's age, and the attitude gap between German and Turkish people in the main regression. None of these crossed effects are significantly different from zero as can be seen in Table 7. 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 in favor of essays with native German names shows that these teachers do not differ from the other teachers in any of the observed characteristics (Table 8).

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 7). 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

Table 7 Interaction effects of ostensible pupil origin on 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*** (0.07)	−0.10** (0.05)	−0.08 (0.05)	−0.10** (0.05)
Interaction effect	0.19* (0.10)	0.01 (0.02)	−0.15 (0.10)	−0.02 (0.09)
Essay and teacher fixed effects	Yes	Yes	Yes	Yes
Nb of observations	880	880	880	880
R squared	0.65	0.65	0.65	0.65

Note: *,**,*** indicate statistical significance at the 10, 5 and 1 % level of confidence. Own data. Dependent variable: grades. Specification including teacher and essay fixed effects. Standard errors in parentheses and clustered by teacher. Grades range from 1 (very insufficient) to 6 (very good)

Table 8 Probit estimates of teacher characteristics on the probability of presenting a grading bias

	Marginal effect	Std. error
Male teacher	−0.08	0.07
Teacher is less than 45 years old	−0.02	0.07
Teacher found it difficult to grade the essays	−0.02	0.07
Holds only the required teacher degree	−0.08	0.10
More than 2 year experience with migrant pupils	−0.07	0.10
Attitude gap	0.00	0.02
Read all essays before grading	−0.06	0.11
Part-time	−0.02	0.08
Followed course in evaluating pupils' written work	−0.03	0.07
Took more than 2 h to correct the essays	0.16∗	0.11
Pseudo R-squared	0.07	

Note: *,**,*** indicate statistical significance at the 10, 5 and 1 % level of confidence. Dependent variable: probability of presenting a grading bias

diminished after they learned to individuate faces of that race showing that the awareness of diversity inside the unknown group reduces unconscious stereotypes. Using German data, Wagner et al. (2006) show that an increase in the percentage of ethnic minority members in a neighborhood reduced the majority's prejudice against them.

Hanna and Linden (2009) find that assigned pupils' origin also affects grades in India, but our results stand in contrast to the results obtained in the Netherlands by Van Ewijk (2011), where no effect of names on grades was found. The design of the experiment is the same in the two 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 awareness with respect to the phenomenon of expectation bias between Dutch and German teachers. Training teachers to evaluate pupils can be expected to transmit such awareness, but both in the Netherlands and in Germany less than 40 % of teachers participated in training to assess pupils' written work. Crossed effects of following a course in evaluating pupils' verbal skills with the name dummy

Table 9 Attitudes towards German versus Turkish people

Attitude towards	German people	Turkish people	Attitude gap	Difference in attitude gaps is significant*
All teachers	58.5 (20.8)	49.9 (19.1)	8.6	
<2 years experience with migrant pupils	69.1 (36.6)	56.1 (30.3)	13	Yes
>2 years experience with migrant pupils	55.4 (23.5)	48.1 (22.7)	7.3	
<5 years experience with migrant pupils	65.3 (26.0)	53.8 (21.3)	11.5	No
>5 years experience with migrant pupils	54.5 (28.2)	47.6 (27.4)	6.9	
Male teacher	55.2 (33.5)	51.3 (22.3)	3.9	No
Female teacher	59.2 (24.1)	49.7 (22.6)	9.6	
Teacher <45 years old	60.4 (33.3)	50.5 (31.9)	9.9	No
Teacher >45 years old	57.6 (26.6)	49.7 (24.0)	7.9	
Teacher has only State examination degree	58.7 (21.9)	49.5 (19.2)	9.2	No
Teacher has further degrees	58.0 (61.6)	52.1 (63.4)	5.9	
Part-time teacher	64.1 (46.5)	54.8 (28.9)	9.3	No
Full-time teacher	57.6 (22.9)	49.1 (21.8)	8.5	

Note: Attitudes measured on a scale of 0 (cold feelings) – 100 (warm feelings). Standard deviations in parentheses

* The difference in the attitude gap between groups is significant at the 10 % confidence level

are not significantly different from zero, meaning 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. The teachers in Germany were not paid a reward, contrary to the Dutch teachers, but there is no evidence that German teachers took the assessment task for the experiment less seriously. For instance, they took much longer on average to correct each essay.

As mentioned in the introduction, both colder feelings towards Turkish persons and lower expectations with regard to the performance of migrant pupils could be the source of differences in behavior in teaching and in grading. The feeling thermometers show that German teachers (just like Dutch teachers, (Van Ewijk (2011)) have less positive attitudes towards Turkish people than towards German people. The average attitude gap amounts to 8.6 points on a 100 point scale (Table 9). Teachers with less than 2 years experience of teaching migrant pupils have warmer feelings towards all groups but the difference in attitudes towards German and Turkish persons is significantly larger. Other teacher characteristics are not correlated with a larger attitude gap.

Our second research objective is to assess differences in teacher expectations by first name origin based on the teachers' secondary school recommendations. The aim

Table 10 Secondary school recommendation by ostensible pupil origin

	Hauptschule	Realschule	Gymnasium	No. of obs.
Name of German origin	34.4 %	38.2 %	27.4 %	532
Name of Turkish origin	38.5 %	39.4 %	22.1 %	348
Difference is significant at the 5 % confidence level	No	No	Yes	

Note: The Hauptschule is the most technical track, the Gymnasium the general track. The Realschule is in between

Table 11 Probit estimates of the recommended secondary school track (average change in probability

	Realschule	Gymnasium
Name of Turkish origin	0.06	−0.11***
Std error	(0.04)	(0.03)
Teacher and essay fixed effects	Yes	Yes
Observations	880	880
Pseudo R squared	0.33	0.33

Note: *,**,*** indicate statistical significance at the 10, 5 and 1 % level of confidence. Dependent variable: expected feasible secondary school track. Standard errors in parentheses and clustered by teacher

is to investigate whether there is a foundation for the observed grading bias in terms of expectations. Descriptive statistics of the recommended secondary school tracks by ostensible pupil origin are presented in Table 10. It appears that essays receive a Gymnasium recommendation more often when they bear a German name.

We expect the recommended feasible secondary school tracks will depend on the grade given to the essays but also on teacher expectations towards the performance of migrant pupils in the different tracks. We do not, however, include the given grade in the estimation of the effect of the names on the recommended secondary school track. This may indeed lead to a reverse causality problem, if the recommended track, as a measure of teacher expectations, affected the grade in turn. The coefficients of such an estimation could therefore not be interpreted in a meaningful way. Table 11 displays the results of probit estimations of the probability of receiving a recommendation to attend the two upper secondary school tracks.[7] We include teacher and essay fixed effects in the estimation, and standard errors are clustered by teacher. Our results indicate that teacher expectations with regard to pupil capacity for attending different secondary school tracks are significantly affected by the origin of the first 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

[7] Estimates from linear regression using the same specification yield 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.

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 %	82.9 %	3.4 %
At the 5 % confidence level	9.1 %	87.5 %	3.4 %
Observations	88	88	88

Table 13 Recommended secondary school tracks: Interaction effects of ostensible pupil origin with teacher characteristics

	Realschule	Gymnasium
Name of Turkish origin	0.07 (0.07)	−0.15* (0.08)
Tu Name* 50 % Turkish names	−0.02 (0.10)	0.06 (0.12)
Name of Turkish origin	0.04 (0.04)	−0.10** (0.04)
Tu Name* Attitude gap	−0.02 (0.02)	0.03 (0.02)
Name of Turkish origin	0.03 (0.05)	−0.03 (0.07)
Tu Name* >5 years experience migrant pupils	0.06 (0.07)	−0.12* (0.06)
Name of Turkish origin	0.06 (0.04)	−0.12*** (0.03)
Tu Name* <2 years experience migrant pupils	0.00 (0.08)	0.06 (0.15)
Name of Turkish origin	0.03 (0.05)	−0.08** (0.04)
Tu Name* Teacher age <45	0.10 (0.07)	−0.10 (0.06)
Teacher and essay fixed effects	Yes	Yes
Nb of observations	880	880

Note: *,**,*** indicate statistical significance at the 10, 5 and 1 % level of confidence. Dependent variable: expected feasible secondary school track. Standard errors in parentheses and clustered by teacher

effects are found for the lower track. These results are in line with the results found for the Netherlands by Van Ewijk (2011).

In order to find out how the expectation bias is distributed among teachers, we rank order the expectation biases of the teachers in the same way that we proceeded for the grading bias in Table 6. That is, a teacher presents an expectation bias if he or she on average recommends secondary school tracks differently for essays with assigned German names as compared to essays with assigned Turkish names. The results are presented in Table 12. It appears that, similarly to the grading bias, only a small fraction of teachers (around 10 %) presents such an expectation bias. The majority of teachers does not have different expectations according to pupil background.

Finally, we looked into whether certain teacher characteristics are related to presenting different expectations by pupil origin. Estimates of the effect of assigning a Turkish first name to an essay on the recommended secondary school track including crossed effects of the name with teacher characteristics are presented in Table 13. Each line represents the estimates from the probit estimation including the specified crossed effect for each of the two upper secondary school types. The results show that the effect of the name on the expected feasible secondary school track does neither vary with the teacher 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. To the contrary, 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 past experience with migrant pupils (as predicted by Bayesian Updating Theory in De Groot (2004)) and with findings on stereotypes from psychological studies (Harari and McDavid 1973).

The different secondary school recommendations for essays that were assigned Turkish first names as well as the difference in attitudes towards Turkish people show that there is a foundation for potentially different teacher behavior towards migrant pupils. Although these effects originate from only a small subgroup of the participating teachers and are relatively small in size, the less favorable secondary school recommendations and lower grades are a direct disadvantage for pupils with Turkish names because attending a lower secondary school track and lower grades often have long-term negative effects on income and employment status.

4 Conclusion

In this paper, we randomly assigned typical German or Turkish names to a set of ten essays to measure the effect of supposed pupil origin on grades. We find that the assignment of a Turkish name has a small but significant negative effect on the perceived quality of an essay as well as on the secondary school recommendation. However, the observed grading bias and lower expectations originate from a small group of teachers only. Most teachers do not grade or recommend different tracks based on ostensible pupil origin. Although this result is in line with findings for essays assigned as being from low caste pupils in India, it contradicts the findings for the Netherlands, where essays with Turkish or Moroccan names did not obtain different grades than those assigned to Dutch names. Interestingly, Dutch teachers do hold lower expectations with respect to the performance of pupils of Turkish or Moroccan origin, but this does not translate into grading discrimination.

Observed teacher characteristics are not correlated with presenting a grading bias, but it is possible that an increased awareness of the existence of expectation bias (which we do not observe) limits this effect in the Netherlands. We do not have information as to teachers' knowledge of expectation bias, but the observed grading bias decreases when a larger share of the essays bears a Turkish name, which points to the potential relevance of being aware of diversity in performance inside the group of migrant pupils. This does not apply for holding lower expectations towards pupils with Turkish names; these do not depend on the share of Turkish names in the name sets.

Furthermore, the less positive teacher attitudes toward Turkish than toward German people are not related to the grading or to the recommendation bias. Only lower performance expectations, not colder feelings, therefore constitute a foundation for grading bias against migrant pupils in Germany. Nevertheless, both colder feelings and lower expectations may affect teacher behavior in class.

The effects we find are relatively small as the grading bias amounts to about 10 % of a standard deviation in essay grades. In contrast, family background including

language skills are thought to account for more than half of the reading test score gap of about one standard deviation between pupils with and without a migrant background in Germany(Ammermüller 2007). However, given that migrant pupils already face lower language skills, lower levels of cultural capital, and statistical discrimination on the labor market, removing the additional penalty resulting from discrimination in grading and from lower secondary school recommendations would be a welcome step forward. Moreover, lower teacher expectations may also affect pupil performance directly through teacher behavior. Finally, considering the teachers' low willingness to participate, we expect the grading bias and recommendation bias to be potentially larger in the actual population of teachers than in our sample of more motivated teachers.

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 and what kind of teacher training can decrease the observed grading bias.

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 many thanks are due to Reyn van Ewijk for his cooperation and feedback, as well as to Katja Coneus, Julia Horstschräer and the anonymous referees for their numerous and constructive comments.

References

Ammermüller A (2007) Poor background or low returns? Why immigrant students in Germany perform so poorly in the Programme for International Student Assessment. Education Economics 15(2):215–230

Anderson-Clark T, Green R, Henley T (2008) The relationship between first names and teacher expectations for achievement motivation. J Lang Soc Psychol 27(1):94–99

Bertrand M, Mullainathan S (2004) Are Emily and Greg more employable than lakisha and Jamal? A field experiment on labour market discrimination. Am Econ Rev 94(4):991–1013

Card D (1999) The causal effect of education on earnings. The Handbook of Labor Economics, vol 3A. Elsevier, Amsterdam

Carlsson M, Rooth DO (2007) Evidence of ethnic discrimination in the Swedisch labor market using experimental data. Labour Econ 14:716–729

Casey T, Dustmann C (2008) Intergenerational transmission of language capital and economic outcomes. J Hum Resour 43(3):660–687

Casteel CA (1998) Teacher-student interactions and race in integrated classrooms. J Educ Res 92(2):115–120

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

De Groot M (2004) Optimal statistical decisions. Wiley Classics Library, New York

Dee T (2005) A teacher like me: does race, ethnicity, or gender matter? Am Econ Rev 95(2):158–165

Ferguson RF (2003) Teachers' perceptions and expectations and the black-white test score gap. Urban Educ 38(4):460–507

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

Hanna R, Linden L (2009) Measuring discrimination in education. NBER Working Paper 15057

Harari H, McDavid J (1973) Name stereotypes and teachers' expectations. J Educ Psychol 65(2):222–225

Harmon C, Oosterbeek H, Walker I (2003) The returns to education: microeconomics. J Econ Surv 17(2):115–155

Hayes A, Krippendorff K (2007) Answering the call for a standard reliability measure for coding data. Commun Methods Meas 1(1):77–89

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

Kaas L, Manger C (2010) Ethnic discrimination in Germany's labour market: a field experiment. IZA Discussion Paper 4741

Kristen C (2006) Ethnische diskriminierung in der grundschule. Kölner Zeitschrift für Soziologie und Sozialpsychologie 58(1)

Lavy V (2008) Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. J Public Econ 92(10–11):2083–2105

Lebrecht S, Pierce L, Tarr M, 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

Murnane R, Willett J, Levy F (1995) The growing importance of cognitive skills in wage determination. Review of Economics and Statistics 77(2):251–266

OECD (2006) Where immigrant students succeed: a comparative review of performance and engagement in PISA 2003, Programme for International Student Assessment edn. Organisation for Economic Cooperation and Development, Paris

Papay J, Murnane R, Willett J (2011) How performance information affects human-capital investment decisions: the impact of test-score labels on educational outcomes. NBER Working Paper 17120

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

Rudman L, Ashmore R, Gary M (2001) 'Unlearning' automatic biases: the malleability of implicit prejudice and stereotypes. J Personality Soc Psychol 81(5):856–868

Seraydaran L, Busse T (1981) First name stereotypes and essay grading. J Psychol 108(2):253–257

Statistisches Bundesamt (2008–2009) Bildung und Kultur: Allgemeinbildende Schulen. Fachserie 11(1)

Tenenbaum HR, Ruck MD (2007) Are teachers' expectations different for racial minority than for european american students? A meta-analysis. J Educ Psychol 99(2):253–273

Van Ewijk R (2011) Same work, lower grade? Student ethnicity and teachers' subjective assessments. Econ Educ Rev 30:1045–1058

Wagner U, Christ O, Pettigrew T, Stellmacher J, Wolf C (2006) Prejudice and minority proportion: contact instead of threat effects. Soc Psychol Q 69(4):380–390