RUHR
ECONOMIC PAPERS

Julia Bredtmann
Sebastian Otten
Christina Vonnahme

Discrimination in Grading? Evidence on Teachers' Evaluation Bias Towards Minority Students

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

RWI

#1122

Imprint

Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Ludger Linnemann
Technische Universität Dortmund, Department of Business and Economics
Economics – Applied Economics
Phone: +49 (0) 231/755-3102, e-mail: Ludger.Linnemann@tu-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83 - 3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Ronald Bachmann, Prof. Dr. Almut Balleer, Prof. Dr. Manuel Frondel,
Prof. Dr. Ansgar Wübker
RWI, Phone: +49 (0) 201/81 49 - 213, e-mail: presse@rwi-essen.de

Editorial Office

Sabine Weiler
RWI, Phone: +49 (0) 201/81 49-213, e-mail: sabine.weiler@rwi-essen.de

Ruhr Economic Papers #1122

Responsible Editor: Ruhr Bachmann

ISSN 1864-4872 (online) – ISBN 978-3-96973-304-2

Ruhr Economic Papers #1122

Julia Bredtmann, Sebastian Otten, and Christina Vonnahme

Discrimination in Grading? Evidence on Teachers' Evaluation Bias Towards Minority Students

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

RWI

Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

https://dx.doi.org/10.4419/96973304
ISSN 1864-4872 (online)
ISBN 978-3-96973-304-2

Julia Bredtmann, Sebastian Otten, and Christina Vonnahme*

Discrimination in Grading? Evidence on Teachers' Evaluation Bias Towards Minority Students

Abstract

We analyze whether teachers discriminate against ethnic minority students in terms of grading. Using comprehensive data on students in German primary and secondary schools, we compare students' scores in standardized, anonymously graded achievement tests with non-anonymous teacher ratings within a difference-in-difference (DiD) framework. We find that, on average, minority students receive lower grades than majority students in both German and Math. However, these differences are not due to discrimination in grading against minority students. Instead, performance gaps between minority and majority students are significantly reduced when being graded by the teacher compared to being assessed through the standardized test. We provide supporting evidence that this finding cannot be explained solely by the fact that minority students face higher barriers on the standardized test due to language difficulties. Rather, our results suggest that teachers have a positive evaluation bias towards ethnic minority students.

JEL-Codes: F22, I24, J15

Keywords: Immigration; discrimination; grading bias

December 2024

* *Julia Bredtmann, RWI, UDE, CReAM, and IZA; Sebastian Otten, UDE, CReAM, and RWI; Christina Vonnahme, RWI.* *– This work was supported by research funding from the German Federal Ministry for Family Affairs, Senior Citizens, Women and Youth. – The authors are grateful to Amanda Agan, Jan Bietenbeck, Anna Bindler, Christina Felfe, Colin Green, Andrea Ichino, Ingo Isphording, Donika Murataj, Hans Henrik Sievertsen, Robert Slonim, and Miriam Wüst as well as to seminar and conference participants at the Lisbon School of Economics, VATT Institute for Economic Research, the University of Bielefeld, the 1st Workshop on Education Economics and Policy, the 9th Leuven Economics of Education Research Conference, the Lisbon Migration Economics Workshop, the 37th Annual Conference of the European Society for Population Economics, and the European Association for Labour Economists Conference 2024 for valuable comments and suggestions.* *– All correspondence to: Julia Bredtmann, RWI, Hohenzollernstraße 1–3, 45128 Essen, Germany, e-mail: julia.bredtmann@rwi-essen.de*

1 Introduction

In most OECD countries, children from immigrant families lag behind their majority peers in terms of educational achievement. Ethnic minority students are less likely to participate in early childhood education, acquire fewer academic skills, and graduate from high school at lower rates than majority students (OECD 2024). These disadvantages are partly due to the lower average socioeconomic resources of immigrant families (e.g., Vonnahme 2021). In addition, migration-specific factors play a role, such as competence in the language of instruction (e.g., Bredtmann et al. 2021). Another possible cause, which has received increasing attention in recent years, is discrimination based on students' ethnic background (e.g., Shi and Zhu 2023; De Benedetto and De Paola 2023; Alesina et al. 2024). Because educational disadvantages may reduce long-term labor market opportunities and impede social participation, understanding the sources of immigrant-native achievement gaps is of paramount policy interest.

In this paper, we focus on one specific source of educational disadvantage: teacher bias towards ethnic majority students. In particular, we analyze whether teachers discriminate against minority students in terms of grading. In doing so, we build on a broader literature on the role of teacher biases. One strand of this literature has focused on analyzing gender stereotyping (e.g., Lavy 2008; Hinnerich et al. 2011; Terrier 2020; Lavy and Megalokonomou 2024). Another important, and recently growing, strand of the literature has focused on teacher bias with respect to students' race or ethnicity. Work investigating racial or ethnic discrimination in schools can be divided into two strands. Experimental studies typically find evidence of discrimination against minority students in exam or essay grading (e.g., Sprietsma 2013; Chowdhury et al. 2020).[1] Observational studies that compare objective measures of achievement, obtained through blindly graded standardized tests, with subjective teacher ratings find mixed results. While Botelho et al. (2015), De Benedetto and De Paola (2023), Rangel and Shi (2023) and Alesina et al. (2024) find evidence of a negative teacher bias towards ethnic minority students, Burgess and Greaves

[1] An exception is van Ewijk (2011), who finds no evidence that teachers grade minority and majority students differently for the same work.

(2013) for England and Shi and Zhu (2023) for the U.S. find evidence that black students are systematically under-assessed by the teacher relative to their white peers, while some ethnic groups are over-assessed. A recent study by Burn et al. (2024) for England further finds mixed evidence of discrimination towards ethnic minority students by subject. When being assessed by the teacher, ethnic minority students receive lower grades in English, but higher grades in Math compared to when grades are assigned through blindly graded exams. Zhu (2024) further highlights the role of measurement error in standard estimates of teacher bias in grading. Analyzing racial bias in teacher evaluations in the U.S. she shows that after correcting for measurement error in standardized test scores, teachers evaluate black students as higher achieving than white students with the same standardized test achievement.

We contribute to this literature by analyzing whether teachers discriminate against ethnic minority students[2] in German primary and secondary schools. Our analysis is based on representative surveys provided by the Institute for Quality Development in Education (IQB), which have been conducted nationwide in grades 4 and 9 since 2008. A key feature of the data is that they include information on both standardized, anonymously graded achievement tests and non-anonymous teacher assessments (i.e., school report grades). Assuming that both types of assessments measure similar skills and cognitive abilities, we compare the results of the two types of achievement measures in a difference-in-difference (DiD) framework. In doing so, we address the problems of (test-unspecific) unobserved heterogeneity between minority and majority students as well as measurement error in test scores that are inherent to standard estimates of teacher bias in grading.

We find that, on average, minority students receive lower grades than majority students in both German and Math. However, these differences are not due to discrimination in grading against minority students. Instead, our DiD results show that minority/majority achievement gaps in German and Math are reduced by 0.26 and 0.20 standard deviations (SD), respectively, when students are graded by the teacher compared to when they are

[2]As ethnic minority students, we define all students who were not born in Germany or who have at least one parent who was not born in Germany.

graded on the standardized test. This represents an improvement of about 5–6% for minority students.[3]

We further explore possible mechanisms for this finding. In particular, we aim to assess whether teachers in fact exhibit a positive evaluation bias towards minority students or whether our finding is merely attributable to the language difficulties faced by ethnic minority students, which may exert a stronger influence on their test performance than on teacher evaluations. Our results indicate that language difficulties can account for some of the observed positive grading bias. However, a positive bias persists even when the sample is restricted to students who were born in Germany and speak only German at home. Furthermore, we also find a positive bias for the subject English, which should not be affected by proficiency in German. We also provide supportive evidence that teachers adjust their assessment standards to account for students' backgrounds. In particular, we show that the positive grading bias is stronger among teachers who report that they demand significantly less from students with low capability. In addition, we find that such a positive teacher bias is not only limited to ethnic minority students, but also extends to (majority) students from low socioeconomic backgrounds. Overall, these findings suggest that teachers adjust their assessment standards to compensate for students' initial disadvantages.

The outline of the paper is as follows. In Section 2, we introduce the data used in the empirical analysis and present descriptive statistics. In Section 3, we describe the empirical framework. In Section 4, we discuss the results of our empirical analysis, explore potential mechanisms, and provide several heterogeneity analyses and robustness checks. Section 5 concludes.

[3]Our estimates of positive grading bias are slightly larger than those reported in Burn et al. (2024), who find that grades are 10–20% of a grade higher for ethnic minority students in Math, and Zhu (2024), who finds that teachers are 2–4% and 1–4% more likely to rate black students as proficient in Math and reading, respectively, than white students with the same test performance.

2 Data and Descriptive Statistics

2.1 Data

Our analysis is based on nationally representative, cross-sectional surveys of 4th- and 9th-grade students provided by the Research Data Center at the Institute for Educational Quality Improvement (FDZ at IQB).[4] The studies are entitled IQB National Assessment Study (*IQB Ländervergleich*) and, as of 2015, IQB Trends in Student Achievement (*IQB Bildungstrend*). Correspondingly, the purpose of the surveys is to measure whether students in the different federal states are meeting the nationwide education targets and whether adjustments are needed in certain states or subjects. Student assessments are mandatory for all selected public schools and partly for private schools. Despite the obligation to participate in the assessment tests, they can be considered "low stakes", meaning that they do not impact students' end-of-semester grades or education degrees. Neither teachers nor students or their parents were informed of individual test results.

The selection process of students started with sampling at the school level. Then, one to two classes per school were randomly selected. In total, we have six waves in the period from 2008 to 2018. The waves from 2011 and 2016 cover grade 4, with students aged 9 to 10 years. In these two waves, tests were conducted in the subjects German and Math. The waves from 2008/9, 2012, 2015 and 2018 cover grade 9, with students aged 14 to 15 years. In these four waves, tests were conducted in either the subjects German and English or the subjects Math and Science. The studies assess the educational performance of students at two important stages of their educational careers. Grade 4 is the last year of primary school in most federal states, after which students are tracked into different types of schools, either academically oriented or preparing for more practical vocational training. Primary schools give their students individual track recommendations, which are binding in some states and non-binding in others. The recommendations are based on the

[4]The data sets are accessible via Köller et al. (2011), Stanat et al. (2014), Pant et al. (2015), Stanat et al. (2018), Stanat et al. (2019) and Stanat et al. (2022) and are described in detail in Sachse et al. (2012), Richter et al. (2015), Lenski et al. (2016), Schipolowski et al. (2018), Schipolowski et al. (2019) and Becker et al. (2022).

school report of the first semester of grade 4. Grade 9 is another important stage, as it is the last year of compulsory lower secondary schooling in most federal states (with some states having longer compulsory schooling).

In addition to assessment tests (described in more detail below), the studies include questionnaires for students, their parents, teachers and principals that provide extensive information on the socioeconomic backgrounds of children and parents, as well as school and teacher characteristics. We use this information to differentiate between ethnic minority and majority students, to control for relevant characteristics in the estimations, to perform heterogeneity and robustness analyses, and to investigate the mechanisms behind our results. The questionnaires also contain information on the students' school report grades in the first semester.[5] End-of-semester grades are given by the teacher who taught the class in a particular subject and are based on both written exams and oral participation in class.[6] They are thus, at least to some extent, a subjective assessment of student performance, which leaves room for bias against certain groups and a potential for discrimination. Grades in Germany range from 1 (very good) to 6 (fail), with 4 (sufficient) being the minimum passing grade. In the empirical analysis, however, we rescale the grades so that higher grades reflect better performance (i.e., 6 is the best and 1 is the worst grade).

The standardized tests were designed, administered, and scored by the *International Association for the Evaluation of Educational Achievement (IEA)* which also conducts other large-scale student assessment studies such as PISA. The interviewers who administered the tests were therefore always external to the school. In addition, and most importantly for our analysis, no information about the students (e.g., name, gender) or the school was disclosed to the administrators who scored the tests. All students in a class took the test at the same time and then completed the survey questionnaire. The duration of the test varied according to the wave (and thus the age of the children), ranging from

[5]The surveys, and thus the standardized tests, were usually administered in the spring, while the end-of-semester grades were given in January, ensuring a short interval between the two types of performance measures.

[6]Written exams and oral participation are considered equally important for determining end-of-semester grades. However, the exact weighting of these components may vary by federal state or even school.

20 to 60 minutes per test block. The tests contained a mixture of closed, semi-open and open questions, and the test language was German (except for the subject English).[7] For German (and English), the tests covered the domains listening and reading. For Math, the tests covered five learning domains: numbers and operations; space and form; patterns and structures; quantities and measures; data, frequencies and probabilities. For 9th graders, the standardized tests are generally similar to the PISA tests. They are, however, more closely aligned with national education targets/curricula and are therefore more comparable to school report grades. Test scores are measured continuously with a mean of 500 and a standard deviation of 100 across the sample. They are calculated as predicted values by the data provider from the raw scores of the test items.[8]

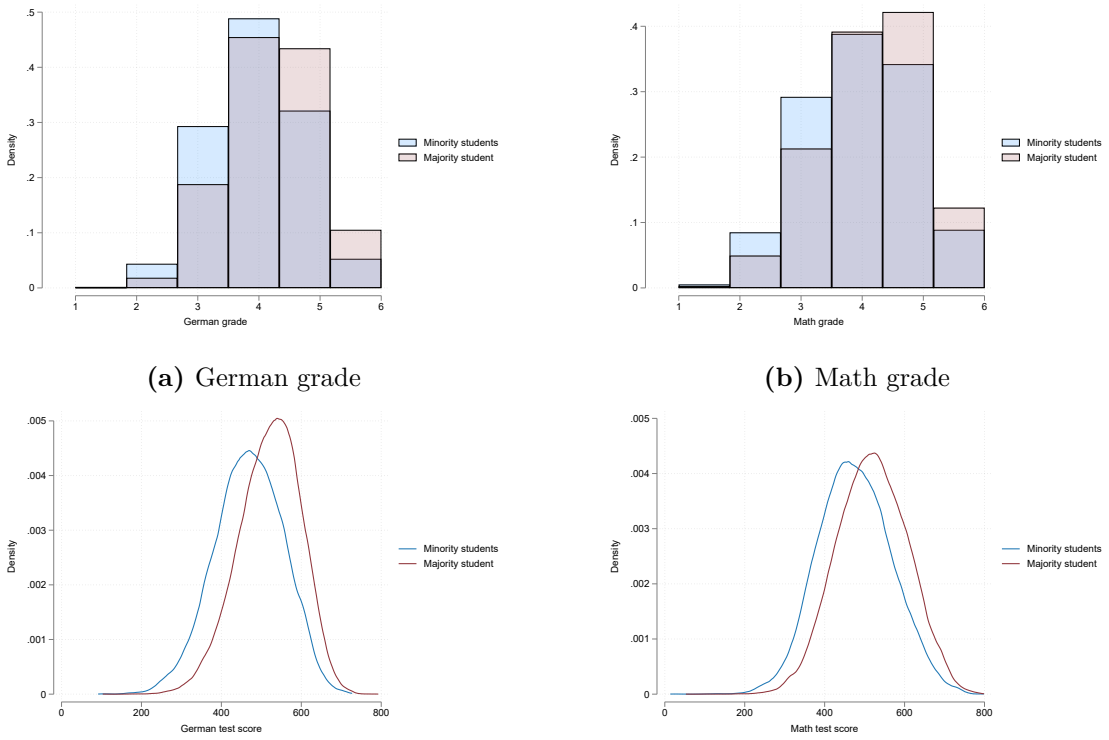
2.2 Sample and Descriptive Statistics

In the following, we present descriptive statistics for our analysis sample. Depending on the subject analyzed, the sample includes 92,937 (German) or 81,022 (Math) students. We use the full sample of students with valid information on school report grades and test scores, excluding only special needs schools (about 3% of the sample). Based on information on children's and parents' country of birth, we define minority students as students who were not born in Germany or who have at least one parent who was not born in Germany.

Figure 1 shows the distribution of grades and test scores in German and Math separately for minority and majority students. In both German and Math, minority students are more likely to receive lower grades and less likely to receive higher grades than majority students. However, a similar picture emerges for test scores. The distribution of test scores for majority students lies to the right of the corresponding distribution for minority students. For both grades and test scores, the distributional differences are more pronounced for German than for Math.

[7] All students took the test regardless of their proficiency in German. An exception was made only for immigrant students who had been in Germany for less than a year.

[8] The calculation of predicted values is based on probabilistic item response theory (IRT) and the resulting values are intended to measure the latent ability of students in each subject or learning domain.



(a) German grade **(b)** Math grade

(c) German test score **(d)** Math test score

Figure 1: Distribution of Grades and Test Scores

Appendix Tables A1 and A2 show descriptive sample statistics for students who took the test in German and Math, respectively. In both samples, around a quarter of the students are ethnic minority students. As can be seen, minority and majority students differ in their observable characteristics. In particular, minority students are more likely than majority students to have parents with a low level of education (ISCED 0-2) and less likely to have parents with a high level of education (ISCED 5-6). The parents of minority students are also less likely to be employed as white-collar workers. The observed differences in academic performance between minority and majority students may therefore be due to differences in socioeconomic characteristics between the two groups. Further statistics for minority students show that 19–20% are of Turkish origin, which is by far the largest country of origin of minority students in Germany, 20% are first-generation immigrants and 39% speak only German at home.

3 Empirical Framework

We start our analysis by regressing students' grades in German and Math on a set of observable characteristic as well as their performance in the standardized test in the respective subject:

$$Grade_{ij} = \alpha + \beta\, M_i + \boldsymbol{X}_i^{'}\, \eta + \theta\, Test\, Score_i + \kappa_j + \varepsilon_{ij}, \tag{1}$$

where $Grade_{ij}$ is the teacher-assigned grade of student i in class j in either German or Math, rescaled and standardized to have a mean of 0 and a standard deviation of 1. M_i is an indicator variable taking the value 1 for minority students, X_i is a set of control variables, including gender, age and its square, parents' education and occupation, and the number of books at home. $Test\, Score_i$ is the student's performance in the standardized test in the respective subject (also standardized to a mean of 0 and a standard deviation of 1). κ_j are class fixed effects that control for all factors varying across classes, such as teacher-specific assessment standards or differences in school quality. We cluster standard errors at the class level and use individual weights in all regressions.

The approach described in Eq. (1), i.e., regressing students' grades on their test scores and further observable characteristics, has two main shortcomings: First, the coefficient of interest, $\hat{\beta}$, may be biased due to unobserved heterogeneity. If minority and majority students differ in unobserved characteristics that are correlated with the grades given to them, such as motivation or language skills, $\hat{\beta}$ will be biased. Second, standardized test scores measure student performance with significant error (see, e.g., Kane and Staiger 2002; Sievertsen 2023). This is, for example, because student performance depends on several factors that are beyond the student's control, such as temperature and air quality, fatigue, and well-being at the time of assessment. In addition, measurement error comes from the test instrument itself. Because each test has a finite number of questions, there is randomness in the selection of questions and thus in the matching of questions to students' individual abilities. As shown by Zhu (2024), in a situation where minority students have lower standardized test scores than majority students, random measurement error in

test scores will bias downward the coefficient estimate for teacher evaluations of minority students ($\hat{\beta}$ in Eq. (1)), suggesting that teachers are more negative in their evaluations of minority students even in a setting without teacher bias.[9]

We address these issues by employing an alternative identification strategy. Following the initial work by Lavy (2008), we estimate a difference-in-difference (DiD) model based on the following equation:

$$Performance_{ib} = \varphi + \lambda\, M_i + \gamma\, NB_b + \delta\, (M_i \times NB_b) + \varepsilon_{ib}, \tag{2}$$

where $Performance_{ib}$ is the academic performance measure (grade or test score) of student i, where b refers to the "non-blind" (i.e., teacher grades) or "blind" (i.e., standardized test scores) performance measure. M_i is an indicator variable for whether the student is a minority student or not and NB_b is an indicator variable for whether the observation is from the "non-blind" or the "blind" performance measure.[10] The estimate for λ accounts for average differences between minority and majority students in the standardized test, while $\hat{\gamma}$ accounts for average differences in the two types of performance measures for majority students. $\hat{\delta}$ represents the treatment effect, i.e., the additional effect of being a minority student on the "non-blind" performance measure (i.e., teacher grades).[11]

The DiD nature of the estimation of Eq. (2) implies that any student-, teacher-, class-, or school-specific effects are implicitly accounted for with respect to the estimated coefficient of interest, $\hat{\delta}$, as long as they have the same effect on the blind and non-blind performance measure. Estimating a DiD model thus eliminates the biases arising from (test-unspecific) unobserved heterogeneity and from measurement error in test scores, which are inherent in the estimates obtained from Eq. (1).

[9]This is because measurement error in $Test\,Score_i$ will bias $\hat{\theta}$ towards zero. If M_i and $Test\,Score_i$ are correlated, then some of the true variation explained by $Test\,Score_i$ will be attributed to M_i instead. In the case where M_i and $Test\,Score_i$ are negatively correlated, $\hat{\beta}$ will be negatively biased, while it will be positively biased in the case of a positive correlation between the two variables (see Zhu 2024).

[10]Our DiD approach differs from standard DiD models in that we compare the outcomes of the treatment and the control group not over time, but over two types of performance measures.

[11]This coefficient estimate is equivalent to an estimate obtained from a first-difference model that regresses the difference between students' grades and their test scores on an indicator for being a minority student.

The underlying identification assumption of such a DiD model is that grades and standardized tests measure comparable competencies and, most importantly, that there are no test-specific systematic differences between minority and majority students. While this identification assumption cannot be tested, we perform several sensitivity and heterogeneity analyses to rule out alternative explanations for our results.

4 Results

4.1 Baseline Results

Table 1 shows the results for Eq. (1), i.e., from regressing students' grades on a set of observable characteristics.[12] The results show that minority students receive on average lower grades than majority students in both German and Math (columns 1 and 4). Even after controlling for class fixed effects and a number of socioeconomic characteristics, the grades given by teachers to minority students are 0.12 SD lower in German (column 2) and 0.05 SD lower in Math (column 5) than the grades given to majority students. However, once controlling for student performance in the standardized test (columns 3 and 6), the coefficient for minority status turns slightly positive. Conditional on test performance, ethnic minority students receive 0.03 and 0.05 SD higher grades in German and Math, respectively, than majority students with the same standardized test achievement in these subjects.

To net out any differences between minority and majority students that similarly affect performance on both types of achievement tests, we next estimate a DiD model as outlined in Eq. (2). Table 2 shows the results of our DiD framework for both German (column 1) and Math (column 3). As shown in columns 2 and 4, the DiD results are equivalent to estimations including individual fixed effects. The results in Table 2 support the evidence from the descriptive analysis (Section 2.2) that minority students, on average, perform worse than majority students on the standardized test in both German and Math. It further reveals that majority students receive significantly lower teacher grades compared

[12] Full estimation results are shown in Appendix Table A3.

Table 1: Estimated Association between Minority Status and Students' Grades

	German			**Math**		
	(1)	(2)	(3)	(4)	(5)	(6)
	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE
Minority student	−0.357†	−0.123†	0.032***	−0.261†	−0.050†	0.049†
	(0.013)	(0.012)	(0.010)	(0.014)	(0.013)	(0.010)
Test score	–	–	0.650†	–	–	0.787†
			(0.006)			(0.006)
Constant	0.093†	5.927†	2.548†	0.063†	5.759†	1.405†
	(0.021)	(0.343)	(0.281)	(0.015)	(0.372)	(0.269)
Controls	no	yes	yes	no	yes	yes
Wave FE	yes	no	no	yes	no	no
Class FE	no	yes	yes	no	yes	yes
Observations	92,937	92,937	92,937	81,022	81,022	81,022
Clusters	4,985	4,985	4,985	6,221	6,221	6,221
Adjusted R^2	0.026	0.263	0.476	0.014	0.184	0.545

Notes: Standard errors are clustered at the class level. Individual weights are applied. Significance level: † 0.1%, *** 1%, ** 5%, * 10%.

to their standardized test performance. However, the differences are very small (0.07 and 0.06 SD in German and Math, respectively), which reassures us that standardized test scores and teacher grades measure comparable competencies.

Table 2: DiD and Fixed Effects Results – German and Math

	German		**Math**	
	(1)	(2)	(3)	(4)
	DiD	**FE**	**DiD**	**FE**
	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE
Minority student	−0.615†	–	−0.462†	–
	(0.017)		(0.015)	
Non-blind performance measure	−0.074†	−0.074†	−0.059†	−0.059†
	(0.012)	(0.012)	(0.010)	(0.010)
Minority × Non-blind	0.259†	0.259†	0.203†	0.203†
	(0.016)	(0.016)	(0.013)	(0.013)
Constant	0.175†	0.000	0.135†	−0.000
	(0.012)	(0.006)	(0.010)	(0.005)
Student FE	no	yes	no	yes
Observations	185,874	185,874	162,044	162,044
Clusters	4,985	4,985	6,221	6,221
Adjusted R^2	0.051	0.015	0.029	0.011

Notes: Standard errors are clustered at the class level. Individual weights are applied. Significance level: † 0.1%, *** 1%, ** 5%, * 10%.

Most importantly, however, the coefficients of the interaction term indicate that minority/majority gaps in test performance are reduced by 0.26 (German) and 0.20 (Math) SD when students are evaluated by their teachers. This represents an improvement of

about 0.2 grade points or 5–6% for minority students. The finding of higher teacher ratings of ethnic minority students compared to their test scores contradicts much of the previous literature, which tends to find negative biases against ethnic minority students (e.g., Botelho et al. 2015; De Benedetto and De Paola 2023; Sahlströhm and Silliman 2024). However, our results are consistent with Zhu (2024), who shows that – after correcting for measurement error in standardized test scores – teachers evaluate black students as higher achieving than white students with the same standardized test achievement.

4.2 Mechanisms

Our baseline results show that minority students are rated more positively by their teachers in relation to their standardized test scores than are majority students. These results may suggest that teachers favor minority students by consciously or unconsciously "inflating" their grades on non-anonymous assessments. Such a positive teacher bias could be explained by teachers adjusting their assessment standards to account for students' backgrounds. As discussed in Zhu (2024), this could be the case because teachers believe that a student from a disadvantaged background who reaches the same level of academic achievement as a student from a more privileged background is demonstrating greater achievement, potentially leading them to adjust their assessments accordingly. In addition, teacher bias may be reflected in lower expectations for ethnic minority students compared to majority students, which may be due to negative biases or stereotypes.[13] Furthermore, results could be influenced by social desirability bias, where teachers may exaggerate their assessments of ethnic minority students to match what they perceive to be socially acceptable responses.

An alternative explanation to bias in teacher ratings could be disadvantages faced by ethnic minority students with respect to standardized test assessments. If minority students have more difficulty excelling on the standardized test for reasons other than differences in the skills being tested, then our finding of relatively higher grades for minority

[13]Evidence that teachers have lower expectations for ethnic minority students is provided, for example, by Gentrup et al. (2020), Papageorge et al. (2020) and Carlana et al. (2022).

students could occur even in the absence of teacher bias.

It is beyond the scope and ability of this paper to determine the relative importance of the various channels that might explain why minority students receive higher teacher grades than majority students with similar standardized test performance. The aim of the following analyses, however, is to assess whether there is evidence that teachers have a positive evaluation bias towards ethnic minority students, or whether the disadvantages faced by minority students on standardized tests are the only explanation for our findings.

One of the main reasons why students from immigrant families may be disadvantaged on the standardized test is language difficulties. While proficiency in the German language should be important for both teacher evaluations and standardized test performance, German proficiency may be less important in normal classroom interactions, where students may be more confident in asking comprehension questions during exams run by their own teachers and where oral participation is assessed as well. Thus, if proficiency in German is more important for students' performance on the standardized test than it is for teachers' ratings, then our finding of relatively higher grades for minority students could occur even in the absence of teacher bias.

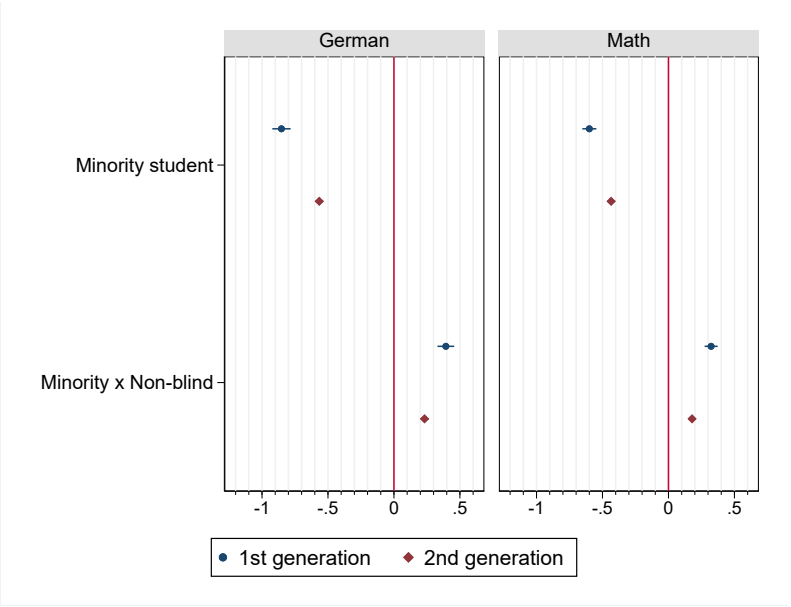


Figure 2: DiD Results – 1st vs. 2nd Generation Immigrants

To test whether language proficiency or lack thereof explains our result, we conduct two types of heterogeneity analyses. First, we examine whether grading bias differs for first-

and second-generation immigrants. As shown in Figure 2, both the negative achievement gap on the standardized test and the improvement in performance when graded by the teacher are higher for first-generation immigrants than for second-generation immigrants, suggesting that language proficiency is important for the estimated achievement gaps. However, even for second-generation immigrants who were born and raised in Germany, there is still a positive grading bias.

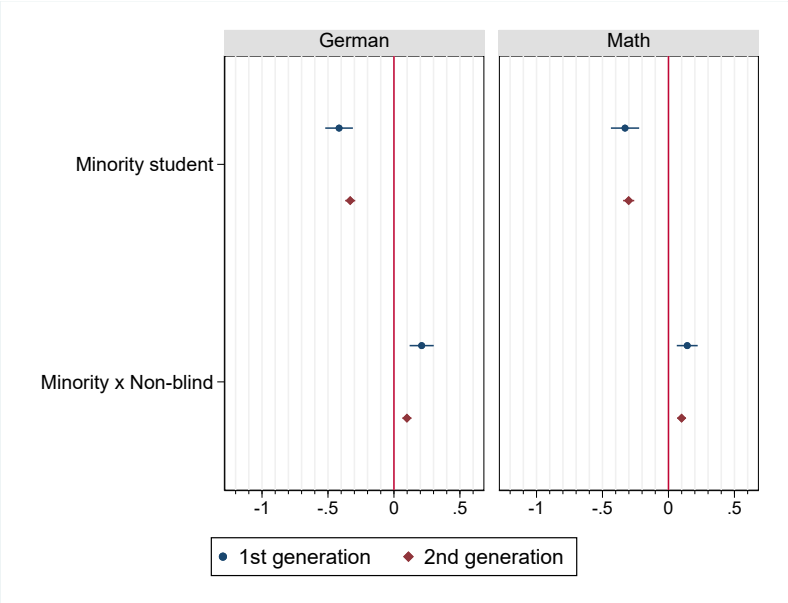


Figure 3: DiD Results – Students Who Only Speak German at Home

Second, we use information on the language spoken at home and estimate our DiD model only for those students who speak only German at home.[14] While both the achievement gap and the grading bias become smaller when restricting the sample to students who only speak German at home, there is still a positive grading bias for both first- and second-generation immigrant students (see Figure 3).

As an additional check on the relevance of proficiency in the test language, we estimate our model for an alternate subject: English. In the 2008 and 2015 waves of the survey, both of which were administered to 9th graders, students also took a standardized test in English. Since the test language for this test is English, proficiency in German should have no effect on students' performance on the test. As can be seen in Figure 4, the

[14]Information on the language spoken at home is obtained from a question that asks students how often they speak German at home. We restrict the sample to children who report that they always speak German at home, while dropping children who never or only sometimes speak German at home.

performance gap between minority and majority students on the standardized test is, as expected, much smaller in English than in German. However, even for English, we see that minority students are graded more favorably by the teacher than majority students relative to their test performance.

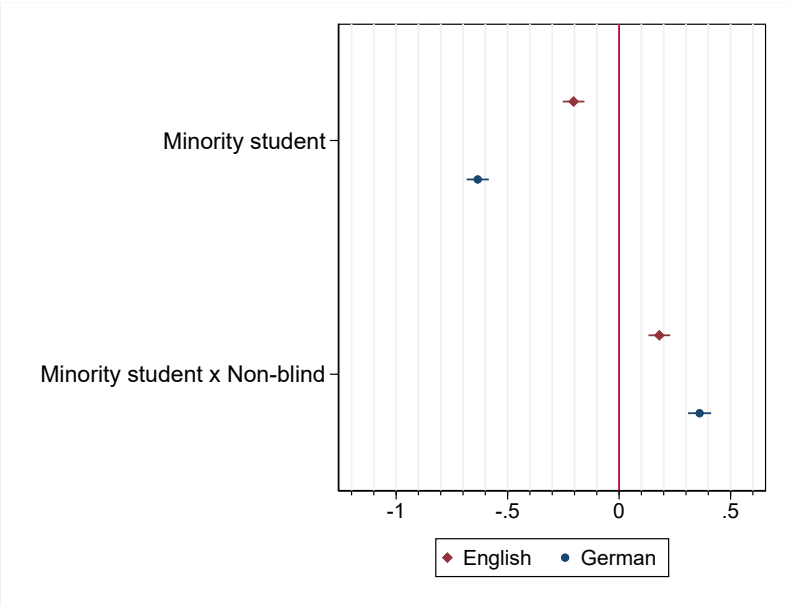


Figure 4: DiD Results – Effect of Minority Status on English vs. German

In sum, the results of the above analyses show that language proficiency plays a role in both the overall minority/majority achievement gap and the grading bias. However, language proficiency does not fully explain the positive grading bias towards minority students.

To further pinpoint mechanisms for our effects, we next analyze whether the positive grading bias exists for students of low socioeconomic status (SES). If teachers adjust their grading standards to account for students' backgrounds, then we would expect them to do so not only in terms of minority status, but also in terms of socioeconomic background.[15] To disentangle the role of minority status and socioeconomic background, we restrict the sample to majority students and estimate a DiD model in which low SES students serve as the treatment group, where low SES is defined as both parents having a low level of education. The results of such regressions are shown in Figure 5. As expected, low SES

[15]Previous evidence has, for example, shown that teachers hold lower expectations of low SES students (Timmermans et al. 2015; Doyle et al. 2023).

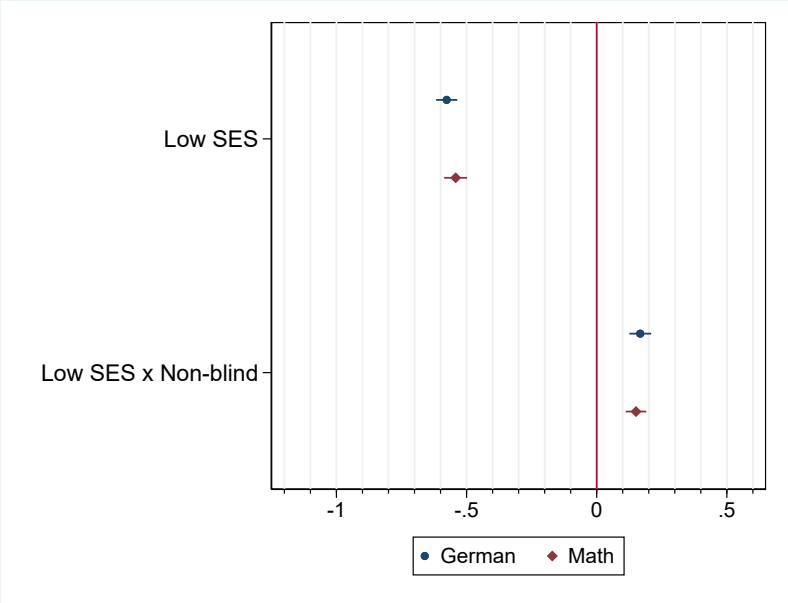


Figure 5: DiD Results – Effect of Low SES for Majority Students

students score lower on the standardized test than higher SES students. However, this gap is reduced by 0.17 (German) and 0.15 (Math) SD when students are graded by the teacher. This represents an improvement of around 0.15 grade points or 5%. Thus, a positive grading bias exists not only for minority students, but also for low SES majority students.[16] This finding suggests that teachers are more favorable to disadvantaged students when grading non-anonymous assessments.

Lastly, we take advantage of the fact that the survey collected information not only on students and their parents, but also on teachers. Unfortunately, most of the questions asked to teachers vary from wave to wave.[17] In waves 2015, 2016 and 2018 of the survey, teachers were asked about the strategies they use to deal with achievement differences among students in their class. In particular, they were asked whether they demand significantly less from students with low capability. Figure 6 shows the results of our DiD model separately for teachers who agree and teachers who disagree with this statement.[18]

[16]We also examine whether the positive grading bias for minority students is driven only by low SES students. As can be seen in Appendix Figure A1, the positive grading bias for minority students is also present in the sample of low SES students and thus is not driven by the fact that minority students are more likely to come from low socioeconomic backgrounds.

[17]In addition, it was only possible to assign teachers uniquely for 79% of the German and 81% of the Math students. We present further heterogeneity analyses by teacher characteristics in Section 4.3.

[18]We define teachers as agreeing with this statement if they say that they sometimes or often demand less from low-capability students and as disagreeing if they say that they rarely or never demand less from low-capability students. Overall, 44% of the German teachers and 29% of the Math teachers agree

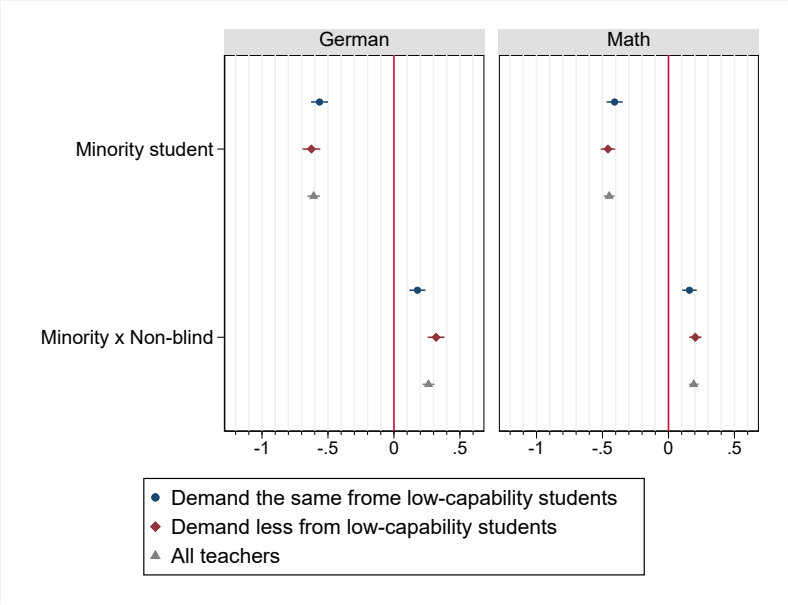


Figure 6: DiD Results – By Whether Teacher Demands Less from Low-Capability Students

For German, we find that the positive grading bias is significantly higher for teachers who demand less from low-capability students than for those who do not report doing so (0.32 vs. 0.18 SD). For Math, we also find a difference in the grading bias between teachers who demand less from low-capacity students and those who do not, which is, however, much smaller (0.2 vs. 0.16 SD) and not statistically significant.[19] While the question used in this heterogeneity analysis is certainly not optimal for testing the hypothesis of teacher grading bias, because it does not explicitly refer to ethnic minority students and because it does not capture differential treatment of students that is unconscious to the teacher, the results of this analysis are still interesting. They show that a non-negligible fraction of teachers report taking students' backgrounds into account, and that this behavior can explain part of the positive grading bias, especially in German.

While the results presented so far do support the hypothesis that at least part of the positive grading bias towards minority students is due to teacher bias, there are two remaining possible channels that we cannot fully rule out based on our empirical analysis. First, the main difference between the standardized test and teacher grades

with this statement.

[19]That German teachers are more likely than Math teachers to say that they demand less from low-capability students, and that the positive grading bias is particularly high for this group of teachers, can be explained by the fact that lack of German proficiency may be one of the main reasons why teachers perceive students as low-achieving.

is that the latter also depend on oral participation in class. Thus, in a scenario where ethnic minority students systematically outperform majority students in oral participation, we may overestimate the positive grading bias based on our DiD model. However, both anecdotal evidence, obtained from discussions with teachers in different types of schools, and analyses of oral participation by student background (e.g., Kiss 2013; Veiga et al. 2021) do not support this hypothesis.[20]

Second, there may be a difference in students' behavior by ethnic background in low- vs. high-stakes tests. Previous evidence has shown that students exert less effort in low-stakes situations, such as standardized tests like PISA, than in high-stakes situations, where performance matters for educational outcomes (e.g., Gneezy et al. 2019; Akyol et al. 2021). If such behavioral differences in low- vs. high-stakes environments vary by student ethnic background, then this could be reflected in our DiD estimate. While we are unable to test whether ethnic minorities behave differently in low versus high-stakes tests, the existing evidence on test-taking tends to suggest that ethnic minority students put relatively more effort into low-stakes tests than majority students (Schlosser et al. 2019). Thus, if differences in effort on the two types of assessments by ethnic background matter in our setting, they would most likely lead us to underestimate a positive assessment bias for ethnic minority students.

4.3 Further Heterogeneity Analyses and Robustness Checks

In this section, we provide further heterogeneity analyses and robustness checks. First, we analyze whether the positive grading bias varies across different ethnic groups. Specifically, we distinguish between two ethnic groups: students from Turkey and students from the EU.[21] Students of Turkish origin represent the largest group of students with migration background in Germany (Federal Statistical Office 2024). They have also been shown

[20] In particular, the study by Kiss (2013) of primary and secondary school children in Germany shows that students from immigrant families are, if anything, less likely to participate frequently in Math classes.

[21] Information on students' (or their parents') origin country is partly summarized to larger groups of countries due to data protection reasons. We are therefore only able to identify a few single origin countries consistently over time. In cases were both the mother and the father are immigrants and the two were born in different origin countries, we assign students the mother's origin country.

to achieve significantly lower than majority students or students of other immigrant backgrounds, and are more likely to come from low socioeconomic backgrounds (Song 2011). We choose students from EU countries (with Poland and Italy being the largest groups) as the comparison group, because they are more similar to students of German origin, both in terms of their cultural and linguistic background and their educational achievement. This is also evident from Figure 7, which shows much larger performance gaps in the standardized test for students of Turkish origin than for students originating from an EU country. As the figure also shows, there is a positive grading bias relative to majority students for both groups, which is, however, much larger for students of Turkish origin. This finding supports our hypothesis that teachers in German schools adjust their assessment standards to account for students' background, by being more favorable to students from disadvantaged backgrounds with similar levels of performance. It is also consistent with previous evidence showing that teachers have lower expectations of students from Turkish origin (Tobisch and Dresel 2017).

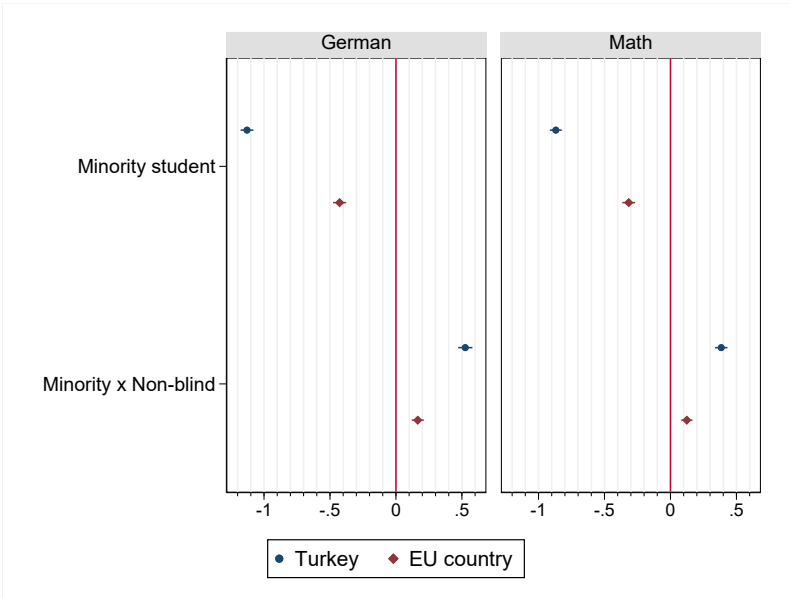


Figure 7: DiD Results – By Students' Country of Origin

Second, we analyze whether the positive grading bias varies with the share of minority students in the class. Specifically, we estimate our DiD model separately for classes with an above-median and classes with a below-median share of minority students.[22] The

[22]We restrict the sample to classes with at least one minority student, which is 88% of the German

results show that the positive grading bias is present only in classes with an above-median share of minority students, while there is no such bias in classes with a below-median share of minority students (see Figure 8). While the share of minority students in the class is correlated with other class or school characteristics and thus reflects socioeconomic differences more generally, this finding shows that it is predominantly teachers in classes with a high share of disadvantaged students who exhibit a positive bias towards ethnic minority students.[23]

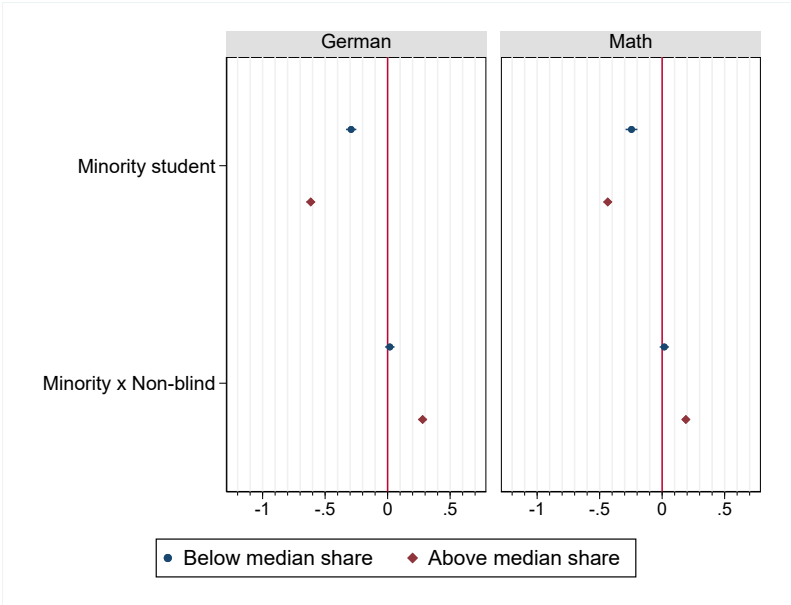


Figure 8: DiD Results – By Share of Minority Students in Class

Third, we analyze whether the grading bias varies along the distribution of grades. As such an analysis is difficult in our DiD model[24], we turn to the basic approach outlined in Eq. (1) and run separate regressions, in which the outcome variable takes the value 1 if the student's grade is above a certain threshold and 0 if it is at or below the threshold. Figure 9 shows results from these regressions for the minority status coefficient. While one has to keep in mind that the estimation coefficients are likely biased downward due

classes and 89% of the Math classes in our sample. The median share of minority students in these classes is 0.23 and 0.24, respectively. Analyses that include all classes yield similar results.

[23]We find similar results if we split the sample according to the average test performance in the class. While there is no grading bias for classes with above-median test performance, we find a strong positive grading bias in classes with a below-median class performance (see Appendix Figure A2).

[24]Estimating separate regressions for different grade thresholds is not a meaningful option as this will lead to inconsistent estimates. Due to the discrete nature of the grades variable, estimating a quantile regression is also not an option.

to unobserved heterogeneity and measurement error in test scores, the pattern shows that the positive grading bias is stronger among the higher (i.e., better) grades. Taken together with the heterogeneous effects by class composition, these findings reveal that teachers appear to be more favorable to high-achieving minority students in classes with high proportions of disadvantaged students. Teachers thus tend to reward those minority students who positively differentiate themselves from their lower-achieving peers. We find, in contrast, no evidence that the grading bias is due to teachers pushing minority students above the pass/fail threshold.

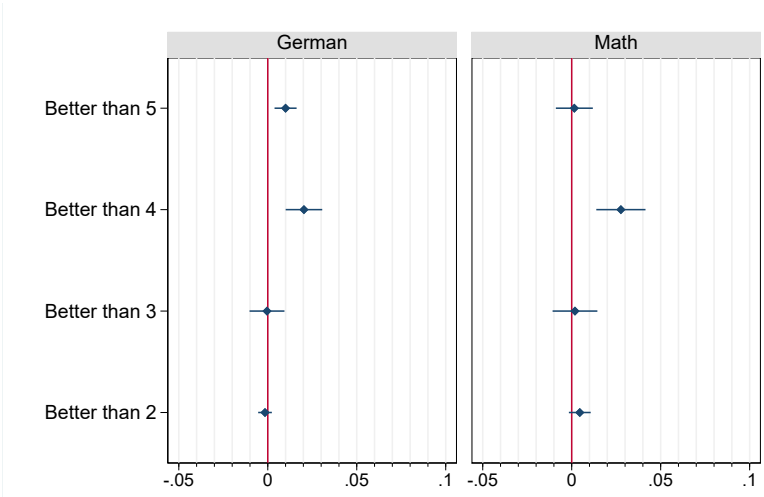


Figure 9: Estimated Association between Minority Status and Passing a Grade Threshold

We also conduct a number of other heterogeneity analyses and robustness checks. In waves 2012 and 2018, students were also administered a standardized test in Science. The estimated effects for Science are comparable to those for Math and are reported in Appendix Figure A3. We also test for gender differences in the positive grading bias towards minority students, for which we find no evidence (see Appendix Figure A4). However, our results show that the grading bias is larger for 9th grade students than for 4th grade students (see Appendix Figure A5).

As shown in Appendix Figures A6 to A10, we also find little difference in grading bias by teacher characteristics. The positive grading bias appears to be more pronounced among teachers with fewer years of experience and tenure, while the observed differences between these groups are only statistically significant for the German sample. However, we observe

no significant variation in grading bias with respect to teachers' gender, age, or migration status. The latter finding is particularly interesting, as one might expect that teachers who themselves belong to a minority group would be more likely to consider student background in their evaluations. Conversely, as shown, e.g., by Ouazad (2014), Gershenson et al. (2016) and Papageorge et al. (2020), teachers may hold higher expectations of students with an identical ethnic background.[25] Thus, it is theoretically unclear whether we would expect the positive bias to be higher or lower for immigrant teachers.

To ensure that our results are not driven by outliers, we check the robustness of our results by excluding the 5% highest and lowest test scores from the sample. The estimates obtained from these regressions are similar to our baseline estimates (see Appendix Figure A11). Lastly, we examine whether our findings are the consequence of relative grading schemes. If teachers, in general, tend to evaluate student performance relative to the average performance of the class, a positive grading bias may occur due to the fact that minority students are more likely to be in low-performing classes. To rule out that such a grading behavior is responsible for the observed effect, we express individual test scores as deviations from the class mean (instead of the cohort mean) and use this alternative test performance measure in our DiD regression. As is evident from Appendix Figure A12, employing class-performance adjusted test scores as outcomes yields estimates that are similar to our baseline estimates.

5 Conclusion

In this paper, we analyze whether teachers discriminate against ethnic minority students in terms of grading. Our analysis is based on representative survey data on students in German primary and secondary schools that have been collected nationwide since 2008. A key feature of the data is that they include information on students' scores on both

[25] Unfortunately, we are unable to match students and teachers of the same ethnicity because we only know whether the teacher was born in Germany or not, but have no information on country of birth (if not Germany) or whether the teacher's parents were born in Germany. Moreover, the proportion of foreign-born teachers in German schools is very small, accounting for only about 2% of the teachers in our sample.

standardized, anonymously graded achievement tests and non-anonymous teacher grades. We compare the results of the two types of achievement measures within a difference-in-difference framework, thereby addressing the problems of (test-unspecific) unobserved heterogeneity between minority and majority students as well as measurement error in test scores.

Although minority students receive, on average, lower grades than majority students, we find no evidence of discrimination in grading against minority students. Instead, our results show that the achievement gap between minority and majority students in German and Math is reduced by 0.26 and 0.20 standard deviations (or 5–6%), respectively, when students are assessed by teachers compared to when they are assessed by the standardized test. This positive grading bias is partly due to the fact that minority students face greater challenges in standardized tests due to language barriers. However, we also provide evidence that teachers adjust their assessment standards to take account of students' backgrounds. In particular, we show that the positive grading bias is stronger among teachers who report that they demand significantly less from students with low capability, and that such a positive teacher bias is not only limited to ethnic minority students, but also extends to (majority) students from low socioeconomic backgrounds.

Our results further reveal that the positive grading bias is particularly strong for students of Turkish origin, who achieve significantly lower than students of other immigrant backgrounds and are more likely to come from low socioeconomic backgrounds. In addition, we show that it is predominantly teachers in classes with a high share of disadvantaged students who exhibit a positive teacher bias towards ethnic minority students. However, we do not find evidence that the grading bias is due to teachers pushing minority students above the pass/fail threshold. Rather, our results suggest that the positive grading bias is stronger among the higher grades. Overall, our findings suggest that teachers adjust their assessment standards to compensate for initial disadvantages, and that such behavior is directed primarily at high-performing minority students in classes with high proportions of disadvantaged students.

It remains unclear, however, whether such a compensatory bias in grading ultimately

serves to reduce or reinforce initial inequalities in educational attainment. From one perspective, a grading bias in favor of minority students can be seen as advantageous for them, as better grades can be decisive for admission to university or when applying for apprenticeships. Consequently, it may contribute towards narrowing existing educational inequalities between minority and majority students. From another perspective, a positive grading bias can be detrimental to students if it is a consequence of lower teacher expectations, which can create self-fulfilling prophecies (Rosenthal and Jacobson 1986). Given that teacher expectations can influence students' academic achievement and attainment (Lavy and Sand 2018; Papageorge et al. 2020; Hill and Jones 2021; Lavy and Megalokonomou 2024), differential assessments for minority students have the potential to negatively impact their performance. This can lead to under-investment in human capital for minority students, which in turn perpetuates longstanding achievement gaps. To avoid such negative consequences and promote unbiased grading, teachers should be sensitized to such biases and their potential consequences, as this can help reduce bias in grading.

References

Akyol, Pelin, Kala Krishna, and Jinwen Wang. 2021. "Taking PISA Seriously: How Accurate are Low-Stakes Exams?" *Journal of Labor Research*, 42: 184–243.

Alesina, Alberto, Michela Carlana, Eliana La Ferrara, and Paolo Pinotti. 2024. "Revealing stereotypes: Evidence from immigrants in schools." *American Economic Review*, 114(7): 1916–1948.

Becker, Benjamin, Johanna Busse, Marlen Holtmann, Sebastian Weirich, Stefan Schipolowski, Nicole Mahler, and Petra Stanat. 2022. *IQB-Bildungstrend 2018.* Berlin: Institut zur Qualitätsentwicklung im Bildungswesen.

Botelho, Fernando, Ricardo A. Madeira, and Marcos A. Rangel. 2015. "Racial Discrimination in Grading: Evidence from Brazil." *American Economic Journal: Applied Economics*, 7(4): 37–52.

Bredtmann, Julia, Sebastian Otten, and Christina Vonnahme. 2021. "Linguistic diversity in the classroom, student achievement, and social integration." *Education Economics*, 29(2): 121–142.

Burgess, Simon, and Ellen Greaves. 2013. "Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities." *Journal of Labor Economics*, 31(3): 535–576.

Burn, Hester, Laura Fumagalli, and Birgitta Rabe. 2024. "Stereotyping and ethnicity gaps in teacher assigned grades." *Labour Economics*, 89: 102577.

Carlana, Michela, Eliana La Ferrara, and Paolo Pinotti. 2022. "Implicit Stereotypes in Teachers' Track Recommendations." *AEA Papers and Proceedings*, 112: 409–414.

Chowdhury, Shyamal, Ilya Klauzner, and Robert Slonim. 2020. "What's in a Name? Does Racial or Gender Discrimination in Marking Exist?" *IZA Discussion Paper No. 13890.*

De Benedetto, Marco A., and Maria De Paola. 2023. "Immigration and teacher bias towards students with an immigrant background." *Economic Policy*, 38(113): 107–154.

Doyle, Lewis, Matthew J. Easterbrook, and Peter R. Harris. 2023. "Roles of socioeconomic status, ethnicity and teacher beliefs in academic grading." *British Journal of Educational Psychology*, 93(1): 91–112.

Federal Statistical Office. 2024. *Statistischer Bericht. Mikrozensus – Bevölkerung mit Migrationshintergrund. Erstergebnisse 2023.* Wiesbaden: Statistisches Bundesamt.

Gentrup, Sarah, Georg Lorenz, Cornelia Kristen, and Irena Kogan. 2020. "Self-fulfilling prophecies in the classroom: Teacher expectations, teacher feedback and student achievement." *Learning and Instruction*, 66: 101296.

Gershenson, Seth, Stephen B. Holt, and Nicholas W. Papageorge. 2016. "Who believes in me? The effect of student–teacher demographic match on teacher expectations." *Economics of Education Review*, 52: 209–224.

Gneezy, Uri, John A. List, Jeffrey A. Livingston, Xiangdong Qin, Sally Sadoff, and Yang Xu. 2019. "Measuring Success in Education: The Role of Effort on the Test Itself." *American Economic Review: Insights*, 1(3): 291–308.

Hill, Andrew J., and Daniel B. Jones. 2021. "Self-Fulfilling Prophecies in the Classroom." *Journal of Human Capital*, 15(3): 400–431.

Hinnerich, Björn Tyrefors, Erik Höglin, and Magnus Johannesson. 2011. "Are boys discriminated in Swedish high schools?" *Economics of Education Review*, 30(4): 682–690.

Kane, Thomas J., and Douglas O. Staiger. 2002. "The Promise and Pitfalls of Using Imprecise Scholl Accountability Measures." *Journal of Economic Perspectives*, 16(4): 91–114.

Kiss, David. 2013. "Are immigrants and girls graded worse? Results of a matching approach." *Education Economics*, 21(5): 447–463.

Köller, Olaf, Michel Knigge, and Bernd Tesch. 2011. "IQB Ländervergleich Sprachen 2008/2009 [IQB National Assessment Study 2008/2009] (IQB-LV 2008-9) (Version 2) [Data set]." Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_LV_2008_v2.

Lavy, Victor. 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.

Lavy, Victor, and Edith Sand. 2018. "On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases." *Journal of Public Economics*, 167: 263–279.

Lavy, Victor, and Rigissa Megalokonomou. 2024. "The Short- and the Long-Run Impact of Gender-Biased Teachers." *American Economic Journal: Applied Economics*, 16(2): 176–218.

Lenski, Anna E., Martin Hecht, Christiane Penk, Felix Milles, Manuel Mezger, Patricia Heitmann, Petra Stanat, and Hans A. Pant. 2016. *IQB-Ländervergleich 2012.* Berlin: Institut zur Qualitätsentwicklung im Bildungswesen.

OECD. 2024. *Education at a Glance 2024: OECD Indicators.* Paris: OECD Publishing.

Ouazad, Amine. 2014. "Assessed by a Teacher Like Me: Race and Teacher Assessments." *Education Finance and Policy*, 9(3): 334–372.

Pant, Hans A., Petra Stanat, Martin Hecht, Patricia Heitmann, Malte Jansen, Anna E. Lenski, Christiane Penk, Claudia Pöhlmann, Alexander Roppelt, Ulrich Schroeders, and Thilo Siegle. 2015. "IQB-Ländervergleich Mathematik und Naturwissenschaften 2012 [IQB National Assessment Study 2012] (IQB-LV 2012) (Version 4) [Data set]." Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_LV_2012_v4.

Papageorge, Nicholas W., Seth Gershenson, and Kyung M. Kang. 2020. "Teacher Expectations Matter." *The Review of Economics and Statistics*, 102(2): 234–251.

Rangel, Marcos A., and Ying Shi. 2023. "First Impressions Matter: Evidence From Elementary-School Teachers." *Journal of Human Resources (forthcoming).*

Richter, Dirk, Katrin Böhme, Jana Bastian-Wurzel, Hans A. Pant, and Petra Stanat. 2015. *IQB-Ländervergleich 2011.* Berlin: Institut zur Qualitätsentwicklung im Bildungswesen.

Rosenthal, Robert, and Lenore Jacobson. 1986. "Pygmalion in the classroom." *The Urban Review*, 3(1): 16–20.

Sachse, Karoline A, Julia Kretschmann, Aleksander Kocaj, Olaf Köller, Michel Knigge, and Bernd Tesch. 2012. *IQB-Ländervergleich 2008/2009.* Berlin: Institut zur Qualitätsentwicklung im Bildungswesen.

Sahlströhm, Ellen, and Mikko Silliman. 2024. "The Extent and Consequences of Teacher Biases against Immigrants." *IZA Discussion Paper No. 16899.*

Schipolowski, Stefan, Johanna Busse, Camilla Rjosk, Nicole Mahler, Benjamin Becker, and Petra Stanat. 2019. *IQB-Bildungstrend 2016.* Berlin: Institut zur Qualitätsentwicklung im Bildungswesen.

Schipolowski, Stefan, Nicole Haag, Felix Milles, Stefanie Pietz, and Petra Stanat. 2018. *IQB-Bildungstrend 2015.* Berlin: Institut zur Qualitätsentwicklung im Bildungswesen.

Schlosser, Analia, Zvika Neeman, and Yigal Attali. 2019. "Differential Performance in High Versus Low Stakes Tests: Evidence from the Gre Test." *The Economic Journal*, 129(623): 2916–2948.

Shi, Ying, and Maria Zhu. 2023. ""Model minorities" in the classroom? Positive evaluation bias towards Asian students and its consequences." *Journal of Public Economics*, 220: 104838.

Sievertsen, Hans Henrik. 2023. "Assessments in Education." In *Oxford Research Encyclopedia of Economics and Finance*, Oxford: Oxford University Press.

Song, Steve. 2011. "Second-generation Turkish youth in Europe: Explaining the academic disadvantage in Austria, Germany, and Switzerland." *Economics of Education Review*, 30(5): 938–949.

Sprietsma, Maresa. 2013. "Discrimination in grading: Experimental evidence from primary school teachers." *Empirical Economics*, 45: 523–538.

Stanat, Petra, Hans A. Pant, Katrin Böhme, Dirk Richter, Sebastian Weirich, Nicole Haag, Alexander Roppelt, Maria Engelbert, and Heino Reimers. 2014. "IQB Ländervergleich Primarstufe 2011 [IQB National Assessment Study 2011] (IQB-LV 2011) (Version 3) [Data set]." Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_LV_2011_v3.

Stanat, Petra, Katrin Böhme, Stefan Schipolowski, Nicole Haag, Sebastian Weirich, Karoline Sachse, Lars Hoffmann, and Felicitas Federlein. 2018. "IQB-Bildungstrend Sprachen 2015 [IQB Trends in Student Achievement 2015] (IQB-BT 2015) (Version 5) [Data set]." Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_BT_2015_v5.

Stanat, Petra, Stefan Schipolowski, Nicole Mahler, Sebastian Weirich, Sofie Henschel, Marlen Holtmann, Benjamin Becker, and Jenny Kölm. 2022. "IQB-Bildungstrend Mathematik und Naturwissenschaften 2018 [IQB Trends in Student Achievement 2018] (IQB-BT 2018)(Version 1) [Data set]." Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_BT_2018_v1.

Stanat, Petra, Stefan Schipolowski, Sebastian Weirich, Nicole Mahler, and Julia Wittig. 2019. "IQB-Bildungstrend Primarstufe 2016 [IQB Trends in Student Achievement 2016] (IQB-BT 2016) (Version 1) [Data set]." Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_BT_2016_v1.

Terrier, Camille. 2020. "Boys lag behind: How teachers' gender biases affect student achievement." *Economics of Education Review*, 77: 101981.

Timmermans, Anneke C., Hans Kuyper, and Greetje van der Werf. 2015. "Accurate, inaccurate, or biased teacher expectations: Do Dutch teachers differ in their expectations at the end of primary education?" *British Journal of Educational Psychology*, 85(4): 459–478.

Tobisch, Anita, and Markus Dresel. 2017. "Negatively or positively biased? Dependencies of teachers' judgments and expectations based on students' ethnic and social backgrounds." *Social Psychology of Education*, 20: 731–752.

van Ewijk, Reyn. 2011. "Same work, lower grade? Student ethnicity and teachers' subjective assessments." *Economics of Education Review*, 30(5): 1045–1058.

Veiga, Feliciano H., Isabel Festas, Óscar F. García, Íris M. Oliveira, Carlota M. Veiga, Conceição Martins, Filomena Covas, and Nuno A. Carvalho. 2021. "Do students with immigrant and native parents perceive themselves as equally engaged in school during adolescence?" *Current Psychology*, 42: 11902–11916.

Vonnahme, Christina. 2021. "Do migrant-native achievement gaps narrow? Evidence over the school career." *Ruhr Economic Papers #932*.

Zhu, Maria. 2024. "New Findings on Racial Bias in Teachers' Evaluations of Student Achievement." *IZA Discussion Paper No. 16185*.

Appendix

Tables

Table A1: Descriptive Statistics – Analysis Sample German

	Minority students		**Majority students**	
	Mean	StdD	Mean	StdD
Female	0.505	0.500	0.498	0.500
Age	13.150	2.608	13.062	2.518
4th grade student	0.466	0.499	0.467	0.499
Mother's education				
Low education	0.236	0.424	0.163	0.369
Medium education	0.244	0.429	0.366	0.482
High education	0.248	0.432	0.356	0.479
Missing	0.273	0.446	0.115	0.319
Father's education				
Low education	0.206	0.405	0.138	0.345
Medium education	0.217	0.412	0.315	0.464
High education	0.266	0.442	0.394	0.489
Missing	0.312	0.463	0.153	0.360
Mother's occupation				
White collar	0.369	0.483	0.642	0.479
Blue collar	0.153	0.360	0.094	0.292
Other	0.199	0.399	0.118	0.323
Missing	0.278	0.448	0.146	0.353
Father's occupation				
White collar	0.297	0.457	0.500	0.500
Blue collar	0.218	0.413	0.154	0.361
Other	0.196	0.397	0.171	0.377
Missing	0.289	0.453	0.175	0.380
Number of books at home	117.151	135.519	178.665	154.210
Low SES family	0.230	0.421	0.122	0.328
Region of origin				
Turkey	0.199	0.399	–	–
EU country	0.204	0.403	–	–
Other country	0.597	0.491	–	–
First-generation immigrant	0.195	0.396	–	–
Speaks only German at home	0.389	0.488	–	–
Observations	23,790		69,147	

Table A2: Descriptive Statistics – Analysis Sample Math

	Minority students		**Majority students**	
	Mean	StdD	Mean	StdD
Female	0.502	0.500	0.500	0.500
Age	12.917	2.646	12.764	2.550
4th grade student	0.520	0.500	0.535	0.499
Mother's education				
Low education	0.281	0.449	0.169	0.375
Medium education	0.228	0.420	0.358	0.479
High education	0.255	0.436	0.353	0.478
Missing	0.236	0.425	0.119	0.324
Father's education				
Low education	0.250	0.433	0.145	0.352
Medium education	0.199	0.399	0.292	0.455
High education	0.266	0.442	0.390	0.488
Missing	0.284	0.451	0.172	0.378
Mother's occupation				
White collar	0.373	0.484	0.651	0.477
Blue collar	0.150	0.357	0.084	0.278
Other	0.194	0.396	0.118	0.323
Missing	0.283	0.450	0.147	0.354
Father's occupation				
White collar	0.305	0.460	0.514	0.500
Blue collar	0.209	0.407	0.135	0.342
Other	0.191	0.393	0.169	0.374
Missing	0.295	0.456	0.182	0.386
Number of books at home	112.199	134.064	169.973	152.531
Low SES family	0.241	0.427	0.117	0.322
Region of origin				
Turkey	0.191	0.393	–	–
EU country	0.223	0.416	–	–
Other country	0.587	0.492	–	–
First-generation immigrant	0.197	0.398	–	–
Speaks only German at home	0.391	0.488	–	–
Observations	21,157		59,865	

Table A3: Estimated Association between Minority Status and Students' Grades – Full Results

	German	**Math**
	Coef/StdE	Coef/StdE
Minority student	0.032***	0.049†
	(0.010)	(0.010)
Test score	0.650†	0.787†
	(0.006)	(0.006)
Female	0.305†	0.109†
	(0.009)	(0.008)
Age	−0.333†	−0.173†
	(0.041)	(0.040)
Age^2	0.009†	0.004***
	(0.001)	(0.001)
Mother's education (Ref.: Low education)		
Medium education	0.057†	0.045†
	(0.014)	(0.013)
High education	0.052†	0.030**
	(0.014)	(0.013)
Missing	0.059***	0.009
	(0.019)	(0.020)
Father's education (Ref.: Low education)		
Medium education	0.007	0.007
	(0.015)	(0.014)
High education	0.072†	0.058†
	(0.015)	(0.013)
Missing	−0.023	−0.018
	(0.019)	(0.018)
Mother's occupation (Ref.: White collar)		
Blue collar	−0.045***	−0.038***
	(0.014)	(0.014)
Other	−0.032***	−0.000
	(0.012)	(0.012)
Missing	−0.028*	0.008
	(0.015)	(0.016)
Father's occupation (Ref.: White collar)		
Blue collar	−0.043†	−0.010
	(0.012)	(0.012)
Other	−0.007	0.006
	(0.010)	(0.011)
Missing	−0.110†	−0.074†
	(0.015)	(0.016)
Number of books at home (in 100)	0.011†	−0.004
	(0.003)	(0.003)
Constant	2.548†	1.405†
	(0.281)	(0.269)
Class FE	yes	yes
Observations	92,937	81,022
Clusters	4,985	6,221
Adjusted R^2	0.476	0.545

Notes: Standard errors are clustered at the class level. Significance level: † 0.1%, *** 1%, ** 5%, * 10%.

Figures

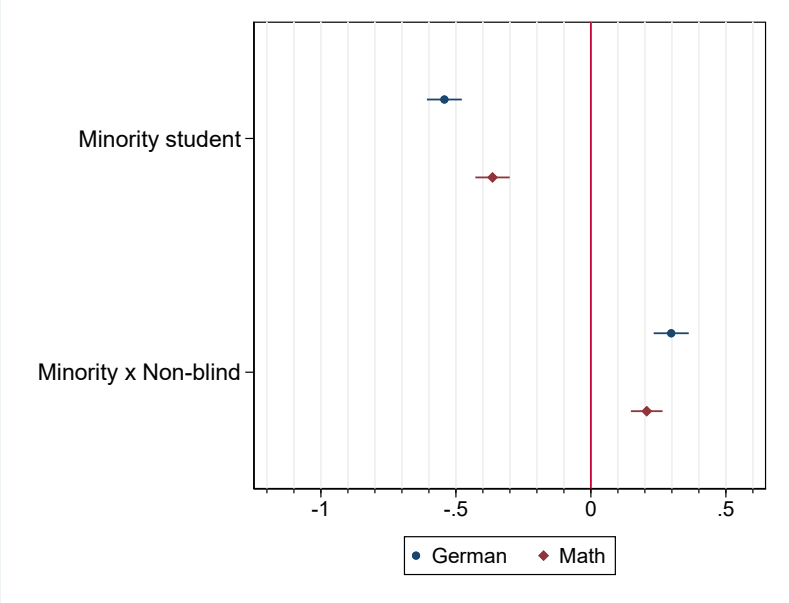


Figure A1: DiD Results – Only Low SES Students

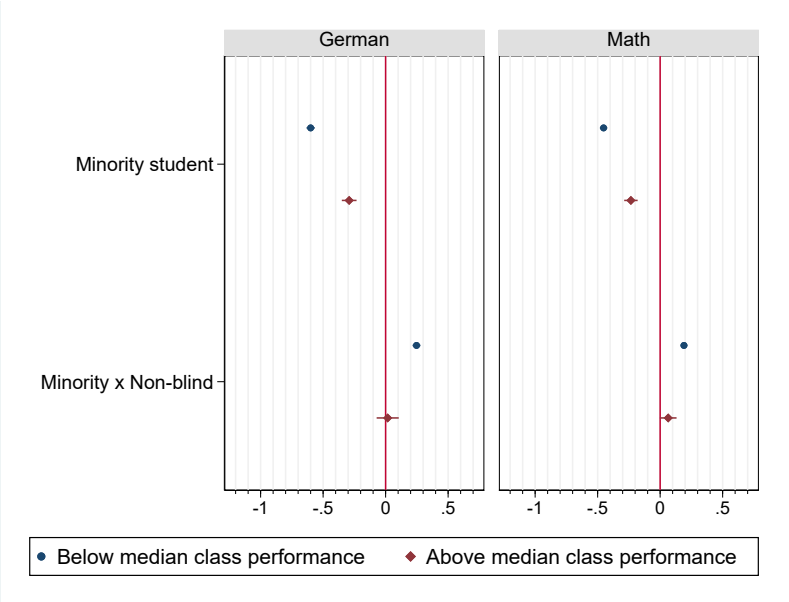


Figure A2: DiD Results – By Average Test Performance in Class

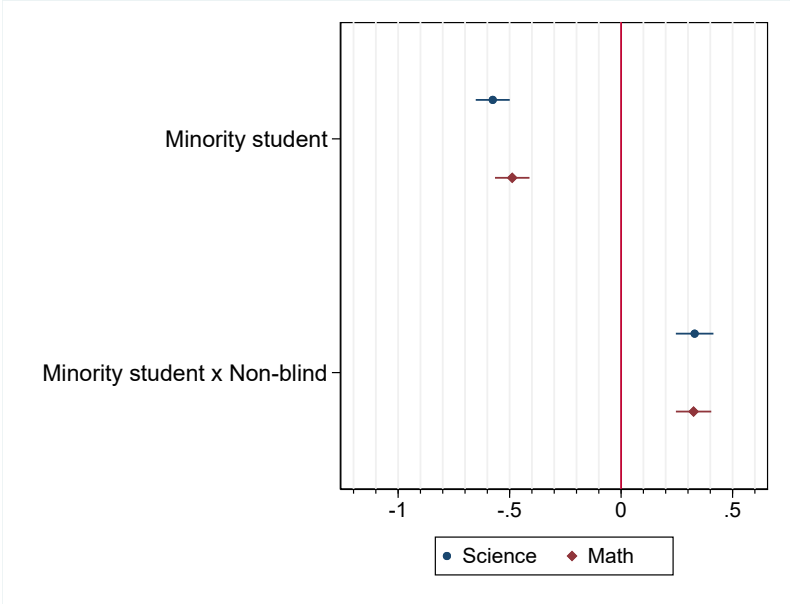


Figure A3: DiD Results – Science and Math

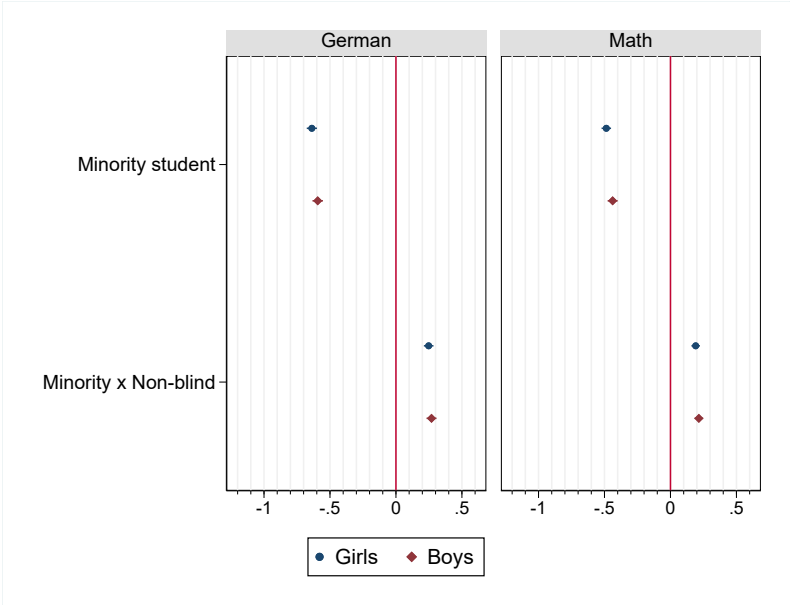


Figure A4: DiD Results – By Gender

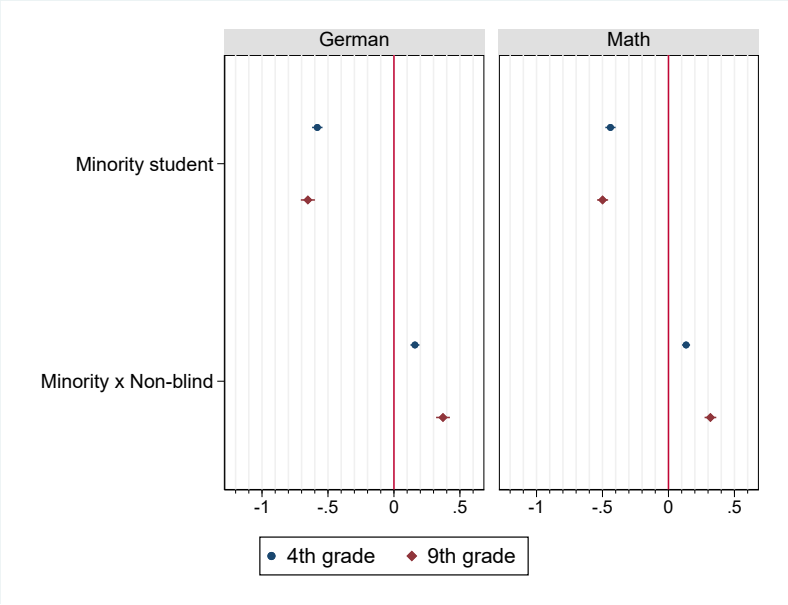


Figure A5: DiD Results – By School Grade

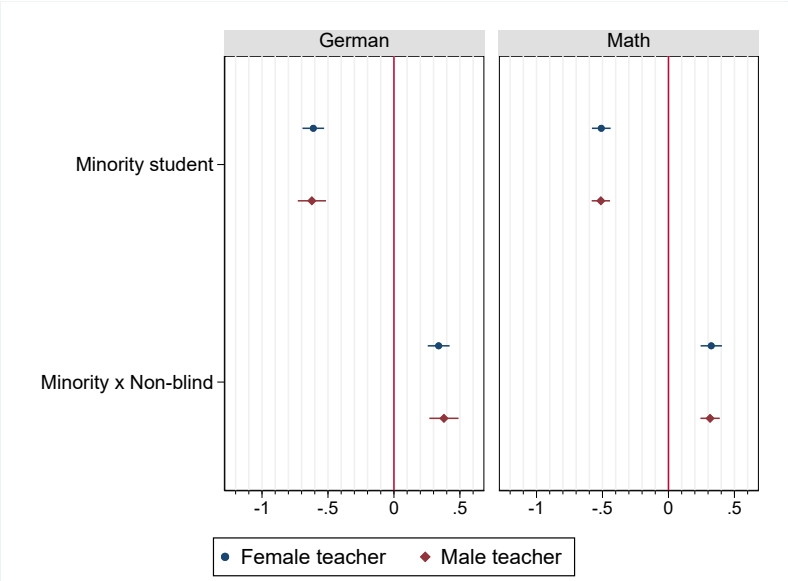


Figure A6: DiD Results – By Teacher's Gender (Only 9th Grade Students)

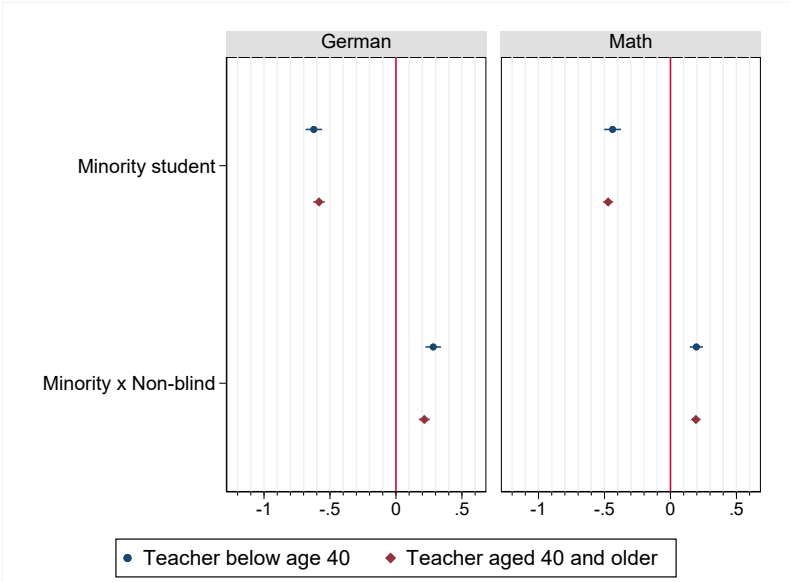


Figure A7: DiD Results – By Teacher's Age

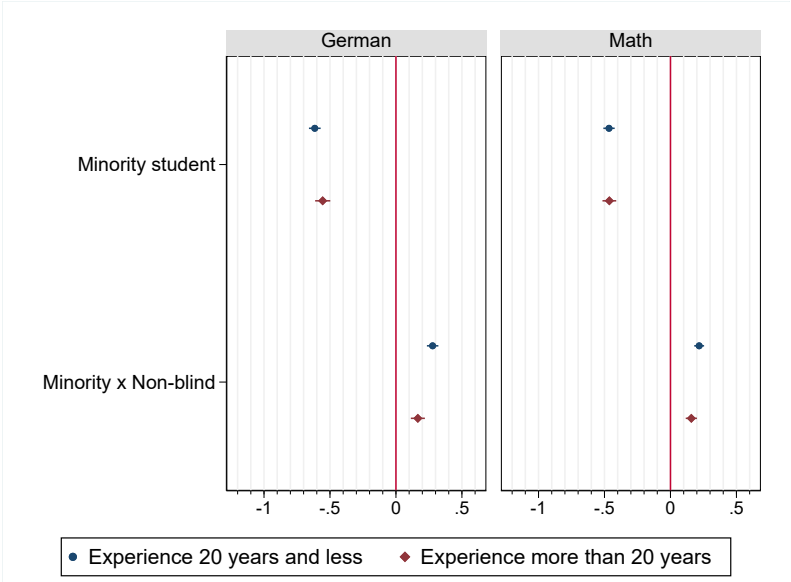


Figure A8: DiD Results – By Teacher's Experience

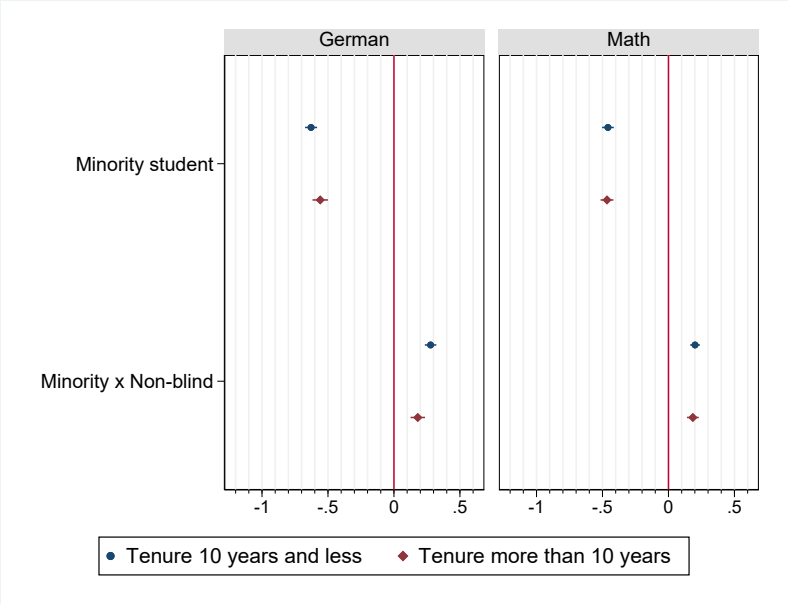


Figure A9: DiD Results – By Teacher's Tenure

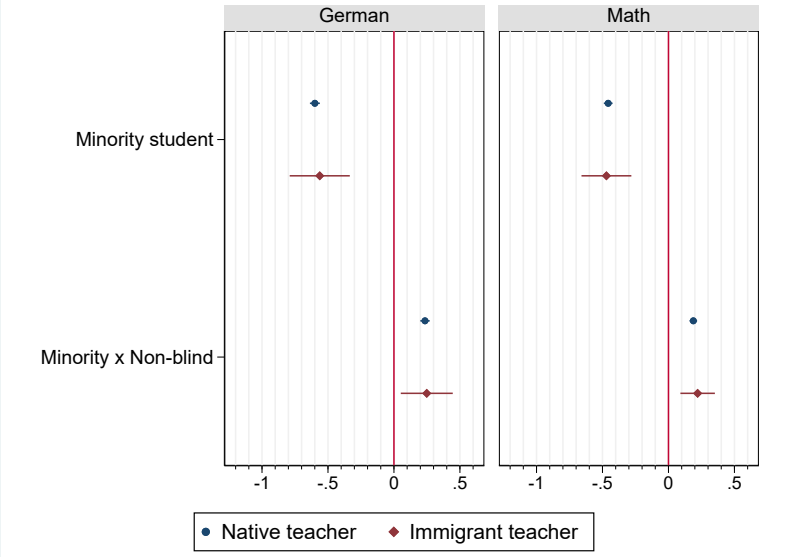


Figure A10: DiD Results – By Teacher's Origin

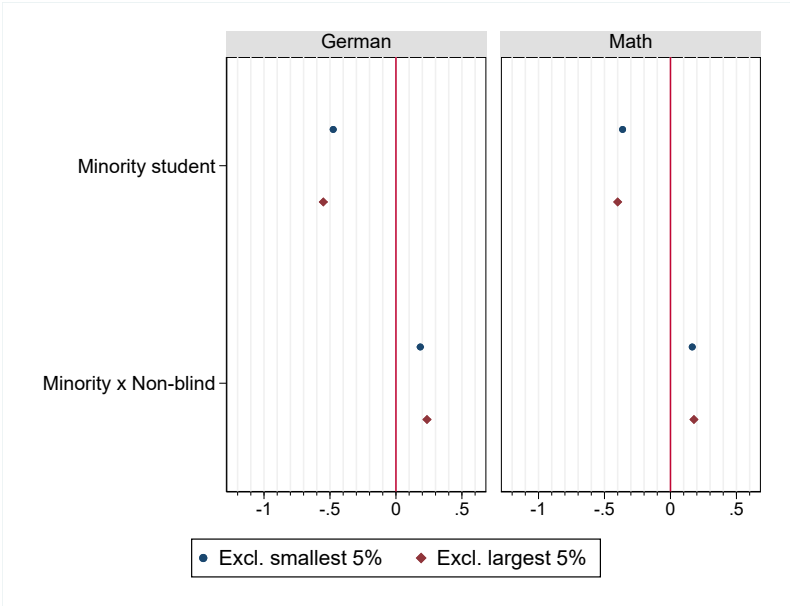


Figure A11: DiD Results – Excluding the 5% Smallest and 5% Largest Test Scores

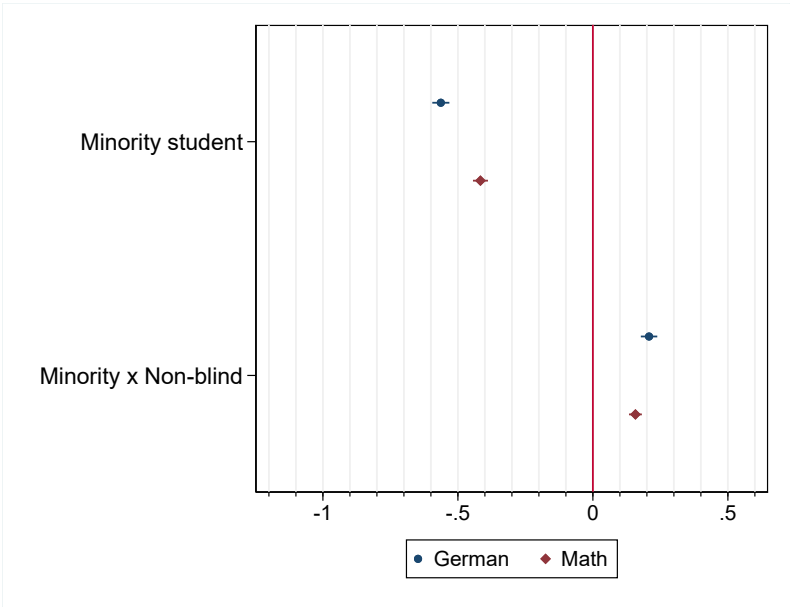


Figure A12: DiD Results – Test Scores as Deviation from the Class Mean