Discussion Paper Series

IZA DP No. 18847

July 2026

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

Julia Bredtmann
RWI, University of Duisburg-Essen,
IZA@LISER and RFBerlin

Sebastian Otten
University of Duisburg-Essen,
RFBerlin and RWI

Christina Vonnahme
Westphalian University of Applied
Sciences, RWI and IZA@LISER

IZA@LISER NETWORK
ISSN: 2365-9793

IZA DP No. 18847 July 2026

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

Abstract

We analyze whether teachers discriminate against ethnic minority students with regard to grading. Using comprehensive data on students in German primary and secondary schools, we compare students' scores on standardized, anonymously graded achievement tests with non-anonymous teacher ratings within a difference-in-differences framework. Although minority students tend to score lower on the standardized test than majority students, this achievement gap is significantly reduced when graded by the teacher. This finding cannot be fully explained by differences in socioeconomic background, language-related differences in standardized test performance, or the sorting of students into educational contexts with different grading standards. Overall, our findings suggest that teachers have a positive evaluation bias toward ethnic minority students.

JEL classification

F22, I24, J15

Keywords

immigration, discrimination, grading bias

Corresponding author

Julia Bredtmann

julia.bredtmann@rwi-essen.de

* 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 Interdisciplinary Centre for Educational Research at the University of Duisburg-Essen, 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, the European Association for Labour Economists Conference 2024, the 6th SOLE/EALE/AASLE World Labor Conference, the Annual Conference of the German Economic Association 2025, and the 40th Conference of the Italian Association of Labour Economics for valuable comments and suggestions.

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; Karna et al. 2025). 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., De Benedetto and De Paola 2023; Shi and Zhu 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 that has far-reaching consequences: teacher bias toward ethnic minority students. In particular, we analyze whether teachers discriminate against minority students in terms of grading. Understanding teacher bias in grading is of key importance given the role that teacher evaluations play in several educational settings, including recommending students for different academic tracks, communicating to students what is expected of them, and ultimately determining admission to university.

In analyzing teacher bias in grading toward ethnic minority students, we build on a broader literature on the role of teacher bias. One strand of this literature has focused on analyzing gender stereotyping (e.g., Lavy 2008; Carlana 2019; Jansson and Tyrefors 2022; Lavy and Megalokonomou 2024). Another important, and recently expanding, 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., Hanna and Linden 2012; 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. Studies that regress teacher-assigned grades on an indicator for being a minority student, standardized test scores, and other observable characteristics, usually document lower teacher ratings for ethnic minority students compared to majority students with the same standardized test performance (e.g., Botelho et al. 2015; De Benedetto and De Paola 2023; Alesina et al. 2024; Sahlström and Silliman 2024; Rangel and Shi 2026). Using a similar approach, both Burgess and Greaves (2013) for England and Shi and Zhu (2023) for the U.S. find evidence that Black students are systematically under-assessed by teachers relative to their White peers, while some ethnic groups are over-assessed. Burn et al. (2024) exploit a change in assessment methods in England to compare grades based on teacher predictions to grades received through blindly marked examinations within a first difference framework. Based on this approach, they find mixed evidence of discrimination toward ethnic minority students by subject. When assessed by the teacher, ethnic minority students receive lower grades in English but higher grades in Math relative to White British students and compared to when grades are assigned blindly. Zhu (2026) further highlights the role of measurement error in standard estimates of teacher bias in grading. Analyzing racial disparities 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 performance.

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 survey data collected nationwide in grades 4 and 9 since 2008. A key feature of the data is that it includes information on both standardized, anonymously graded achievement tests and non-anonymous teacher assessments (i.e., school report

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

[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.

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-differences (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, we find no evidence that these differences are due to discrimination in grading against minority students. Instead, our results show that minority students receive higher grades than majority students relative to their standardized test performance. In particular, our DiD results reveal 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 graded on the standardized test. Based on the average grades in these subjects, this represents an improvement of about 0.2 grade points or 5–6% for minority students.[3]

Next, we explore possible mechanisms for our finding. Our results reveal that about half of the positive grading differential is due to minority students being more likely to end up in classes or schools where teachers apply more lenient grading standards. However, a significant grading differential remains even after accounting for student sorting. We also provide supportive evidence that teachers adjust their assessment standards based on students' backgrounds. In particular, we show that the grading differential is stronger among teachers who report that they demand significantly less from students with low capability. In addition, we find that such a positive grading differential is not only limited to ethnic minority students, but also extends to (majority) students from low socioeconomic backgrounds. We perform several robustness checks to rule out the possibility that our results are driven by distributional differences between grades and test scores, particularly

[3] Although direct comparisons of effect sizes across studies are difficult due to variations in outcome variables and research designs, our estimates appear slightly larger in relative terms, yet still within the same range as previous studies documenting positive grading differentials between minority and majority students. Burn et al. (2024) find that ethnic minority students receive grades that are 2–4% higher in Math than those received by White British students. Zhu (2026) finds that teachers are 3–4% and 1–4% more likely to rate Black students as proficient in Math and reading, respectively, than White students with the same standardized test performance.

ceiling effects.

Our findings reveal a crucial aspect of teacher grading behavior: teachers may not be neutral to students' characteristics when assessing their scholastic skills. Rather, our results suggest that teachers may adjust their grading standards to compensate for students' initial disadvantages. This has important implications for educational outcomes and equity. On the one hand, such adjustments can be seen as beneficial because they help to reduce existing inequalities. Given that grades are a key determinant of educational trajectories—affecting track recommendations, admission to higher education, and future career opportunities—bias in favor of disadvantaged students may help to level the playing field. By compensating for structural disadvantages, teachers can provide ethnic minority students and those from lower socioeconomic backgrounds with enhanced access to educational opportunities they might otherwise have difficulty attaining.

However, the presence of positive grading bias also raises concerns. As emphasized by Ferman and Fontes (2022), biased grades can send misleading signals about students' true abilities, thereby introducing efficiency costs by misaligning grades with actual skills. Achievement grades are a primary form of feedback that teachers provide to students, shaping their perceptions of their own abilities and guiding their learning strategies. When grades are distorted by teacher bias, students and their parents may misinterpret their true academic standing, leading to suboptimal decisions about investments in human capital. Such distortions align with principal-agent models of strategic information transmission, wherein biased information can induce agents—in this case, students and their parents—to adopt inefficient learning strategies and effort levels.

These divergent impacts underscore the fundamental equity-efficiency trade-off central to public economics, where efforts to promote fairness frequently conflict with accurate resource allocation (Okun 1975). Adjusting grades to reduce disparities among disadvantaged students enhances equity but potentially compromises efficiency by diminishing the informational content of grades as indicators of true student ability. Understanding this trade-off is vital for comprehensively evaluating the broader implications of teacher grading bias.

In addition, positive bias in grading may be symptomatic of lower teacher expectations for certain groups of students, which can have long-term negative consequences. Research has shown that teacher expectations are a critical input into the educational process and can vary based on student characteristics such as race and gender (Lavy and Sand 2018; Papageorge et al. 2020; Hill and Jones 2021; Lavy and Megalokonomou 2024). Lower expectations may lead teachers to set less demanding academic standards, which in turn may limit student motivation and performance. This is consistent with the concept of self-fulfilling prophecies (Rosenthal and Jacobson 1968), whereby students internalize teachers' expectations and adjust their behavior accordingly. If teachers consistently hold lower expectations for minority students, minority students may be less motivated to develop the skills necessary for long-term academic success. This may have lasting consequences, as teacher expectations have been found to influence not only immediate academic achievement, but also post-secondary outcomes (Papageorge et al. 2020).

The remainder of the paper is structured 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.

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 titled IQB National Assessment Study (*IQB Ländervergleich*) and, as of 2015, IQB Trends in Student Achievement (*IQB*

[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).

Bildungstrend). Accordingly, 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/or 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 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 data also contains 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] Although teachers have no particular incentive to be anything other than accurate when grading students[7], teacher-assigned grades are, at least to some extent, subjective assessments of student performance, which leave 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. However, to facilitate interpretation in the empirical analysis, we rescale the grades so that higher grades reflect better performance (i.e., 6 is the best grade and 1 is the worst).

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 administrators of 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 40 to 60 minutes per subject. The tests contained a mixture of closed, semi-open and open questions, and the test language was German (except for the subject English).[8] 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,

[5]End-of-semester grades were given in January, while the surveys, and thus the standardized tests, were usually administered the following spring term, ensuring a short interval between the two types of performance measures. Information on students' end-of-semester grades is provided by the schools, and is therefore unaffected by recall or reporting bias on the part of students or teachers.

[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.

[7]Since students do not choose their teachers, teachers have little reason to cultivate a reputation for leniency. Moreover, schools do not appear to face strong external incentives to inflate average grades, as information on students' grades is not publicly reported. If anything, the institutional setting suggests that teachers are expected to assign grades they view as professionally defensible and internally consistent.

[8]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.

the standardized tests are generally similar to the PISA tests. They are, however, 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.[9]

2.2 Sample and Descriptive Statistics

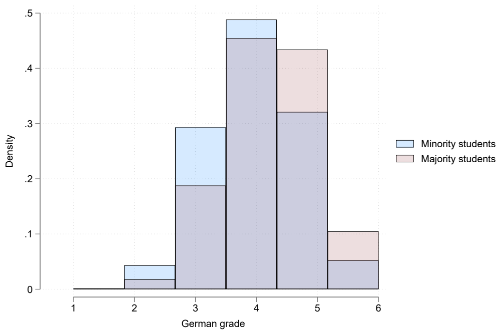
In the following, we present descriptive statistics for our analysis sample. Depending on the subject analyzed, the sample includes 92,937 (German) or 81,022 (Math) students. We use the full sample of students with valid information on school report grades and test scores, excluding only special needs 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.[10]

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 are 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.

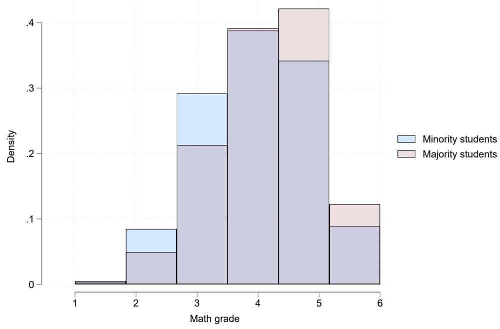
Appendix Tables A1 and A2 show descriptive sample statistics for students who took the test in German and Math, respectively. In both samples, around a quarter of the students are ethnic minority students. As can be seen, minority and majority students differ in their observable characteristics. In particular, minority students are more likely

[9]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.

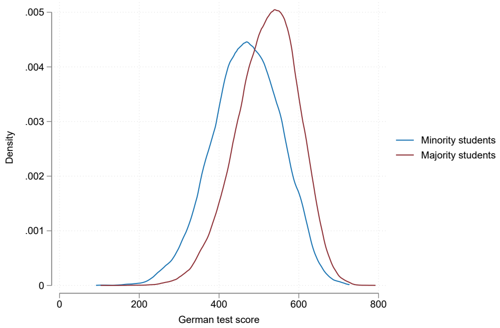
[10]Information on students' (or their parents') origin country is partly summarized to larger groups of countries due to data protection reasons. We are therefore only able to identify a few single origin countries consistently over time. In cases where both the mother and the father are immigrants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



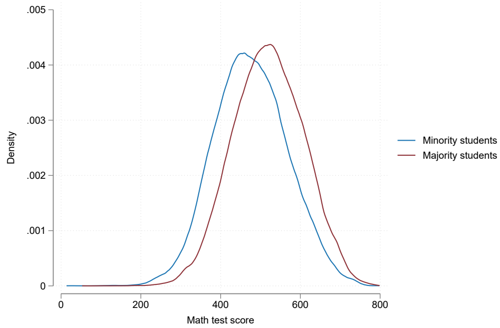
(a) German grade



(b) Math grade



(c) German test score



(d) Math test score

Figure 1: Distribution of Grades and Test Scores

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, 17% are first-generation immigrants and 42% 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 characteristics as well as their performance in the standardized test in the respective subject:

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

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 and its square. $Test\, Score_i$ 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).[11] 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 student weights in all regressions.[12]

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 (2026), 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

[11] Both grades and test scores are standardized separately by subject and survey wave.

[12] The student weight indicates how many students in the population each individual student in the sample represents. Factors at the school, class, and individual levels are included in the calculation of student weights.

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.[13]

We address these issues by employing an alternative identification strategy. Following the initial work by Lavy (2008), we estimate a difference-in-differences (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.[14] 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).

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).

Our empirical framework exploits the fact that each student is observed with two distinct performance measures: a teacher-assigned grade and a standardized test score that

[13] This is because measurement error in $Test\,Score_i$ will bias $\hat{\theta}$ toward 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 2026).

[14] 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. This DiD approach is equivalent to a first-difference model that regresses the difference between students' grades and their test scores on an indicator for being a minority student.

is graded blindly. The identifying assumption is that performance on the standardized test provides a noisy but unbiased proxy for latent achievement. Under this assumption, systematic differences between teacher-assigned grades and test scores by minority status, as estimated by $\hat{\delta}$ in Eq. (2), reflect differences in grading standards rather than differences in underlying ability. Importantly, any such estimated grading differential may encompass both differential grading treatment within teachers and differences arising from the sorting of students into teachers, classes, or schools with different grading stringency. While the identification assumptions underlying our approach cannot be tested, we perform several heterogeneity and sensitivity analyses to explore the mechanisms behind our findings and assess their robustness.[15]

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.[16] 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 0.04 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.

[15] In particular, we aim to determine whether our finding of relatively higher grades for ethnic minority students is indicative of teachers having a positive evaluation bias toward minority students, or if it can fully be explained by the disadvantages faced by minority students on standardized tests or the sorting of minority students into classes or schools with teachers who grade more generously. In addition, we address concerns that our results are driven by distributional differences between grades and test scores.

[16] 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	$-0.357^{\dagger}$	$-0.115^{\dagger}$	0.033^{***}	$-0.261^{\dagger}$	-0.043^{***}	$0.050^{\dagger}$
	(0.013)	(0.012)	(0.010)	(0.014)	(0.013)	(0.010)
Test score	–	–	$0.649^{\dagger}$	–	–	$0.786^{\dagger}$
			(0.006)			(0.006)
Constant	$0.093^{\dagger}$	$5.639^{\dagger}$	$2.521^{\dagger}$	$0.063^{\dagger}$	$5.433^{\dagger}$	$1.365^{\dagger}$
	(0.021)	(0.343)	(0.281)	(0.015)	(0.370)	(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	0.266	0.476	0.014	0.188	0.545

Notes: Standard errors are clustered at the class level. Student weights are applied. Significance level: † 0.1%, *** 1%, ** 5%, * 10%. Full estimation results are shown in Appendix Table A3.

As outlined in Section 3, the estimates obtained by regressing students' grades on their test scores and other observable characteristics may be biased due to unobserved heterogeneity and measurement error in standardized test scores. That measurement error may be relevant in our context is apparent from the coefficient estimates for the test scores in Table 1, which demonstrate that a 1 SD increase in the test scores in German and Math increases the respective grades by only 0.65 and 0.79 SD, respectively. A common approach to addressing measurement error in test scores is to estimate an instrumental variables (IV) regression. For example, contemporaneous test scores in another subject or lagged test scores (from the same or a different subject) can serve as instruments for contemporaneous test performance in a given subject (e.g., Botelho et al. 2015; Terrier 2020; Zhu 2026). While the cross-sectional nature of the data precludes the use of lagged achievement as an instrument, Table 2 shows the estimated effect of minority status on German grades when the German test score is instrumented by the contemporaneous Math (for 4th-grade students) or English (for 9th-grade students) test score.[17] The first-stage results show that

[17] Ideally, we would estimate a similar IV regression for Math, using German (for 4th-grade students) or Science (for 9th-grade students) test performance as an instrument for Math test performance. However, there are two main limitations to this approach. First, a valid instrument must not only be correlated with the test score, but also influence teacher-assigned grades solely through its impact on test scores. In settings that analyze grading disparities by migration background, this exclusion restriction is likely not met when Math test performance is instrumented by German test performance. This is because for ethnic

test performance in Math/English is a strong predictor of test performance in German (see column 1). Furthermore, the second-stage results in column 2 show that the coefficient for the test score in German increases to 0.93, compared to 0.65 in the OLS model. In addition, the effect of minority status on German grades increases to 0.10 SD, compared to an effect of 0.03 SD in the OLS model.[18] Although the IV results may still be affected by measurement error due to potential correlations in errors across subjects (e.g., if a student is unwell during the testing period), the results in Table 2 reveal that measurement error in test scores significantly biases the positive effect of minority status on students' grades downward.

Table 2: IV Regression: Effect of Minority Status on Students' German Grades

	IV-1	**IV-2**	**OLS**
	(1)	(2)	(3)
	Coef/StdE	Coef/StdE	Coef/StdE
Minority student	$-0.224^{\dagger}$	$0.097^{\dagger}$	$0.034^{\dagger}$
	(0.008)	(0.011)	(0.010)
Test score Math/English	$0.601^{\dagger}$	–	–
	(0.004)		
Test score German	–	$0.931^{\dagger}$	$0.649^{\dagger}$
		(0.009)	(0.006)
Controls	yes	yes	yes
Class FE	yes	yes	yes
Observations	90,749	90,749	90,749
Clusters	4,924	4,924	4,924
Adjusted R^2	0.480	0.304	0.352
F-statistic (instrument)	18,703	–	–

Notes: Column (1) shows the results of the first stage of the IV regression, in which the standardized German test score is regressed on the Math (for 4th-grade students) or English (for 9th-grade students) test score and other observable characteristics. Column (2) shows the results of the second stage of the IV regression, where the German grade is regressed on the instrumented German test score and other observable characteristics. For comparison, column (3) shows the results of the respective OLS regression. The slightly smaller sample compared to our baseline model results from missing values in the instrumental variable. Standard errors are clustered at the class level. Student weights are applied. Significance level: † 0.1%, *** 1%, ** 5%, * 10%.

To address this issue and additionally account for test-unspecific unobserved heterogeneity

minority students, German language proficiency, as measured by the German test, likely affects overall school performance, including Math grades. Second, only a small fraction of students in our sample are tested in both Science and Math. This means that we cannot use contemporaneous test performance in either German or Science to instrument the Math test score.

[18] Separate regressions for 4th- and 9th-grade students yield similar results (see Appendix Table A4).

between minority and majority students, we next estimate a DiD model as outlined in Eq. (2). Table 3 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 3 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. They further reveal that majority students receive significantly lower teacher grades compared 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 3: DiD and Fixed Effects Results – German and Math

	German		**Math**	
	(1)	(2)	(3)	(4)
	DiD Coef/StdE	**FE** Coef/StdE	**DiD** Coef/StdE	**FE** Coef/StdE
Minority student	$-0.615^{\dagger}$	–	$-0.462^{\dagger}$	–
	(0.017)		(0.015)	
Non-blind performance measure	$-0.074^{\dagger}$	$-0.074^{\dagger}$	$-0.059^{\dagger}$	$-0.059^{\dagger}$
	(0.012)	(0.012)	(0.010)	(0.010)
Minority × Non-blind	$0.259^{\dagger}$	$0.259^{\dagger}$	$0.203^{\dagger}$	$0.203^{\dagger}$
	(0.016)	(0.016)	(0.013)	(0.013)
Constant	$0.175^{\dagger}$	0.000	$0.135^{\dagger}$	−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. Student 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. Based on the average grades in these subjects, 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., Burgess and Greaves 2013; Botelho et al.

2015; De Benedetto and De Paola 2023). It is, however, consistent with several recent studies documenting higher teacher ratings for ethnic minority students (e.g., Schuessler and Sønderkov 2023; Burn et al. 2024; Zhu 2026).

4.2 Potential Mechanisms and Explanations

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 (2026), several forms of teacher bias may contribute to this pattern. For example, teachers may 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 stereotypes.[19] 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 sorting of students into schools or classes taught by teachers with different grading standards. Although students cannot choose their teachers, minority students may be more likely to end up in schools, classes, or with teachers with more lenient grading standards, creating a positive correlation between grades and minority status even in the absence of teacher bias. In addition, disadvantages faced by ethnic minority students with respect to standardized test assessments could play a role. If minority students have more difficulty excelling on standardized tests for reasons other than differences in the skills being tested, then our

[19] 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).

finding of relatively higher grades for minority students could occur even in the absence of teacher bias.

It is beyond the scope 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 toward ethnic minority students, or whether the grading differential is exclusively attributable to alternative factors.

We start our analysis of the potential mechanisms behind the positive minority/majority grading differential by analyzing whether such a grading differential also exists for students of low socioeconomic status (SES). If teachers adjust their grading standards in response to students' disadvantaged backgrounds, this pattern should not be limited to ethnic minority students but should also be observable among other disadvantaged groups, such as low SES students.[20] 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 2. As expected, low SES 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 differential exists not only for minority students, but also for low SES majority students. Importantly, this finding should not be interpreted as implying that the positive minority/majority grading differential is driven by the fact that minority students are more likely to come from disadvantaged socioeconomic backgrounds. As shown in Appendix Figure A1, the positive minority/majority grading differential persists even among low SES students. Rather, the finding of a similar grading differential for low-SES majority students suggests that the positive minority grading differential may be part of a

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

broader pattern in which teachers adjust their grading standards in response to students' perceived disadvantages.[21]

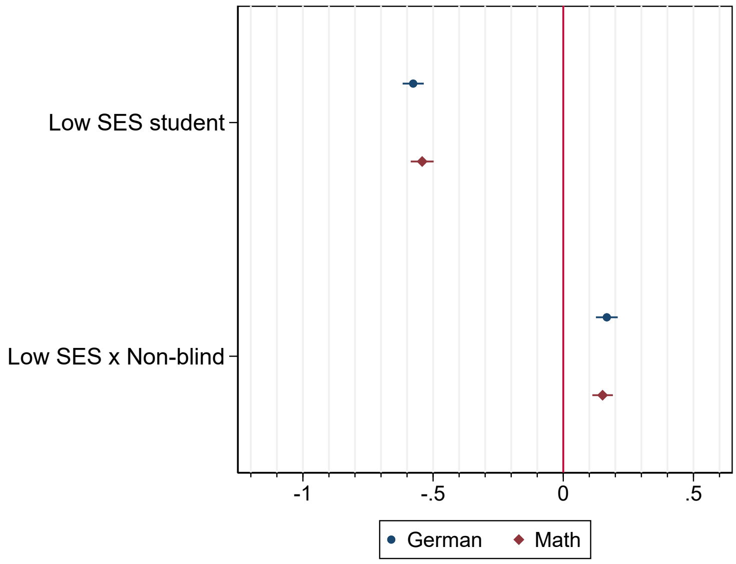


Figure 2: DiD Results – Effect of Low SES for Majority Students

In addition, we take advantage of the fact that the survey collected information not only on students and their parents, but also on teachers.[22] 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 3 shows the results of our DiD model separately for teachers who agree and teachers who disagree with this statement.[23] For German, we find that the positive grading differential 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

[21] In addition to exploring the role of SES, we further analyze whether the grading differential varies across different ethnic groups by distinguishing between students from Turkey and those from the EU. Students of Turkish origin have been shown 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). Appendix Figure A2 shows a positive grading differential relative to majority students for both groups. However, the discrepancy is much larger for students of Turkish origin.

[22] Unfortunately, most of the questions asked to teachers vary from wave to wave. 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.

[23] 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 with this statement.

differential between teachers who demand less from low-capability students and those who do not, which is, however, much smaller (0.20 vs. 0.16 SD) and not statistically significant.[24] 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 differential, especially in German.[25]

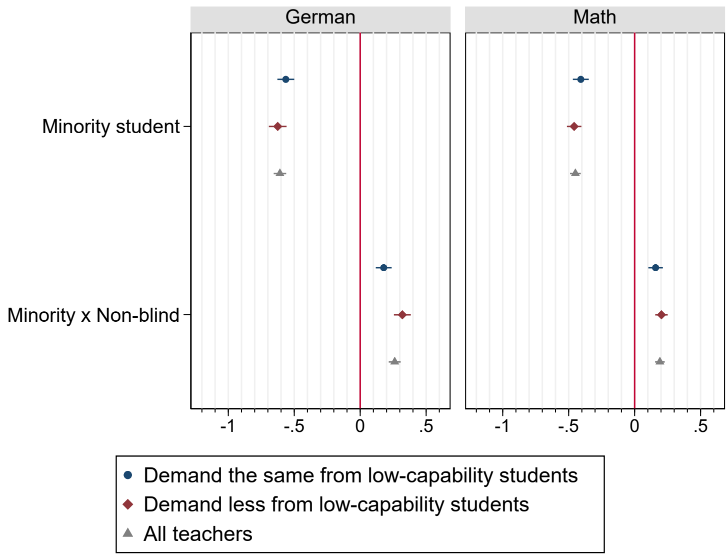


Figure 3: DiD Results – By Whether Teacher Demands Less from Low-Capability Students

As discussed above, an alternative explanation to teacher bias is sorting of students into teachers, classes, or schools with different grading standards. To assess the role of sorting, we re-estimate our main specification by successively including interactions between the dummy variable for the non-blind performance measure and fixed effects at the school, class,

[24] That teachers of German are more likely than Math teachers to say that they demand less from low-capability students, and that the positive grading differential 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.

[25] Importantly, a heterogeneity analysis based on a similar question asking teachers whether they demand significantly more from high-achieving students reveals no difference in the grading differential (see Appendix Figure A3). This finding suggests that it is not differential treatment of students by teachers per se, but the treatment of low-capability students in particular that is able to explain the grading differential between minority and majority students.

and teacher levels.[26] This specification allows us to account for all systematic differences in grading stringency across these dimensions. The remaining minority/majority grading differential therefore reflects differences in the relative assessment of minority and majority students by the same teacher (or, analogously, within the same class or school). Figure 4 shows that including teacher (or class or school) fixed effects reduces the estimated grading differential by about one half compared to the baseline specification. However, the effect remains positive and statistically significant. The persistence of a substantial grading differential after accounting for teacher-specific grading standards indicates that the results cannot be explained by sorting alone.

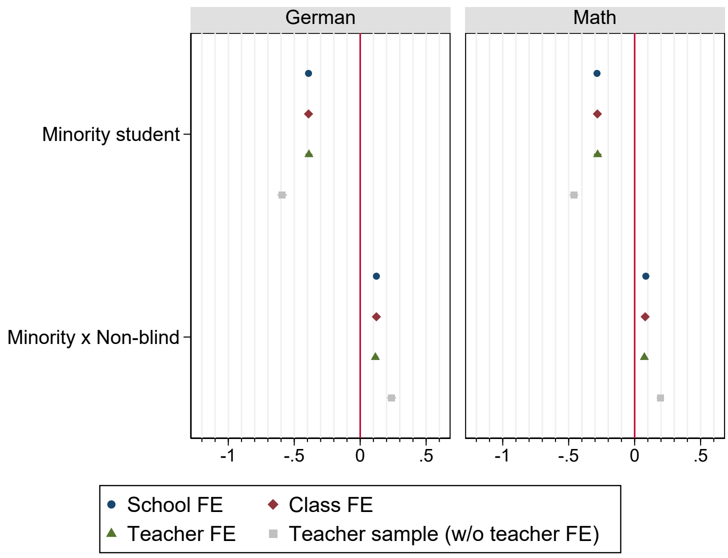


Figure 4: DiD Results – Interaction with School, Class, and Teacher Fixed Effects

A second, alternative explanation for the positive grading differential is the disadvantages that ethnic minority students may face on standardized tests. The main reason students from immigrant families may perform poorly on standardized tests is language difficulties. Although 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 comfortable

[26] Note that in our setting, where one class per school is usually surveyed and all students in a class are typically taught by the same (subject-specific) teacher, there is significant overlap between school, class, and teacher FE. Therefore, we are unable to determine the relative importance of each of these dimensions.

asking comprehension questions during exams administered by their own teachers and where oral participation is also assessed. Therefore, if proficiency in German is more important for students' performance on standardized tests than for teacher evaluations, our finding of relatively higher grades for minority students could occur in the absence of teacher bias. To determine whether German proficiency is relevant in explaining the positive grading differential, we estimate our model using an alternative subject: English. In the 2008 and 2015 waves of the survey, which were both administered to 9th graders, students took a standardized test in the subject English. Since English was the language of this test, proficiency in German should not affect students' performance. As Figure 5 shows, the 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 observe that minority students receive more favorable grades from teachers relative to their test performance than majority students.[27]

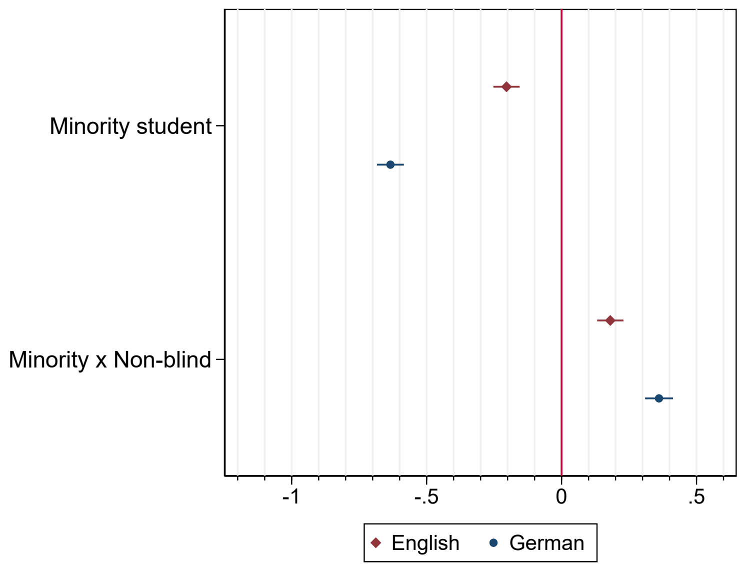


Figure 5: DiD Results – Effect of Minority Status on English vs. German

As an additional check on the relevance of language proficiency, we analyze the heterogeneity of the grading differential by whether the student is a first- or second-generation immigrant, and by whether the student speaks only German at home. As Figure 6 shows, both the negative achievement gap on the standardized test and the improvement in performance

[27]This finding holds when including teacher FE (see Appendix Figure A4).

when graded by the teacher are higher for first-generation immigrants than for second-generation immigrants. In addition, both estimates decrease when the sample is restricted to students who only speak German at home. These results suggest that language proficiency is important for the estimated achievement gaps, but does not fully explain the grading differential between minority and majority students.[28]

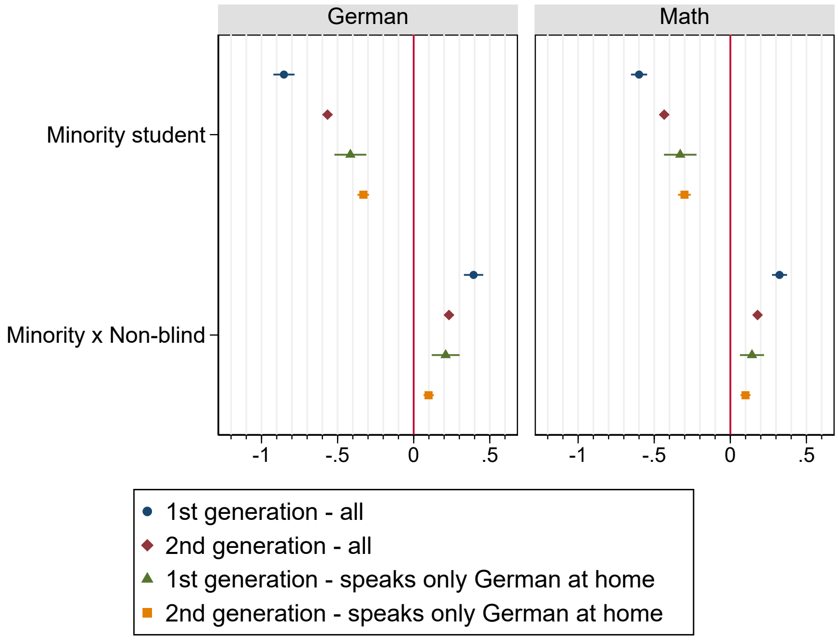


Figure 6: DiD Results – 1st vs. 2nd Generation Immigrants

While the results presented thus far support the hypothesis that teacher bias contributes to the positive grading differential between minority and majority students, two possible channels remain that our empirical analysis cannot fully rule out. First, the main difference between standardized tests and teacher grades is that the latter also depend on oral participation in class. Thus, if ethnic minority students systematically outperform majority students in oral participation, this could explain part of the estimated grading differential. 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.[29]

[28] Importantly, even if language proficiency fully explained the differential, it would not refute teacher bias, as a lack of language proficiency may be one—if not the most important—reason why teachers perceive minority students as disadvantaged.

[29] In particular, the study by Kiss (2013) of primary and secondary school children in Germany shows

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; Hvidman and Sievertsen 2021). If such behavioral differences in low- vs. high-stakes environments varied by student ethnic background, this could be reflected in our DiD estimate. Although we are unable to test whether ethnic minorities behave differently in low- versus high-stakes tests, the existing evidence on test-taking suggests 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 mattered in our setting, they would most likely lead us to underestimate the positive grading differential between minority and majority students.

Taken together, the analyses in this section suggest that the positive minority/majority grading differential documented in the baseline results cannot be fully explained by differences in socioeconomic background, language-related differences in standardized test performance, or the sorting of students into educational contexts with different grading standards. Since the differential persists after accounting for these factors, the results are consistent with the presence of a positive teacher evaluation bias in favor of ethnic minority students.

The additional analyses further provide evidence on how this positive evaluation bias may operate. First, the differential is not specific to ethnic minority students, but also appears among majority students from low socioeconomic backgrounds. This finding suggests that the positive grading differential observed for minority students is part of a broader grading pattern that also extends to other disadvantaged groups. Second, the differential is more pronounced among teachers who report adapting their expectations to low-capability students. Together, these findings are consistent with the interpretation that teachers partly adjust their grading standards in response to students' backgrounds. In this sense, teacher bias appears to operate in a compensatory manner, resulting in

that students from immigrant families are, if anything, less likely to participate frequently in Math classes.

more favorable grades for students who perform well despite disadvantaged starting conditions. However, although our empirical design identifies systematic differences between teacher-assigned grades and standardized test outcomes, it cannot distinguish the precise behavioral mechanisms underlying such compensatory grading. For example, it cannot determine whether such grading arises from a conscious desire to promote equity through benevolent compensation or whether it reflects lower expectations or implicit stereotypes. Therefore, the compensatory grading interpretation should be understood as the behavioral explanation that best fits the overall pattern of evidence rather than as a mechanism that can be directly identified by our empirical approach.

4.3 Further Heterogeneity Analyses and Robustness Checks

In this section, we present additional heterogeneity analyses and robustness checks. As revealed by our results in Section 4.2, part of the positive grading differential between minority and majority students stems from differential grading standards across teachers, classes, or schools. One possible explanation for this variation is that teachers with different characteristics have different grading standards and are sorted into different schools. To explore the role of teacher characteristics in explaining the grading differential, we conduct heterogeneity analyses based on teachers' gender, age, experience, tenure, and migration status. As shown in Appendix Figure A5, we find little difference in the grading differential by teacher characteristics. The positive grading differential 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 the grading differential with respect to teachers' gender, age, or migration status.[30]

[30]The latter finding is particularly interesting, because one might expect that teachers who belong to a minority group themselves 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. Unfortunately, we are unable to match students and teachers of the same ethnicity because we only know whether the teacher was born in Germany. We have no information on the teacher's country of birth or the birthplace of their parents. Moreover, the proportion of foreign-born teachers in German schools is very small, accounting for only about 2% of the teachers in our sample. Thus, it is theoretically unclear whether we should expect the grading differential to be higher or lower for immigrant teachers.

Another possible explanation for differential grading standards across classes or schools is that teachers adjust their grading standards depending on their students' characteristics. To explore the role of class composition, we conduct heterogeneity analyses by the share of minority students and by the average test performance in the class.[31] The results reveal strong differences in the grading differential by class composition (see Appendix Figure A6). A positive grading differential 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. Likewise, a strong positive grading differential is found in classes with a below-median class performance, while there is no such grading differential in classes with above-median test performance. These findings suggest that minority/majority grading differentials across classes (or teachers and schools) can mostly be explained by differential grading standards across high- and low-performing classes.

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 A7. We also test for student gender differences in the positive grading bias toward minority students, for which we find no evidence (see Appendix Figure A8). However, our results show that the grading bias is larger for 9th-grade students than for 4th-grade students (see Appendix Figure A9). Appendix Figure A10, which estimates our baseline DiD based on a sample including students with missing information on any of the sociodemographic characteristics controlled for in Eq. (1), further shows that our findings are not affected by selective non-response.

Next, we account for the fact that test scores offer a much finer scale than teacher-assigned grades, which are inherently discrete. The restricted scope of the teacher assessments implies that teachers may be unable to differentiate precisely between various students, a circumstance that could differ by minority status. Therefore, as a robustness check, we aggregate the test score distribution to the level of the grade distribution and

[31] When analyzing heterogeneity by the share of minority students in a class, we restrict the sample to classes with at least one minority student, which applies to 88% of the German classes and 89% of the Math classes in our sample. Analyses that include all classes yield similar results.

estimate a DiD model that compares the actual teacher-assigned grades with the "grades" based on test performance.[32] As Appendix Figure A11 shows, the results are robust and thus not driven by distributional differences between grades and test scores.

Lastly, we examine a further consequence of the discrete and bounded nature of our outcome variable: ceiling effects (e.g., Gorry 2017; Burn et al. 2024). Ceiling effects may arise when the outcome variable is bounded from above. In our setting, the upper bound of the grade distribution at a value of 6, combined with the higher average performance of majority students, could in principle limit observable variation in grades at high achievement levels, thereby constraining teachers' ability to differentiate among top-performing students. If ceiling effects were driving the estimated grading differential, we would expect to observe heterogeneity across the grade–test score gap distribution, with larger minority-related differences in the middle of the distribution, where ceiling-induced compression would mechanically generate small grade-test score discrepancies. As a descriptive check, we therefore estimate quantile regressions of the grade–test score difference on minority status. Appendix Figure A12 shows a positive grading differential at all examined quantiles, with a modest increase toward higher quantiles. This upward-sloping pattern is difficult to reconcile with a simple mechanical ceiling-effect explanation, because ceiling-induced compression would be expected to reduce the grade–test score gap among top-achieving students. More generally, since the estimated differential arises across the entire distribution, it does not appear to be driven by a small subset of students with unusual grade–test discrepancies.[33]

In addition, to ensure that our results are not driven by particularly high- or low-performing students, we assess robustness by excluding the 5% highest and lowest standardized test scores from the sample. The resulting estimates are very similar to the

[32] We aggregate test scores separately for each wave and use the observed grade distribution to map standardized test scores into six grade categories. For instance, if 10% of students receive the highest grade (6) in Math in a given wave, we assign the top 10% of students on the standardized Math test to grade category 6, and so on.

[33] As an alternative, and more direct test of ceiling effects, we would like to assess the heterogeneity of the grading differential across students' grade or test score distributions. However, since grades and test scores are both outcomes in our DiD framework, the results of such sample-split regressions are difficult to interpret in our setting because they would be distorted by endogenous selection into grade and test-score bins.

baseline estimates (see Appendix Figure A13), indicating that our findings are not driven by extreme performers, for whom ceiling and floor effects, respectively, would be most likely to matter. Taken together, both robustness checks are inconsistent with a mechanism in which the grading differential is primarily generated by ceiling-induced compression at the top of the grading scale. Rather, they suggest that the estimated grading differential operates broadly across the distribution and is not driven by extreme test results.

5 Conclusion

In this paper, we analyze whether teachers discriminate against ethnic minority students with respect to grading. Based on nationally representative survey data on German primary and secondary school students, we compare students' scores on standardized, anonymously graded achievement tests with non-anonymous teacher ratings within a difference-in-differences 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.

Contrary to common concerns about ethnic discrimination in grading, teachers in German schools appear to grade ethnic minority students more favorably than their performance on standardized tests would suggest. Specifically, we find that the minority/majority achievement gap in German and Math is reduced by 0.26 and 0.20 standard deviations (or 5–6%), respectively, when students are evaluated by their teachers. Exploring the mechanisms behind the positive grading differential, we find that about half of the differential is driven by the sorting of minority students into classes or schools where teachers apply more lenient grading standards. However, a significant within-teacher grading differential remains after accounting for student sorting. In addition, we demonstrate that the disadvantages ethnic minority students may face on standardized tests due to language barriers cannot account for the positive grading differential, as a similar differential occurs for the subject English. Finally, we provide evidence that teachers adjust their assessment standards to account for students' backgrounds. Specifically, we show that the positive grading differential is stronger among teachers who report that they demand significantly less from students with low capability, and that such a grading differential is not limited to ethnic minority students but also extends to (majority) students from low socioeconomic backgrounds. Our results are robust to a range of robustness checks. In particular, we show that our findings are not driven by distributional differences between grades and test scores, particularly ceiling effects. Overall, our findings suggest that at least part of the positive grading differential between minority and majority students can be explained by a positive grading bias toward ethnic minority students. Specifically, the findings are consistent with the interpretation that teachers in German schools adjust their grading standards based on perceived student disadvantage. This compensatory behavior appears to apply not only to ethnic minority students but also to those from low-socioeconomic backgrounds.

From a policy perspective, these findings highlight the trade-off between equity and efficiency in education. Given the considerable influence that grades exert on educational trajectories, a grading bias that favors minority students has the potential to contribute to the narrowing of existing ethnic disparities in education, thereby promoting equity. However, the inaccurate signals of actual achievement associated with biased grades may impede students' development of essential skills or result in suboptimal investments in human capital, thereby promoting inefficiency. A further potential downside of such a positive grading bias stems from, or coexists with, lower underlying teacher expectations for minority students, which can create self-fulfilling prophecies (Rosenthal and Jacobson 1968). As demonstrated in the literature, negative teacher expectations have the potential to adversely affect students' human capital investments and long-run performance (Lavy and Sand 2018; Papageorge et al. 2020; Hill and Jones 2021; Lavy and Megalokonomou 2024), potentially perpetuating longstanding achievement gaps. In light of these findings, teacher training programs should prioritize the development of strategies that promote equitable support and accurate assessment for students from diverse backgrounds. Additionally, policies aimed at reducing socioeconomic disparities and language barriers could serve to mitigate the need for teachers to adjust grading standards as a compensatory mechanism.

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. 2019. "Implicit Stereotypes: Evidence from Teachers' Gender Bias." *The Quarterly Journal of Economics*, 134(3): 1163–1224.

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.

Ferman, Bruno, and Luiz Felipe Fontes. 2022. "Assessing knowledge or classroom behavior? Evidence of teachers' grading bias." *Journal of Public Economics*, 216: 104773.

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.

Gorry, Devon. 2017. "The impact of grade ceilings on student grades and course evaluations: Evidence from a policy change." *Economics of Education Review*, 56: 133–140.

Hanna, Rema N., and Leigh L. Linden. 2012. "Discrimination in Grading." *American Economic Journal: Economic Policy*, 4(4): 146–68.

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

Hvidman, Ulrik, and Hans Henrik Sievertsen. 2021. "High-Stakes Grades and Student Behavior." *Journal of Human Resources*, 56(3): 821–849.

Jansson, Joakim, and Björn Tyrefors. 2022. "Grading bias and the leaky pipeline in economics: Evidence from Stockholm University." *Labour Economics*, 78: 102212.

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

Karna, Uditi, Min Sok Lee, John A. List, Andrew Simon, and Haruka Uchida. 2025. "Measuring racial educational disparities over time amongst top achievers." *Nature*, 639: 976–984.

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.

Okun, Arthur M. 1975. *Equality and Efficiency: The Big Tradeoff.* Washington, D.C.: Brookings Institution Press.

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. 2026. “First Impressions Matter: Evidence From Elementary-School Teachers.” *Journal of Human Resources*, 61(1): 123–159.

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. 1968. “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.

Sahlström, 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.

Schuessler, Julian, and Kim M. Sønderkov. 2023. “Compensating Discrimination: Behavioral Evidence from Danish School Registers.” *https://doi.org/10.31235/osf.io/5zm87.*

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.

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. 2026. "Teacher Perceptions, Test Scores, and Racial Disparities in the Classroom: Evidence from the US." *Journal of Human Resources (forthcoming).*

Appendix

Tables

Table A1: Descriptive Statistics – Analysis Sample German

	Minority students		**Majority students**	
	Mean	StdD	Mean	StdD
Female	0.500	0.500	0.498	0.500
Age	12.830	2.614	12.770	2.535
4th grade student	0.531	0.499	0.522	0.500
Mother's education				
Low education	0.248	0.432	0.168	0.374
Medium education	0.258	0.438	0.363	0.481
High education	0.252	0.434	0.372	0.483
Missing	0.242	0.428	0.097	0.296
Father's education				
Low education	0.217	0.412	0.138	0.345
Medium education	0.232	0.422	0.306	0.461
High education	0.275	0.447	0.426	0.495
Missing	0.276	0.447	0.130	0.336
Mother's occupation				
White collar	0.375	0.484	0.655	0.475
Blue collar	0.178	0.383	0.090	0.286
Other	0.185	0.389	0.122	0.327
Missing	0.261	0.439	0.134	0.341
Father's occupation				
White collar	0.315	0.464	0.525	0.499
Blue collar	0.254	0.435	0.149	0.356
Other	0.164	0.370	0.171	0.377
Missing	0.267	0.442	0.155	0.362
Number of books at home	114.910	131.309	181.676	152.223
Low SES family	0.232	0.422	0.119	0.324
Region of origin				
Turkey	0.196	0.397	–	–
EU country	0.218	0.413	–	–
Other country	0.586	0.493	–	–
First-generation immigrant	0.169	0.375	–	–
Speaks only German at home	0.415	0.493	–	–
Observations	23,790		69,147	

Notes: The table shows the means and standard deviations of observable student characteristics for the sample of students who participated in the German test. Student weights are applied.

Table A2: Descriptive Statistics – Analysis Sample Math

	Minority students		**Majority students**	
	Mean	StdD	Mean	StdD
Female	0.499	0.500	0.500	0.500
Age	12.410	2.577	12.249	2.476
4th grade student	0.620	0.485	0.632	0.482
Mother's education				
Low education	0.292	0.455	0.181	0.385
Medium education	0.237	0.425	0.350	0.477
High education	0.255	0.436	0.366	0.482
Missing	0.216	0.412	0.104	0.305
Father's education				
Low education	0.254	0.435	0.149	0.356
Medium education	0.205	0.404	0.272	0.445
High education	0.278	0.448	0.429	0.495
Missing	0.263	0.441	0.150	0.357
Mother's occupation				
White collar	0.378	0.485	0.659	0.474
Blue collar	0.166	0.372	0.081	0.272
Other	0.181	0.385	0.123	0.329
Missing	0.274	0.446	0.138	0.344
Father's occupation				
White collar	0.320	0.467	0.534	0.499
Blue collar	0.238	0.426	0.133	0.340
Other	0.161	0.368	0.168	0.374
Missing	0.280	0.449	0.165	0.371
Number of books at home	111.037	129.346	173.465	149.075
Low SES family	0.240	0.427	0.119	0.323
Region of origin				
Turkey	0.190	0.393	–	–
EU country	0.235	0.424	–	–
Other country	0.575	0.494	–	–
First-generation immigrant	0.167	0.373	–	–
Speaks only German at home	0.424	0.494	–	–
Observations	21,157		59,865	

Notes: The table shows the means and standard deviations of observable student characteristics for the sample of students who participated in the Math test. Student weights are applied.

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

	German	**Math**
	Coef/StdE	Coef/StdE
Minority student	0.033^{***}	$0.050^{\dagger}$
	(0.010)	(0.010)
Test score	$0.649^{\dagger}$	$0.786^{\dagger}$
	(0.006)	(0.006)
Female	$0.304^{\dagger}$	$0.109^{\dagger}$
	(0.009)	(0.008)
Age	$-0.331^{\dagger}$	$-0.169^{\dagger}$
	(0.041)	(0.040)
Age^2	$0.009^{\dagger}$	0.004^{***}
	(0.001)	(0.001)
Mother's education (Ref.: Low education)		
Medium education	$0.056^{\dagger}$	$0.043^{\dagger}$
	(0.014)	(0.013)
High education	$0.051^{\dagger}$	0.029^{**}
	(0.014)	(0.013)
Missing	0.059^{***}	0.009
	(0.019)	(0.020)
Father's education (Ref.: Low education)		
Medium education	0.007	0.006
	(0.015)	(0.014)
High education	$0.071^{\dagger}$	$0.057^{\dagger}$
	(0.015)	(0.013)
Missing	−0.023	−0.018
	(0.019)	(0.018)
Mother's occupation (Ref.: White collar)		
Blue collar	-0.044^{***}	-0.037^{***}
	(0.014)	(0.014)
Other	-0.031^{***}	−0.000
	(0.012)	(0.012)
Missing	-0.027^{*}	0.009
	(0.015)	(0.016)
Father's occupation (Ref.: White collar)		
Blue collar	$-0.043^{\dagger}$	−0.009
	(0.012)	(0.012)
Other	−0.007	0.007
	(0.010)	(0.011)
Missing	$-0.109^{\dagger}$	$-0.074^{\dagger}$
	(0.015)	(0.016)
Number of books at home (in 100)	0.028^{***}	0.024^{**}
	(0.010)	(0.011)
Number of books at home $(\text{in } 100)^2$	-0.004^{*}	-0.006^{***}
	(0.002)	(0.002)
Constant	$2.521^{\dagger}$	$1.365^{\dagger}$
	(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. Student weights are applied. Significance level: † 0.1%, *** 1%, ** 5%, * 10%.

Table A4: IV Regression: Effect of Minority Status on German Grade for 4th- and 9th-Grade Students

	4th-grade students			**9th-grade students**		
	IV-1	IV-2	OLS	IV-1	IV-2	OLS
	(1)	(2)	(3)	(4)	(5)	(6)
	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE
Minority student	$-0.157^{\dagger}$	$0.085^{\dagger}$	0.024^{*}	$-0.274^{\dagger}$	$0.096^{\dagger}$	0.037^{**}
	(0.011)	(0.014)	(0.013)	(0.010)	(0.017)	(0.016)
Test score Math	$0.644^{\dagger}$	–	–	–	–	–
	(0.006)					
Test score English	–	–	–	$0.524^{\dagger}$	–	–
				(0.006)		
Test score German	–	$0.945^{\dagger}$	$0.670^{\dagger}$	–	$0.854^{\dagger}$	$0.587^{\dagger}$
		(0.010)	(0.007)		(0.018)	(0.010)
Controls	yes	yes	yes	yes	yes	yes
Class FE	yes	yes	yes	yes	yes	yes
Observations	42,497	42,497	42,497	48,252	48,252	48,252
Clusters	2,461	2,461	2,461	2,463	2,463	2,463
Adjusted R^2	0.552	0.429	0.484	0.360	0.178	0.211
F-statistic (instrument)	13,685	–	–	6,699	–	–

Notes: Columns (1) and (4) show the results of the first stage of an IV regression, in which the standardized German test score is regressed on the Math (column (1)) and English (column (4)) test score and other observable characteristics. Columns (2) and (5) show the results of the second stage of the IV regression, where the German grade is regressed on the instrumented German test score and other observable characteristics. For comparison, columns (3) and (6) show the results of the respective OLS regressions. Standard errors are clustered at the class level. Student weights are applied. Significance level: † 0.1%, *** 1%, ** 5%, * 10%.

Figures

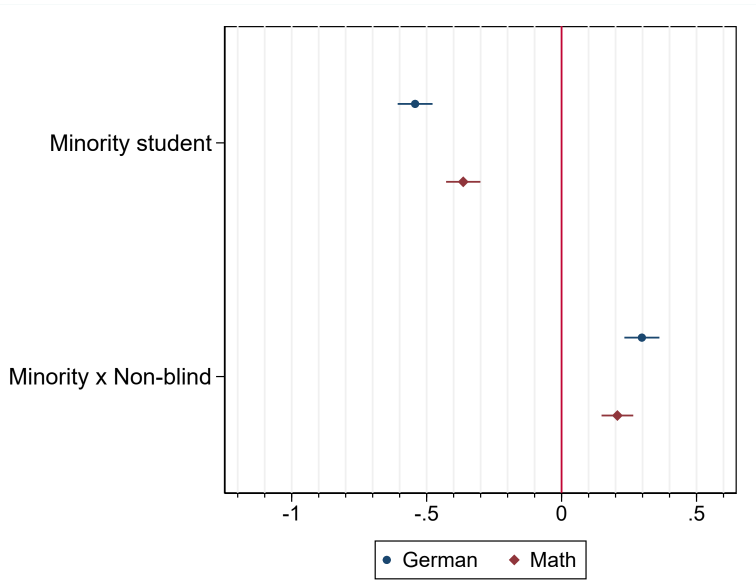


Figure A1: DiD Results – Only Low SES Students

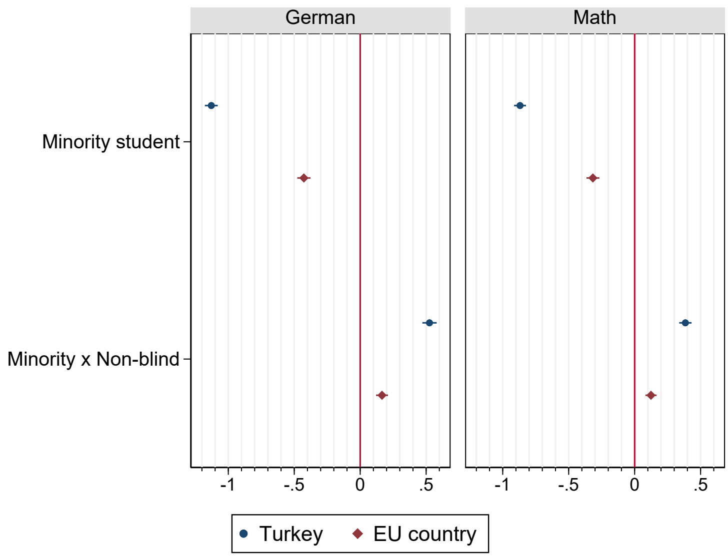


Figure A2: DiD Results – By Students' Country of Origin

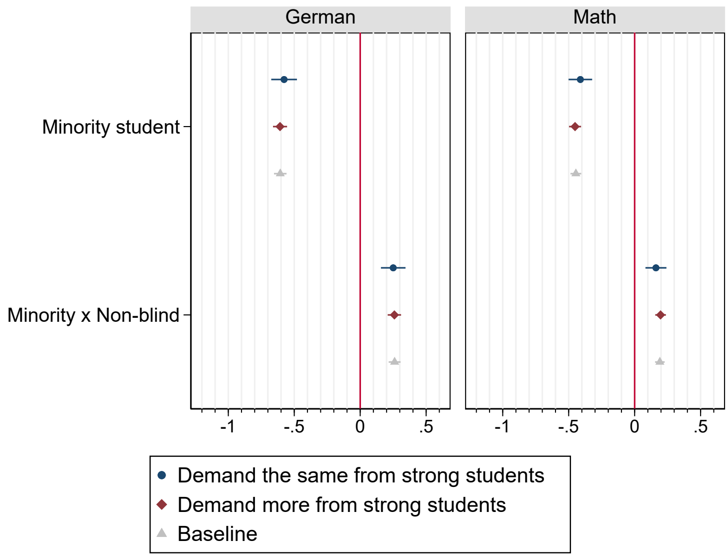


Figure A3: DiD Results – By Whether Teacher Demands More from High-Achieving Students

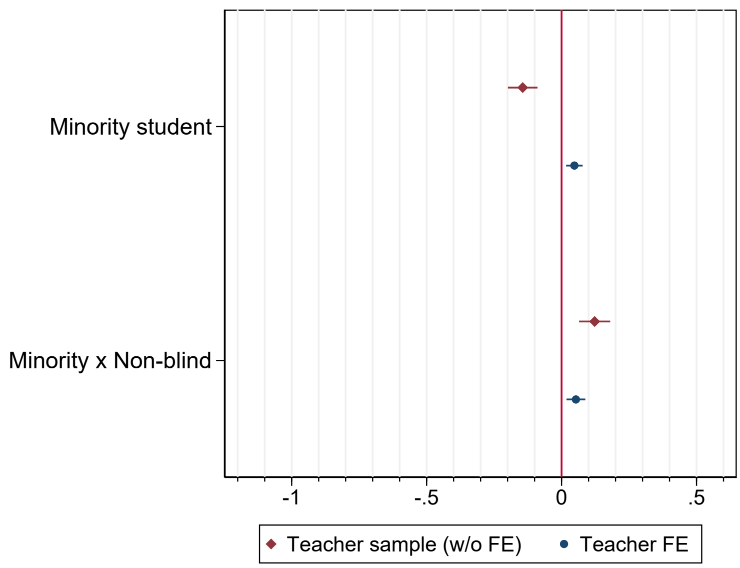


Figure A4: DiD Results – English, Interaction with Teacher FE

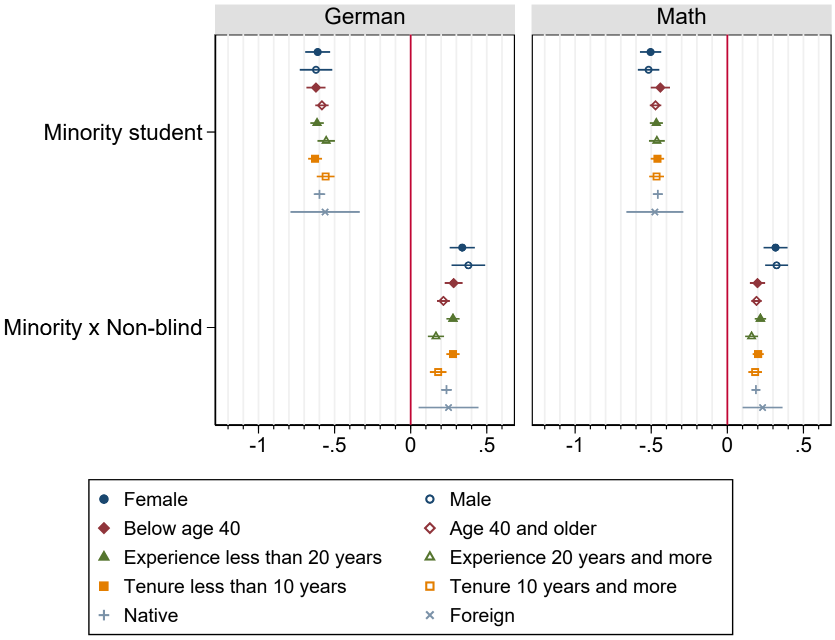


Figure A5: DiD Results – By Teacher Characteristics

Note: Due to the small fraction of male teachers in 4th grade, the sample for the heterogeneity analysis by gender is restricted to 9th-grade students.

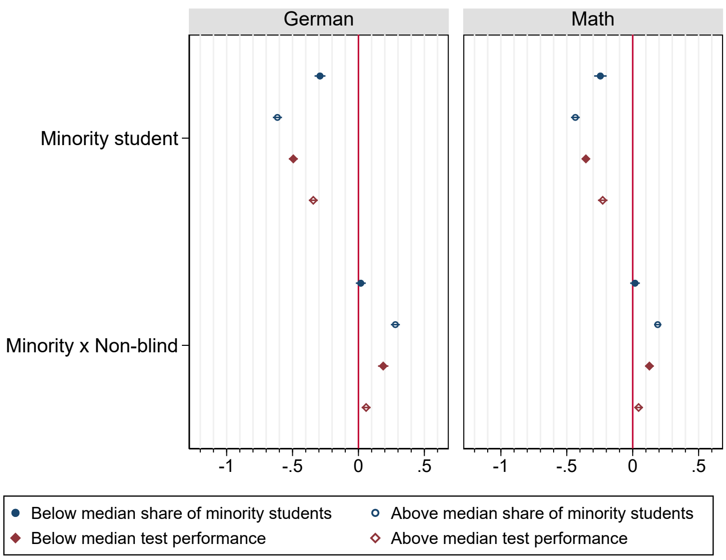


Figure A6: DiD Results – By Class Composition

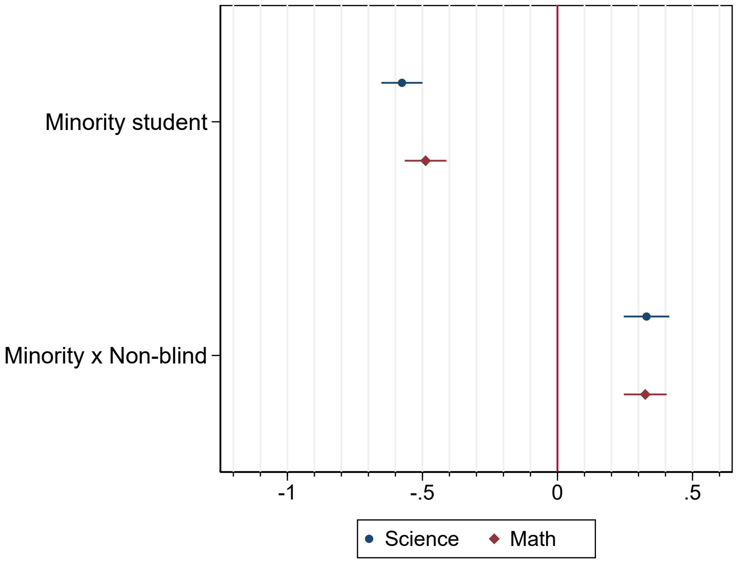


Figure A7: DiD Results – Science and Math

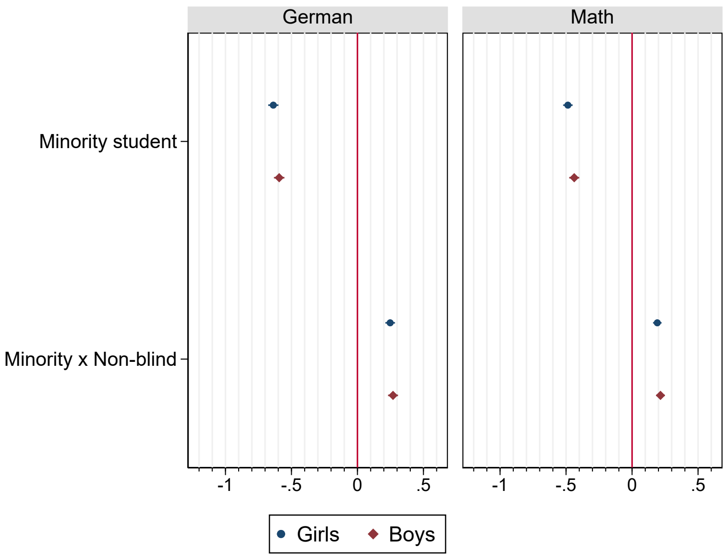


Figure A8: DiD Results – By Student Gender

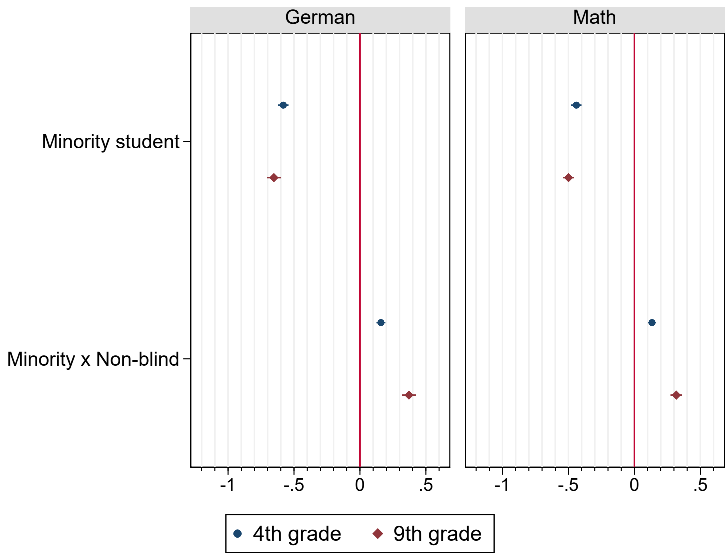


Figure A9: DiD Results – By School Grade

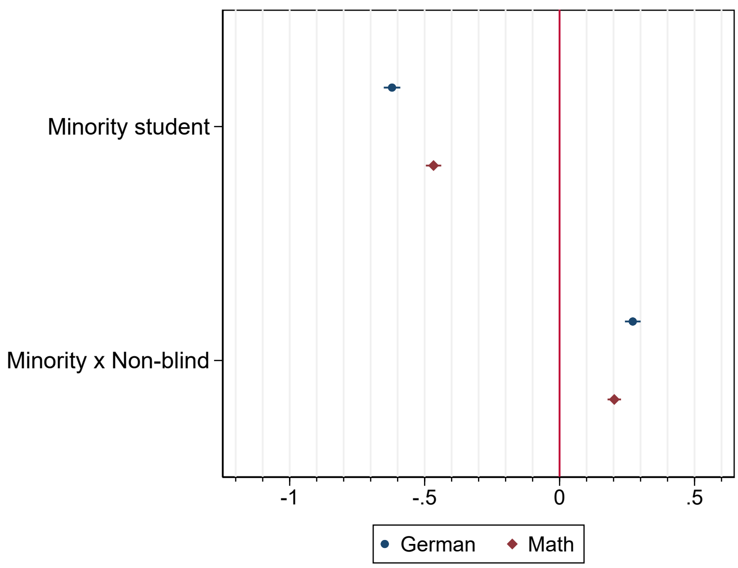


Figure A10: DiD Results – Sample Including Students with Missing Information on Sociodemographic Characteristics

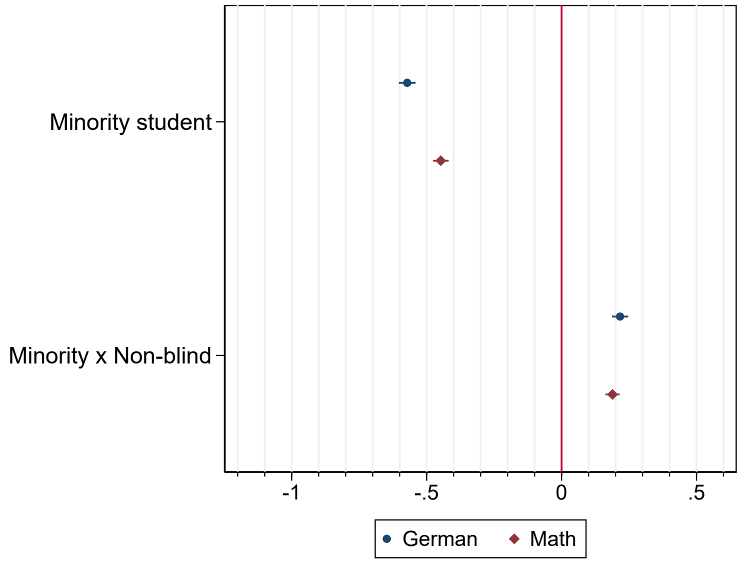


Figure A11: DiD Results – Test Scores Aggregated to Grade Distribution

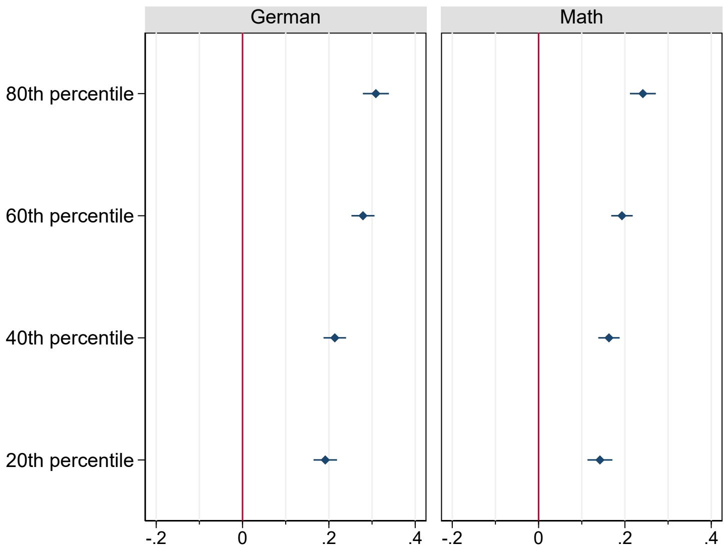


Figure A12: Grading Differential Across the Distribution of the Grade–Test Score Gap

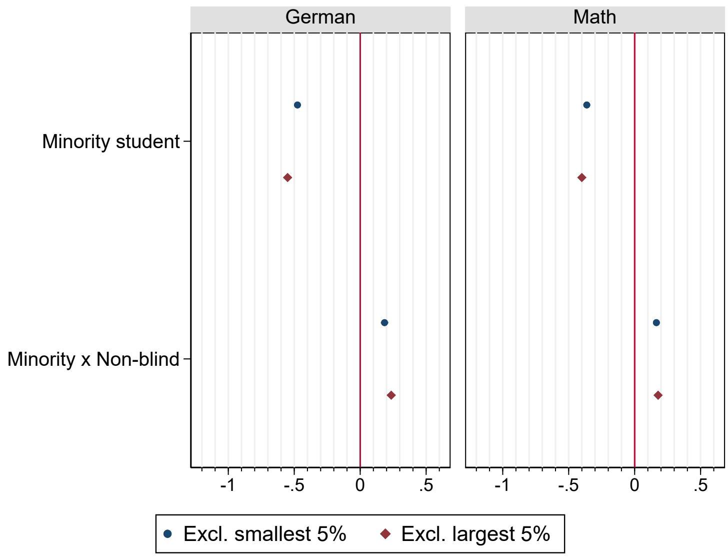


Figure A13: DiD Results – Excluding the 5% Smallest and 5% Largest Test Scores