RUHR

ECONOMIC PAPERS

Gender Differences in Teacher Grading Across School Subjects: Evidence From Germany

Sebastian Otten

#1222

Imprint

Ruhr Economic Papers

Ruhr Economic Paper #1222 "Gender Differences in Teacher Grading Across School Subjects: Evidence From Germany"

Responsible Editor: Volker Clausen

Jointly published by

RWI – Leibniz-Institut für Wirtschaftsforschung e.V.
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

Bergische Universität Wuppertal, Schumpeter School of Business and Economics
Gaußstraße 20, 42119 Wuppertal, Germany

Series Coordination:
RWI Press Office, presse@rwi-essen.de, phone +49 (0) 201 8149-213
RWI – Leibniz Institute for Economic Research, Hohenzollernstr. 1–3, 45128 Essen, Germany
www.rwi-essen.de

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.
The Institute has the legal form of a registered association; Vereinsregister, Amtsgericht Essen VR 1784

The working papers published in the series constitute work in progress circulated to stimulate discussion and critical comments. Views expressed represent exclusively the authors' own opinions and do not necessarily reflect those of the editors and institutions.

ISSN 1864-4872 (online)
ISBN 978-3-96973-407-0
DOI https://dx.doi.org/10.4419/96973407

Gender Differences in Teacher Grading across School Subjects: Evidence from Germany*

Sebastian Otten[a,b,c]

[a] *University of Duisburg-Essen*
[b] *ROCKWOOL Foundation Berlin*
[c] *RWI – Leibniz Institute for Economic Research*

July 2026

Abstract

This study examines whether gender differences in teacher-assigned grades vary systematically across subjects. Using data from German secondary schools, I compare teacher-assigned grades with anonymously graded standardized test scores within a difference-in-differences framework. Across six core subjects spanning the humanities and STEM, I find systematic grading differentials favoring female students. The estimated differential is largest in German and Math and substantially smaller in English and the sciences, indicating considerable cross-subject heterogeneity. The differential also varies markedly by school track and teacher experience, with systematically larger estimates in the academic track and particularly pronounced differentials among novice teachers.

JEL classification: I21; I24; J16

Keywords: Gender differences; Teacher grading; Grade–test differential; Educational inequality

*The author is grateful to Julia Bredtmann and Christina Vonnahme for their valuable comments and suggestions. – Contact details: Sebastian Otten, University of Duisburg-Essen (UDE), Universitätsstraße 12, 45141 Essen, Germany, E-mail: sebastian.otten@uni-due.de.

1 Introduction

School grades shape students' educational trajectories, self-perceptions, and later opportunities. In most OECD countries, girls outperform boys in several educational outcomes, including average grades and university entrance qualifications (OECD 2025a). Yet international achievement tests show a different pattern, with girls scoring higher in reading and boys scoring higher in Math (Machin and Pekkarinen 2008; OECD 2025b). This divergence raises the question of whether teacher assessments reflect only subject-specific achievement or also gender-related differences in how performance is perceived and rewarded.

This paper examines whether gender differences in teacher-assigned grades vary systematically across core subjects spanning the humanities and STEM. Using student-level data from German secondary schools, I compare teacher-assigned grades with anonymously graded standardized test scores in a difference-in-differences (DiD) framework. I find female-favoring grade–test differentials across all analyzed subjects. The magnitude of these differentials varies substantially across subjects, school tracks, and teacher experience.

Most studies comparing teacher-graded and externally graded assessments document gender-related grading patterns that tend to favor girls. Focusing primarily on language and Math, this literature finds that girls receive higher grades than boys with comparable achievement (e.g., Lavy 2008; Cornwell et al. 2013; Falch and Naper 2013; Terrier 2020; Di Liberto et al. 2022; Contreras 2024). Yet the evidence is not uniform. Hinnerich et al. (2011) find no grading bias in Swedish language exams, Jansson and Tyrefors (2025) document an opposite gender differential in a university setting,[1] and Nennstiel and Gilgen (2024) find heterogeneous gender differentials across subjects in Germany. Despite this growing body of work, evidence remains limited on whether these patterns generalize across a wider range of subjects and which features of the learning environment shape these differentials.

This study makes two main contributions. First, whereas prior work typically concen-

[1]Their setting differs from that of the present study, as they examine a university-wide anonymization reform and find that anonymous grading improves female students' grades relative to those of male students.

trates on Math and/or language, I provide subject-spanning evidence on gender-related grade–test differentials across the humanities and STEM in German secondary schools. Second, I document novel heterogeneity in the grade–test differential by school track and teachers' years of experience, dimensions that are rarely examined in the literature.

2 Data and Empirical Strategy

2.1 Data

The analysis uses nationally representative repeated cross-sectional surveys of ninth-grade students provided by the Institute for Educational Quality Improvement (IQB). I use the 2008/09, 2012, 2015, and 2018 waves, covering students aged 14–15. Achievement in German and English was assessed in 2008/09 and 2015, whereas Math and the science subjects (Biology, Chemistry, and Physics) were assessed in 2012 and 2018. The sample is drawn in two stages: schools are selected first, followed by one or two randomly chosen classes per school. Participation is mandatory for sampled public schools.

For each subject, two performance measures are observed: (i) teacher-assigned grades from the first semester and (ii) standardized achievement test scores in the corresponding subject.[2] Grades are assigned by the subject teacher and are based on written exams and oral participation. In Germany, grades range from 1 (very good) to 6 (fail); I rescale them so that higher values indicate better performance.

The standardized tests are externally administered and scored, and closely resemble PISA in format. They are aligned with national curricula and therefore capture the same subject-specific competencies as school grades. The tests do not affect students' grades, and neither teachers nor students receive individual-level test results. Test scorers receive no information about students or schools. Linked student, teacher, and principal questionnaires provide information on students' socioeconomic background as well as

[2]First-semester grades are awarded in January, and standardized tests follow in spring, implying a short interval during which substantial changes in relative gender performance are unlikely. Grades are school-reported, avoiding recall and reporting bias.

teacher and school characteristics.[3]

Depending on the subject, the sample comprises between 31,590 (Biology) and 49,548 (German) students, yielding twice as many observations in the stacked DiD estimation. Appendix Table A.1 reports summary statistics, and Appendix Figure B.1 plots the gender-specific distributions of grades and test scores. Across all six subjects, girls are more likely than boys to receive higher grades. Test scores, however, are more evenly distributed: female students outperform male students in most subjects, with the exception of Math and Physics.

2.2 Empirical Strategy

Following Lavy (2008), I exploit within-student variation between teacher-assigned grades (non-blind) and anonymously scored standardized tests (blind) in a difference-in-differences (DiD) framework. Pooling both measures for each student, I estimate:

$$Y_{ib} = \alpha + \beta F_i + \gamma NB_b + \delta(F_i \times NB_b) + \varepsilon_{ib}, \tag{1}$$

where Y_{ib} denotes student i's outcome on assessment type $b \in \{\text{blind, non-blind}\}$.[4] F_i is an indicator equal to one for female students, and NB_b is an indicator equal to one for non-blind teacher-assigned grades and zero for blind test scores. The coefficient β captures the average female–male gap in blind test scores, while γ captures the average difference between non-blind and blind assessments for male students. The coefficient of interest is δ, which measures the differential gender gap in non-blind teacher-assigned grades relative to blind standardized test scores. A positive δ implies that girls receive higher teacher-assigned grades than boys, relative to the gender gap in blind test scores. This gap may reflect either teacher bias in grading or teachers rewarding non-cognitive traits not captured by blind standardized tests, such as effort or compliance.

[3] For a more detailed description of the data and the two performance measures, see Bredtmann et al. (2024). Information on how to obtain access to the underlying data is provided in Appendix C.

[4] To ensure comparability across assessment types, I standardize grades and test scores separately by subject and survey wave, using the full sample for each subject-wave combination, so that each measure has a mean of zero and a standard deviation of one.

Identification relies on the assumption that, absent differential grading, the female–male gap would not differ systematically between blind test scores and non-blind teacher-assigned grades. This assumption is plausible because the blind tests are scored without knowledge of students' identity, ensuring that blind scores are free of gender-related assessment effects, and because they are closely aligned with national curricula, so that both assessments measure the same underlying competencies. I estimate Eq. (1) separately by subject. Standard errors are clustered at the class level, and all regressions are weighted using student sampling weights.

3 Results

3.1 Baseline Results

Table 1 reports estimates of Eq. (1) separately by subject. Female students outperform male students on the blind standardized test in all subjects except Math, where boys score significantly higher, and Physics, where the gender gap is small and statistically insignificant. Male students, in turn, receive significantly lower teacher-assigned grades relative to their blind test performance in all subjects except Biology.

Table 1: Grading Differential by Subject – DiD Results

	Humanities Subjects		**STEM Subjects**			
	German (1)	**English** (2)	**Math** (3)	**Biology** (4)	**Chem.** (5)	**Physics** (6)
Female	0.229***	0.159***	−0.162***	0.196***	0.038**	−0.021
	(0.018)	(0.018)	(0.017)	(0.019)	(0.019)	(0.019)
Non-blind	−0.133***	−0.078***	−0.120***	−0.037	−0.058**	−0.052**
	(0.024)	(0.025)	(0.022)	(0.023)	(0.026)	(0.024)
Female × Non-blind	0.268***	0.157***	0.238***	0.073***	0.114***	0.102***
	(0.019)	(0.019)	(0.018)	(0.022)	(0.022)	(0.022)
Observations	99,096	96,630	76,000	63,180	66,836	64,794
Clusters	2,517	2,469	3,755	3,140	3,282	3,202
Adjusted R^2	0.037	0.016	0.004	0.014	0.003	0.001

Notes: Coefficient estimates are obtained by estimating Eq. (1) separately by subject. Standard errors, reported in parentheses, are clustered at the class level. All regressions are weighted using student sampling weights. Significance levels: *** 1%, ** 5%, * 10%.

The coefficient of primary interest is that on the interaction term *Female* × *Non-blind*.

The coefficient is positive and statistically significant across all six subjects, indicating that the female–male gap is larger in teacher-assigned grades than in anonymously graded tests. This finding is consistent with a systematic grading differential in favor of girls (e.g., Lavy 2008; Kiss 2013; Terrier 2020; Contreras 2024). The magnitude of this differential varies markedly across subjects: it is largest in German (0.27 SD) and Math (0.24 SD), more moderate in English (0.16 SD), and smaller in the science subjects (0.07 to 0.11 SD). On the original grade scale, these estimates correspond to an advantage of about 0.07 to 0.25 grade points for female students.[5] Robustness checks using alternative OLS and IV specifications yield similar qualitative patterns (see Appendix Tables A.2 and A.3).

The pronounced differential in German may reflect gender stereotypes regarding verbal proficiency: if teachers associate language ability more strongly with female students, they may unconsciously evaluate girls' work more favorably. By contrast, the large female-favoring differential in Math, a traditionally male-stereotyped domain, is particularly noteworthy. This pattern may reflect two related mechanisms. First, Math achievement depends more heavily on continuous practice and homework completion, areas in which female students tend to display stronger non-cognitive skills (Cornwell et al. 2013; Contreras 2024). Second, teachers may evaluate girls particularly favorably when they succeed in a counter-stereotypical domain (Carlana 2019; Lievore and Triventi 2023).

3.2 Heterogeneity Analysis

Table 2 reports heterogeneity in the DiD coefficient on *Female* × *Non-blind* by school track (Panel A) and teacher experience (Panel B).[6] Each cell reports the coefficient from regressions estimated separately by subgroup and subject.

Panel A shows that the positive female–male grading differential is consistently larger in the academic track (*Gymnasium*) than in the other school tracks across all six subjects. The difference ranges from about 0.08 to 0.13 SD (or 0.07–0.12 grade points) and is

[5]Grade-point effects are calculated by multiplying the coefficients in SD units by the subject-specific standard deviation of the rescaled grade variable.

[6]Additional heterogeneity by teacher gender, teacher age, and classroom gender composition is reported in Appendix Table A.4.

statistically significant according to the equality tests reported in the table. This finding highlights an institutional dimension rarely examined in prior work, possibly reflecting more demanding and multidimensional assessment standards in *Gymnasium* that leave greater scope for subjective evaluation. This result is substantively important because grades shape students' subsequent educational trajectories at multiple stages, including school-track choice, access to higher education, and the choice of field of study (Terrier 2020; Lavy and Megalokonomou 2024).

Table 2: Heterogeneity in Grading Differential by School Track and Teacher Experience

	Humanities Subjects		**STEM Subjects**			
	German (1)	**English** (2)	**Math** (3)	**Biology** (4)	**Chem.** (5)	**Physics** (6)
Panel A: School track						
Gymnasium	0.359***	0.275***	0.334***	0.154***	0.190***	0.176***
	(0.024)	(0.024)	(0.023)	(0.025)	(0.024)	(0.027)
Other tracks	0.275***	0.147***	0.227***	0.062**	0.084***	0.083***
	(0.025)	(0.024)	(0.021)	(0.027)	(0.027)	(0.026)
Diff. (Gym. − Other)	0.083**	0.128***	0.107***	0.093**	0.106***	0.093**
Panel B: Teachers' experience						
Novice	0.359***	0.233***	0.455***	0.169**	0.134**	0.104
	(0.055)	(0.076)	(0.057)	(0.078)	(0.061)	(0.084)
Experienced	0.250***	0.175***	0.239***	0.092***	0.132***	0.139***
	(0.024)	(0.022)	(0.021)	(0.028)	(0.028)	(0.028)
Diff. (Novice − Exp.)	0.110*	0.058	0.216***	0.077	0.003	−0.036

Notes: Each cell reports the coefficient on *Female* × *Non-blind* from separate DiD regressions estimated for the indicated subgroup. Standard errors, reported in parentheses, are clustered at the class level. The rows labeled Diff. report the difference in coefficients between the two subgroup estimates within each subject. Statistical significance in these rows refers to the null hypothesis that the coefficients are equal across subgroups. All regressions are weighted using student sampling weights. Full regression results are reported in Appendix Tables A.5–A.8. Significance levels: *** 1%, ** 5%, * 10%.

Panel B shows additional heterogeneity by teachers' years of experience. I distinguish between teachers with less than two years of experience and those with at least two years, capturing the earliest career stage when grading practices may still be developing. The estimated female–male grading differential is positive for both groups, but it is larger among novice teachers, especially in German and Math. The pattern is most pronounced in Math, where the differential increases from 0.24 SD among experienced teachers to 0.46 SD among novice teachers, a difference of 0.22 SD (roughly 0.23 grade points). By contrast, experience-related differences are smaller and statistically insignificant in English and the science subjects. These results are consistent with Terrier (2020), who also finds larger gender grade–test differentials among less experienced Math teachers, but no corresponding difference in French. One possible interpretation is that grading practices become more standardized with teaching experience, although the estimates may also capture other differences between novice and experienced teachers.

4 Conclusion

Using a difference-in-differences design, I show that the female–male gap is systematically larger in teacher-assigned grades than in anonymously graded test scores across six core subjects in German secondary schools. The estimated differential is largest in German and Math, consistently larger in the academic track, and particularly pronounced among novice teachers.

These findings suggest that non-blind assessment may contribute to gender differences in recorded achievement beyond what is reflected in anonymously graded test performance. While the analysis cannot fully disentangle teacher bias from the differential rewarding of non-cognitive skills such as effort and compliance, the systematic pattern across subjects, tracks, and experience levels points to a structural feature of non-blind assessment. More broadly, the results highlight early-career teachers as a potentially important margin for mentoring, grading guidance, and other efforts to standardize assessment practices, especially in educational contexts where grades are highly consequential.

References

Bredtmann, J., Otten, S., Vonnahme, C., 2024. Discrimination in grading? Evidence on teachers' evaluation bias towards minority students. Ruhr Economic Papers No. 1122. doi:10.4419/96973304.

Carlana, M., 2019. Implicit stereotypes: Evidence from teachers' gender bias. The Quarterly Journal of Economics 134, 1163–1224. doi:10.1093/qje/qjz008.

Contreras, D., 2024. Gender differences in grading: Teacher bias or student behaviour? Education Economics 32, 762–785. doi:10.1080/09645292.2023.2252620.

Cornwell, C., Mustard, D.B., Van Parys, J., 2013. Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. Journal of Human Resources 48, 236–264. doi:10.3368/jhr.48.1.236.

Di Liberto, A., Casula, L., Pau, S., 2022. Grading practices, gender bias and educational outcomes: Evidence from Italy. Education Economics 30, 481–508. doi:10.1080/09645292.2021.2004999.

Falch, T., Naper, L.R., 2013. Educational evaluation schemes and gender gaps in student achievement. Economics of Education Review 36, 12–25. doi:10.1016/j.econedurev.2013.05.002.

Hinnerich, B.T., Höglin, E., Johannesson, M., 2011. Are boys discriminated in Swedish high schools? Economics of Education Review 30, 682–690. doi:10.1016/j.econedurev.2011.02.007.

Jansson, J., Tyrefors, B., 2025. The effect of an anonymous grading reform for male and female university students. Economics Letters 248, 112241. doi:10.1016/j.econlet.2025.112241.

Kiss, D., 2013. Are immigrants and girls graded worse? Results of a matching approach. Education Economics 21, 447–463. doi:10.1080/09645292.2011.585019.

Lavy, V., 2008. Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. Journal of Public Economics 92, 2083–2105. doi:10.1016/j.jpubeco.2008.02.009.

Lavy, V., Megalokonomou, R., 2024. The short- and the long-run impact of gender-biased teachers. American Economic Journal: Applied Economics 16, 176–218. doi:10.1257/app.20210052.

Lievore, I., Triventi, M., 2023. Do teacher and classroom characteristics affect the way in which girls and boys are graded? British Journal of Sociology of Education 44, 97–122. doi:10.1080/01425692.2022.2122942.

Machin, S., Pekkarinen, T., 2008. Global sex differences in test score variability. Science 322, 1331–1332. doi:10.1126/science.1162573.

Nennstiel, R., Gilgen, S., 2024. Does chubby Can get lower grades than skinny Sophie? Using an intersectional approach to uncover grading bias in German secondary schools. PLOS ONE 19, 1–23. doi:10.1371/journal.pone.0305703.

OECD, 2025a. Education at a Glance 2025: OECD Indicators. OECD Publishing, Paris. doi:10.1787/1c0d9c79-en.

OECD, 2025b. Gender Equality in a Changing World: Taking Stock and Moving Forward, Gender Equality at Work. OECD Publishing, Paris. doi:10.1787/e808086f-en.

Terrier, C., 2020. Boys lag behind: How teachers' gender biases affect student achievement. Economics of Education Review 77, 101981. doi:10.1016/j.econedurev.2020.101981.

Appendix

A. Tables

Table A.1: Descriptive Statistics for the German Analysis Sample

	Male students		Female students	
	Mean	SD	Mean	SD
German grade	3.774	0.825	4.201	0.836
English grade	3.774	0.903	4.069	0.937
Math grade	3.864	1.040	3.944	1.054
Biology grade	4.060	0.951	4.315	0.929
Chemistry grade	4.021	1.016	4.173	0.992
Physics grade	4.018	0.988	4.097	0.980
Age	15.467	0.665	15.347	0.628
Migration background	0.276	0.447	0.286	0.452
Mother's education				
Low education	0.153	0.360	0.159	0.365
Medium education	0.399	0.490	0.420	0.493
High education	0.332	0.471	0.317	0.465
Missing	0.115	0.319	0.105	0.306
Father's education				
Low education	0.131	0.338	0.142	0.349
Medium education	0.373	0.484	0.387	0.487
High education	0.374	0.484	0.352	0.478
Missing	0.121	0.326	0.118	0.323
Mother's occupation				
White collar	0.578	0.494	0.595	0.491
Blue collar	0.128	0.335	0.130	0.336
Other	0.159	0.365	0.157	0.364
Missing	0.135	0.342	0.118	0.323
Father's occupation				
White collar	0.471	0.499	0.479	0.500
Blue collar	0.199	0.399	0.202	0.401
Other	0.200	0.400	0.199	0.399
Missing	0.131	0.337	0.120	0.325
Number of books at home	179.648	168.097	184.502	165.006
Share Gymnasium	0.357	0.479	0.399	0.490
Share of novice teachers	0.096	0.295	0.112	0.315
Observations	24,751		24,797	

Notes: The table reports means and standard deviations for the variables in the German analysis sample. Because the analysis sample varies by subject, summary statistics for grades in the other subjects are based on the respective subject-specific analysis samples. Summary statistics are weighted using student sampling weights.

Table A.2: Gender Differential in Teacher Grades – OLS and IV Estimates I

	German			English			Math		
	OLS	1^{st} St.	2^{nd} St.	OLS	1^{st} St.	2^{nd} St.	OLS	1^{st} St.	2^{nd} St.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	0.399***	0.112***	0.362***	0.227***	−0.030***	0.212***	0.309***	−0.243***	0.369***
	(0.014)	(0.008)	(0.014)	(0.013)	(0.008)	(0.013)	(0.043)	(0.022)	(0.036)
Test score	0.586***	–	0.853***	0.737***	–	1.043***	0.842***	–	1.051***
	(0.010)		(0.018)	(0.010)		(0.017)	(0.032)		(0.047)
Instrument	–	0.525***	–	–	0.576***	–	–	0.647***	–
		(0.006)			(0.007)			(0.018)	
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes
Class FE	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	48,252	48,252	48,252	48,252	48,252	48,252	6,462	6,462	6,462
Clusters	2,463	2,463	2,463	2,463	2,463	2,463	2,414	2,414	2,414
Adjusted R^2	0.379	0.748	0.220	0.414	0.728	0.256	0.431	0.777	0.325
F-stat. (IV)		6,771			7,563			1,251	

Notes: The table reports coefficient estimates from OLS and IV regressions for each subject separately. The OLS columns report estimates from regressions of teacher-assigned grades on *Female*, the corresponding external test score, the full set of controls, and class fixed effects. The 1^{st} Stage columns report estimates from first-stage regressions of the subject-specific test score on *Female*, the excluded instrument, the full set of controls, and class fixed effects. The 2^{nd} Stage columns report IV estimates from the corresponding second-stage regressions of teacher-assigned grades on *Female* and the instrumented test score. The excluded instruments are the English test score for German, the German test score for English, and the Chemistry test score for Math. The smaller Math sample reflects limited instrument availability, as students typically took either the external Math test or the science test, but not both. In addition, the language tests were administered in different survey waves and therefore do not provide a better alternative instrument. The OLS estimates are reported for the same sample as the corresponding IV regressions. The row "F-stat. (IV)" reports the first-stage F-statistic for the excluded instrument. Standard errors, reported in parentheses, are clustered at the class level. All regressions are weighted using student sampling weights. Significance levels: *** 1%, ** 5%, * 10%.

Table A.3: Gender Differential in Teacher Grades – OLS and IV Estimates II

	Biology			Chemistry			Physics		
	OLS	1^{st} St.	2^{nd} St.	OLS	1^{st} St.	2^{nd} St.	OLS	1^{st} St.	2^{nd} St.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	0.177***	0.147***	0.171***	0.130***	0.056***	0.132***	0.090***	−0.067***	0.094***
	(0.018)	(0.004)	(0.017)	(0.017)	(0.004)	(0.016)	(0.017)	(0.004)	(0.016)
Test score	0.542***	–	0.604***	0.644***	–	0.688***	0.648***	–	0.688***
	(0.013)		(0.014)	(0.013)		(0.013)	(0.012)		(0.013)
Instrument	–	0.906***	–	–	0.928***	–	–	0.920***	–
		(0.004)			(0.004)			(0.004)	
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes
Class FE	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	29,743	29,743	29,743	30,741	30,741	30,741	30,741	30,741	30,741
Clusters	2,909	2,909	2,909	2,976	2,976	2,976	2,976	2,976	2,976
Adjusted R^2	0.340	0.936	0.184	0.359	0.940	0.224	0.361	0.943	0.228
F-stat. (IV)		66,657			58,393			64,421	

Notes: The table reports coefficient estimates from OLS and IV regressions for each subject separately. The OLS columns report estimates from regressions of teacher-assigned grades on *Female*, the corresponding external test score, the full set of controls, and class fixed effects. The 1^{st} Stage columns report estimates from first-stage regressions of the subject-specific test score on *Female*, the excluded instrument, the full set of controls, and class fixed effects. The 2^{nd} Stage columns report IV estimates from the corresponding second-stage regressions of teacher-assigned grades on *Female* and the instrumented test score. The excluded instruments are the Chemistry test score for Biology, the Physics test score for Chemistry, and the Chemistry test score for Physics. The OLS estimates are reported for the same sample as the corresponding IV regressions. The row "F-stat. (IV)" reports the first-stage F-statistic for the excluded instrument. Standard errors, reported in parentheses, are clustered at the class level. All regressions are weighted using student sampling weights. Significance levels: *** 1%, ** 5%, * 10%.

Table A.4: Heterogeneity in Grading Differential by Teacher Characteristics and Class Composition

	Humanities Subjects		**STEM Subjects**			
	German (1)	**English** (2)	**Math** (3)	**Biology** (4)	**Chem.** (5)	**Physics** (6)
Panel A: Teachers' gender						
Female	0.287***	0.203***	0.286***	0.108***	0.174***	0.122***
	(0.027)	(0.023)	(0.027)	(0.031)	(0.031)	(0.045)
Male	0.198***	0.110**	0.234***	0.082	0.074	0.151***
	(0.040)	(0.047)	(0.030)	(0.051)	(0.046)	(0.033)
Diff. (Female − Male)	0.089*	0.093*	0.052	0.027	0.100*	−0.029
Panel B: Teachers' age						
Age below 35 years	0.312***	0.175***	0.333***	0.174***	0.224***	0.101*
	(0.039)	(0.042)	(0.035)	(0.051)	(0.057)	(0.061)
Age above 35 years	0.230***	0.188***	0.234***	0.073**	0.107***	0.144***
	(0.028)	(0.025)	(0.024)	(0.030)	(0.029)	(0.029)
Diff. (Below − Above)	0.082*	−0.013	0.099**	0.100*	0.117*	−0.043
Panel C: Share of girls in class						
Below 40% girls	0.359***	0.229***	0.260***	0.112***	0.200***	0.100**
	(0.033)	(0.032)	(0.036)	(0.041)	(0.037)	(0.039)
Above 40% girls	0.284***	0.166***	0.268***	0.092***	0.107***	0.114***
	(0.020)	(0.020)	(0.018)	(0.023)	(0.023)	(0.024)
Diff. (Below − Above)	0.075*	0.062*	−0.008	0.020	0.094**	−0.014

Notes: Each cell reports the coefficient on *Female* × *Non-blind* from separate DiD regressions estimated for the indicated subgroup. Standard errors, reported in parentheses, are clustered at the class level. The rows labeled Diff. report the difference in coefficients between the two subgroup estimates within each subject. Statistical significance in these rows refers to the null hypothesis that the coefficients are equal across subgroups. All regressions are weighted using student sampling weights. Significance levels: *** 1%, ** 5%, * 10%.

Table A.5: Heterogeneity in Grading Differential by School Track – Full Results I

	German		English		Math	
	Gym. (1)	**Other** (2)	**Gym.** (3)	**Other** (4)	**Gym.** (5)	**Other** (6)
Female	0.166***	0.182***	0.083***	0.118***	−0.229***	−0.181***
	(0.015)	(0.022)	(0.018)	(0.022)	(0.018)	(0.018)
Non-blind	−0.740***	0.203***	−0.690***	0.257***	−0.817***	0.307***
	(0.025)	(0.026)	(0.026)	(0.029)	(0.022)	(0.022)
Female × Non-blind	0.359***	0.275***	0.275***	0.147***	0.334***	0.227***
	(0.024)	(0.025)	(0.024)	(0.024)	(0.023)	(0.021)
Observations	39,598	59,498	38,594	58,036	28,560	47,440
Clusters	899	1,618	878	1,591	1,222	2,533
Adjusted R^2	0.140	0.067	0.117	0.043	0.126	0.057

Notes: The table reports coefficient estimates from Eq. (1), estimated separately by subject and school track. Within each subject, coefficients are shown for students attending Gymnasium and for students attending other school tracks. Standard errors, reported in parentheses, are clustered at the class level. All regressions are weighted using student sampling weights. Significance levels: *** 1%, ** 5%, * 10%.

Table A.6: Heterogeneity in Grading Differential by School Track – Full Results II

	Biology		Chemistry		Physics	
	Gym. (1)	**Other** (2)	**Gym.** (3)	**Other** (4)	**Gym.** (5)	**Other** (6)
Female	0.103***	0.191***	−0.053***	0.071***	−0.119***	0.006
	(0.018)	(0.020)	(0.019)	(0.023)	(0.019)	(0.023)
Non-blind	−0.541***	0.405***	−0.623***	0.435***	−0.589***	0.433***
	(0.024)	(0.025)	(0.025)	(0.029)	(0.024)	(0.026)
Female × Non-blind	0.154***	0.062**	0.190***	0.084***	0.176***	0.083***
	(0.025)	(0.027)	(0.024)	(0.027)	(0.027)	(0.026)
Observations	25,602	37,578	27,988	38,848	27,264	37,530
Clusters	1,096	2,044	1,198	2,084	1,168	2,034
Adjusted R^2	0.081	0.068	0.082	0.070	0.075	0.066

Notes: The table reports coefficient estimates from Eq. (1), estimated separately by subject and school track. Within each subject, coefficients are shown for students attending Gymnasium and for students attending other school tracks. Standard errors, reported in parentheses, are clustered at the class level. All regressions are weighted using student sampling weights. Significance levels: *** 1%, ** 5%, * 10%.

Table A.7: Heterogeneity in Grading Differential by Teacher Experience – Full Results I

	German		English		Math	
	Novice (1)	**Exp.** (2)	**Novice** (3)	**Exp.** (4)	**Novice** (5)	**Exp.** (6)
Female	0.228***	0.256***	0.163**	0.152***	−0.239***	−0.166***
	(0.056)	(0.021)	(0.068)	(0.021)	(0.052)	(0.020)
Non-blind	−0.537***	−0.140***	−0.296***	−0.155***	−0.535***	−0.166***
	(0.077)	(0.030)	(0.076)	(0.032)	(0.064)	(0.026)
Female × Non-blind	0.359***	0.250***	0.233***	0.175***	0.455***	0.239***
	(0.055)	(0.024)	(0.076)	(0.022)	(0.057)	(0.021)
Observations	6,040	62,206	6,530	58,518	4,560	50,370
Clusters	140	1,558	158	1,452	230	2,462
Adjusted R^2	0.088	0.041	0.033	0.018	0.035	0.005

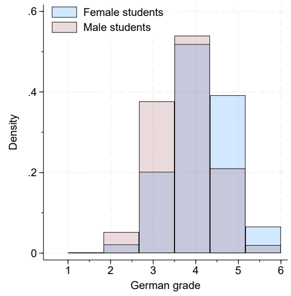
Notes: The table reports coefficient estimates from Eq. (1), estimated separately by subject and teacher experience. Within each subject, coefficients are shown for classes taught by teachers with less than 2 years of experience and for classes taught by teachers with 2 or more years of experience. Standard errors, reported in parentheses, are clustered at the class level. All regressions are weighted using student sampling weights. Significance levels: *** 1%, ** 5%, * 10%.

Table A.8: Heterogeneity in Grading Differential by Teacher Experience – Full Results II

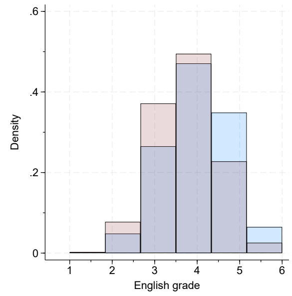
	Biology		Chemistry		Physics	
	Novice (1)	**Exp.** (2)	**Novice** (3)	**Exp.** (4)	**Novice** (5)	**Exp.** (6)
Female	0.139**	0.209***	−0.008	0.032	−0.013	−0.017
	(0.056)	(0.025)	(0.052)	(0.024)	(0.081)	(0.024)
Non-blind	−0.191***	−0.113***	−0.108	−0.164***	−0.061	−0.141***
	(0.071)	(0.030)	(0.074)	(0.033)	(0.095)	(0.029)
Female × Non-blind	0.169**	0.092***	0.134**	0.132***	0.104	0.139***
	(0.078)	(0.028)	(0.061)	(0.028)	(0.084)	(0.028)
Observations	4,226	37,904	3,950	41,352	3,148	39,894
Clusters	211	1,845	194	1,979	157	1,923
Adjusted R^2	0.017	0.018	0.002	0.006	0.000	0.003

Notes: The table reports coefficient estimates from Eq. (1), estimated separately by subject and teacher experience. Within each subject, coefficients are shown for classes taught by teachers with less than 2 years of experience and for classes taught by teachers with 2 or more years of experience. Standard errors, reported in parentheses, are clustered at the class level. All regressions are weighted using student sampling weights. Significance levels: *** 1%, ** 5%, * 10%.

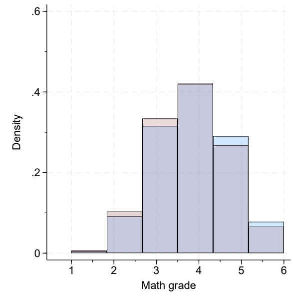
B. Figures



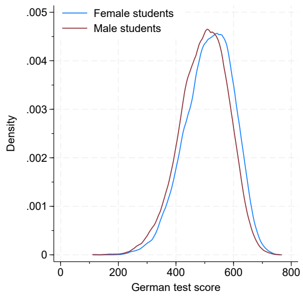
(a) German grade



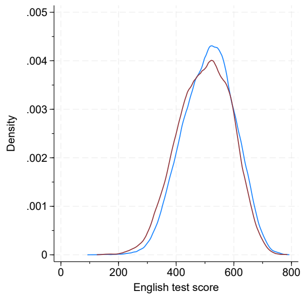
(b) English grade



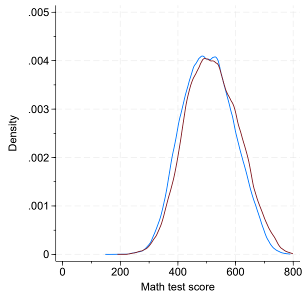
(c) Math grade



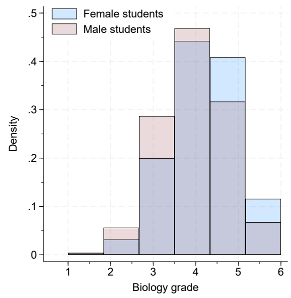
(d) German test score



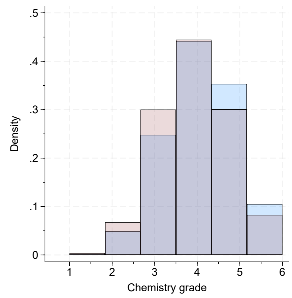
(e) English test score



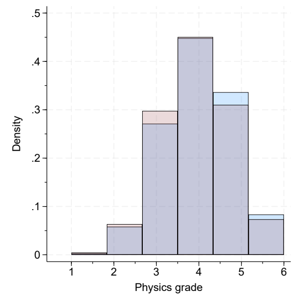
(f) Math test score



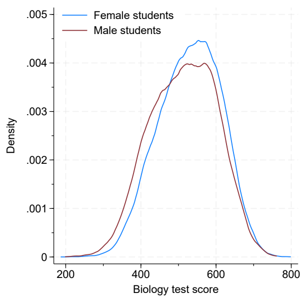
(g) Biology grade



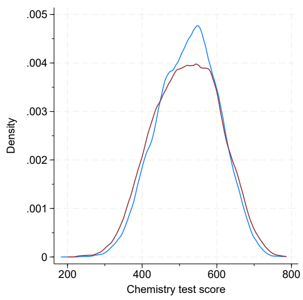
(h) Chemistry grade



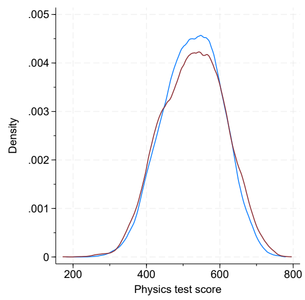
(i) Physics grade



(j) Biology test score



(k) Chemistry test score



(l) Physics test score

Figure B.1: Distribution of Grades and Test Scores

Notes: The figure shows the distribution of grades and test scores for each of the six subjects, separated by gender. While teacher-assigned grades are measured on a scale from 1 to 6, with higher values indicating better performance, test scores are measured on a continuous scale with a mean of 500 and a standard deviation of 100.

C. Data Access

This paper uses data provided by the Research Data Center at the Institute for Educational Quality Improvement (FDZ at IQB). The analysis draws on two waves of the IQB National Assessment Study (*IQB Ländervergleich*: 2008/2009 and 2012) and two waves of the IQB Trends in Student Achievement (*IQB Bildungstrend*: 2015 and 2018). These data are available from the FDZ at IQB as anonymized Scientific Use Files. Access requires a formal application to the FDZ at IQB and is subject to the applicable data use agreement.

The corresponding data set references are Köller et al. (2011), Pant et al. (2015), Stanat et al. (2018), and Stanat et al. (2022). Detailed study documentation is provided in Sachse et al. (2012), Lenski et al. (2016), Schipolowski et al. (2018), and Becker et al. (2022).

References for the Appendix

Becker, B., Busse, J., Holtmann, M., Weirich, S., Schipolowski, S., Mahler, N., Stanat, P., 2022. IQB-Bildungstrend 2018. Institut zur Qualitätsentwicklung im Bildungswesen, Berlin. doi:10.18452/25472.

Köller, O., Knigge, M., Tesch, B., 2011. IQB Ländervergleich Sprachen 2008/2009 [IQB National Assessment Study 2008/2009] (IQB-LV 2008-9) (Version 2) [Data set]. Institut zur Qualitätsentwicklung im Bildungswesen, Berlin. doi:10.5159/IQB_LV_2008_v2.

Lenski, A.E., Hecht, M., Penk, C., Milles, F., Mezger, M., Heitmann, P., Stanat, P., Pant, H.A., 2016. IQB-Ländervergleich 2012. Institut zur Qualitätsentwicklung im Bildungswesen, Berlin. doi:10.20386/HUB-42547.

Pant, H.A., Stanat, P., Hecht, M., Heitmann, P., Jansen, M., Lenski, A.E., Penk, C., Pöhlmann, C., Roppelt, A., Schroeders, U., Siegle, T., 2015. IQB-Ländervergleich Mathematik und Naturwissenschaften 2012 [IQB National Assessment Study 2012] (IQB-LV 2012) (Version 4) [Data set]. Institut zur Qualitätsentwicklung im Bildungswesen, Berlin. doi:10.5159/IQB_LV_2012_v4.

Sachse, K.A., Kretschmann, J., Kocaj, A., Köller, O., Knigge, M., Tesch, B., 2012. IQB-Ländervergleich 2008/2009. Institut zur Qualitätsentwicklung im Bildungswesen, Berlin. doi:10.18452/3126.

Schipolowski, S., Haag, N., Milles, F., Pietz, S., Stanat, P., 2018. IQB-Bildungstrend 2015. Institut zur Qualitätsentwicklung im Bildungswesen, Berlin. doi:10.18452/19997.

Stanat, P., Böhme, K., Schipolowski, S., Haag, N., Weirich, S., Sachse, K., Hoffmann, L., Federlein, F., 2018. IQB-Bildungstrend Sprachen 2015 [IQB Trends in Student Achievement 2015] (IQB-BT 2015) (Version 5) [Data set]. Institut zur Qualitätsentwicklung im Bildungswesen, Berlin. doi:10.5159/IQB_BT_2015_v5.

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S., Henschel, S., Holtmann, M., Becker, B., Kölm, J., 2022. IQB-Bildungstrend Mathematik und Naturwissenschaften 2018 [IQB Trends in Student Achievement 2018] (IQB-BT 2018) (Version 1) [Data set]. Institut zur Qualitätsentwicklung im Bildungswesen, Berlin. doi:10.5159/IQB_BT_2018_v1.