Gender Differentials in Test Scores and Teacher Assessments: Evidence from Germany

Bernhard Enzi*

Preliminary draft. Do not cite or print without permission.

Abstract

Are there systematic differences in the way teachers grade their male and female students conditional on the same performance? Using rich micro-level data from the German National Educational Panel Study (NEPS) and applying fixed-effects estimators to account for unobserved heterogeneity, I can show that girls are graded worse in mathematics, while boys are in German conditional on performing the same in standardized tests. No such gender gap exists for science. The findings are robust to several specifications and cannot be explained by non-cognitive skills, teacher characteristics or in-class activities.

JEL-Codes: I2, J16

Keywords: gender, grading

*Ifo Institute – Leibniz Institute for Economic Research at the University of Munich. Center for the Economics of Education and Innovation. Poschingerstrasse 5, 81679 Muenchen. Email: enzi@ifo.de

1 Introduction

Are there systematic differences in the way teachers grade their male and female students conditional on the same performance? Experimental[1] and observational[2] studies have shown that boys and children with a migration background tend to be graded worse conditional on the same performance.

Investigating school grade differentials conditional on the same performance is important for various reasons. First, Altonji and Pierret (2001) have shown that high school grades are highly correlated with wages at labor market entry. Hence, systematic differences in grading schemes that are not caused by actual performance differences may induce wages that are not reflecting productivity discrepancies but factors that an employer might not want to take into account at the employment decision. These avoidable uncertainties might induce inefficiencies. Second, systematic grading differences by a student's gender may explain the gender role reversal in education over the past decade, as Goldin, Katz and Kuziemko (2006) show that there exist advantages for females in the US school environment.

4th and 5th class grades can be even more consequential than the ones from higher school years in a tracked school system. In most German states,[3] grades in these years determine the secondary school track and, thereby, future career paths as only the highest secondary school track provides direct, unrestricted access to tertiary education.

Using the rich data set of the German National Educational Panel Study (NEPS), I investigate the relationship of 5th and 6th class students' genders and their previous year's grades in math, science and German. NEPS consists of extensive questionnaires for students, parents, teachers and school principals that allow me to control for many determining factors of grades. Besides grades and basic characteristics (e.g. gender, age, migration status and socio-economic status), it includes information about life satisfaction, intelligence, leisure time activities and non-cognitive skills. Most importantly, it includes objective measures of performances in math, science and German.

Using fixed effects estimators to account for unobserved heterogeneities, I find indications of subject specific grading by a student's gender. While girls are, conditional on all controls and a classroom fixed effect (FE), advantaged by 22.2% of a standard deviation (SD) in German, they are disadvantaged by 19.7% of a SD in math relative to boys. No significant gender gap exists for science. These findings are robust to many different specifications. Investigating whether the gender gap can be explained by heterogeneous teacher effects, I find no altering effects by teacher characteristics (e.g. migrant status, gender) or different in-class time use.

The remainder of the paper is structured as follows: the following section presents the data. Section 3 describes the estimation strategy and presents headline results and their discussion. Section 4 concludes.

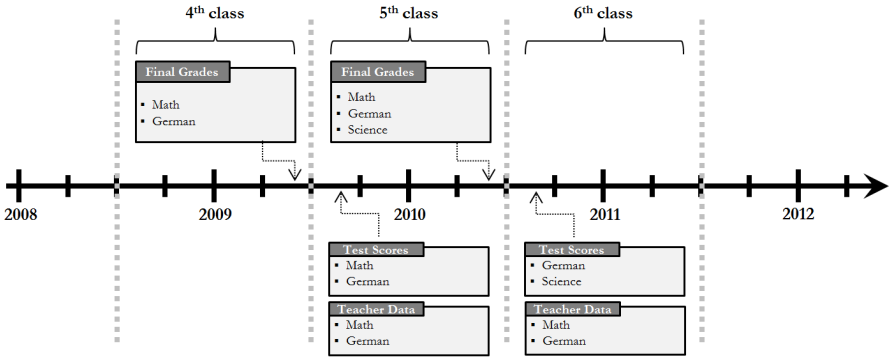
[1] E.g. Hanna and Linden (2009), Hinnerich, Hoeglin and Johannesson (2011a, 2011b) and Sprietsma (2013).
[2] E.g. Burgess and Greaves (2013), Cornwell, Mustard and Van Parys (2013) and Lavy (2008).
[3] See Lohmar and Eckhardt (2010) for a detailed description of the German school system.

2 Data

To identify and explain gender grade differentials, I use data from the 2010 and 2011 waves of the National Educational Panel Study (NEPS) on a cohort that was first tracked in 5th class. NEPS data was collected via a stratified sampling procedure: At first, a random sample of schools was drawn. Within those schools, up to two random classes were selected to participate.

Therefore, the data used in this study consists of student observations in 5th and 6th class. The students were tested in mathematics, science and German and were also asked about their last final grades in these subjects. Figure (1) illustrates the timing and availability of the respective data in each class for this cohort.

Figure 1: Illustration of data availability and timeline of testing and grading



To investigate the relationship between a student's gender and his or her grades while conditioning on performance, I use pairs of last final grades and test scores from the beginning of the subsequent year. As grade and test score are solely divided by school holidays, they are based on the same underlying performance. One of these pairs is available for math and science, two for German. The only pair that can be linked to the respective teacher is the pair of 5th class final grade in German and the 6th class test score.

I limit the analysis to students who were taught by one teacher in each subject and to those that did not require any form of special education. Apart from testing, students, parents, principals, as well as German and math teachers were extensively asked about background information, which allows me to control for many other determining factors of grades. Participation for each of these individuals was voluntary and about 5% of the students did not participate in the testing.[4]

All grades and test scores are standardized to have a mean of zero and a standard deviation of one with higher values indicating better performance. Table (1) presents some first descriptive evidence, showing firstly, that boys are on average better in all standardized tests while they are only graded favorably in mathematics. Secondly, boys tend to show less beneficial social behavior compared to girls.

[4]Testing and questioning took place at the same time for this cohort, which is why I cannot investigate if non-participation was random. Conducting such a test for a different cohort shows that test participation was not random conditional on observables.

Table 1: Descriptive statistics on students' test scores, grades and background

	Female				Male			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Test scores and grades								
Math grade	-0.108	(0.981)	-3.8	1.5	0.106	(1.004)	-3.8	1.5
Math test score	-0.127	(0.994)	-4	3.5	0.128	(0.991)	-3	3.5
German grade	0.135	(0.959)	-4.1	1.7	-0.127	(1.018)	-4.1	1.7
German test score	-0.017	(0.904)	-3.9	2.7	0.025	(0.893)	-5.1	3.1
Science grade	0.020	(0.958)	-4.3	1.4	-0.016	(1.025)	-4.3	1.4
Scientific test score	-0.096	(0.964)	-2.6	6.7	0.096	(1.025)	-3.3	6.7
Student background								
DGCF (perceptual speed)	0.106	(0.981)	-3.2	3.6	-0.083	(1.007)	-3.3	3.6
DGCF (reasoning)	-0.025	(0.987)	-2.6	1.9	0.055	(1.006)	-2.6	1.9
Age	11.735	(0.779)	6.2	15	11.836	(0.807)	9.3	17
Born in Germany	0.964	(0.187)	0	1	0.955	(0.207)	0	1
Household size	4.424	(1.384)	0	15	4.463	(1.408)	0	15
PC availability	1.344	(0.527)	0	2	1.408	(0.531)	0	2
Life satisfaction	7.877	(2.301)	0	10	7.856	(2.336)	0	10
Freq: Sports	3.785	(1.051)	1	5	4.061	(1.085)	1	5
Freq: Reading	3.270	(1.345)	1	5	2.852	(1.405)	1	5
Number of books	3.859	(1.356)	1	6	3.824	(1.447)	1	6
Share of migrants in school	3.153	(1.290)	1	7	2.976	(1.326)	1	7
Friends' care for school	2.896	(1.147)	1	5	2.970	(1.220)	1	5
Non-cognitive skills								
SDQ-Scale: Prosocial behavior	7.867	(1.701)	0	10	6.857	(2.009)	0	10
SDQ-Scale: Problem behavior	2.275	(1.813)	0	10	2.548	(1.913)	0	10
Considerate	2.567	(0.516)	1	3	2.331	(0.564)	1	3
Likes to share things	2.600	(0.531)	1	3	2.393	(0.603)	1	3
Loner	1.557	(0.662)	1	3	1.615	(0.708)	1	3
Helpful	2.718	(0.485)	1	3	2.466	(0.589)	1	3
Has friends	2.845	(0.396)	1	3	2.816	(0.443)	1	3
Popular	2.262	(0.636)	1	3	2.305	(0.626)	1	3
Nice to younger children	2.661	(0.518)	1	3	2.457	(0.590)	1	3
Is teased	1.319	(0.580)	1	3	1.426	(0.645)	1	3
Helps others voluntarily	2.319	(0.565)	1	3	2.212	(0.601)	1	3
Gets along better with adults	1.508	(0.619)	1	3	1.642	(0.675)	1	3
Observations	4941				5216			

Note: All variables of the category 'Test scores and grades' are standardized to have a mean of zero and standard deviations of one with higher values indicating better performance. All variables of the category 'Student background' are when not stated otherwise self-explanatory. DGCF variables represent basic cognitive skills with higher values indicating better performance. PC availability is scaled with 0 (No PC), 1 (Shared PC) and 2 (Own PC). Life satisfaction is scaled from 0 (Very unhappy) to 10 (Very happy). Frequency of sports, frequency of reading, number of books, share of migrants and friends' care for school are scaled from 1 (Low, low frequency or low number) to 5, 6 or 7 (High, high frequency or high number). All variables in the 'Non-cognitive skills' section are reported from 1 (Not applicable) to 3 (Clearly applicable), except for the SDQ measures in prosocial and problem behavior that are reported from 0 (Clearly Applicable) to 10 (Not applicable).

3 Empirical framework & results

To investigate the relationship between grades, test scores and gender, I follow the approach of Cornell, Mustard and Van Parys (2013) and model the grade production function in a more general form:

$$grade_i^k = \alpha^k \times female_i + \gamma^k \times testscore_i^k + \lambda^k X_i + \mu_i^k + \tau_j^k + \xi_m^k + \varepsilon_i^k \qquad (1)$$

The grade of student i in subject k is a linear, additive function of a gender indicator, the respective test score and control variables X with their subject specific coefficients α, γ and λ. The unobserved factors comprise a teacher (τ), school (ξ), individual (μ) and idiosyncratic (ε) component. To account for unobserved teacher and school effects I use classroom and school fixed-effects estimators. However, as it is impossible to account for unobserved student heterogeneity with this data, I will use a large set of control variables to minimize omitted variable bias. Table (2) presents the main estimation results.

Only using within school variation and conditioning on standardized test scores in mathematics, science and German, estimation results from setting (1) show that girls are graded less favorably than their male counterparts in mathematics while the opposite is true for German. No gender difference is revealed for science through all settings. These first results may be driven by unobserved

teacher factors, but including a classroom fixed effect in specification (2) does not substantially change the estimates. Additionally controlling for student background characteristics in setting (3) does not alter the results noticeably either.

Setting (4) presents the headline results and also controls for non-cognitive skills as in Cornwell, Mustard and Van Parys (2013). In contrast to their findings, the addition of non-cognitive skills does not explain the gender-grade gap. The set of variables measuring non-cognitive skills includes the students' results in two SDQ questionnaires[5] and information about behavioral features as seen in Table (1). These non-cognitive skill measures plausibly add to the explained variation and take away explanatory power of the test scores in grades, as non-cognitive skills are important for classroom activities weakly correlated with test score performance. However, the magnitude and direction of the gender effect remains the same.

Table 2: Estimated gender gaps in math, science and German grades

	(1)			(2)			(3)			(4)		
	Math	German	Science	Math	German	Science	Math	German	Science	Math	German	Science
Female	-0.185***	0.237***	0.0708	-0.193***	0.238***	0.0612	-0.159***	0.265***	0.0529	-0.197***	0.222***	0.0368
	(0.0271)	(0.0214)	(0.0445)	(0.0285)	(0.0218)	(0.0486)	(0.0381)	(0.0273)	(0.0681)	(0.0446)	(0.0289)	(0.0804)
Test score	0.334***	0.547***	0.222***	0.328***	0.528***	0.225***	0.284***	0.485***	0.191***	0.265***	0.429***	0.194***
	(0.0169)	(0.0212)	(0.0245)	(0.0173)	(0.0226)	(0.0270)	(0.0243)	(0.0282)	(0.0407)	(0.0259)	(0.0317)	(0.0397)
School FE		Yes			Yes			Yes			Yes	
Classroom FE		No			Yes			Yes			Yes	
Student info		No			No			Yes			Yes	
Non-cognitive		No			No			No			Yes	
Observations	4786	9313	2415	4786	9313	2415	2788	5790	1577	2342	4858	1322
R^2	0.419	0.298	0.277	0.461	0.367	0.364	0.534	0.430	0.475	0.571	0.481	0.536

Note: Dependent variable: standardized grade with a mean of zero and a standard deviation of one in the respective subject. The regressions were sparately run with standard errors clustered by school and including the respective fixed effect indicated by 'School FE' or 'Classroom FE'. Standard errors are shown in parenthesis below the coefficients. 'Student info' includes two measures of intelligence as well as age, friends' care for schooling, migrant and socio-economic status, frequency of sports and reading, share of migrants in the classroom, computer availability at home and life satisfaction. 'Non-cognitive' includes results from two SDQ questionnaires and measures of being 'considerate', 'helpful', 'popular', 'nice to younger children' and 'teased by other children', as well as 'getting along better with adults than children', 'being a loner', 'having friends' and 'likes to share'. Stars denote significance of the estimates as follows: ***1%, **5%, and *10%.

These results may still be driven by omitted student variables. A first-difference approach across subjects would account for this unobserved student heterogeneity and, thus, yield unbiased results due to the omission of student variables, but α^{mat} and α^{ger} could not be individually identified anymore. However, their difference, $\Delta\alpha$, still is: Conducting this FD approach as a robustness check yields no statistically different results from the difference of the two gender coefficients in specification (4), evidence for the robustness of this finding.

However, there are a few remaining potential threats to this identification strategy. First, it might be the case that male and female students participate differently in the classroom conditional on the same test score. If classroom participation is determined by performance, but not as a simple linear function of the respective test score, many potential bias scenarios are imaginable. Consider the case in which male students, no matter their actual performance, do not participate in the German classroom, while female students do according to their performance. Relative to girls, boys would get worse grades conditional on the same test score, as classroom participation is an important

[5]See Goodman et al. (2000) for a brief description of SDQ questionnaires, that test for peculiar behavioral traits.

determinant of grades. The estimator related to the female indicator variable would, therefore, be confounded by different classroom participation patterns. Checking for this by including gender - test score interactions does not reveal any significant gender specific test score effects.

Secondly, it may be the case that teachers' grading patterns confound these estimates. As well as in the classroom participation case, many potential biases may arise. Consider for example the case in which an average teacher in German grades on a curve[6] and girls outperform boys. Boys are therefore, holding everything else constant, relatively pushed down the grade distribution, although the underlying performance gap might not suggest so. The female effect in a regression conditioning on test scores would therefore be overestimated. Running regressions on the small subset of students who are taught by the same teacher, thus implicitly assuming that a teacher would use the same grading scheme in both subjects, shows that coefficients remain at the same magnitude, although they are not significantly different from zero anymore due to the small sample size.

Further using teacher data with the German test score - grade combination in 5th class, I cannot find substantially altering effects for the gender estimator. These analyses include differences by teachers' basic traits like age, gender and origin (East and West Germany), as well as teachers' self-reported determinants of final grades in the form of classroom participation, essay writing, dictation, written tests or homework assignments.

4 Conclusion

Conducting an analysis of grade determinants, I find that gender plays a crucial role in grade production. Accounting for several potential identification threats and testing various specifications to explain the gender gap, I find no factor that can do so.

As these results are not driven by unobserved teacher traits and - due to the large set of control variables - may not be by unobserved student heterogeneity, one could, if omitted student variable bias is truly accounted for, interpret them as quasi-causal effects: Solely based on his or her gender, a student might be assessed differently for example through gender stereotype grading of teachers. However, further research is necessary to support this claim.

Future research should a) examine the influence of student-teacher interactions on grade production more thoroughly, b) find ways to account for unobserved student heterogeneity while keeping the gender coefficient identifiable and c) supplementary investigate other potential grading gaps (e.g. migrant status).

[6] According to Becker and Rosen (1992) grading on a curve means that teachers assess students' performances relative to the performances of their classmates.

References

[1] Joseph G. Altonji and Charles R. Pierret. Employer learning and statistical discrimination. *The Quarterly Journal of Economics*, 116(1):313–350, 2001.

[2] William E. Becker and Sherwin Rosen. The Learning Effect of Assessment in High School. *Economics of Education Review*, 11(2):107–118, 1992.

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

[4] Christopher Cornwell, David B. Mustard, and Jessica Van Parys. Noncognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School. *Journal of Human Resources*, 48(1):236–264, 2013.

[5] Claudia Goldin, Lawrence F. Katz, and Ilyana Kuziemko. The Homecoming of American College Women: The Reversal of the College Gender Gap. *NBER Working Paper Series*, 12139, 2006.

[6] Robert Goodman, Tamsin Ford, Helen Simmons, Rebecca Gatward, and Howard Meltzer. Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *The British Journal of Psychiatry*, 177(6):534–539, December 2000.

[7] Rema Hanna and Leigh Linden. Measuring discrimination in education. *NBER Working Paper Series*, 15057, 2009.

[8] Björn Tyrefors Hinnerich, Erik Höglin, and Magnus Johannesson. Are boys discriminated in Swedish high schools? *Economics of Education Review*, 2011.

[9] Björn Tyrefors Hinnerich, Erik Höglin, and Magnus Johannesson. Ethnic Discrimination in High School Grading: Evidence from a Field Experiment. *SSE/EFI Working Paper Series in Economics and Finance*, 733, 2011.

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

[11] B. Lohmar and T. Eckhardt. *The Education System in the Federal Republic of Germany 2008: A Description of the Responsibilities, Structures and Developments in Education Policy for the Exchange of Information in Europe.* Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs, Bonn, 2010.

[12] Maresa Sprietsma. Discrimination in grading: experimental evidence from primary school teachers. *Empirical Economics*, 45(1):523–538, June 2012.