

Piopiunik, Marc; Schlotter, Martin

Working Paper

Identifying the Incidence of "Grading on a Curve": A Within-Student Across-Subject Approach

ifo Working Paper, No. 121

Provided in Cooperation with:

Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Piopiunik, Marc; Schlotter, Martin (2012) : Identifying the Incidence of "Grading on a Curve": A Within-Student Across-Subject Approach, ifo Working Paper, No. 121, ifo Institute - Leibniz Institute for Economic Research at the University of Munich, Munich

This Version is available at:

<https://hdl.handle.net/10419/73836>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

ifo Working Papers

Identifying the Incidence of "Grading on a Curve": A Within-Student Across-Subject Approach

Marc Piopiunik
Martin Schlotter

Ifo Working Paper No. 121

January 2012

An electronic version of the paper may be downloaded from the Ifo website
www.cesifo-group.de.

Identifying the Incidence of "Grading on a Curve": A Within-Student Across-Subject Approach*

Abstract

Theoretical work shows that grading on a curve, i.e., teachers assessing students relative to their classmates, can negatively affect students' learning effort. However, little is known about its empirical incidence. To overcome bias from non-random sorting and omitted variables like teachers' grading standards, we exploit within-student across-subject variation observing both teacher-assigned grades and test scores of German 4th-graders in reading and math. We find that having classmates with one standard deviation higher test scores lowers a student's grade by about 10 percent of a standard deviation. Importantly, only female teachers grade on a curve and there is no association between students' learning effort and relative grading.

JEL Code: I2.

Keywords: Teacher grading, grading on a curve, learning incentives.

Marc Piopiunik
Ifo Institute – Leibniz Institute for
Economic Research
at the University of Munich,
Poschingerstr. 5
81679 Munich, Germany
Phone: +49(0)89/9224-1312
piopiunik@ifo.de

Martin Schlotter
Bavarian Ministry of Economic Affairs,
Infrastructure, Transport and Technology
Prinzregentenstr. 28
81679 Munich, Germany
Phone: +49(0)89/2160-2275
martin.schlotter@stmwivt.bayern.de

* We thank Oliver Falck, David Figlio, Harry Patrinos, Till von Wachter, Joachim Winter, Ludger Woessmann, seminar participants at the University of Munich, the Ifo Institute, the World Bank, as well as participants at the workshop in Trondheim, the EEA meeting in Oslo, the IIPF meeting in Ann Arbor, the German Economic Association meeting in Frankfurt, and at the EALE meeting in Paphos for helpful discussions and comments. All remaining errors are our own.

1 Introduction

This paper is about grading on a curve. Following the grading scheme definition of Becker and Rosen (1992), *grading on a curve*, or *relative grading*, means that teachers assess student performance relative to the performance of their classmates.¹ In contrast, *absolute grading* means that teachers grade student performance based on specific learning criteria, which typically are set by the individual teacher.² Teacher-assigned grades, which are intended to provide information about students' knowledge and skills of a specific subject, are a relevant educational outcome. Labor market entry and educational career might depend on school grades, rather than on (typically unobservable) test scores—a frequently studied outcome in the economics of education literature (see Hanushek (2002) and Hanushek and Woessmann (2011) for overviews). For example, school grades contained in résumés might be an important productivity signal of individuals entering the labor market because employers must assess the quality of their applicants with limited information from résumés, personal interviews, and references. Indeed, employers have been shown to discriminate among workers on the basis of easily observable characteristics that are correlated with productivity, such as years of schooling or educational attainment: at the beginning of the labor market career, wages are strongly related to years of schooling, while they are not related to unobservable cognitive ability (Altonji and Pierret, 2001). Furthermore, school grades might affect the educational career if, for example, the secondary school track choice depends on the grades obtained in primary school (like in Germany). Besides students' ability, students' learning effort, and teachers' grading standards³, it is the grading scheme of the teacher that greatly determine students' school grades. Moreover, theoretical work suggests that the grading scheme can affect students' learning effort (see Landeras, 2009; Bishop, 1999).

In this paper, we provide causal evidence of the incidence of relative grading (as opposed to absolute grading). Employing a within-student across-subject approach allows eliminating unobservable school-, teacher-, and student-level characteristics that may confound the results of existing studies on this topic. We identify teachers' grading scheme using

¹Throughout the paper, the terms "grading on a curve" and "relative grading" are used interchangeably.

²*Relative* grading, for example, means that the best 10 percent of students in the classroom receive the best grade, the next 10 percent receive the second-best grade, and so on. Under an *absolute* grading scheme, to receive a certain grade, a student must pass a threshold score, or absolute standard, that has been set by the teacher in advance. For example, in a test with a maximum of 100 points, teachers assign the best grade for [100,91] points, the second-best grade for [90,81] points, and so on.

³Empirical studies consistently show that high grading standards—that is, teachers assigning relatively low grades for given performance levels—have large positive effects on student achievement (see Figlio and Lucas, 2004; Betts, 1995; Betts and Grogger, 2003; Bonesrønning, 1999, 2004).

data from the German extension of the Progress in International Reading Literacy Study (PIRLS-E) 2001, which contains both test scores and teacher-assigned grades in reading and math for fourth-grade students. Given the definition above, teachers grade on a curve if a student earns a *lower* grade in the subject in which her classmates' performance is relatively *better*, holding constant the student's own performance.⁴ One could estimate the association between a student's grade and her classmates' performance, controlling for the student's own performance, in an OLS model with a sample that pools data from the two subjects. However, the resulting coefficient on classmates' performance might suffer from bias due to unobserved student and teacher characteristics as well as non-random sorting into schools and classrooms. Such biases occur, for example, when able teachers elicit better student performance but also have higher grading standards (see Figlio and Lucas, 2004), or when principals assign teachers with high grading standards to classrooms with low-performing students. To overcome these potential biases, we restrict the sample to students who have been taught both subjects by the same teacher and difference both grades and test scores across the two subjects reading and math. First-differencing across subjects eliminates all unobserved student-specific and teacher-specific factors that do not differ between subjects. This approach identifies the impact of classmates' academic performance on a student's grade in the same subject.⁵ In addition, using a correlated random effects model, we verify the validity of the overidentification restriction of the first-differenced model, that the impact of classmates' performance on a student's grade is the same in both subjects. Finally, we provide first descriptive evidence on the relationship between relative grading and students' learning effort, measured by students' participation in class.

Our main results show that having classmates with one standard deviation higher average test scores lowers a student's grade by about 10 percent of a standard deviation, given his or her own performance. The effect size is very similar for alternative measures of classmates' performance, such as the median value or the student's rank in the performance distribution of the class. We find that only female teachers, but not male teachers grade on a curve. Exploiting the difference in grading schemes across teacher gender, we find that relative

⁴Note that our setting differs from the traditional peer effects literature. In our model, we include a student's performance as an explanatory variable, whereas peer effects models estimate the effect of classmates' performance on a student's performance. More intuitively, we estimate the effect of classmates' performance on an individual student's grade *after* peers may already have affected the student's performance.

⁵The empirical strategy follows Metzler and Woessmann (2010), who estimate the effect of teacher subject knowledge on student achievement in reading and math, extending the approach of Dee (2005, 2007), who uses variation across different teachers to analyze the effects of teacher characteristics that do not vary across subjects.

grading is not related to lower students' learning effort: students' effort, as measured by students' participation in class, is very similar in classes with female teachers and in classes with male teachers. Even though this finding might suffer from bias due to unobservable gender-specific teacher characteristics, it is in contrast with theoretical reasoning that relative grading provides an incentive for lower average class performance because all students will receive the same grades at less effort (see Bishop, 1999).

To date, the empirical incidence of relative grading has mainly been studied by pedagogical scholars and psychologists, who consistently find that classmates' average test scores are significantly negatively related to a student's grade in the same subject, controlling for the student's own test score. Based on this finding, the authors conclude that teachers grade on a curve (see Milek et al., 2009; Trautwein and Baeriswyl, 2007). However, these studies exploit variation in classmates' performance across teachers, classrooms, and even schools. Thus, the resulting estimates might suffer from biases due to unobserved student and teacher characteristics and non-random sorting into schools and classrooms.

Even if relative grading and absolute grading have the same impact on students' learning effort, as suggested by our descriptive results, teachers' grading schemes are important for other reasons. Grades—and not unobservable test scores—are likely to be relevant for the educational career of an individual. In several German states, for example, the grade point average (GPA) in German and math at the end of primary school—exactly what we observe in our data—has a considerable impact on the secondary school track choice (see Kropf et al., 2010). The latter, in turn, strongly determines the final school degree and has long-lasting effects on access to tertiary education and future labor market outcomes (Dustmann, 2004). Moreover, grades are crucial academic outcomes if a student's GPA must be above a certain threshold to progress to the next grade or to graduate from school. Likewise, admission to post-secondary institutions might be selective and based on the GPA received in secondary school. Furthermore, systematic differences in the way teachers assess students' academic performance also raise a fairness issue that should be of interest for policy makers: students whose teachers grade relatively (female teachers) are assessed with respect to their classmates' performance, whereas students whose teachers grade absolutely (male teachers) receive grades that are independent of the performance of their classmates, given the students' own performance.

The remainder of the paper proceeds as follows. Section 2 provides a brief overview of the school system and teacher grading in Germany. Section 3 discusses our empirical strategy

and Section 4 presents the student dataset and descriptive statistics. Section 5 shows results from OLS and first-differenced models and discusses effect heterogeneity. In Section 6, we provide descriptive evidence on the association between teachers' grading scheme and student's learning effort. Section 7 concludes.

2 The German School System and Teacher Grading

In Germany, children start school in the year after they turn six and attend four grades in primary school (*Grundschule*).⁶ At about age 10, students are allocated to one of three types of secondary school which differ by both duration and curriculum. Basic schools (*Hauptschule*) provide basic general education and lead to a certificate after grade 9.⁷ Middle schools (*Realschule*) provide a more extensive general education and last six years. High schools (*Gymnasium*) offer the most academic curriculum and cover nine grades. A lower-level certificate can be obtained after eight years (*Fachhochschulreife*) that qualifies students to attend a polytechnic (*Fachhochschule*). Basic schools and middle schools are more vocational-oriented. High schools are the only type of secondary school that provide direct entry into tertiary education. The high school leaving certificate (*Abitur*) is a prerequisite for attending university or other institutions of higher education.

Parents' decision as to their child's secondary school track is to a large extent based on a recommendation by primary school teachers. At the end of primary school, students do not take ability tests, which could provide information as to their academic potential, nor are there any centralized examinations, which could facilitate secondary school track decisions. Instead, primary school teachers recommend a secondary school track for each student. This recommendation is mostly based on the student's grades in the two major subjects German and math. The grades in these subjects are primarily based on the results of written exams taken during the school year and graded by the teacher. However, a student's class participation typically also contributes to the final grade in a subject. Importantly, note that educational authorities do not instruct teachers on how they should grade student performance. Therefore, teachers in Germany are free to grade either relatively or absolutely.

⁶In two states, Brandenburg and Berlin, primary school lasts six years.

⁷There are no basic schools in the eastern states of the former German Democratic Republic. In some states, an additional fourth school type—comprehensive schools (*Gesamtschule*)—offers all lower and upper secondary education levels. Where comprehensive schools exist, only a minor fraction of students attends this school type. See Lohmar and Eckhardt (2010) for a more detailed description of the German school system.

The GPA-based secondary school recommendations are binding in some, but not all, German states. School authorities usually define a cutoff for the average grade in German and math that students must achieve to receive a recommendation for a certain secondary school track.⁸ In states where the recommendation is not binding, parents are free to choose the child’s secondary school track. However, parents are likely to be strongly influenced by their children’s grades when making this decision, leading to a very strong correspondence between teacher recommendation and the secondary school track actually chosen. For all states, both those with binding and non-binding teacher recommendations, the secondary school track chosen by the parents coincides with the teacher recommendation 83 percent of the time (see Pietsch and Stubbe, 2007).

Changing secondary school type is possible in theory, but very rare in practice (see Jürges and Schneider, 2007). Therefore, the secondary school track decision is a crucial point in a student’s educational career since it strongly determines his or her final school degree, which, in turn, affects post-secondary education opportunities and future earnings (Dustmann, 2004). Moreover, the track choice affects the formation of cognitive skills: students in high schools accumulate more skills than students in basic schools (during the same time) due to a more beneficial learning environment and a more demanding curriculum (Baumert et al., 2009).

3 The Within-Student Across-Subject Model

Our empirical strategy exploits the fact that the students in our sample are taught by the same teacher in both German and math:

$$y_{ig} = \beta_1 * \bar{S}_g^{(-i)} + \alpha_1 * S_{ig} + \delta_1 * X_{ig} + \gamma_1 * Z_i + \theta_1 * T_t + \mu_i + \tau_t + \epsilon_{ig} + \rho_{tg} \quad (1)$$

$$y_{im} = \beta_2 * \bar{S}_m^{(-i)} + \alpha_2 * S_{im} + \delta_1 * X_{im} + \gamma_1 * Z_i + \theta_1 * T_t + \mu_i + \tau_t + \epsilon_{im} + \rho_{tm} \quad (2)$$

The dependent variables y_{ig} and y_{im} are the teacher-assigned grades of student i in German and math, respectively. We model grades as the outcomes of a set of different inputs:

⁸School grades in Germany range from 1 (very good) to 6 (fail). Students in Baden-Württemberg and Saxony, for example, need an average grade of 2.5 in German and math to receive a recommendation for high school, while in Bavaria students need an average grade of 2.0 (see Kropf et al., 2010, for more details).

$\overline{S}_g^{(-i)}$ and $\overline{S}_m^{(-i)}$ are the average actual performance (measured by PIRLS test scores)⁹ of student i 's classmates in reading and math, the determinant we are most interested in.¹⁰ S_{ig} and S_{im} represent student i 's performance in reading and math. X_{ig} and X_{im} are observable subject-specific characteristics of student i that affect grades. Z_i is a vector of observable subject-invariant characteristics of student i . Apart from student i 's own characteristics, several teacher-specific characteristics might affect students' grades. Restricting our sample to classrooms taught by the same teacher in both subjects, T_t captures all observable teacher characteristics (such as gender, age, and experience), which do not differ across subjects. The error term contains the following components: μ_i and τ_t represent unobservable student- and teacher-specific factors that are identical across subjects; ϵ_{ig} and ϵ_{im} are subject-specific student influences; and ρ_{tg} and ρ_{tm} are unobservable subject-specific teacher terms.

A straightforward way of identifying the effect of student i 's classmates' performance on student i 's grades would be differencing Equations (1) and (2), thereby eliminating confounding unobservable subject-invariant student and teacher characteristics.¹¹ However, this implicitly assumes that $\beta_1 = \beta_2$, that is, classmates' performance affects the grade of student i in German the same way it does in math. This means in our context that the strength of relative grading is the same for both subjects.

Building on the work of Chamberlain (1982), several studies test this assumption by estimating a correlated random effects model (see, for example, Ashenfelter and Krueger, 1994; Ashenfelter and Zimmerman, 1997; Metzler and Woessmann, 2010). We follow this approach and model the subject-invariant student- and teacher-specific error terms μ_i and τ_t , which are potential sources of bias in standard OLS models, as follows:

$$\mu_i = \eta_1 * \overline{S}_g^{(-i)} + \eta_2 * \overline{S}_m^{(-i)} + \alpha_2 * S_{ig} + \alpha_3 * S_{im} + \delta_2 * X_{ig} + \delta_3 * X_{im} + \gamma_2 * Z_i + \theta_2 * T_t + \omega_i \quad (3)$$

$$\tau_t = \psi_1 * \overline{S}_g^{(-i)} + \psi_2 * \overline{S}_m^{(-i)} + \alpha_3 * S_{ig} + \alpha_4 * S_{im} + \delta_3 * X_{im} + \delta_4 * X_{ig} + \gamma_3 * Z_i + \theta_3 * T_t + \sigma_i \quad (4)$$

⁹While the school subject is called *German*, the PIRLS data contain students' performance in *reading*. We assume that this performance measure also reflects students' writing skills since the PIRLS score is based on written answers.

¹⁰In other specifications, we replace classmates' average performance by other measures of classmates' performance, such as the median test score or the rank of student i in the classroom's performance distribution.

¹¹Note that including student and teacher fixed effects yields the same results as first-differencing across the two subjects. Identification with fixed effects across subjects was first introduced by Dee (2005, 2007), who estimates the effect of student-teacher demographic matches and student-teacher gender interactions on student outcomes.

Equations (3) and (4) represent the correlation of μ_i and τ_t with observable student and teacher characteristics. Student-subject-specific characteristics for both subjects appear in both equations. Only the η and ψ coefficients are allowed to vary across subjects. Student i 's test scores S_{ig} and S_{im} and all other subject-specific variables are assumed to have the same influence on μ_i and τ_t , respectively. Plugging Equations (3) and (4) into Equations (1) and (2) yields:

$$\begin{aligned}
y_{ig} = & (\beta_1 + \eta_1 + \psi_1) * \bar{S}_g^{(-i)} + (\eta_2 + \psi_2) * \bar{S}_m^{(-i)} + (\alpha_1 + \alpha_2 + \alpha_3) * S_{ig} + (\alpha_2 + \alpha_3) * \\
& * S_{im} + (\delta_1 + \delta_2 + \delta_3) * X_{ig} + (\delta_2 + \delta_3) * X_{im} + (\gamma_1 + \gamma_2 + \gamma_3) * Z_i + \\
& + (\theta_1 + \theta_2 + \theta_3) * T_t + \epsilon_{ig} + \rho_{tg} + \omega_i + \sigma_i
\end{aligned} \tag{5}$$

$$\begin{aligned}
y_{im} = & (\beta_2 + \eta_2 + \psi_2) * \bar{S}_m^{(-i)} + (\eta_1 + \psi_1) * \bar{S}_g^{(-i)} + (\alpha_1 + \alpha_2 + \alpha_3) * S_{im} + (\alpha_2 + \alpha_3) * \\
& * S_{ig} + (\delta_1 + \delta_2 + \delta_3) * X_{im} + (\delta_2 + \delta_3) * X_{ig} + (\gamma_1 + \gamma_2 + \gamma_3) * Z_i + \\
& + (\theta_1 + \theta_2 + \theta_3) * T_t + \epsilon_{im} + \rho_{tm} + \omega_i + \sigma_i
\end{aligned} \tag{6}$$

In these models, β_1 and β_2 are selection-corrected effects of classmates' performance in the two subjects (see Ashenfelter and Krueger, 1994, p. 1162). The terms $\eta_1 + \psi_1$ and $\eta_2 + \psi_2$ reflect the bias due to unobserved subject-invariant student and teacher characteristics. Such biases arise, for example, in OLS models that relate individual grades to classmates' average performance and to a set of additional control variables (as in existing studies on relative grading; see Milek et al. 2009 and Trautwein and Baeriswyl 2007).

Given the assumption of the first-differenced model that $\beta_1 = \beta_2$, the model of Equations (5) and (6) is overidentified since there are two reduced-form parameters to estimate (β_1 and β_2), but only one structural parameter of interest, β (see Ashenfelter and Zimmerman, 1997, p. 2). This allows us to test the overidentification restriction implicitly integrated in that model, that is, we test whether $\beta_1 = \beta_2$. Moreover, we can test whether $\eta_1 = \eta_2$ and $\psi_1 = \psi_2$, which tells us whether biases in standard OLS models due to unobserved teacher and student characteristics are identical in both subjects. If these conditions hold, β_1 and β_2 can be replaced by β , η_1 and η_2 by η and ψ_1 and ψ_2 by ψ to obtain:

$$\begin{aligned}
y_{ig} = & (\beta + \eta + \psi) * \bar{S}_g^{(-i)} + (\eta + \psi) * \bar{S}_m^{(-i)} + (\alpha_1 + \alpha_2 + \alpha_3) * S_{ig} + (\alpha_2 + \alpha_3) * S_{im} + \\
& + (\delta_1 + \delta_2 + \delta_3) * X_{ig} + (\delta_2 + \delta_3) * X_{im} + (\gamma_1 + \gamma_2 + \gamma_3) * Z_i + (\theta_1 + \theta_2 + \theta_3) * \\
& * T_t + \epsilon_{ig} + \rho_{tg} + \omega_i + \sigma_i
\end{aligned} \tag{7}$$

$$\begin{aligned}
y_{im} = & (\beta + \eta + \psi) * \bar{S}_m^{(-i)} + (\eta + \psi) * \bar{S}_g^{(-i)} + (\alpha_1 + \alpha_2 + \alpha_3) * S_{im} + (\alpha_2 + \alpha_3) * S_{ig} + \\
& + (\delta_1 + \delta_2 + \delta_3) * X_{im} + (\delta_2 + \delta_3) * X_{ig} + (\gamma_1 + \gamma_2 + \gamma_3) * Z_i + (\theta_1 + \theta_2 + \theta_3) * \\
& * T_t + \epsilon_{im} + \rho_{tm} + \omega_i + \sigma_i
\end{aligned} \tag{8}$$

Equations (7) and (8) allow us to identify the parameter of interest (β), which is the difference between the coefficient on classmates' performance in the same subject ($\beta + \eta + \psi$) and the coefficient on classmates' performance in the other subject ($\eta + \psi$). Again, $\eta + \psi$ represents the bias due to unobserved subject-invariant student and teacher effects that might plague OLS models. Since the restricted correlated random effects model of Equations (7) and (8) is just another representation of the first-differenced model, we can identify β in the first-differenced model:

$$y_{ig} - y_{im} = \beta * (\bar{S}_g^{(-i)} - \bar{S}_m^{(-i)}) + \alpha * (S_{ig} - S_{im}) + \delta * (X_{ig} - X_{im}) + (\epsilon_{ig} - \epsilon_{im}) + (\rho_{tg} - \rho_{tm}) \tag{9}$$

The results of the overidentification test are valid only if we assume that the unobserved subject-specific student and teacher factors, $(\epsilon_{ig}, \epsilon_{im})$ and (ρ_{tg}, ρ_{tm}) , are random. In other words: if unobserved subject-specific student or teacher characteristics are correlated with classmates' performance, β cannot be identified in Equation (9). In this case, the overidentification test from the unrestricted correlated random effects model of Equations (5) and (6) is not informative either.

Potential threats to the randomness of the unobserved subject-specific student and teacher factors could arise from several sources. Metzler and Woessmann (2010), for example, provide evidence that teachers' subject-specific knowledge positively affects students' test scores. If we assume, for example, that teachers who are better in math are also more demanding in this subject, in the sense of stricter grading standards, our identification would be hampered: instead of only identifying teachers' grading schemes, β would also reflect teachers' grading standards. Although we cannot completely rule out this possibility, Figlio and Lucas (2004, p. 1820) find that U.S. teachers of fourth-grade students have similar grading standards in reading and math, thus providing evidence against a potential bias due to subject-specific grading standards.

Subject-specific differences in student motivation could be another reason why the coefficient of interest might be biased in the first-differenced model. Given her own performance, a student might be less motivated in the subject in which her classmates' perform quite well. If students' school grades are also affected by motivation and not only by performance,

β would be biased. In some specifications, we take into account students' subject-specific motivation with an indicator for participation in German and math. Our results are robust to the inclusion of this additional control variable.

4 Data on German Fourth-Grade Students from PIRLS

For our empirical analysis, we use data from the German extension of the Progress in International Reading Literacy Study (PIRLS-E) 2001. The international PIRLS was initiated and organized by the International Association for the Evaluation of Educational Achievement (IEA), an independent international cooperative of national research institutions and governmental research agencies.¹² The objective of PIRLS is to study trends in reading achievement for fourth-grade students (9 and 10 year olds). While students from all 16 German states participated in the international reading assessment on the first day of testing, 12 German states expanded the international study (PIRLS-E) by using a national questionnaire and testing students in math and science during a second day of testing.¹³ Importantly, the knowledge and skills tested in PIRLS-E are significant goals in the curriculum of fourth-grade students in Germany (Bos et al., 2004, p. 16). The target population of fourth graders is particularly interesting in the German context because the test takes place while students are in the last grade of primary school, that is, when grades are decisive for the students' secondary school track choice.

In addition to the objective measures of student performance in German and math (i.e., the PIRLS test scores in reading and math), the dataset also contains the teacher-assigned grades in the two subjects German and math. The grades are reported by the teachers and refer to the grades assigned to the students in the mid-year report card of the fourth grade.¹⁴ PIRLS-E also contains a measure of students' cognitive ability and extensive information on student characteristics and family background.¹⁵ The cognitive ability measure (*Kognitiver Fähigkeitstest*, *KFT*) should, however, not be interpreted as a measure of invariant, innate

¹²See Mullis et al. (2003) for a description and results of the international study.

¹³The four states that did not participate in the math test (Brandenburg, Mecklenburg-Western Pomerania, Lower Saxony, and Saxony-Anhalt) were excluded from the sample.

¹⁴Although PIRLS-E also provides test scores and grades in science, we limit our analysis to the subjects German and math since these are the major subjects in primary school and because secondary school recommendations are largely based on grades in these two subjects.

¹⁵We use multiple imputation by chained equations (MICE) to impute missing values of family background characteristics. This approach provides valid inferences under the assumption that data are missing at random.

ability, but as a measure that reflects both innate ability and cognitive abilities accumulated through education (Heller and Perleth, 2000, p. 54).

Our identification strategy requires that students are taught both German and math by the same teacher. Since the dataset does not contain an explicit teacher ID, we use three teacher characteristics—gender, age (in years), and teaching experience (in years)—to determine whether a classroom has the same teacher in both subjects. We exclude any classroom for which one of these three characteristics differs across German and math teachers. Furthermore, few students were dropped from the sample because the grade was missing for one or both subjects.¹⁶

Table 1 reports descriptive statistics of the estimation sample consisting of 2,550 students from 129 classrooms and 81 schools. Individual test scores and school grades are standardized to have a mean of 0 and a standard deviation of 1. *Classmates' average test scores* is the simple average of all test scores in the class, excluding a student's own test score. We rescaled all grades linearly such that higher values represent better grades, now ranging from 1 (fail) to 6 (very good). 36 percent of the students received a recommendation for high school, the most academic secondary school track. 79 percent of the students were born in Germany and 89 percent speak always or almost always German at home. The majority of teachers (77 percent) are female, which is typical for German primary schools.¹⁷ On average, teachers are about 47 years old, have 22 years of teaching experience, and teach about 23 students in a classroom.¹⁸

Table 2 presents descriptive statistics of teacher characteristics by teacher gender for both the full sample and the estimation sample. Teachers in the estimation sample have characteristics similar to those of teachers in the full sample, except that teachers in the full sample are slightly older and therefore have slightly more teaching experience. While

¹⁶In case of missing grades, typically an entire classroom was excluded from the sample because teachers did not report grades for any student in the class. Because primary school lasts six years in Berlin, teacher recommendations are not available for these students.

¹⁷Official statistics from the German Federal Statistical Office show that in the school year 2000/2001 (the school year of the PIRLS-E 2001 testing), 83.7 percent of primary school teachers were female in those German states that are part of our sample (Federal Statistical Office, 2002, p. 362). The difference of about 7 percentage points might be due to male teachers predominantly teaching primary school students in the higher grade levels, for example, fourth-graders that were tested in PIRLS (unfortunately, we have no official data on that).

¹⁸Table A1 provides descriptive statistics of the full sample, which consists of all teachers who reported their gender and all students with test scores in both reading and math. In contrast, the estimation sample excludes all classrooms having different teachers for German and math. Furthermore, students with missing information on the school grade in German or math are excluded from the estimation sample. Average test scores in reading and math are slightly lower in the full sample than in the estimation sample. All other covariates are very similar in both samples.

education level—which was reported only by teachers of German—is almost identical among male teachers across the two samples, female teachers in the estimation sample, on average, have slightly higher education levels. Comparisons between female and male teachers show that female teachers are on average a bit younger and therefore have less teaching experience than male teachers. Female teachers also have slightly fewer students in their classrooms.

5 Results

As a baseline, we first present OLS regressions with pooled data across the two subjects German and math, followed by the results of the correlated random effects model. In Section 5.2, we provide estimates of the incidence of relative grading using the first-differenced models. Section 5.3 analyzes potential effect heterogeneities.

5.1 Pooled OLS and Correlated Random Effects Model

Table 3 presents the association between classmates’ average performance and own teacher-assigned grade, using different sets of control variables, in OLS models that pool the two subjects German and math. As expected, a student’s grade is positively associated with his or her classmates’ average performance if the student’s own performance is not taken into account, indicating that the student belongs to a high-performing class (Column 1). However, once the student’s own performance is accounted for, having better-performing classmates is associated with a lower grade (Column 2). The next two columns reveal that own grade and classmates’ average performance remain strongly negatively correlated when numerous student-, teacher-, and class-level characteristics are added as controls. These OLS results basically replicate the findings of existing studies (see Milek et al., 2009; Trautwein and Baeriswyl, 2007). To eliminate potential biases that might arise from differences across schools or from non-random sorting of students and teachers into schools, Column (6) introduces school fixed effects thereby only using variation across classrooms within schools for identification. For comparison, we reestimate the specification of Column (4) with a sample that keeps only schools with more than one classroom (see Column 5). Comparing Columns (5) and (6) shows that introducing school fixed effects does not change results. In sum, given her own performance, a student’s grade is negatively associated with her

classmates' average performance, even if student and teacher characteristics are controlled for. This negative association is in line with teachers grading relatively.¹⁹

As noted above, OLS models are likely to suffer from biases due to unobserved student or teacher characteristics, or due to non-random sorting into classrooms within schools. Using only within-student variation across subjects taught by the same teacher in a first-differenced model eliminates any bias due to these sources. As discussed in Section 3, the validity of the assumption underlying the first-differenced model can be tested. In the unrestricted correlated random effects model (see Equations (5) and (6)), we test whether the intensity of relative grading is the same for both German and math (see Table 4). While the point estimates on classmates' average performance in the same subject (-0.267 and -0.149) differ across German and math (cf. Columns 1 and 2), this difference is not statistically significant. The same is true for the estimates on classmates' average performance in the other subject.

These results support the use of the restricted correlated random effects model (cf. Equations (7) and (8)). We can now estimate a single coefficient on classmates' performance both in the same subject and in the other subject (Column 3). We find a significantly negative estimate of classmates' average performance in the *same* subject and a coefficient close to zero of classmates' average performance in the *other* subject. The coefficient of interest, β , is the difference between the coefficient on classmates' average performance in the same subject and the coefficient on classmates' average performance in the other subject. This difference implies a statistically significant estimate for β of -0.226.²⁰

The restricted model allows us to estimate the bias due to unobserved subject-invariant student and teacher characteristics, which is represented by the coefficient on classmates' average performance in the *other* subject. The negative point estimate on classmates' average performance in other subject of -0.060 in Column (3) suggests that the OLS models might slightly, if at all, overestimate the true β . Note that the point estimate is not even statistically significant. Thus, potential biases in OLS models due to unobserved school-, teacher-, and student-level characteristics or non-random sorting seem not to be an

¹⁹Given the findings of Table 3 and considering that secondary school recommendations are primarily based on the average grade across the two subjects German and math, we should also observe that a teacher is more likely to give a recommendation for high school, the lower the performance of a student's classmates. Indeed, Table A2 shows that the teacher recommendation for a student is negatively associated with his or her classmates' average performance (measured by PIRLS test scores), controlling for the student's own performance (test scores averaged across reading and math) and additional factors.

²⁰The first-differenced model implies that subject-specific covariates have the same effect on outcomes, whereas they might have different effects in the unrestricted correlated random effects model. This leads to the small difference from the first-differenced estimate in Column (2) of Table 5.

issue. Surprisingly, factors such as within-school selection of teachers with different grading standards to classrooms seem to play only a minor role.

5.2 First-Differenced Model

The restricted correlated random effects model is a special representation of the first-differenced model. Therefore, the implied β in Column (3) of Table 4 should be similar to the coefficient in the first-differenced model. This is indeed the case: the coefficient on classmates' average performance in the first-differenced model fairly coincides with the implied β of the correlated random effects model (see Column (2) in Table 5).²¹ The negative effect of classmates' performance on own grade indicates that teachers grade a student's performance by comparing it with that of his or her classmates. The point estimate on classmates' average performance (-0.225) is one third smaller in magnitude than the respective OLS coefficient (cf. Column (2) in Table 3), indicating that the OLS model is biased by omitted variables. The magnitude of the coefficient of the first-differenced model means that an increase in classmates' average performance by one standard deviation lowers a student's own grade by about 10 percent of a standard deviation (=coefficient*SD classmates' performance/SD individual grade: $-0.225*0.46/1=-0.104$).

The last three columns in Table 5 investigate whether the negative effect of classmates' average performance on a student's grade is driven by the student's class participation. This is a potentially important channel since class participation—along with the results of written exams—might affect the grade of a subject. Especially, better-performing classmates could lower the grade of a student not only when the teacher grades on a curve, but also because more able classmates might negatively affect a student's motivation to participate in class. A student with top-performing classmates might participate less in class because her classmates are rather active themselves and/or because she considers herself as less able. Thus, a low-performing student might perceive herself as "a little fish in a big pond," a phenomenon often analyzed in social sciences (see, for example, Gerlach et al., 2007). Therefore, we control for students' class participation with an index of subject-specific participation, computed as the simple average of four survey questions that are identical for both German and math. Specifically, students were asked to indicate how strongly they agreed with the following statements: (1) "I usually participate intensively in class." (2) "I secretly often do other

²¹Results of the first-differenced model are identical if we instead pool the observations across the two subjects and include student fixed effects (which implicitly implies including teacher fixed effects as well).

things in class." (3) "In class, my thoughts are often somewhere else." (4) "I frequently raise hands in class." Students could indicate whether they agreed with the statement completely, almost, a little, or not at all. We coded as 1 the answer indicating the lowest participation and 4 for the answer indicating the highest participation in class.²²

Column (4) reveals that the negative association between a student's grade and his or her classmates' average performance is not driven by the student's class participation.²³ As expected, stronger class participation is associated with a better grade, even when the impact of individual and classmates' performance is taken into account. However, including a student's class participation barely diminishes the effect of classmates' performance on own grade. This indicates that having better-performing classmates leads to a lower grade not because the classmates induce a student to lower her participation, but because teachers compare students' performance when assigning grades. Column (5) furthermore shows that controlling for classmates' average participation does not change the effect of classmates' performance either. This suggests that it is indeed the performance of classmates and not their level of class participation that affects a student's grade. The last Column shows that the coefficient estimate on classmates' average performance remains unchanged when we control for both individual and classmates' participation in class.

Instead of using classmates' *average* performance, one may also consider alternative measures of classmates' performance. If teachers grade on a curve, they might, for example, compare a student's performance with the performance of the median student rather than with an artificial average performance. Therefore, we reestimate all specifications with classmates' *median* performance and find that the coefficients on classmates' performance barely change (see Table A3 in the appendix).

Another possibility is that teachers rank the performance of all students in a classroom from best to worst and assign grades according to a student's position in the performance distribution. To investigate this possibility, we compute a student's percentage rank in the classroom's test score distribution separately for reading and math, ranging from 0 (no classmate performs better) to 1 (all classmates perform better). Column (4) in Table A4 shows that being the worst-performing instead of the best-performing student in class lowers

²²Alternatively, we also counted how often a student indicated the highest participation in the four sub-categories. The results do not change qualitatively.

²³The sample size decreases by slightly more than half because the participation questions for math were randomly asked of only half the students in each classroom and some students did not answer (while all students were asked these questions for the subject German). The basic specification for the smaller sample (Column 3) shows that the coefficient estimate on classmates' average performance is slightly smaller than with the full sample, but not statistically different.

the student's grade by about half a standard deviation, even when own performance and individual class participation are taken into account. This is a much larger effect than the respective point estimate in the specification with classmates' average performance (Table 5). However, consider that moving from the very bottom to the very top of the performance distribution equals about four standard deviations and recall that an increase in classmates' average performance by about four standard deviations is associated with a lower grade by 40 percent of a standard deviation. This means that the magnitude of the effect in a model with performance rank (48 percent of a standard deviation) is quite similar to the effect size in a model with classmates' average performance (40 percent of a standard deviation).

In sum, the results are very similar when we use alternative measures of classmates' performance instead of classmates' average performance. Therefore, we use only the average test score as the measure for classmates' performance in all subsequent analyses.²⁴ We still wait for information on the most common test statistic for the reliability ratio, Cronbach's α , of the PIRLS-E German and math test to provide measurement-error corrected estimates. Therefore, the coefficient on classmates' average performance in Column (2) of Table 5 remains our main effect size. In order to get a first idea of the magnitude of the measurement error, we used the Cronbach's α from the German sample of the international PIRLS 2001 reading test for the reading *and* math test (see Mullis et al., 2003, p. 298, for Cronbach's alpha values of PIRLS 2001 reading tests). Using this imperfect estimate of the reliability ratios, we find that the measurement-error corrected effect is about 50 percent higher than the coefficient on classmates' average performance in Column (2) of Table 5.

5.3 Effect Heterogeneity

To assess whether the intensity of relative grading differs with teacher or class characteristics, we add interaction terms between classmates' average performance and various sub-group indicators. As a reference point, Column (1) of Table 6 presents the basic first-differenced specification of Table 5. Column (2) includes an interaction term between classmates' average performance and a dummy for whether the teacher is female. The coefficient on the interaction term is large in magnitude and highly statistically significant, whereas the coefficient on classmates' average performance is not statistically different from zero and even

²⁴Classmates' average test scores in math and reading are likely to be subject to measurement error that attenuates the coefficient of interest. The resulting bias can be consistently estimated with the reliability ratio that indicates how much the true coefficient β of classmates' average performance is asymptotically attenuated due to classical measurement error in our explanatory variable (see Metzler and Woessmann, 2010)

slightly positive.²⁵ These results indicate that only female teachers grade on a curve, whereas male teachers grade absolutely.²⁶ In contrast to teacher gender, the intensity of relative grading is independent of teacher experience (Columns 3) and class size (Columns 4). The coefficient estimates on these interaction terms are virtually zero, while the point estimates on classmates' average performance are negative and slightly larger in magnitude than in the basic specification without interaction terms.²⁷

Work by Dee (2005, 2007) shows that the interaction between student's gender and teacher's gender affects not only student test scores but also teacher perceptions of student performance. Therefore, we investigate whether teachers assign grades differently depending on whether the student is of the same gender as the teacher or not. The coefficient on the interaction term between classmates' average performance and a dummy variable indicating that student and teacher are the same gender is slightly positive, but not statistically different from zero (Column 5). Thus, there is no evidence that the interaction between teacher gender and student gender affects the teacher's grading scheme.²⁸

Finally, we investigate whether relative grading is more relevant for the better-performing or the worse-performing students. It might be that (female) teachers compare the performance of the best students, but not of the worst students, when assigning grades. Alternatively, teachers might rather compare the performance of the weakest students. The first three columns of Table 7 present results for all classrooms; the last two columns contain only classrooms with female teachers. Columns (2) and (4) include an interaction term

²⁵The coefficients hardly change when a dummy variable for female teacher is added to the specification to account for differences in grading standards across subjects. Given that female teachers, but not male teachers, grade on a curve, we reestimate Table 5 and add interaction terms between all explanatory variables and a binary indicator for female teacher. In line with Column (2) of Table 6, the interaction between female teacher and classmates' average performance is significantly negative and reveals that only female teachers grade on a curve. All other interaction terms are not statistically different from zero.

²⁶Since secondary school recommendations are strongly based on the grade point average in German and math, differential grading schemes should also translate into differential relationships between a student's recommendation and her classmates' performance. We test this hypothesis by re-estimating the specifications of Table A2 separately for female and male teachers. Consistent with female teachers grading on a curve, we find that a student with a female teacher is less likely to receive a recommendation for high school if she has better-performing classmates, given her own performance (Panel A of Table A5). And consistent with male teachers grading absolutely, the likelihood of receiving a student's high school recommendation is not associated with her classmates' performance in classes with male teachers, given the students' own performance (Panel B).

²⁷The class size dummy equals 1 for classrooms with more than 23 students, about the mean class size, and is 0 for smaller classrooms. Interacting classmates' average performance with a discrete class size variable instead of the binary dummy similarly yields a coefficient very close to 0.

²⁸We also experimented with including two additional interaction terms in all specifications of Table 6. We interacted a dummy for female student with both the individual test score difference and with the interaction term. All results remain qualitatively unchanged and neither of the two additional interaction terms is statistically significant in any specification.

between classmates' average performance and an individual student's average performance across reading and math. The point estimates on these interaction terms are positive, but not statistically different from zero, indicating that the intensity of grading on a curve is similar for weak and strong students. In Columns (3) and (5) we introduce interactions of classmates' average performance with a binary indicator for whether a student belongs to the top 50 percent of the class. The positive, but imprecisely estimated coefficients suggest that relative grading might be weaker among the better-performing students. In sum, there is small evidence that the intensity of relative grading differs between well- and low-performing students.

6 The Association between 'Grading on a Curve' and Students' Learning Effort

Theoretical work suggests that relative grading can negatively affect students' learning effort. In a relative grading system, classmates' studying efforts can have a detrimental effect on a student because rewards for learning (that is, grades) depend on a student's rank in class. Therefore, learning becomes a zero-sum game, which could lead to negative effects on students' learning effort because students might persuade each other "not to study too much" (Bishop, 2006). Learning incentives might be lower in the presence of relative grading even if students do not cooperate on effort levels because incentives also depend on the random grade component, that is, the performance uncertainty of students' test scores in exams (Landeras, 2009). Performance uncertainty arises because a student never knows with certainty how well he or she will perform on the exam, given her exam preparation. For example, the student might feel ill on the day of testing or perhaps he or she did not prepare for the exact questions asked. However, performance uncertainty could affect all students in the classroom. For example, noise in the classroom during the test might lower the performance of all students alike. The uncertainty of outperforming a classmate (relative grading) is therefore larger than the uncertainty of meeting a fixed standard (absolute grading) when the random component of test performance is mainly student-specific (and not class-specific). This is the case since the performance uncertainty of two competing students adds up with relative grading but not with absolute grading.

The previous section provides evidence that female teachers grade on a curve, whereas male teachers use an absolute grading scheme (see Column (2) in Table 6). We exploit this

gender difference in grading schemes to provide descriptive evidence that relative grading does not lower students' learning effort. We use students' class participation as a proxy for general learning effort, assuming that students who actively participate in class also are more likely to do their homework properly and to prepare well for exams at home (see p. 14 for questions on class participation). If relative grading leads to less learning effort than absolute grading, we should find differences in students' class participation between classrooms with female teachers and those with male teachers. To this end, we use OLS models with pooled data across all classrooms and across the two subjects German and math.²⁹ Column (1) of Table 8 shows that students with female teachers do not engage in less class participation than students of male teachers. We interpret this as first descriptive evidence that relative grading by female teachers does not lower students' learning effort. Adding a student's grade as a control variable (Column 2) shows that, as expected, class participation and grade are positively correlated. Furthermore, the insignificant interaction term between student's grade and female teacher indicates that the association between participation and grade is independent of the teacher's gender. Including further control variables in the remaining specifications of Table 8 does not affect the coefficient of interest. The coefficient on the female teacher dummy always remains statistically insignificant and close to zero.

If class participation was the same in classrooms with female and male teachers for teachers with the same grading scheme, this finding suggests that students' learning is not affected by relative grading. Of course, there might be other differences between female and male teachers that could lead to differences in students' participation. For example, students might generally participate more in classes taught by men if male teachers are more able than female teachers to elicit student effort. In the presence of such gender-specific teacher differences, the results of Table 8 tell us nothing about the effects of the grading scheme on students' learning effort.

In the remainder of this section, we investigate potential reasons for gender differences in grading schemes. This also implies that we investigate whether there exist two important gender-specific differences that could lead to differences in students' learning effort: whether female and male teachers assess students' performance differently and whether there are gender-specific differences in teaching styles and subject motivation.

²⁹We include only those students who provided information on class participation for both German and math.

Potential Reasons for Gender Differences in Grading Schemes

To better understand why male and female teachers use different grading schemes, we provide descriptive evidence on whether teachers assess students' performance differently, whether teaching style differs, and whether female and male teachers are differently motivated in their teaching of German and math. One reason for why grading schemes differ could be that female teachers use methods to assess student performance that are different from methods used by male teachers. For example, an absolute grading system may be more likely if teachers use written exams instead of oral exams. Assessing performance on an oral exam may be a more subjective process, making it more likely that a teacher (perhaps unconsciously) compares one student's performance with that of another. However, the first rows of Table 9 show that female and male teachers have a similar likelihood of assessing students' reading performance by means of an oral exam. Female and male teachers also put similar emphasis on classroom tests and both often use multiple-choice tests to assess students' reading performance (not shown in table). This provides some evidence that female and male teachers do not differ in the way they assess students' performance.

The remainder of Table 9 presents descriptive statistics of gender-specific teacher characteristics that are *not* directly related to students' grades, but do show that female and male teachers are similar in several dimensions. First, female and male teachers have similar styles of teaching reading. Most teachers of either gender always or often teach the whole class together; most teachers, regardless of gender, very seldom, if ever, group students by ability level. Second, female and male teachers report similar frequencies of teaching reading per week, and the same percentage of teachers report that they usually spend more time practicing reading with a student individually if the student is lagging behind his or her classmates. Furthermore, students were asked to report the amount of time they usually need to do their homework in German and math. Table 9 shows that students spend about the same amount of time on German homework as they do on math homework, irrespective of whether they are taught by a female or by a male teacher. These answers indicate that female teachers put very similar emphasis on one subject than do their male counterparts as measured by the time students spend studying at home.³⁰

In sum, the descriptive statistics show that female and male teachers behave similarly in several important dimensions. Therefore, these dimensions are unlikely to explain why

³⁰While two out of three differences for time spent on German homework are statistically significant, the overall distribution is quite similar.

female teachers grade on a curve, while male teachers use absolute grading schemes. Also, the similarity suggests that students' learning effort should not differ because female and male teachers are different with respect to these dimensions.

7 Conclusion

Understanding how teachers assign grades is important since school grades might affect both the educational career and the labor market entry. In the German tracked school system, for example, the type of secondary school attended is strongly affected by the grade point average for the major subjects German and math achieved in primary school. Similarly, a student's GPA must often meet a certain threshold to progress to the next grade or to graduate from school. Finally, theoretical studies show that the grading scheme might affect students' learning effort.

This paper investigates the empirical incidence of grading on a curve, using data on fourth-grade students from the German extension of the Progress in International Reading Literacy Study 2001. Our identification strategy uses variation in students' test scores and teacher-assigned grades across the subjects German and math, along with variation in classmates' test scores across the two subjects. We identify the effect of classmates' performance on a student's grade in a within-student across-subject model by differencing grades and test scores across subjects and restricting the sample to classrooms in which both subjects are taught by the same teacher. This approach likely eliminates biases due to non-random sorting and omitted student as well as unobservable teacher traits, such as teachers' grading standards.

We find that having classmates with one standard deviation higher average test scores lowers a student's grade by about 10 percent of a standard deviation. Further specifications show that effects are very similar when alternative measures for classmates' performance, such as the median value or a student's position in the performance distribution of the class, are used. We find that only female teachers, but not male teachers grade on a curve. Additional results suggest that the intensity of relative grading does not differ between high-performing and low-performing students and that relative grading does not depend on teacher experience or class size. Exploiting the difference in grading schemes between female and male teachers, we find no association between relative grading and students' learning effort. Future studies might try to provide more causally interpretable results of this link.

Overall, the results indicate that grading schemes differ between female and male teachers. While a student's grades—a potentially important academic outcome—depend only on the student's own performance when the student is taught by a male teacher, grades additionally depend on the performance of classmates when the student is taught by a female teacher. The difference in grading schemes across teacher gender remains a puzzle; understanding its causes and potential consequences is an interesting topic for future research.

References

- ALTONJI, J. AND C. PIERRET (2001): "Employer Learning and Statistical Discrimination," *Quarterly Journal of Economics*, 116, 313–350.
- ASHENFELTER, O. AND A. KRUEGER (1994): "Estimates of the Economic Returns to Schooling from a New Sample of Twins," *American Economic Review*, 84, 1157–1173.
- ASHENFELTER, O. AND D. ZIMMERMAN (1997): "Estimates of the Returns to Schooling from Sibling Data: Fathers, Sons, and Brothers," *Review of Economics and Statistics*, 79, 1–9.
- BAUMERT, J., M. BECKER, M. NEUMANN, AND R. NIKOLOVA (2009): "Frühübergang in ein grundständiges Gymnasium - Übergang in ein privilegiertes Entwicklungsmilieu? Ein Vergleich von Regressionsanalyse und Propensity Score Matching," *Zeitschrift für Erziehungswissenschaft*, 12, 189–215.
- BECKER, W. AND S. ROSEN (1992): "The Learning Effect of Assessment and Evaluation in High School," *Economics of Education Review*, 11, 107–118.
- BETTS, J. (1995): "Do Grading Standards Affect the Incentive to Learn?" Working paper, University of California-San Diego.
- BETTS, J. AND J. GROGGER (2003): "The Impact of Grading Standards on Student Achievement, Educational Attainment, and Entry-Level Earnings," *Economics of Education Review*, 22, 343–352.
- BISHOP, J. (1999): "Are National Exit Examinations Important for Educational Efficiency?" *Swedish Economic Policy Review*, 6, 349–398.
- (2006): "Drinking from the Fountain of Knowledge: Student Incentive to Study and Learn - Externalities, Information Problems and Peer Pressure," in *Handbook of the Economics of Education*, ed. by E. Hanushek and F. Welch, Elsevier, vol. 2, 909–944.
- BONESRØNNING, H. (1999): "The Variation in Teachers' Grading Practices: Causes and Consequences," *Economics of Education Review*, 18, 89–106.
- (2004): "Do the Teachers' Grading Practices Affect Student Achievement?" *Education Economics*, 12, 151–167.
- BOS, W., A. VOSS, E. LANKES, K. SCHWIPPERT, O. THIEL, AND R. VALTIN (2004): "Schullaufbahneempfehlungen von Lehrkräften für Kinder am Ende der 4. Jahrgangsstufe," in *IGLU: Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich*, ed. by W. Bos, E. Lankes, M. Prenzel, K. Schwippert, R. Valtin, and G. Walther, Münster: Waxmann, 191–228.
- CHAMBERLAIN, G. (1982): "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18, 5–46.
- DEE, T. (2005): "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?" *American Economic Review*, 95, 158–165.

- (2007): “Teachers and the Gender Gaps in Student Achievement,” *Journal of Human Resources*, 42, 528–554.
- DUSTMANN, C. (2004): “Parental Background, Secondary School Track Choice, and Wages,” *Oxford Economic Papers*, 56, 209–230.
- FEDERAL STATISTICAL OFFICE (2002): *Statistical Yearbook 2002*, Stuttgart: Metzler-Poeschel.
- FIGLIO, D. AND M. LUCAS (2004): “Do High Grading Standards Affect Student Performance?” *Journal of Public Economics*, 88, 1815–1834.
- GERLACH, E., U. TRAUTWEIN, AND O. LUEDTKE (2007): “Referenzgruppeneffekte im Sportunterricht: Kurz- und langfristig negative Effekte sportlicher Klassenkameraden auf das sportliche Selbstkonzept,” *Zeitschrift für Sozialpsychologie*, 38, 73–83.
- HANUSHEK, E. (2002): “Publicly Provided Education,” in *Handbook of Public Economics*, ed. by A. Auerbach and M. Feldstein, Elsevier, vol. 4, 2046–2141.
- HANUSHEK, E. AND L. WOESSMANN (2011): “The Economics of International Differences in Educational Achievement,” in *Handbook of the Economics of Education*, ed. by E. Hanushek, S. Machin, and L. Woessmann, Elsevier, vol. 3, 89–200.
- HELLER, K. AND C. PERLETH (2000): *KFT 4-12+R - Kognitiver Fähigkeits-Test fuer 4. bis 12. Klassen*, Göttingen: Beltz.
- JÜRGES, H. AND K. SCHNEIDER (2007): “What Can Go Wrong Will Go Wrong: Birthday Effects and Early Tracking in the German School System,” MEA Discussion Paper No. 138.
- KROPF, M., C. GRESCH, AND K. MAAZ (2010): “Überblick über die rechtlichen Regelungen des Übergangs in den beteiligten Ländern,” in *Der Übergang von der Grundschule in die weiterführende Schule - Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten*, ed. by Bundesministerium für Bildung und Forschung (BMBF), Bonn, 399–429.
- LANDERAS, P. (2009): “Student Effort: Standards vs. Tournaments,” *Applied Economics Letters*, 16, 965–969.
- LOHMAR, B. AND T. ECKHARDT (2010): “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,” Bonn: Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs.
- METZLER, J. AND L. WOESSMANN (2010): “The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation,” IZA Discussion Paper No. 4999.
- MILEK, A., O. LÜDTKE, U. TRAUTWEIN, K. MAAZ, AND T. STUBBE (2009): “Wie konsistent sind Referenzgruppeneffekte bei der Vergabe von Schulformempfehlungen? Bundeslandspezifische Anaylsen mit Daten der IGLU-Studie,” *Zeitschrift für Erziehungswissenschaft*, Special Issue 12, 282–301.

- MULLIS, I., M. MARTIN, E. GONZALEZ, AND A. KENNEDY (2003): *PIRLS 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary School in 35 Countries*, Chestnut Hill, MA: International Study Center, Boston College.
- PIETSCH, M. AND T. STUBBE (2007): "Inequality in the Transition from Primary to Secondary School: School Choices and Educational Disparities in Germany," *European Educational Research Journal*, 6, 424–445.
- TRAUTWEIN, U. AND F. BAERISWYL (2007): "Wenn leistungsstarke Klassenkameraden ein Nachteil sind: Referenzgruppeneffekte bei Übertrittsentscheidungen," *Zeitschrift für Pädagogische Psychologie*, 21, 119–133.

Tables

Table 1
Descriptive Statistics: Estimation Sample

Variable	Mean	Std. dev.	Min	Max
Individual school grades (reversed scale)				
German (non-standardized)	4.21	0.93	1	6
Math (non-standardized)	4.30	0.98	1	6
German (z-standardized)	0	1	-3.45	1.93
Math (z-standardized)	0	1	-3.36	1.73
Difference (German-math)	0	0.83	-3.36	3.25
High school recommendation	0.36		0	1
Individual test scores				
Reading (non-standardized)	545.21	65.07	263.78	728.33
Math (non-standardized)	506.93	96.79	181.54	806.12
Reading (z-standardized)	0	1	-4.33	2.81
Math (z-standardized)	0	1	-3.36	3.09
Difference (German-math)	0	0.90	-3.89	3.71
Classmates' average test scores				
Reading	0	0.45	-2.79	0.97
Math	0	0.48	-2.62	0.82
Difference (German-math)	0	0.31	-0.88	0.81
Individual participation in class^a				
German	3.12	0.62	1	4
Math	3.26	0.60	1	4
Difference (German-math)	-0.14	0.46	-2.25	2.50
Classmates' avg. participation in class^a				
German	3.12	0.18	2.63	3.72
Math	3.26	0.26	2.20	4.00
Difference (German-math)	-0.14	0.20	-0.99	0.55
Student characteristics				
Cognitive ability	0	1	-2.84	3.03
Age (in months)	125.9	5.6	102	159
Female	0.48		0	1

(continued on next page)

Table 1 (continued)

Variable	Mean	Std. dev.	Min	Max
Family background characteristics				
School degree of father				
ISCED 2 or lower	0.19		0	1
ISCED 3 or 4	0.09		0	1
ISCED 5 or higher	0.12		0	1
School degree of mother				
ISCED 2 or lower	0.20		0	1
ISCED 3 or 4	0.10		0	1
ISCED 5 or higher	0.22		0	1
Annual household income (in \$)				
Less than 20.000	0.16		0	1
20.000-29.999	0.22		0	1
30.000-39.999	0.24		0	1
40.000-49.999	0.17		0	1
50.000-59.999	0.10		0	1
60.000 or more	0.11		0	1
Number of books at home				
0-25	0.06		0	1
26-100	0.13		0	1
>100	0.37		0	1
Born in Germany				
Student	0.79		0	1
Mother	0.80		0	1
Father	0.80		0	1
German spoken at home				
Always or almost always	0.89		0	1
Sometimes	0.10		0	1
Never	0.01		0	1

(continued on next page)

Table 1 (continued)

Variable	Mean	Std. dev.	Min	Max
Variables at class level				
Female teacher	0.77		0	1
Missing values	0		0	0
Teacher's age (in years)	46.8	10.2	26	63
Missing values	0.01		0	1
Teacher's education				
ISCED 3	0.21		0	1
ISCED 4	0.02		0	1
ISCED 5 or higher	0.71		0	1
Missing values	0.07		0	1
Teacher's experience (in years)	21.5	12.1	1	42
Missing values	0.04		0	1
Class size (reported by teacher)	22.9	4.4	9	32
Number of observations				
Students	2,550			
Classrooms	129			
Students per class	19.8	4.9	8	31
Schools	81			

Notes: Observations are weighted with the inverse of students' sampling probabilities. Std. Dev.: Standard deviations are reported only for continuous and discrete variables. Test scores are the first plausible values of the respective test domain. *Classmates' average test scores* are simple averages of all individual test scores in the class, excluding own test score. *School grades* were rescaled, ranging from 1 (fail) to 6 (very good). Means and standard deviations of the following variables include imputed values: student/mother/father born in Germany (5.8% missing values/7.7%/9.3%), mother's education (36.2%), father's education (37.9%), number of books at home (9.5%), and household income (22.1%). The ISCED education levels combine school, vocational, and university degrees. ISCED 2 or lower: not more than lower secondary education; ISCED 3 and 4: upper secondary education and non-tertiary postsecondary education; ISCED 5 or higher: tertiary education and higher.

^a Since international PIRLS was designed to test students in reading, only half the students in each classroom were asked to provide information on participation in math, whereas all students were asked to provide information on participation in German.

Table 2
Descriptive Statistics of Teacher Characteristics

Variable	Female teachers		Male teachers	
	Full sample	Estimation sample	Full sample	Estimation sample
	(1)	(2)	(3)	(4)
Teacher's age (in years)	47.6 (9.4)	45.9 (10.7)	50.8 (8.1)	49.8 (7.9)
Missing values	0.011	0.010	0.000	0.000
Teacher's education				
ISCED 3	0.245	0.182	0.292	0.300
ISCED 4	0.014	0.020	0.000	0.000
ISCED 5 or higher	0.677	0.717	0.688	0.667
Missing values	0.059	0.081	0.021	0.033
Teacher's experience (in years)	22.2 (11.7)	21.1 (12.7)	24.4 (9.9)	22.9 (9.9)
Missing values	0.035	0.040	0.016	0.033
Class size (reported by teacher)	22.6 (3.5)	22.6 (4.2)	23.2 (3.9)	23.7 (4.8)
Teachers	282	99	64	30
Classrooms	229	99	63	30
Students	4,369	1,920	1,270	630
Schools	141	67	55	24

Notes: Sample means reported; standard deviations of continuous variables in parentheses. The full sample contains all students with test scores in reading and math and teachers with reported gender. The estimation sample excludes students with missing information on the school grade in German or math and excludes all classrooms in which the two subjects were taught by different teachers. Teachers' education level was reported only by German teachers.

Table 3
Pooled OLS Regressions: Relationship between Own Grade and Classmates' Performance

	Full sample			Schools > 1 class		
	(1)	(2)	(3)	(4)	(5)	(6)
Classmates' average performance	0.123*** (0.045)	-0.345*** (0.046)	-0.410*** (0.043)	-0.413*** (0.046)	-0.303*** (0.062)	-0.330*** (0.057)
Individual performance		0.604*** (0.014)	0.410*** (0.014)	0.403*** (0.014)	0.402*** (0.018)	0.401*** (0.018)
Cognitive ability			0.205*** (0.019)	0.212*** (0.018)	0.199*** (0.020)	0.209*** (0.019)
Student's age			-0.015***	-0.016***	-0.015***	-0.015***
Female student			(0.003)	(0.002)	(0.003)	(0.003)
Family background			0.067*	0.068*	0.034	0.057
Teacher characteristics			(0.032)	(0.032)	(0.036)	(0.034)
Class size (+squared)			Yes	Yes	Yes	Yes
School dummies				Yes	Yes	Yes
Classrooms	129	129	129	129	95	95
Students	2,550	2,550	2,550	2,550	1,880	1,880
Student-subject observations	5,100	5,100	5,100	5,100	3,760	3,760

Dependent variable: individual grades in German and math. The sample is pooled across both subjects German and math. Robust standard errors (in parentheses) clustered at classroom level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 4
Correlated Random Effects Model

Dependent variable:	German grade	Math grade	Grades
	(1)	(2)	(3)
Classmates' average performance in same subject	-0.267** (0.133)	-0.149 (0.128)	-0.286*** (0.056)
Classmates' average performance in other subject	-0.038 (0.144)	-0.234 (0.154)	-0.060 (0.058)
Individual performance in same subject	0.371*** (0.023)	0.283*** (0.026)	0.356*** (0.013)
Individual performance in other subject	0.179*** (0.026)	0.221*** (0.025)	0.172*** (0.016)
Classrooms	129	129	129
Students	2,550	2,550	2,550
Chi-squared (coeff. on same subject equal)		0.028	
Prob>chi-squared		0.599	
Chi-squared (coeff. on other subject equal)		0.620	
Prob>chi-squared		0.432	
Implied beta			-0.226***
Prob>chi-squared			0.0002

Dependent variables: grade in German (Column 1), grade in math (Column 2), and grade in German and math (Column 3). Models (1) and (2) are estimated by seemingly unrelated regressions (SUR). All regressions control for student characteristics, family background, teacher characteristics, and class size. Robust standard errors (in parentheses) are clustered at the classroom level; standard errors are estimated by maximum likelihood in the SUR models. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5
First-Differenced Model: Within-Student Across-Subject Approach

	All students			Dropping students without information on participation in math		
	(1)	(2)	(3)	(4)	(5)	(6)
Classmates' average performance		-0.225*** (0.085)	-0.311*** (0.093)	-0.272*** (0.094)	-0.321*** (0.094)	-0.291*** (0.094)
Individual performance	0.172*** (0.021)	0.184*** (0.021)	0.225*** (0.029)	0.199*** (0.030)	0.224*** (0.029)	0.196*** (0.030)
Classmates' mean participation					-0.114 (0.153)	-0.241 (0.149)
Individual participation				0.375*** (0.057)		0.396*** (0.057)
Classrooms	129	129	129	129	129	129
Students	2,550	2,550	1,219	1,219	1,219	1,219

Dependent variable: difference in grade between German and math. All explanatory variables are differences between German/reading and math. Robust standard errors (in parentheses) are clustered at the classroom level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 6
Effect Heterogeneity

CHAR:	Female teacher		Teacher experience		Large class		Same gender	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Classmates' average performance	-0.225*** (0.085)	0.095 (0.110)	-0.306* (0.173)	-0.241** (0.117)	-0.265** (0.113)			
Classmates' avg. performance x CHAR		-0.403*** (0.145)	0.003 (0.007)	0.032 (0.171)	0.079 (0.191)			
Individual performance	0.184*** (0.021)	0.184*** (0.021)	0.188*** (0.021)	0.184*** (0.021)	0.183*** (0.021)			
Classrooms	129	129	124	129	129			
Students	2,550	2,550	2,442	2,550	2,550			

Dependent variable: difference in grade between German and math. All explanatory variables are differences between German/reading and math. Large class equals 1 for classrooms with more than 23 students (about mean class size) and 0 for smaller classrooms. Same gender equals 1 if gender of teacher equals gender of student; 0 otherwise. Robust standard errors (in parentheses) are clustered at the classroom level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 7
Grading on a Curve More Relevant for Top- or Low-Performers?

	All classrooms		Only classrooms with female teachers		
	(1)	(2)	(3)	(4)	(5)
Classmates' average performance	-0.225*** (0.085)	-0.230*** (0.087)	-0.340** (0.137)	-0.323*** (0.098)	-0.460*** (0.153)
Individual performance	0.184*** (0.021)	0.184*** (0.021)	0.184*** (0.021)	0.202*** (0.023)	0.201*** (0.023)
Classmates' avg. performance x individual performance		0.070 (0.065)		0.094 (0.072)	
Classmates' avg. performance x above median performance			0.230 (0.156)		0.291 (0.181)
Classrooms	129	129	129	99	99
Students	2,550	2,550	2,550	1,920	1,920

Dependent variable: difference in grade between German and math. All explanatory variables are differences between German/reading and math. Robust standard errors (in parentheses) are clustered at the classroom level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 8
The Association between Relative Grading and Students' Learning Effort

	Full sample			Schools > 1 class		
	(1)	(2)	(3)	(4)	(5)	(6)
Female teacher	0.032 (0.048)	0.014 (0.049)	0.016 (0.048)	0.018 (0.051)	0.008 (0.064)	0.042 (0.063)
Grade		0.189*** (0.036)	0.160*** (0.037)	0.154*** (0.035)	0.125** (0.044)	0.145*** (0.044)
Grade x female teacher		0.029 (0.042)	0.028 (0.042)	0.036 (0.040)	0.100* (0.049)	0.085 (0.048)
Individual performance			0.057*** (0.019)	0.040** (0.018)	0.040 (0.023)	0.038 (0.023)
Student characteristics				Yes	Yes	Yes
Family background				Yes	Yes	Yes
Teacher characteristics				Yes	Yes	Yes
Class size (+squared)				Yes	Yes	Yes
School dummies						
Classrooms	129	129	129	129	95	95
Students	1,219	1,219	1,219	1,219	898	898
Student-subject observations	2,438	2,438	2,438	2,438	1,796	1,796

Dependent variable: class participation in German and math, respectively. The sample is pooled across both subjects. Only students included who provided information on participation in both German and math. Robust standard errors (in parentheses) clustered at classroom level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 9
Performance Assessment, Teaching Style, and Motivation by Teacher Gender

Teacher questions	Female teachers	Male teachers	Difference
Assess students' reading performance orally			
At least once a week	0.617	0.500	0.117 (0.106)
Once or twice a month	0.340	0.429	-0.088 (0.104)
Once or twice a year	0.043	0.071	-0.029 (0.047)
Teach reading the whole class together			
Always or almost always	0.351	0.310	0.040 (0.101)
Often	0.412	0.483	-0.070 (0.105)
Sometimes	0.216	0.207	0.010 (0.088)
Never	0.021	0.000	0.021 (0.027)
Group students with similar abilities when teach reading			
Always or almost always	0.043	0.000	0.043 (0.038)
Often	0.202	0.207	-0.005 (0.086)
Sometimes	0.617	0.655	-0.038 (0.104)
Never	0.138	0.138	0.000 (0.074)
Teach reading			
Every day	0.577	0.517	0.060 (0.106)
Three or four days a week	0.237	0.310	-0.073 (0.093)
Fewer than three days a week	0.186	0.172	0.013 (0.082)
Usually spend more time practicing reading with a student individually if she lags behind	0.844	0.828	0.016 (0.078)
Teaching math generally means fun	0.889	0.833	0.056 (0.069)
Teaching math generally means interesting topics	0.670	0.767	-0.097 (0.097)
Student questions			
Time usually needed for homework in German			
15 minutes	0.531	0.596	-0.066 (0.023)
30 minutes	0.364	0.323	0.041 (0.022)
45 minutes or more	0.105	0.081	0.024 (0.014)
Time usually needed for homework in math			
15 minutes	0.537	0.561	-0.023 (0.023)
30 minutes	0.338	0.325	0.014 (0.022)
45 minutes or more	0.124	0.115	0.010 (0.015)

Notes: Samples are the estimation samples which consist only of classrooms that are taught by the same teacher in German and math. The sample size of female teachers varies between 94 and 99, and the male teacher sample varies between 28 and 30. The last column reports differences between female and male teachers, with standard errors in parentheses. The variables *Teaching math generally means fun* and *Teaching math generally means interesting topics* equal 1 if the teacher answered "completely true" or "almost true" and equal 0 for the answers "a little bit true" and "not true at all."

Appendix

Table A1
Descriptive Statistics: Full Sample

Variable	Mean	Std. dev.	Min	Max
Individual school grades (reversed scale)				
German (z-standardized)	0	1	-3.57	1.89
German (non-standardized)	4.26	0.92	1	6
German missing	0.22		0	1
Math (z-standardized)	0	1	-3.42	1.78
Math (non-standardized)	4.28	0.97	1	6
Math missing	0.22		0	1
Difference (German-math)	0	0.87	-3.49	3.23
High school recommendation	0.36		0	1
High school recommendation missing	0.09		0	1
Individual test scores				
Reading (z-standardized)	0	1	-4.18	2.89
Reading (non-standardized)	539.25	66.54	263.78	729.50
Math (z-standardized)	0	1	-5.37	3.54
Math (non-standardized)	500.47	99.95	-30.70	850.33
Difference (German-math)	0	0.96	-3.81	5.22
Individual participation in class^a				
German	3.11	0.61	1	4
Math	3.24	0.60	1	4
Difference (German-math)	-0.14	0.47	-2.50	2.50
Classmates' participation in class^a				
German	3.11	0.19	2.63	3.56
Math	3.23	0.26	2.27	3.88
Difference (German-math)	-0.12	0.21	-0.75	0.65
Individual characteristics				
Cognitive ability	0	1	-3.68	3.08
Age (in months)	126.4	6.0	102	159
Female	0.49		0	1
Family background characteristics				
School degree of father				
ISCED 2 or lower	0.22		0	1
ISCED 3 or 4	0.56		0	1
ISCED 5 or higher	0.22		0	1
School degree of mother				
ISCED 2 or lower	0.23		0	1
ISCED 3 or 4	0.63		0	1
ISCED 5 or higher	0.14		0	1
Annual household income (in \$)				
Less than 20.000	0.17		0	1
20.000-29.999	0.21		0	1
30.000-39.999	0.24		0	1
40.000-49.999	0.16		0	1
50.000-59.999	0.10		0	1
60.000 or more	0.12		0	1
Number of books at home				
0-25	0.20		0	1
26-100	0.36		0	1
>100	0.44		0	1

(continued on next page)

Table A1 (continued)

Variable	Mean	Std. dev.	Min	Max
Born in Germany				
Student	0.79		0	1
Mother	0.80		0	1
Father	0.79		0	1
German spoken at home				
Always or almost always	0.89		0	1
Sometimes	0.10		0	1
Never	0.01		0	1
Variables at class level				
Female teacher	0.63		0	1
Missing values	0.25		0	1
Teacher's Age (in years)	48.2	9.4	26	63
Missing values	0.12		0	1
Teacher's education				
ISCED 3	0.15		0	1
ISCED 4	0.00		0	1
ISCED 5 or higher	0.39		0	1
Missing values	0.45		0	1
Teacher's experience (in years)	22.6	11.4	0	42
Missing values	0.14		0	1
Class size (reported by teacher)	25.1	8.9	9	60
Number of observations				
Students	5,856			
Classrooms	308			
Students per class	19.1	4.9	6	31
Schools	166			

Notes: Observations are weighted with the inverse of students' sampling probabilities. Std. Dev.: Standard deviations are reported only for continuous and discrete variables. Test scores are the first plausible values of the respective test domain. *Classmates' average test scores* are simple averages of all individual test scores in the class, excluding own test score. *School grades* were rescaled, ranging from 1 (fail) to 6 (very good). Means and standard deviations of the following variables include imputed values: student/mother/father born in Germany, mother's education, father's education, number of books at home, and household income. The ISCED education levels combine school, vocational, and university degrees. ISCED 2 or lower: not more than lower secondary education; ISCED 3 and 4: upper secondary education and non-tertiary postsecondary education; ISCED 5 or higher: tertiary education and higher.

^a Since international PIRLS was designed to test students in reading, only half the students in each classroom were asked to provide information on participation in math, whereas all students were asked to provide information on participation in German.

Table A2
Teacher Recommendation for Secondary School Track (OLS)

	(1)	(2)	(3)	(4)
Classmates' performance	-0.131*** (0.027)	-0.154*** (0.026)	-0.152*** (0.026)	-0.049* (0.024)
Individual performance	0.285*** (0.009)	0.243*** (0.012)	0.242*** (0.012)	0.073*** (0.013)
German grade				0.198*** (0.013)
Math grade				0.117*** (0.012)
Student characteristics		Yes	Yes	Yes
Teacher characteristics			Yes	Yes
Class size (+squared)			Yes	Yes
Classrooms	124	124	124	124
Students	2,338	2,338	2,338	2,338

Dependent variable: high school recommendation by primary school teacher (yes=1, no=0). All columns estimated with linear probability models. Grades and performance measures are z-standardized. Performance measures refer to average performance across reading and math. Student characteristics include student's age and gender, educational degrees of parents, household income, number of books at home, indicators whether student, mother or father were born in Germany and whether German is spoken at home. Teacher characteristics include teacher's age, gender, teaching experience, and binary indicators for whether these characteristics are missing. Robust standard errors (in parentheses) clustered at classroom level. Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Table A3
Alternative Measure for Classmates' Performance: Median Test Score

	All students		Dropping students without information on participation in math			
	(1)	(2)	(3)	(4)	(5)	(6)
Classmates' median performance		-0.198** (0.078)	-0.276*** (0.088)	-0.234*** (0.089)	-0.275*** (0.089)	-0.238*** (0.089)
Individual performance	0.172*** (0.021)	0.181*** (0.020)	0.222*** (0.029)	0.196*** (0.030)	0.222*** (0.029)	0.195*** (0.030)
Classmates' median participation					0.010 (0.112)	-0.049 (0.108)
Individual participation				0.373*** (0.056)		0.376*** (0.056)
Classrooms	129	129	129	129	129	129
Students	2,550	2,550	1,219	1,219	1,219	1,219

Dependent variable: difference in grade between German and math. All explanatory variables are differences between German/reading and math. Robust standard errors (in parentheses) are clustered at the classroom level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table A4
Alternative Measure for Classmates' Performance: Position in Class Performance Distribution

	All students			Dropping students without information on participation in math		
	(1)	(2)	(3)	(4)	(5)	(6)
Position in class performance distribution		-0.362** (0.147)	-0.527*** (0.172)	-0.484*** (0.171)	-0.522*** (0.170)	-0.485*** (0.168)
Individual performance	0.172*** (0.021)	0.068 (0.049)	0.055 (0.060)	0.044 (0.060)	0.046 (0.058)	0.039 (0.058)
Position in class participation distribution					-0.373*** (0.101)	-0.249*** (0.106)
Individual participation				0.380*** (0.056)		0.337*** (0.059)
Classrooms	129	129	129	129	129	129
Students	2,550	2,550	1,219	1,219	1,219	1,219

Dependent variable: difference in grade between German and math. All explanatory variables are differences between German/reading and math. Position in class performance distribution ranges from 0 (no classmate performs better) to 1 (all classmates perform better); position in class participation distribution similarly. Robust standard errors (in parentheses) are clustered at the classroom level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table A5
Teacher Recommendation for Secondary School Track by Teacher Gender (OLS)

	(1)	(2)	(3)	(4)
Panel A: Female Teachers				
Classmates' average performance	-0.150*** (0.030)	-0.168*** (0.030)	-0.171*** (0.029)	-0.071** (0.025)
Individual performance	0.289*** (0.010)	0.242*** (0.014)	0.240*** (0.014)	0.072*** (0.016)
German grade				0.199*** (0.015)
Math grade				0.117*** (0.014)
Student characteristics		Yes	Yes	Yes
Teacher characteristics			Yes	Yes
Class size (+squared)			Yes	Yes
Classrooms	95	95	95	95
Students	1,776	1,776	1,776	1,776
Panel B: Male Teachers				
Classmates' average performance	-0.022 (0.049)	-0.084* (0.041)	-0.028 (0.060)	0.067 (0.064)
Individual performance	0.275*** (0.017)	0.248*** (0.023)	0.247*** (0.023)	0.073*** (0.018)
German grade				0.194*** (0.022)
Math grade				0.124*** (0.027)
Student characteristics		Yes	Yes	Yes
Teacher characteristics			Yes	Yes
Class size (+squared)			Yes	Yes
Classrooms	29	29	29	29
Students	562	562	562	562

Dependent variable: high school recommendation by primary school teacher (yes=1, no=0). All columns estimated with linear probability models. Grades and performance measures are z-standardized. Performance measures refer to average performance across reading and math. Student characteristics include student's age and gender, educational degrees of parents, household income, number of books at home, indicators whether student, mother or father were born in Germany and whether German is spoken at home. Teacher characteristics include teacher's age, gender, teaching experience, and binary indicators for whether these characteristics are missing. Robust standard errors (in parentheses) clustered at classroom level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Ifo Working Papers

- No. 120 Kauppinen, I. and P. Poutvaara, Preferences for Redistribution among Emigrants from a Welfare State, January 2012.
- No. 119 Aichele, R. and G.J. Felbermayr, Estimating the Effects of Kyoto on Bilateral Trade Flows Using Matching Econometrics, December 2011.
- No. 118 Heid, B., J. Langer and M. Larch, Income and Democracy: Evidence from System GMM Estimates, December 2011.
- No. 117 Felbermayr, G.J. and J. Gröschl, Within US Trade and Long Shadow of the American Secession, December 2011.
- No. 116 Felbermayr, G.J. and E. Yalcin, Export Credit Guarantees and Export Performance: An Empirical Analysis for Germany, December 2011.
- No. 115 Heid, B. and M. Larch, Migration, Trade and Unemployment, November 2011.
- No. 114 Hornung, E., Immigration and the Diffusion of Technology: The Huguenot Diaspora in Prussia, November 2011.
- No. 113 Riener, G. and S. Wiederhold, Costs of Control in Groups, November 2011.
- No. 112 Schlotter, M., Age at Preschool Entrance and Noncognitive Skills before School – An Instrumental Variable Approach, November 2011.
- No. 111 Grimme, C., S. Henzel and E. Wieland, Inflation Uncertainty Revisited: Do Different Measures Disagree?, October 2011.
- No. 110 Friedrich, S., Policy Persistence and Rent Extraction, October 2011.
- No. 109 Kipar, S., The Effect of Restrictive Bank Lending on Innovation: Evidence from a Financial Crisis, August 2011.
- No. 108 Felbermayr, G.J., M. Larch and W. Lechthaler, Endogenous Labor Market Institutions in an Open Economy, August 2011.

- No. 107 Piopiunik, M., Intergenerational Transmission of Education and Mediating Channels: Evidence from Compulsory Schooling Reforms in Germany, August 2011.
- No. 106 Schlotter, M., The Effect of Preschool Attendance on Secondary School Track Choice in Germany, July 2011.
- No. 105 Sinn, H.-W. und T. Wollmershäuser, Target-Kredite, Leistungsbilanzsalden und Kapitalverkehr: Der Rettungsschirm der EZB, Juni 2011.
- No. 104 Czernich, N., Broadband Internet and Political Participation: Evidence for Germany, June 2011.
- No. 103 Aichele, R. and G.J. Felbermayr, Kyoto and the Carbon Footprint of Nations, June 2011.
- No. 102 Aichele, R. and G.J. Felbermayr, What a Difference Kyoto Made: Evidence from Instrumental Variables Estimation, June 2011.
- No. 101 Arent, S. and W. Nagl, Unemployment Benefit and Wages: The Impact of the Labor Market Reform in Germany on (Reservation) Wages, June 2011.
- No. 100 Arent, S. and W. Nagl, The Price of Security: On the Causality and Impact of Lay-off Risks on Wages, May 2011.
- No. 99 Rave, T. and F. Goetzke, Climate-friendly Technologies in the Mobile Air-conditioning Sector: A Patent Citation Analysis, April 2011.
- No. 98 Jeßberger, C., Multilateral Environmental Agreements up to 2050: Are They Sustainable Enough?, February 2011.
- No. 97 Rave, T., F. Goetzke and M. Larch, The Determinants of Environmental Innovations and Patenting: Germany Reconsidered, February 2011.
- No. 96 Seiler, C. and K. Wohlrabe, Ranking Economists and Economic Institutions Using RePEc: Some Remarks, January 2011.
- No. 95 Itkonen, J.V.A., Internal Validity of Estimating the Carbon Kuznets Curve by Controlling for Energy Use, December 2010.