mea – Mannheimer Forschungsinstitut Ökonomie und Demographischer Wandel
Gebäude L 13, 17_D-68131 Mannheim_Sekretariat +49 621 181-2773/1862_Telefax +49 621 181-1863_www.mea.uni-mannheim.de

Teacher quality and incentives

Theoretical and empirical effects of standards on teacher quality

Hendrik Jürges, Wolfram F. Richter and
Kerstin Schneider

91-2005

May 2005

Teacher quality and incentives

Theoretical and empirical effects of standards on teacher quality[+]

Hendrik Jürges[*]

MEA, University of Mannheim, and DIW Berlin

Wolfram F. Richter[†]

University of Dortmund, CESifo, Munich and IZA, Bonn

Kerstin Schneider[#]

University of Wuppertal and CESifo, Munich

31 May 2005

Abstract: Applying the theory of yardstick competition to the schooling system, we show that it is optimal to have central tests of student achievement and to engage in benchmarking because it raises the quality of teaching. This is true even if teachers' pay (defined in monetary terms) is not performance related. If teachers value reputation, and if teaching output is measured so that it becomes comparable, teachers will increase their effort. The theory is tested using the German PISA-E data. Use is made of the fact that central exams exist in some federal states of Germany but not in all. The empirical evidence suggests that central exams have a positive effect on the quality of teaching.

Keywords: education, teacher quality, central examinations, yardstick competition, matching

JEL-Code: I28

[+] We are grateful for helpful comments from John D. Wilson, participants of the CESifo area conference on public economics, and three anonymous referees.

[*] Email: juerges@mea.uni-mannheim.de

[†] Email: wolfram.richter@uni-dortmund.de

[#] Email: kerstin.schneider@wiwi.uni-wuppertal.de

1. Introduction

Until recent years, the idea of reforming the school system has been virtually unheard of in Germany. However, that has changed dramatically after the results of two recent international studies on student achievement – TIMSS (Third International Mathematics and Science Study) and PISA (OECD Programme for International Student Assessment) – have gained the attention of a broader public and initiated intense political discussions about the need to reform the German school system. Part of the discussion has focused on too little financial resources flowing into the school system. However, from an economists' point of view, changing the institutional setup of the school system appears to be a more cost efficient approach. Although the short-term costs of changing the environment in which students, teachers, and schools act can be high, it is unlikely that new rules cost much once they are firmly established. One change in the institutional setup as a reaction to PISA has already been implemented: standardized tests will be more widely adopted in Germany in the future. In the present paper we look at whether and under what conditions standardized tests can create incentives for teachers to deliver higher quality teaching.

Teacher quality is often viewed as one of the most important inputs in an education production function. Hence, there is a broad consensus that academic achievement of students can be raised if the quality of teachers improves. Teacher quality has several dimensions: formal qualifications, teaching experience, the quality of teacher education itself, or the teaching effort made by teachers given their formal qualification.[1] There are several routes for education policy to improve the quality of school teachers. In this paper we concentrate on the effort of teachers and the way politics is challenged to improve the incentives for teachers to perform better.

One way to create incentives for teachers is performance related pay. However, the empirical evidence on the relationship between teacher salaries and teacher quality is surprisingly mixed. For example, Lavy (2002, 2004) evaluates a rank-order tournament in Israel. The results suggest positive incentive effects of both performance related salaries and performance related resources given to schools. Moreover, monetary incentives in form of teacher salaries are found to be more cost effective than awarding more resources to the teacher's school. Hanushek, Kain and Rivkin (2004), using data on Texas elementary schools, show that teacher salaries and teacher supply are only weakly related. The composition of teachers within a school district appears to be more affected by characteristics of students than by salary schedules. Apart from tying teacher's pay to the quality of teaching, higher quality

[1] For a survey of the literature see Hanushek and Rivkin (2004).

could be enforced by stricter certification and licensing provisions. Angrist and Guryan (2003) argue that this strategy can fail: the introduction of state-mandated teacher testing in the US has increased teacher wages with no corresponding increase in quality.

Strengthening non-monetary incentives in the schooling system is yet another alternative. This could be simply done by setting common standards, testing students against this standard, and finally making the results public. Teachers will then be motivated to perform well in order to gain non-monetary rewards like reputation or acceptance among colleagues, parents, and students.

In Section 2, we show in a theoretical model that it is optimal to let a teacher's reward (with monetary and non-monetary components) depend on the absolute and relative performance of the teacher's class. To measure performance as an indicator of teacher quality, common standards are needed and students have to be tested against these standards. Measuring the performance of a teacher's class raises the effort put forth by teachers and hence also the academic performance of the students. It is argued that controlling for the socio-economic background of the school or the students can reap additional efficiency gains.[2] An intelligent benchmarking yields the maximal efficiency gains. In the present paper, we use data from the German PISA-E study (PISA-extension) to estimate the effects of external standards on teacher quality. Two types of variables are used to measure teacher quality. First, we use subjective measures of teacher and school quality from the student and parents questionnaires. Second, we analyze student performance as measured by the PISA test score to estimate the effect of external standards on achievement.

Estimating the effects of central exams is not straightforward, because it is typically decided on the country level whether to have or not to have central exams. Thus within a country there is hardly any variation in exam types which makes it difficult to estimate the effects of central standards using national data[3]. Germany is an exception because, due to its federal structure, there has been a long-standing tradition of testing against external standards at the end of secondary schooling in some federal states and of having no standardized tests in

[2] Benchmarking in the school system is already practiced, for instance in the US state of California. Public schools are evaluated based on a so-called Academic Performance Index (API). Each school has to meet a target and is either rewarded for achieving the target or sanctioned for failing to reach the target. Schools are ranked according to their API value. They are ranked within schools of their type, but also, and this picks up the idea of benchmarking, schools are compared to 100 other schools that are similar with respect to demographic characteristics (Similar Schools Rank).

[3]Using international data, the effects could theoretically be estimated (Bishop 1997, 1999; Wößmann 2002) but the drawbacks are manifold (Jürges and Schneider, 2004; Jürges, Schneider and Büchel, 2003). Another example of a country with a federal structure and central exams in some but not all provinces is Canada. Bishop (1997, 1999) exploits the Canadian data.

others.[4] Moreover, the federal states differ with respect to the practice of publishing the data from standardized tests. Saxony, for example publishes the results from the central exit exams on the school's home page together with information about the endowment and profile of the school. In other states the data is published on a more aggregated level and more detailed results are primarily reported back to the school. In most states with standardized tests poor performance of schools causes the supervisory school authority to search for the cause of problem. Hence the German schooling system in its federal diversity is suitable to test for the effects of external standards on teacher quality measures and indirect measures like test scores in international tests.

Jürges, Schneider and Büchel (2005) use data from TIMSS-Germany and estimate the effect of central exit exams (CEE) on test scores in Germany with a difference-in-difference estimator. The estimate is positive and significant but smaller than previous studies had suggested. While Jürges et al. (2005) estimate the effect to be at least one third of a school year equivalent using German data only, Wößmann (2002) uses the international TIMSS micro data and estimates the effect to be as much as about one school year equivalent. Here we present a complementary approach to estimate the effect of central standards, focusing on the quality of teachers. We use data from the PISA-E study to show that teachers' performance is in fact better when standards are enforced through central exit exams. In order to estimate the CEE effects, students in CEE-states and non-CEE states are matched on the basis of the propensity score. The results support the predictions from the theoretical model. Teacher quality is higher in states with CEEs.

The paper proceeds as follows: The theoretical argument is developed in Section 2. Section 3 describes the data. In Section 4 we discuss the empirical model and the results, and in Section 5 we briefly summarize the main findings and conclude.

2. The Model

The theoretical literature almost unanimously argues that CEEs and hence central standards improve student performance and might even raise welfare (Costrell, 1997, Effinger and Polborn, 1999). Central exit examinations are purported to function better as incentives for students, teachers and schools than decentralized examinations (e.g. Bishop, 1997, 1999). Students, for example, benefit because the results of CEEs are more valuable as signals on the

[4] However, as a result of the unsatisfactory performance of German students in international student achievement studies like PISA or TIMSS, standardized tests will be adopted in almost all of the remaining federal states in the near future.

job market than the results of non-central examinations, simply because the former are better comparable. Furthermore, students who have to meet an external standard at the end of their school career have no incentive to establish a low-achievement cartel in class, possibly with the tacit consent of the teachers. Student test results can be used to monitor teacher and teaching quality on a regular basis. Whether incentives to improve teaching quality, arguably an important factor in the education production function, should come solely from reputation effects on the teacher or school level, or in form of higher pay for better teachers is open to discussion (Hanushek et al., 1999; Lavy, 2002, 2004; Glewwe, Ilias and Kremer, 2003).

The following model focuses on the incentives for teachers and describes how teachers determine effort and how a social planner chooses the components of the teacher's reward to maximize a social welfare function. The basic idea is that the planner is interested in setting the right incentives for teachers to put forth effort, which is unobservable. The outcome of teaching, the academic achievement of students, reflects effort to some degree, but achievement is an imperfect measure of effort when classes are not homogenous with respect to their average ability. With heterogeneous classes the planner does not know for sure how much effort the teacher has invested. The literature on yardstick competition shows how a first-best level of welfare can be obtained by competing away the asymmetry of information (Armstrong, Cowan and Vickers, 1994). The following model is an application of yardstick competition to the schooling system.

First consider the teachers decision on teaching effort. Each teacher is allocated to one class *i*. The index *i* thus uniquely identifies teachers and their classes. The average ability of students in class *i* is $\tilde{\theta}_i$ (the tilde denotes stochastic variables). Average ability of the students differs between classes, but we assume that there is no sorting of students by ability and that the average ability of the class is only known to the teacher but not to the planner. Let $\tilde{\theta}$ be the benchmark for $\tilde{\theta}_i$. The benchmark $\tilde{\theta}$ could be the average ability of a set of classes against which $\tilde{\theta}_i$ is compared. Since average ability is stochastic and students are not sorted by ability we get $E\tilde{\theta} = E\tilde{\theta}_i = \mu$, $\operatorname{var}\tilde{\theta}_i = \sigma^2$, $\operatorname{cov}(\tilde{\theta}_i, \tilde{\theta}) = r\sigma^2$, and $r>0$. The positive covariance between the ability of students in class *i* and the benchmark ensures that there is no systematic sorting of teachers and students. Thus low ability students cannot always be an excuse for the poor performance of teachers' classes. This relationship is crucial for the argument. Only if the ability of students in a class and its benchmark are positively related, it is meaningful to compare academic achievement and to condition the teacher's reward on the relative academic achievement of the students.

Student achievement, $\tilde{a}_i$ (as measured in e.g. central exams or standardized tests like PISA) depends on the ability of the students and teacher's effort, e_i. In particular we choose an additive structure

$$\tilde{a}_i = \tilde{\theta}_i + e_i .$$

The achievement of the benchmark is denoted by $\tilde{a} = \tilde{\theta} + \tilde{e}$.

The teacher's reward, $\tilde{W}_i$, consists of a basic salary and a bonus that depends on the performance of the own class and also on the performance of the benchmark

$$\tilde{W}_i(a_i, a\ ) = \overline{W} + \alpha\tilde{a}_i - \delta\tilde{a}\ \text{, with } \alpha, \delta \geq 0 .$$

Note that the bonus does not have to be a monetary bonus but could be reputation or recognition by students, parents or colleagues. Being in a school with a high reputation can be quite valuable for a teacher. Similarly, being assessed as a (relatively) bad teacher can cause disutility and might set strong incentives to improve by working harder. We choose the interpretation of $\alpha\tilde{a}_i - \delta\tilde{a}$ as non-monetary components of the teacher reward to apply the model to the German schooling system. Teacher's pay in Germany is basically not related to performance but rises with the age of the teacher. Thus, the career profile of a German teacher is fairly flat. Nevertheless, some federal states decided to make the quality of teaching visible and hence comparable by testing students centrally, thereby allowing the reputation of a teacher to depend directly on the quality of the output: student achievement.

The parameters α and δ are policy parameters in this model. If they assume strictly positive values, the teacher's reward depends on the absolute and relative performance of her class. If only α is positive, the reward depends on the performance of the own class only, but it is not feasible to compare the performance of the teacher's own class to the performance of the benchmark. Positive values of δ indicate that recognition depends also on the performance of the benchmark. Put differently, if my class performs well, I gain recognition. However, if the benchmark performs well, my results are worth less than if the benchmark performs poorly. If α and δ are both zero, teachers receive a basic, performance independent salary only. This is the case if performance is not measured and no benchmark exists against which to compare the achievement of the teacher or the students, respectively. Benchmarking requires a common standard for measuring achievement, which is enforced by means of central exams.

In the following we show that a social planner would optimally choose positive values for both parameters, α and δ. The choice of some positive α is a direct means to elicit teacher's effort. The choice of a positive value for δ is less obvious and needs to be proven. As we will show, δ is smaller than α in the optimum. However, it is larger the stronger the correlation of the average ability $\widetilde{\theta}_i$ and the benchmark $\widetilde{\theta}$. Thus benchmarking is socially desirable only to the extent to which comparability of abilities is given.

Teachers derive utility from the expected reward, but utility also depends negatively on the work effort. Reward and effort have to be traded off. Moreover, if teachers are risk averse, they do not like uncertain rewards. We write the teachers expected utility function as

$$E(\widetilde{U}_i) = E(\widetilde{W}_i - \tfrac{1}{2}e_i^2) - \tfrac{1}{2}\gamma \operatorname{var}(\widetilde{W}_i).$$

Using the expressions for $\widetilde{W}_i$ and e_i, we get

$$E(\widetilde{W}_i - \tfrac{1}{2}e_i^2) = \overline{W} + (\alpha - \delta)\mu + \alpha e_i - \delta e \; - \tfrac{1}{2}e_i^2 \text{ and} \qquad (1)$$

$$\operatorname{var}(\widetilde{W}_i) = (\alpha^2 + \delta^2)\sigma^2 - 2\alpha\delta\sigma^2 r.$$

Since the variance does not depend on the effort e_i, teachers determine optimal effort by maximizing (1), which results in $e_i^* = \alpha$. Setting $\widetilde{W}_i^* := \widetilde{W}_i|_{e_i = e_i^*}$ and assuming symmetry, $e_i^* = e^*$, we obtain

$$E(\widetilde{W}_i^* - \tfrac{1}{2}e_i^{*2}) = \overline{W} + (\alpha - \delta)\mu + \tfrac{1}{2}\alpha^2 - \delta\alpha.$$

The social planner decides on the policy parameters, i.e. the structure of the teacher reward. In decentralized systems, the social planner could be the principal of the school, in centralized systems it could be the ministry of education. The social planner maximizes a welfare function of the type

$$G = G(\widetilde{a}_i^*, \widetilde{W}_i^*)$$

with $\frac{\partial G}{\partial \widetilde{a}_i^*} > 0$ and $\frac{\partial G}{\partial \widetilde{W}_i^*} < 0$, i.e., the social planner is interested in the academic performance of the students but wants to keep the rewards low. Assuming additivity yields

$$G = E(a_i^* - \widetilde{W}_i^*) \quad = \mu + \alpha - \left[\overline{W} + (\alpha - \delta)\mu + \alpha^2 - \delta\alpha\right]$$

The planner maximizes the welfare function by determining the optimal structure of teachers' reward, respecting the participation constraint. Thus she

$$\max_{\overline{W},\alpha,\delta} G \quad s.t. \quad E(\widetilde{W}_i^* - \tfrac{1}{2}e^{*2}_i) - \tfrac{1}{2}\gamma \operatorname{var}(\widetilde{W}_i^*) = const.$$

The corresponding Lagrangean is

$$\begin{aligned} \Lambda = & [\mu + \alpha - \overline{W} - (\alpha - \delta)\mu - \alpha^2 + \delta\alpha] \\ & + \lambda\left[\overline{W} + (\alpha - \delta)\mu + \tfrac{1}{2}\alpha^2 - \delta\alpha - \gamma\sigma^2(\tfrac{1}{2}\alpha^2 + \tfrac{1}{2}\delta^2 - r\alpha\delta)\right] \end{aligned} \tag{2}$$

Partial differentiation with respect to $\overline{W}$ yields $\lambda = 1$.

Using $\lambda = 1$ in (2) gives

$$\Lambda = (\mu + \alpha) - \tfrac{1}{2}\alpha^2 - \gamma\sigma^2(\tfrac{1}{2}\alpha^2 + \tfrac{1}{2}\delta^2 - r\alpha\delta). \tag{3}$$

Differentiating (3) with respect to δ yields

$$\delta = r\alpha, \tag{4}$$

and finally from the first-order condition with respect to α we get

$$\alpha = \frac{1}{1 + \gamma\sigma^2(1 - r^2)}. \tag{5}$$

Note that $\alpha > 0$, whereas $\delta > 0$ only if $r > 0$. Hence it is always optimal to reward teachers according to the absolute performance of the class. However, it is only optimal to reward teachers also according to relative academic achievement if comparability can be ensured. The better the comparability as measured by a large value of *r*, the better the benchmark. In case of perfect correlation, *r*=1, the first best, $1 = \delta = \alpha = e_i^*$, is obtained. This raises the issue on how to choose the benchmark against which to compare the achievement of class *i*. Clearly, if $\widetilde{\theta}$ is the average ability of all students in the country, $\widetilde{\theta}$ is non-stochastic and *r* vanishes. As a result $\delta = 0$. The interpretation is that there is less gain in social welfare from benchmarking when comparability cannot be ensured. But even if *r*<1, it still pays to reward teachers according to absolute academic achievement, although the first best is not achieved. The reason is that teachers are assumed to be risk averse, and the social planner has to account for this as the structure of the teacher reward affects the participation constraint. The more risk averse teachers are, or the larger the variance of students' average ability, the more costly it is to reward teachers according to student achievement.

To summarize the main results of the theoretical model: We have demonstrated that it is efficiency enhancing to let teachers' reward depend on absolute and relative performance measures based on the academic achievement of students. The requirement for this is a standardized evaluation of student achievement in form of centralized high-stakes testing,

e.g., a central exam. Thus we expect the quality of teachers to be higher and the performance of students to be better when achievement is measured and published. Moreover, efficiency gains can be realized if the performance of classes as an indicator of teacher quality is evaluated relative to a good benchmark. This can be achieved by controlling for observables like the socio-economic background of students. In the following empirical part of the paper, we apply the model to the German schooling system and use the institutional variation to test whether teacher quality is in fact higher with central exams.

3. The practice of CEEs in Germany and the German PISA-E data

The data used in the empirical analysis are drawn from the German PISA 2000 extension study (PISA-E)[5]. The OECD-Programme for International Student Assessment – PISA – aims to assess the knowledge and skills of students approaching the end of compulsory schooling in the basic fields of reading, mathematics, and science.[6] A total of 32 countries participated in the first assessment in 2000 with the focus of the testing being on the reading literacy of 15 year olds students. In 2003 the major domain was mathematical literacy and in 2006 the focus will be scientific literacy. Students are not only tested in the respective fields, but they are also asked to complete a detailed questionnaire on their teachers and schools and on their general background. The data are augmented by "home questionnaires", to be completed by parents and "school questionnaires", completed by school principals. In each country tested between 4,500 and 10,000 students. In Germany, 5,000 students from 219 schools participated in the first PISA test.

Germany complemented PISA 2000 by a national extension, called PISA-E, which was conducted simultaneously with the PISA test. About 40,000 students from 1,300 schools participated in PISA and PISA-E combined. Students eligible for the test were drawn in a two-stage probability sample design. Within each federal state and school type, 25 schools were drawn with sampling probability proportional to school size.[7] Within each school, two samples were drawn. The first was a sample of 15 year olds (target n = 25), independent of the grade. This was the main sample, with the same eligibility criteria as the international

[5] The data are available freely on the website of the German *Kultusministerkonferenz* (Ministries of education of the federal states; http://www.kmk.org/).

[6] One argument against central (high-stakes) tests is the possibility of teaching-to-the-test (for a discussion see e.g. Lazear, 2004). Our data, however, stems from the PISA-study and not from exams based on a national curriculum. Thus teachers in Germany were not familiar with the test content, and, if teaching-to-the-test is a relevant problem, it should not affect the PISA results. Another difference between low- and high-stakes testing is student motivation. It is presumably higher in high-stakes so that low-stakes test results might be a downward-biased estimate of the true student ability.

[7] In Thuringia and Saxony, 75 middle schools were sampled.

PISA study. The second was a supplementary sample of 9th graders who are not 15 years old (target n = 10).[8] In the present study, we work with the main sample of 15 year olds only, for two important reasons. First, only in this sample we are able to identify schools as the primary sampling units, i.e. only in this sample we are able to tell which students belong to the same school. This is important for the computation of correct standard errors. Second, only in this sample we have available test results that are comparable across federal states.[9] Although the data include information on all 16 German federal states and all school types, we further restricted our sample by excluding all students from Berlin and Hamburg as well as all students from comprehensive schools (*Gesamtschule*, see below). In each case, non-participation rates are considered as too high to yield reliable results (Baumert et al. 2002).

Before we discuss the practice of CEEs in Germany, we briefly describe the German school system in Figure 1[10]. All children in Germany attend primary school, which covers grades 1 to 4, or in some states grades 1 to 6. There is no formal exit examination at the end of primary schooling. Rather, students are generally allocated to one of the three secondary school types on the basis of the primary school's recommendation. If the primary school's recommendation conflicts with the parents' wishes, however, the final decision about the future course of education lies either with the parents, the secondary school, or the school supervisory authority, depending on the federal state. Thus it is the idea of the German schooling system to sort students according to ability. Note that it follows from the theoretical model that comparison of student achievement across school types is undesirable but that a comparison of teachers within a school type is preferable on efficiency grounds.

<about here Figure 1>

The *Hauptschule*, *Realschule* and *Gymnasium* are the three main types of secondary school; each leads to a specific leaving certificate. The *Hauptschule* provides its students with basic general education, and usually comprises grades 5 to 9 (or 10 in some states). The *Realschule* provides a more extensive general education, usually comprising grades 5 to 10. The *Gymnasium* provides an in-depth general education covering both lower and upper secondary level, and usually comprises grades 5 to 13 (or 12 in states in eastern Germany). Depending on their academic performance, students can switch between school types. A fourth type of school is the *Gesamtschule* (comprehensive school). This type of secondary school offers all lower secondary level leaving certificates, as well as providing upper

[8] For a detailed description of PISA-E and its sample design see Baumert et al. (2002).
[9] In the sample of 9th graders it is not possible to identify 15 year olds and non-15 year olds belonging in the same school. Moreover, the data release only contains test results that have been standardized by federal state.
[10] A detailed description of the German school system can be found in Jonen and Boene (2001).

secondary education. It only plays a minor role in most federal states with less than 10 percent of all students attending a comprehensive school.

As mentioned at the outset, decisions concerning the institutional settings of the schooling systems are largely determined on the level of the federal states in Germany. One prominent example of state-specific institutions is the existence of external standards in form of central exit exams (CEE) that allow to compare the quality of teachers by comparing test results, i.e. the academic achievement of the students.

Central exit examinations are most common at the end of upper-secondary education (see Table 1). In 2000, seven out of the sixteen German federal states had a central *Abitur* (high-school diploma) at the state level. These states are concentrated in the south (Baden-Württemberg, Bavaria, Saarland) and east (Mecklenburg-Western Pomerania, Saxony, Saxony-Anhalt, Thuringia). The other states had decentralized systems, where teachers design problems for exit examinations individually subject to the approval of the school supervisory authority. Six states had central exit examinations at the end of *Realschule* and only four had them at the end of *Hauptschule.* Note that different CEEs are designed for different school types within a federal state, such that comparisons of exam results across school types are not possible. According to our theoretical argument made in the previous section such comparisons are neither needed nor wanted to achieve efficiency.

<about here Table 1>

This institutional variation found in Germany allows to test empirically for the effects of central exit exams on the quality of teachers. However, estimating the effect of CEE is not straightforward for various reasons. Teachers' effort or the quality of teaching is unobservable. PISA-E contains a large set of items that can be used to construct indices of teaching quality, which are not necessarily unrelated. Students evaluated their classes and teachers with respect to several dimensions such as achievement pressure, teacher support, disciplinary climate, clarity of instruction, excessive demands, and teachers' individual orientation. Parents were asked to evaluate teachers' demands and efforts, and their overall satisfaction with the school. In addition to these subjective indicators we also use student test results in PISA-E as a more objective indicator of teacher effort. Unlike in TIMSS, teachers were not interviewed in PISA-E, so that we have no self-assessed measures of teacher effort.

<about here Table 2>

The qualitative teacher variables derived from the students' reports are listed in Table 2. Here, we only mention the number of items used to construct the indices and their

reliability (measured by Cronbach's α). Overall, the reliability of the indices is at acceptable to good levels. A detailed list of all items can be found in the Appendix. Here, we only give a short description:

- *Achievement pressure* measures the frequency with which teachers tell their students to work harder.
- *Teacher support* measures the frequency with which teachers help students when they have problems of understanding.
- *Bad disciplinary climate* measures the frequency with which bad discipline among students undermines teaching.
- *Clarity of instruction* measures the frequency with which lessons and exercises are clearly structured.
- *Excess demand* measures the frequency with which students think that teachers ask too much of them.
- *Individual orientation* measures the frequency with which teachers commend below-average students who make progress.

Students were asked to evaluate teachers in both mathematics and German classes. For mathematics classes, we have two additional indicators, the frequency of repetitive exercises and the frequency of innovative exercises (i.e. exercises that require to apply skills in changing contexts). Parents' evaluations are measured by answers to single questions on academic level, teachers' effort and overall satisfaction with the school.

Besides the subjective judgements of students and parents we use PISA-E test results as more objective indicators of teacher effort. Note however that, since central exams create incentives for both teachers and students (Jürges et al. 2005), test score differences do not identify the pure effect of CEEs on teacher effort but the combined incentive effect on teachers and students. PISA-E does not contain a single test score for each student but five so-called plausible values (Mislevy 1991). This method, closely related to multiple imputation, has been developed for large-scale low-stakes testing such as PISA (originally developed for the US National Assessments of Educational Progress (NAEP), see e.g. Allen, Donoghue & Schoeps (2001)). Because PISA aims at testing broad areas of proficiency in limited testing time, it is not possible to give each student enough items to obtain a precise estimate of their proficiency in each area. Instead, students are given a curtailed set of items (test books), so that test results are not directly comparable across students with different test books. The solution to this problem is to compute plausible values. Loosely speaking, these are informed

guesses about the test score a student would achieve if he or she was tested on all items and took the test seriously. Each single plausible value is an unbiased estimate of student ability. All results reported below are based on all five plausible values, with differences being averaged across regressions including different plausible values and standard errors computed as the square root of the sum of average within and between regression variances (Mislevy 1991).

<about here Table 3>

Table 3 summarizes raw differences in student and parent-assessed teacher effort and student achievement between states with and without central exams (states are classified according to Table 1 throughout the empirical part of the paper). We have standardized all multi-item scales to mean zero and variance one, so that differences can be interpreted in terms of standard errors. The parents' variables are dichotomised and indicate whether parents think that the academic level is too low or far too low, whether they think that teachers exert themselves not at all or only a little bit for their students, and whether parents are dissatisfied or very dissatisfied with the school of their children.

Note also that we report separate results for the three school types *Hauptschule*, *Realschule*, and *Gymnasium*. Students in *Haupt*- and *Realschule* take central exams at the end of lower secondary schooling, i.e., at the end of grade 9 or 10. We thus expect stronger effects of CEEs in these types of schools than in *Gymnasium*, where the central exams are still three to four years in the future.

Interestingly, there are only two variables that are significantly different between CEE and non-CEE states in all three school types: mathematics performance and the parents' assessment of the school's academic level. Students in CEE-states have higher test scores in mathematics, and the difference is largest in *Hauptschule*. CEEs could thus be interpreted as particularly beneficial for weaker students who are typically allocated to this type of school. Although the similar pattern for reading performance seems to substantiate this interpretation, there are two reasons why it might be wrong. First, the share of students attending *Hauptschule* is typically higher in CEE states, i.e. the student population differs systematically between both states. Since students in *Hauptschule* represent the bottom of the ability distribution, the average ability of students will be systematically higher the larger the proportion of students in this type of school. Second, the raw difference in test scores is most likely not an unbiased estimate of the causal effect of external standards. Both issues will be

taken up in the remainder of the paper, when we control for observed differences using a propensity score approach.

The range of score differences in mathematics is from .129 to .379 standard deviations. These differences are somewhat smaller than those found in Jürges et al. (2005) for the German TIMSS middle school sample, where the raw difference in mathematics scores was .433 standard deviations (*Hauptschule* and *Realschule* combined).

The fact that a larger proportion of parents in non-CEE states than in CEE-states consider the academic level of their children's schools as too low is interesting. If schools in CEE-states were "better" only because parents in CEE-states had on average stronger preferences for better education, one would not expect such a difference. In a sense, schools in CEE-states are thus "overshooting" in relative terms, or rather, schools in non-CEE states are "undershooting". This could mean that differences between both regimes not only reflect differences in tastes for education.

Although none of our direct teacher quality indicators is statistically significantly different between CEE and non-CEE states across *all* school types, there are several systematic patterns. Teachers in CEE states exert more pressure on their students to perform well, they create a more disciplined climate in class (in particular in *Hauptschule*), and exercises are less repetitive and have more variety. In Gymnasium, teachers in CEE states are less supportive when students have trouble understanding and students think more often that demands are excessive. When their children are in *Realschule* or *Gymasium*, parents in CEE states have a more favorable view of their children's schools. They think less often that teachers exert too little effort and they are less often dissatisfied with the school as a whole.

All differences discussed so far are raw differences between schools with and without CEEs. The socio-economic background of the students varies across states and school types. Below, we will control for these variations. Table 4 describes background variables by CEE status and school type. The table neatly shows how students from different backgrounds are sorted into school types. Consider parental education, which we measure as the highest degree obtained by either parent. The proportion of parents with a high education level (i.e. with a college degree) is about 50 percent in *Gymnasium* but less than 10 percent in *Hauptschule*. Differences between CEE and non-CEE states are rather small, except in *Hauptschule*, where we find about 10 percent more parents with medium education (i.e. who finished upper secondary school) at the expense of low education. Other indicators, like the number of books at home, whether there is classic literature at home, or whether parents have read to the child before it was able to read by itself are often found to be better indicators of student

background than formal education. Here, we also find striking differences between schools but only small differences across states. Family structure, measured by the percentage of children living with single parents, is rather uniform across states and school types. Household wealth, coarsely measured by number of cars per adult in the student's household, shows the expected variation across school types but only little variation across states. Thus the heterogeneity with respect to the students' background which would make it - according to the theoretical model - more problematic to assess teacher quality based on student performance, is in fact significantly reduced by allowing comparisons only within a school type.

The main difference between CEE and non-CEE states is the proportion of students with a migration background, here measured by the language spoken at home. In *Realschule* and *Gymnasium*, the proportion of students who don't speak German at home is only half as large in CEE than in non-CEE states. One reason for these differences is that CEEs are more common in the East than in the West, a heritage of the former GDR's school system, while at the same time the proportion of migrants is much lower. More than one half of all *Realschule* students with CEE are from East Germany, whereas less than 10 percent of those without CEEs are from the East.

<about here Table 4>

4. Estimation and results

In the following we estimate the effect of external standards on teacher quality. Using German PISA data, the most basic approach to identify the effect of CEE on student achievement would seem to estimate *simple differences* between average achievement in CEE states and non-CEE states, controlling for student background and other variables of interest. Simple differences, however, have only limited value because they ignore a potentially confounding effect: the endogeneity of CEEs because of self-selection.

Although it cannot be ruled out completely that parents vote with their feet and move between federal states in order to send their children to schools with or without a central exit examination, this seems to be rather unlikely. We therefore assume that the treatment status is exogenous given the institutional arrangement in each federal state. However, in the long run institutions can change. The existence of CEEs might reflect unobserved variables such as the electorate's preferences for education, i.e. parental attitudes towards education and

achievement in school. When CEEs are correlated with such attitudes, simple differences between CEE and non-CEE states are a biased measure of the CEE effect.

The attempt to estimate the effect of CEE is subject to the fundamental problem of causal inference, namely that it is impossible to observe the individual treatment effect (Holland, 1986). One cannot observe the same student at the same time as being student in a state with and without CEE. In the present paper, we estimate the effect of CEEs using an econometric matching estimator. Matching estimators have recently gained much attention in the labor market literature, in particular in the context of program evaluation (for overviews see e.g. Heckman *et al*. (1998) or Blundell and Costa Dias (2000)). They provide an alternative to instrumental variables when there are no good or convincing instruments. However, every attempt to identify causal effects must make use of generally untestable assumptions. In the case of matching estimators the assumption is that the selection into a treatment is completely determined by observable variables and that given the observable variables the selection into the treatment is random (unconfoundedness assumption). Provided that the unconfoundedness assumption holds, we can interpret the assignment of students into CEE and non-CEE states as a randomized experiment (given all observed characteristics), which in turn enables us to interpret our estimates as causal effects of external standards. If the unconfoundedness assumption does not hold, we still estimate correlations conditional on background characteristics. The simplest form of matching proceeds as follows: For each combination of student characteristics compare the quality of teachers in non-CEE states (the controls). Then compute some average difference with respect to the joint distribution of student characteristics. Of course, the larger the number of variables and the larger the number of possible values, the higher the probability of not having a non-CEE student to compare to a CEE student or vice versa. One solution to this dimensionality problem is to condition the comparison on the propensity score (Rosenbaum and Rubin, 1983), which is just the conditional probability of receiving the treatment given the pre-treatment variables. Rosenbaum and Rubin (1983) show that when the selection into treatment is random given the observables, it is also unconfounded given the propensity score. It is thus possible to compute treatment effects conditional on a one-dimensional index.

Still, when the variables are of high dimensionality, it is often not possible to find members of the treatment group and of the control group with exactly the same propensity score. In order to make propensity score matching feasible, we apply nearest neighbor matching, i.e. each treated individual is matched with the non-treated individual with the "nearest" propensity score.

<about here Table 5>

The variables used to calculate the propensity score are the same covariates as described in Table 4. We compute the propensity score as the linear prediction of a probit regression of being subject to a CEE on these covariates. In order to show that the matching procedure has indeed produced a balanced sample of treated (CEE) and control (non-CEE) students, we calculate the means of all covariates in the matched sample and test whether these are different (see Table 5). First, note that the control group in each school type consists of relatively few different non-CEE students. For instance, in Hauptschule, each of these students contributes $1{,}374/699 \approx 1.97$ observations to the control group. The corresponding number in Realschule and Gymnasium are 4.0 and 2.37, respectively. t-values in Table 5 account for this fact. Overall, the matching procedure was successful in creating a balanced sample. The only notable difference between treatment and control group seems to be the proportion of children with 51-100 books at home in the Hauptschule sample. However, a test that the overall distribution of the number of books is equal across groups does not reject the null hypothesis. We believe that our choice of educational background variables captures heterogeneity with respect to tastes for education in CEE and non-CEE states to a very large extent. Of course, since we are lacking school or teacher data, we cannot totally exclude the possibility that there is still some unobserved heterogeneity left, i.e., unmeasured differences between CEE and non-CEE states that might account for any differences observed after matching. In that case, our estimated would have to be interpreted as conditional correlations.

The matching estimates are displayed in Table 6. The most stable results are similar to those already discussed in the context of the raw differences: First, students in CEE states show significantly better test results than those in non-CEE states. Interestingly, the CEE effect estimated here is often larger than the raw difference. We conclude that students in CEE states perform better because of external standards that are enforced by central exit exams. The results confirm earlier studies by Jürges et al. (2005) and Wößmann (2002). Second, the estimated effect is largest in *Hauptschule* and smallest in *Gymnasium*, which is plausible considering the students' different time horizon. Students in *Haupt-* and *Realschule* will pass their exam within a shorter period than those in Gymnasium, so our result is consistent with the idea that the effect of central exams is stronger when the exams are in the near future. Third, we consider the size of the estimated CEE effect. Size effects are usually reported in terms of school year equivalents. Although this would be possible in principle (we have data on three different grades), it is not very useful because the sample of 8^{th} graders clearly is a negative selection and the 10^{th} graders are a positive selection of all students in 8^{th}

and 10th grade, respectively. Grade differences will thus greatly overestimate average school year differences.

Let us now turn to other, subjective, quality measures. Here, the picture is rather mixed. Apparently, although CEEs have rather consistent effects on student performance, they can have quite different effects on various aspects of teacher performance, depending on the school type. First, achievement pressure is generally perceived to be higher in CEE-states than in non-CEE states; the only significant (and sizeable) difference can be found for *Hauptschule* mathematics. Teacher support is perceived to be worse in CEE states *Realschule* and *Gymnasium*. In *Hauptschule*, one gets the opposite – although insignificant – result. Here, the disciplinary climate is better in CEE states, whereas it tends to be worse in *Gymnasium*. According to the students' reports, instruction is significantly clearer only in German classes in Hauptschule.

<about here Table 6>

Demands are mostly not perceived to be more excessive in CEE states. The differences to non-CEE states are somewhat larger in mathematics than in German classes, but they are either insignificant or only marginally significant. Hauptschule teachers in CEE states are generally more oriented towards individual achievement, that is they show interest in and support the progress of all students, independent of their abilities. For Gymnasium teachers, exactly the opposite holds. According to their students, they are less oriented towards individual achievement, in particular in mathematics lessons.

Critics of central exams often claim that students are taught to the test. If that were the case in Germany, one would expect significantly more repetitions, in particular of exercises relevant for the central exam, and less innovations. We find no empirical support for such critique. In particular students in Hauptschule seem to benefit from more innovative exercises and students in Realschule report less repetitions.

Let us now turn to the parents' view. As already mentioned, parents in CEE states are less likely to say that the academic level of their children's school is too low. At the same time they less often think that teachers exert too little effort and they are less often dissatisfied with the school as a whole. However, in contrast to their views on the academic level, differences between CEE and non-CEE states are not significant – with one exception: Parents whose children are in Hauptschule are significantly more satisfied with the school in CEE states.

Overall, the students' and parents' view of teachers' behavior suggest that at least Hauptschule teachers in CEE-states are perceived to be better teachers. This is less clear for

teachers in other school types. Thus, CEEs appear to primarily benefit the weaker students by raising the quality of teachers. In order to reconcile these findings with the strong and consistent student performance differential, one could argue that the subjective quality indicators from the student questionnaire do not capture teacher quality reliably. Subjective quality indicators across states with different school systems might not be as comparable as objective indicators like test scores. Furthermore, some of the differential in achievement is certainly due to incentives on students rather than teachers.

Finally, the German states with CEEs differ quite substantially with respect to how the data from CEEs are used. However, in none of the German states formal rankings of schools or teachers exist. While in most states the school level CEE data is only known to the school authority and the school itself, Saxony practices an open information policy. Schools publish the data from the CEEs in the Internet together with information about endowment and profile.[11] Thus in Saxony the information for a simple ranking of schools based on exam results and also a ranking which controls for the socio-economic background, is available to the public. According to our theoretical model in Section 2, CEEs should work even better in Saxony than in other federal states with central exams. Unfortunately, it is not possible to seriously estimate this effect since only one federal state is committed to providing the school level information to the public and we have to confine ourselves to some informal discussion. Doing so, we restrict our attention to the federal states in eastern Germany that are more homogenous with respect to the economic situation and the traditions in schooling. The matching estimates suggest that compared to other students in eastern German federal states with CEEs, students in Saxony perform better in all school types. The difference is significant for students in Realschule or Gymnasium.[12] Thus there is at least some indication of better academic performance in Saxony, the only federal state that publishes data from CEEs on the school level in addition to other information and thereby allows for a comparison of schools.

5. Conclusions

The paper has made two contributions to the literature on teacher quality. First we argue that it is optimal to reward teachers depending on the absolute and relative academic achievement of students, because this raises the (unobservable) effort of teachers and efficiency. This is true even if the pay (in monetary terms) is not performance related. If

[11] http://ssdb2.inf.tu-dresden.de/output/

[12] Note that while it is very common in eastern Germany to combine Haupt- and Realschule in a middle school we have information in PISA-E about the school track of the students.

teachers value reputation, they increase effort if the output of teaching – academic achievement of the students – is measured and published. Consequently, academic achievement of students should be tested centrally and made comparable by using a benchmark. The reward mechanism works best if the benchmark is chosen carefully, controlling for observables like the socio-economic background of the school or the students. Second, we used the German PISA-E data to test whether teacher quality is higher when academic achievement of students is evaluated according to a central standard. One particularity of the German schooling system is its federal structure. Some federal states test students centrally whereas others do not. In the German system of CEEs, the benchmarks as a basis for comparison are not perfect. However, the allocation of students into three school types reduces the heterogeneity of student background substantially, thereby improving the quality of the benchmark. Our matching estimates suggest that the quality of teaching tends to be higher in federal states with CEEs. We explain this finding by teachers' response to non-monetary rewards like reputation. Only in CEE states it is possible to measure and to compare student achievement and hence (indirectly) teachers' effort.

6.References

Allen, N.L. Donoghue, J.R., Schoeps, T.L., 2001, The NAEP 1998 Technical Report. Washington, DC: National Center for Education Statistics.

Armstrong, M., Cowan, S., Vickers, J., 1994, Regulatory Reform, Economic Analysis and British Experience, MIT Press.

Adams, R. /Wu, M., 2002, PISA 2000 Technical Report. OECD: Paris.

Angrist J., J. Guryan, 2003, Does Teacher Testing Raise Teacher Quality? Evidence from State Certification Requirements, NBER Working Paper 9545.

Bishop, J.H., 1997, The Effect of National Standards and Curriculum-Based Exams on Achievement. *The American Economic Review 87*, 260-264.

Bishop, J.H., 1999, Are national exit examinations important for educational efficiency? *Swedish Economic Policy Review 6*, 349-401.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Tillmann, K.-J., Weiß, M., 2002, PISA 2000 – Die Länder der Bundesrepublik Deutschland im Vergleich. Opladen: Leske+Budrich.

Blundell R, Costa Dias M., 2000, Evaluation Methods for Non-Experimental Data. *Fiscal Studies* 21, 427-268.

Costrell, R.M., 1997, Can Educational Standards Raise Welfare?, *Journal of Public Economics* 65, 271-293.

Effinger M.R., Polborn, M.K., 1999, A Model of Vertically Differentiated Education, *Journal of Economics* 69, 53-69.

Glewwe, P., Ilias, N., Kremer, M., 2003, Teacher Incentives, NBER Working Paper 9671.

Hanushek, E.A., Kain, J.F. ,Rivkin, S.G., 2004, Why Public Schools Lose Teachers, *Journal of Human Resources 39*, 326-354.

Hanushek, E.A., Rivkin, S.G., 2004, How to Improve the Supply of High Quality Teachers, *Brookings Papers on Education Policy,* 7-25.

Holland, P.W., 1986, Statistics and Causal Inference. *J Am Stat Assoc* 81; 945-960.

Heckman J.J., Ichimura H., Todd P., 1998, Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program. *Rev Econ Stud* 64: 605-654.

Jonen, G., Boele, K., 2001, *The Education System in the Federal Republic of Germany 2000.* German EURYDICE Unit, Bonn. http://www.kmk.org/dossier/dossier_2000_engl_e-book.pdf [2002, July 1].

Jürges, H., Schneider, K., 2004, International Differences in Student Achievement: An Economic Perspective, *German Economic Review* 5, 357-380.

Jürges, H., Schneider, K., Büchel, F., 2005, The Effect of Central Exit Examinations on Student Achievement: Quasi-Experimental Evidence from TIMSS Germany, *Journal of the European Economic Association 3 (5),* forthcoming.

Lavy, V., 2002, Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement, *Journal of Political Economy* 110, 1286-1317.

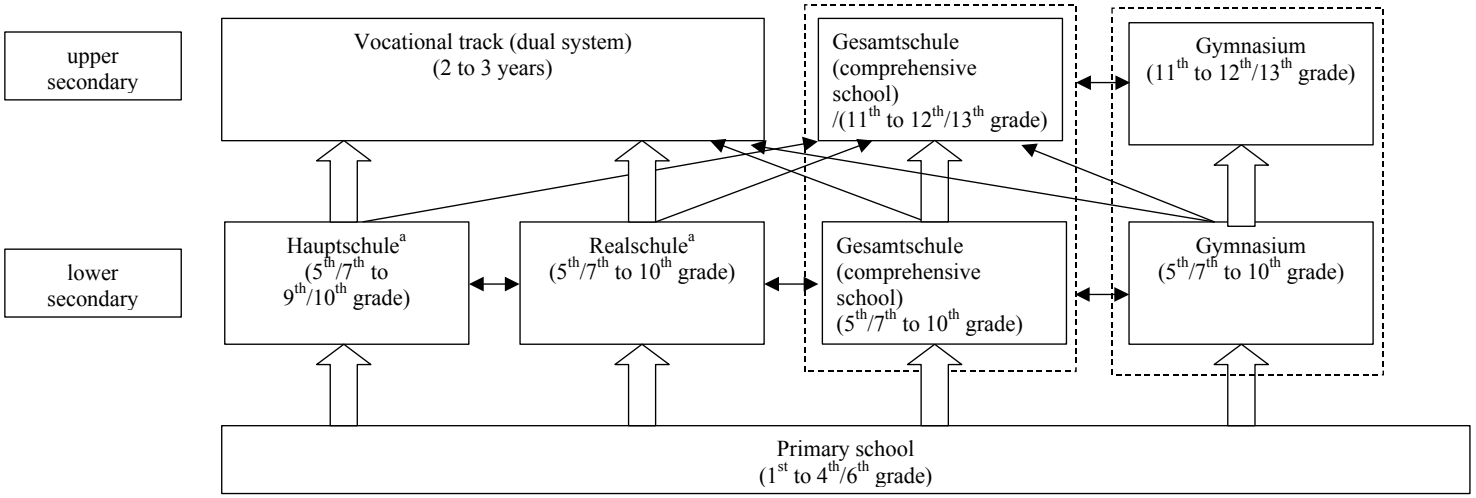
Lavy, V., 2004, Performance Pay and Teachers' Effort, Productivity and Grading Ethics, NBER Working Paper 10622.

Lazear, E. P., 2004, Speeding, tax fraud, and teaching to the test. NBER Working Paper 10932.

Mislevy, R. J. 1991, Randomization-based inference about latent variables from complex samples. *Psychometrika*. 56, 177-196.

Rosenbaum, P.R. and D.B. Rubin, 1983, The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70: 41-55.

Wößmann, L., 2002, Central Exams Improve Educational Performance: International Evidence. Kiel Discussion Papers 397.



[a] Some Eastern German states integrate Haupt- and Realschule in a middle school.

Figure 1: A model of the German school system

Table 1: Federal States with CEE by degree

	Lower Secondary					Upper Secondary	
	Hauptschule	*Realschule*	Middle school (Hauptschule + Realschule)	Comprehensive school	Gymnasium	Comprehensive School	High school diploma (Abitur)
Baden-Württemberg (BW)	+	+			+		+
Bavaria (BY)	+	+			–		+
Mecklenburg-W. Pomerania (MV)	–	+		+	–	+	+
Saarland (SA)	–	–		–	–	+	+
Saxony			+		–	+	+
Saxony-Anhalt (ST)			+	+	–		+
Thuringia (TH)			+	+	–	+	+

No CEEs in Berlin, Brandenburg, Bremen, Hamburg, Hesse, Lower Saxony, North Rhine Westphalia, Rhineland-Palatine, and Schleswig-Holstein. Grey cells: school type does not exits; +: CEE; –: no CEE.

Table 2: Indicators of teacher effort

	Mathematics		German		General	
	# of items	alpha	# of items	alpha	# of items	alpha
Students evaluations of						
achievement pressure	3	.65	3	.58		
teacher support	7	.90	6	.85		
bad disciplinary climate	6	.86	6	.80		
clarity of instruction	5	.65	5	.78		
excess demand	4	.74	4	.74		
individual orientation	3	.77	3	.84		
repetitive exercises	2	.57				
innovative exercises	3	.60				
Parents evaluations of						
school's academic level					1	
teachers' efforts					1	
overall satisfaction with school					1	

Table 3: Differences in outcome variables between CEE and non-CEE states

	Hauptschule		*Realschule*		*Gymnasium*	
	mean CEE minus mean non-CEE	t-value difference	mean CEE minus mean non-CEE	t-value difference	mean CEE minus mean non-CEE	t-value difference
Mathematics						
achievement pressure	0.087	1.737+	0.039	1.049	0.024	0.439
teacher support	-0.015	-0.271	0.067	1.519	-0.109	-2.213*
bad disciplinary climate	-0.127	-2.156*	-0.167	-3.315**	-0.014	-0.248
clarity of instruction	0.001	0.025	0.081	1.872+	0.055	1.127
excess demand	0.030	0.656	-0.047	-1.266	0.075	1.694+
individual orientation	0.071	1.290	0.044	1.109	-0.081	-1.792+
repetitive exercises	0.003	0.050	-0.164	-4.416**	-0.142	-3.350**
innovative exercises	0.106	2.416*	0.094	2.740**	-0.024	-0.552
mathematics score	0.379	6.568**	0.129	2.665**	0.156	4.600**
German						
achievement pressure	0.057	1.135	0.071	2.079*	0.037	0.811
teacher support	0.007	0.145	-0.018	-0.415	-0.097	-2.344*
bad disciplinary climate	-0.216	-3.875**	0.034	0.814	0.037	0.793
clarity	-0.026	-0.510	-0.017	-0.408	0.014	0.320
excess demand	0.001	0.013	-0.008	-0.214	0.059	1.689+
individual orientation	0.073	1.358	-0.030	-0.761	-0.027	-0.649
reading score	0.315	6.028**	0.080	1.831+	0.045	1.463
General						
school's academic level too low	-0.138	-7.041**	-0.111	-6.130**	-0.117	-6.817**
teachers' exert little effort	0.013	0.838	-0.042	-2.900**	-0.036	-2.279*
dissatisfied with school overall	-0.035	-1.255	-0.063	-2.628**	-0.048	-2.036*

+ p<0.10; * p<0.05; ** p<0.01, t-values adjusted for clustering.

Table 4: Description of covariates (prior to matching).

	Hauptschule		*Realschule*		*Gymnasium*	
	Mean Non-CEE	Mean CEE	Mean Non-CEE	Mean CEE	Mean Non-CEE	Mean CEE
Boy	0.538	0.579	0.476	0.475	0.456	0.420
Age	15.374	15.309	15.373	15.355	15.371	15.351
Grade						
8th	0.319	0.216	0.133	0.114	0.058	0.052
9th	0.570	0.754	0.637	0.673	0.608	0.664
10th	0.111	0.029	0.230	0.213	0.335	0.284
Single parent	0.288	0.303	0.270	0.298	0.269	0.274
Parents' education						
low	0.488	0.405	0.333	0.337	0.121	0.130
medium	0.417	0.504	0.488	0.472	0.389	0.381
high	0.095	0.092	0.179	0.191	0.490	0.489
Books at home						
0-10	0.152	0.126	0.048	0.055	0.015	0.007
11-50	0.290	0.285	0.203	0.222	0.069	0.081
51-100	0.257	0.282	0.263	0.250	0.137	0.146
101-250	0.164	0.174	0.249	0.255	0.249	0.265
251-500	0.081	0.078	0.146	0.135	0.275	0.265
500+	0.057	0.056	0.091	0.082	0.255	0.236
Classic Literature	0.224	0.241	0.369	0.441	0.679	0.720
Read to child						
rarely/never	0.226	0.195	0.109	0.097	0.055	0.055
once/month	0.118	0.125	0.079	0.096	0.045	0.049
once/week	0.319	0.366	0.315	0.358	0.244	0.260
daily	0.392	0.357	0.518	0.463	0.665	0.646
East Germany	0.029	0.153	0.042	0.533	0.083	0.379
Speaks no German at home	0.282	0.216	0.127	0.063	0.100	0.046
Cars per adult	0.737	0.760	0.809	0.813	0.855	0.841
N (unweighted)	2,587	1,374	3,758	4,821	3,623	3,682

Table 5: Covariate Differences in Matched Samples

	Hauptschule			*Realschule*			*Gymnasim*		
	Control	Treated	t-value	Control	Treated	t-value	Control	Treated	t-value
Propensity Score	0.348	0.348	0.000	0.582	0.582	-0.000	0.516	0.516	-0.000
Boy	0.593	0.579	-0.488	0.470	0.475	0.193	0.426	0.420	-0.240
Age	15.297	15.309	0.748	15.353	15.355	0.171	15.344	15.351	0.608
Grade									
8	0.221	0.216	-0.176	0.099	0.114	0.996	0.047	0.052	0.523
9	0.750	0.754	0.130	0.691	0.673	-0.754	0.667	0.664	-0.132
10	0.029	0.029	0.059	0.210	0.213	0.160	0.286	0.284	-0.107
Single parent	0.284	0.303	0.712	0.314	0.298	-0.568	0.274	0.274	-0.008
Parental education									
low	0.431	0.405	-0.790	0.312	0.337	0.904	0.116	0.130	0.982
medium	0.479	0.504	0.732	0.483	0.472	-0.391	0.391	0.381	-0.484
high	0.091	0.092	0.060	0.205	0.191	-0.714	0.492	0.489	-0.139
Books at home									
0-10	0.111	0.126	0.902	0.055	0.055	-0.079	0.010	0.007	-0.787
11-50	0.261	0.285	0.950	0.212	0.222	0.505	0.066	0.081	1.240
51-100	0.340	0.282	-2.114*	0.263	0.250	-0.661	0.152	0.146	-0.450
101-250	0.163	0.174	0.488	0.253	0.255	0.141	0.274	0.265	-0.544
251-500	0.063	0.078	1.134	0.134	0.135	0.094	0.263	0.265	0.112
500+	0.063	0.056	-0.494	0.083	0.082	-0.036	0.235	0.236	0.071
Classic Literature	0.242	0.241	-0.045	0.458	0.441	-0.747	0.723	0.720	-0.176
Read to child									
never/rarely	0.134	0.152	0.846	0.074	0.082	0.763	0.044	0.045	0.169
once/month	0.125	0.125	0.035	0.090	0.096	0.536	0.045	0.049	0.440
once/week	0.342	0.366	0.915	0.375	0.358	-0.863	0.253	0.260	0.348
daily	0.400	0.357	-1.562	0.460	0.463	0.106	0.658	0.646	-0.493
Speaks no German at home	0.220	0.216	-0.109	0.056	0.063	0.665	0.047	0.046	-0.148
East	0.152	0.153	0.011	0.533	0.533	0.000	0.379	0.379	-0.000
Cars per adult	0.745	0.760	0.535	0.816	0.813	-0.137	0.844	0.841	-0.149
N (unweighted)	699	1,374		1,205	4,821		1,552	3,682	
N (weighted)	1,374	1,374		4,821	4,821		3,682	3,682	

Note – Statistics computed on the weighted sample; t-values account for multiple uses of observations.
+ p<0.10; * p<0.05; ** p<0.01

Table 6: Differences between CEE and non-CEE states, nearest neighbor matching estimates

	Hauptschule		*Realschule*		*Gymnasium*	
	mean CEE minus mean non-CEE	t-value difference	mean CEE minus mean non-CEE	t-value difference	Mean CEE minus mean non-CEE	t-value difference
Mathematics						
achievement pressure	0.168	2.391*	0.009	0.155	0.041	0.765
teacher support	0.014	0.199	-0.030	-0.503	-0.181	-3.438**
bad disciplinary climate	-0.104	-1.436	-0.038	-0.678	0.079	1.317
clarity of instruction	0.047	0.656	0.053	0.967	-0.019	-0.357
excess demand	0.069	1.121	-0.010	-0.242	0.078	1.671+
individual orientation	0.108	1.476	0.057	1.120	-0.107	-1.926+
repetitive exercises	0.037	0.546	-0.168	-3.568**	-0.074	-1.439
innovative exercises	0.144	2.365*	0.020	0.365	-0.120	-2.211*
mathematics score	0.311	4.805**	0.255	4.301**	0.164	3.493**
German						
achievement pressure	0.059	0.942	0.054	1.166	0.004	0.082
teacher support	0.081	1.276	-0.085	-1.731+	-0.123	-2.649**
bad disciplinary climate	-0.148	-2.123*	0.022	0.334	0.048	1.008
clarity	0.129	1.896+	-0.070	-1.302	-0.027	-0.514
excess demand	-0.021	-0.309	0.010	0.226	-0.015	-0.330
individual orientation	0.144	2.020*	-0.042	-0.956	-0.016	-0.301
reading score	0.309	4.875**	0.252	4.177**	0.160	3.485**
General						
school's academic level too low	-0.140	-4.805**	-0.097	-4.268**	-0.110	-5.624**
teachers' exert little effort	-0.020	-0.859	-0.000	-0.017	-0.024	-1.225
dissatisfied with school overall	-0.047	-2.223*	-0.010	-0.764	-0.011	-0.814

+ p<0.10; * p<0.05; ** p<0.01, t-values adjusted for clustering.

Table A1: Items used to generate teacher effort indices

Student variables	
Achievement pressure	The teacher wants students to work hard. The teacher tells students that they can do better. The teacher does not like it when students deliver sloppy work.
Teacher support	The teacher shows an interest in every student's learning. The teacher gives students an opportunity to express opinions. The teacher helps students with their work. The teacher continues teaching until the students understand. The teacher does a lot to help students. The teacher helps students with their learning. [Mathematics only:] The teacher gives helpful advice for my work
Bad disciplinary climate	The teacher has to wait a long time for students to quiet down. Students cannot work well. Students don't listen to what the teacher says. Students don't start working for a long time after the lesson begins. There is noise and disorder. At the start of class, more than five minutes are spent doing nothing.
Clarity	The teacher gives clear instructions what to do Everything that we do is well planned There are specific rules that we must adhere to The teacher tells us at the beginning of the lesson what to do The teacher summarizes what was done in the previous lesson
Excess demand	Time is too short to finish my work The things we do are too difficult for me The teacher tells us things that I do not understand You stop listening because you do not understand anything
Individual orientation	Our teacher acknowledges improvements even if students are below average When I really make an effort, the teacher commends me even if others are still better than me Our teacher also commends weak students when they make improvements.
Repetitive exercises	We make little progress because we repeat so much We always do the same exercises
Innovative exercises	By some of our exercises, you can really see if you have understood the topic We often apply what we learn to new topics You have to pay close attention because the exercises are similar but always a bit different
Parent variables	
School's academic level	How do you rate the academic level of your child's school – far too low, too low, about right, too high or far too high?
Teachers' efforts	How much do teachers exert themselves for their students – not at all, a little bit, somewhat, much, or very much?
Overall satisfaction with school	How satisfied are you with your child's school – very dissatisfied, dissatisfied, neither dissatisfied nor satisfied, satisfied, or very satisfied?

Discussion Paper Series

Mannheim Research Institute for the Economics of Aging Universität Mannheim

To order copies, please direct your request to the author of the title in question.

Nr.	Autoren	Titel	Jahr
78-05	Daniel Schunk Cornelia Betsch	Explaining heterogeneity in utility functions by individual differences in decision modes	05
79-05	Franz Rothlauf Daniel Schunk Jella Pfeiffer	Classification of Human Decision Behavior: Finding Modular Decision Rules with Genetic Algorithms	O5
80-05	Lothar Essig	Methodological aspects of the *SAVE* data set	05
81-05	Lothar Essig	Imputing total expenditures from a non-exhaustive list of items: An empirical assessment using the SAVE data set	05
82-05	Mathias Sommer	Trends in German households' portfolio behavior – assessing the importance of age- and cohort-effects	05
83-05	Lothar Essig	Household Saving in Germany: Results from SAVE 2001-2003	05
84-05	Lothar Essig	Precautionary saving and old-age provisions: Do subjective saving motive measures work?	05
85-05	Axel Börsch-Supan Lothar Essig	Personal assets and pension reform: How well prepared are the Germans?	05
86-05	Lothar Essig	Measures for savings and saving rates in the German SAVE data set	05
87-05	Felix Freyland ed. by Axel Börsch-Supan	Household Composition and Savings: An Overview	05
88-05	Felix Freyland ed. by Axel Börsch-Supan	Household Composition and Savings: An Empirical Analysis based on the German SOEP Data	05
89-05	Hendrik Jürges	Unemployment, restrospective error, and life satisfaction	05
90-05	Hendrik Jürges	Gender Ideology, Division of Housework, and the Geographic Mobility Families	05
91-05	Hendrik Jürges Wolfram F. Richter Kerstin Schneider	Teacher quality and incentives – Theoretical and empirical effects of standards on teacher quality	05