CESifo Working Papers
www.cesifo.org/wp

Total Instructional Time Exposure and Student Achievement:
An Extreme Bounds Analysis Based on German State-Level Variation

Philipp Mandel
Bernd Süssmuth

CESIFO WORKING PAPER NO. 3580
CATEGORY 5: ECONOMICS OF EDUCATION
SEPTEMBER 2011

CESifo *Working Paper No. 3580*

Total Instructional Time Exposure and Student Achievement:
An Extreme Bounds Analysis Based on German State-Level Variation

Abstract

Using pooled data on instructional time and student performance by subject, our study finds evidence for the school inputs-student achievement relationship for German states. This finding is robust both to the inclusion of state fixed effects and in an extensive extreme bounds analysis. It stands in contrast to the majority of related studies. We argue that this is due to an error-in-variables problem and implied misinterpretation of existing studies that disregard the fact of learning being a cumulative process by relying on rather poor proxies for instructional time. Highschool ninth graders from the OECD Programme of International Student Assessment (PISA-E) tests' bottom percentiles bene.t most from extra-instructional time measured in cumulated form from first up to ninth grade. Besides total instructional time exposure, we identify eight further social environment and institutional variables with robust impact on student performance. In contrast to instructional time hardly any of these factors can be affected by policy in the short run.

JEL-Code: I210, I280, L380.

Keywords: education production function, student performance, school resources.

Philipp Mandel
Institute for Empirical Research in Economics (IEW) / Econometrics
University of Leipzig
Grimmaische Strasse 12
Germany – 04109 Leipzig
mandel@wifa.uni-leipzig.de

Bernd Süssmuth
Institute for Empirical Research in Economics (IEW) / Econometrics
University of Leipzig
Grimmaische Strasse 12
Germany – 04109 Leipzig
suessmuth@wifa.uni-leipzig.de

We thank Carolin Amann, Fabian Feierabend, Constantin Tabor, Stephanie Najort, Marcus Strobel, Bastian Gawellek, and Alexander Mandel for excellent research assistance, particularly, in assembling the data. Thanks are also due to Marco Sunder for many helpful comments and suggestions.

1 Introduction

This paper is in the tradition of the seminal study by Card and Krueger (1992) in that it relies on cross-state variation in education inputs and institutions. There is a continuing debate on whether schooling resources have a bearing on student outcomes (Krueger 2003; Hanushek 2003, 2004, 2006a). Todd and Wolpin (2003) see econometric misspecification and failure to account for major determinants of student achievement as the central problem in correctly identifying the relationship. Recently, a little studied input receives growing attention: In Coates (2003), Eren and Millimet (2007), Marcotte (2007), Marcotte and Hemelt (2008), and Lavy (2010) the focus is on instructional time by subject.

Our study is unique in using data of instructional time *cumulated* from all academic years leading up to the test date in each of the two subjects math and reading. We rely on cross-state variation in Germany, where 16 states share the same cultural and legal system but pursue different education policies (Schulte 2004). The fact that German states have responsibility for both primary and secondary education, makes our data particularly suited to analyze the impact cumulative instruction has on student achievements. As for the educational instruction–performance relationship, Marcotte (2007) and Marcotte and Hemelt (2008) are the only studies that focus on and consider the cumulative nature of instruction as determinant of student performance. They make use of intra-state school level and snowfall (unscheduled closings) data for students in grades 3, 5 and 8 in Maryland. Their approximation of a cumulated effect is based on the hypothesis that the lower the grade, the less room exists for making up and the higher the relative weight of lost instruction. Therefore, the instructional time shortfall effect decreases with grade. However, this cumulated effect is of second order as only measures of total snowfall in the academic year of the test date (Marcotte 2007) or in the preceding 3 years (Marcotte and Hemelt 2008) are considered. Coates (2003) relying on district-level data for Illinois and considering uncumulated daily instruction in third grade classes, finds that a 10 percent increase in mathematics instruction per week raises the average math score by about 0.4 percent. Similar small effects are found for English instruction. Eren and Millimet (2007) analyze the joint effect the daily number of class periods and the average class length (in minutes) has on cognitive test results of US public schools 10th graders (National Education Longitudinal Study of 1988). Only uncumulated 10th grade instructional conditions are considered. Their reform-type finding is that changing the

system from one with ≤ 6 daily classes lasting ≥ 51 minutes to another one with seven 45-minutes classes increases test scores by 2 percent.

In sociology of education, sociolinguistics, and the neurosciences, a recent body of literature is concerned with the structured processing of knowledge ascribing it to different modes of learning. An essential element of these modes is to conceptualize knowledge-building through cumulative learning. Accordingly, learning is found to be a cumulative process during which new knowledge is dependent and based on a precedingly acquired stock of knowledge; see, among others, Freebody *et al.* (2008), Maton (2009), and Yew *et al.* (2011). Yet, to the present, empirical work in the economics of education literature does not take these findings into account and relies on rather poor measures of instructional time as independent variables. This is not to say that the relevance of current and past inputs and the deficiencies of approaches abstracting from input histories is not recognized in the literature (Card and Krueger 1996, Todd and Wolpin 2003). It is simply and mostly due to data limitations not done. Typically, estimates are based on instructional time proxies such as students' self-reported hours of instruction per week as they relate to the respective test year. Todd and Wolpin (2003) refer to these measures as "contemporaneous inputs." Given insights from sociology and neuro-sciences or, in general, from "multidisciplinary empirical literature studies" (Todd and Wolpin 2003, p. F3), however, a cumulative measure such as cumulated instructional time (henceforth, CIT) from first grade to test year for each observed cohort is required. This becomes all the more obvious if we look at an arbitrarily chosen mathematics sample task from PISA (OECD 2009, p. 125) that reads as follows:

Mathematics Unit 27: *A result of global warming is that the ice of some glaciers is melting. Twelve years after the ice disappears, tiny plants, called lichen, start to grow on the rocks. Each lichen grows approximately in the shape of a circle. The relationship between the diameter of this circle and the age of the lichen can be approximated with the formula:*

$$d = 7.0 \times \sqrt{(t - 12)} \quad \text{for} \quad t \geq 12,$$

where d represents the diameter of the lichen in millimetres, and t represents the number of years after the ice has disappeared.

In the two questions that followed students are asked to calculate (Q1) the diameter of the lichen, 16 years after the ice disappeared and (Q2) the number of years that the ice disappeared at a spot, where the diameter of some lichen is found to be 35

millimetres. Most obviously (Q1) and (Q2) can be answered based on knowledge on subtracting, multiplying, taking roots, and the technique of substitution or trial and error that students acquired over several years starting from the very first grade. These skills might have barely something to do with instruction in the ninth grade.

The vast majority of studies analyzing data from international student assessments like PISA or TIMSS (Third International Math and Science Study) only considers instructional time at the relevant grade level as an explanatory for test performance. Usually, these "snap-shot"-type measures of instructional time are drawn either from students' self-reporting or from test add-ons such as principals' questionnaires as, for example, in Baker *et al.* (2004), Lavy (2010), and Wössmann (2010). Given the serious error-in-variables problem contained in these measures, it does not come as a surprise that their impact on student test scores is mostly estimated as not statistically different from zero (as, for example, in Wössmann 2010). Besides studies that use snap-shot query-based measures, there are few studies that rely on the length of a school day and/or the length of a school year to proxy instructional time as input in an education production or more general Mincer-type framework (Dewey *et al.* 2000, Lee and Barro 2001, Pischke 2007). Two exceptional studies that try to consider, at least, partially the cumulative nature of learning are Afonso and St. Aubyn (2006) and Moser and Angelone (2009). The first of these studies makes use of a variable "intended instruction time in public institutions in hours per year for the 12 to 14-year-olds" cumulated for the three years preceding the PISA 2003 tests for 25 different countries. Similarly, Moser and Angelone (2009) also partially accumulate instructional time in Swiss cantons from seventh to ninth grade to estimate its impact on PISA 2006 scores. Again, as expected given the rough approximation of total instructional time for both studies there is no clear-cut evidence for the input–achievement relationship. Afonso and St. Aubyn (2006) find no significant evidence. The evidence reported in Moser and Angelone (2009) is mixed and depends on the subject studied. A significant positive association is found for instructional time and test scores in math.

Our study contributes along two lines to the literature. First, it addresses the outlined serious error-in-variables problem and implied shortcoming in empirical work on education inputs and outcomes measured by international student test scores by using data on *total* instructional time that students were exposed to from first to ninth (i.e. test date) grade by subject. Secondly, besides quantifying the actual impact of cumulative instructional time on PISA test scores, we address model uncertainty and robustness,

which are also issues that are widely ignored in the literature, by using extreme bound analysis (EBA) techniques for our estimates.

The rest of the paper is organized as follows. Section 2 outlines our data and used methods. Section 3 reports and discusses our findings. Finally, Section 4 concludes.

2 Data and methodology

2.1 Data

Data on student achievement are drawn from the national extension of the PISA studies in 2000, 2003, 2006 (PISA-E) and from the first so-called "*Ländervergleich* 2009," that is, the follow-up study of PISA-E. PISA as well as the *Ländervergleich* test representative samples of 15-year-old students in math, science and reading literacy (in 2003 also in problem solving – in 2009, exclusively in reading and English). PISA-E used the same tests as the international PISA study. Apart from high schools, profound variation in the tracking and tracking systems among the remaining school types makes a comparison of student achievement across German states for these types of schools virtually impossible (Prenzel *et al.* 2008). Some of the remaining school types actually not even exist in each German state. This system heterogeneity is not an issue for high schools, on which we will focus in the following. The PISA-E test's sample size is several times the one of the international tests comprising two overlapping samples of 15-year-olds and ninth graders. Each sample covers about 40,000 students made of state samples ranging from 1,600 to 5,000 students for the 16 German federal states. Since German confidentiality requirements preclude the use of student-level data across states, one is restricted to use pooled state-level data.[1] German states mean performance is measured on a standardized scale: Just like for any PISA and/or PISA-E participating OECD country, state or province, scores for each subject and year are centered to an OECD mean of 500 and a standard deviation of 100. All our regressions include dummy variables for year and test subject, respectively.

Our data on instructional time are compiled from the respective state by-laws, taking

[1] Pooling is a common practice in the literature. An example for pooling subjects is Eren and Millimet (2007). Pooling German states and merging in data on aggregate countries is done in Wössmann (2010). Coates (2003) also considers pooling three test years.

into account amendments of ordinances, correcting for festivities and celebrations, such as state-wide holidays and any changes over time in these regulations for the period of observation. We followed the respective PISA cohorts from first to ninth grade. First to fourth grade concerns elementary school, fifth to ninth grade concerns high school (referred to as "*Gymnasium*" in German). For each of the 16 states, we construct an annual instructional time variable that is summed up to CIT. The data used in the construction of the variable come from several sources. The major part is drawn from administrative regulations, which can be found in official ministerial and/or school administration documents or in law and ordinance gazettes of the states. For some federal states, information on instructional time is given in special by-laws, so-called "*Stundentafelverordnungen*," as well as in regulations and mandates concerning training, examination and school rules such as the Bavarian "*Volksschulordnung* (VSO)" for elementary schools and the "*Gymnasialschulordnung* (GSO)" for high schools. In case of doubt and missing data, we obtained the information from the respective ministry of education and cultural affairs. Data on weighting schemes for a different intra-state distribution of teaching focus at the high school level (natural sciences, modern languages, math, etc.) is drawn from the respective statistical office's database.

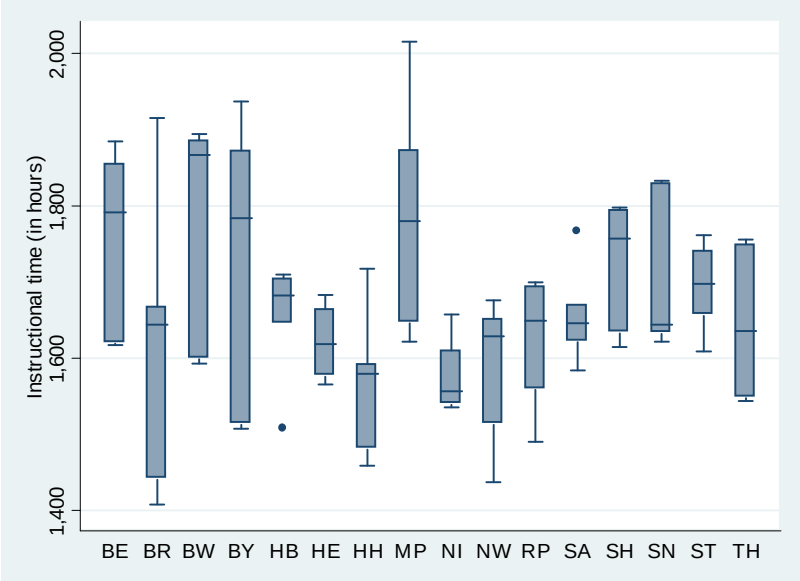

Figure 1. Math and reading CIT across states; PISA-E cohorts 2000/03/06/09

..

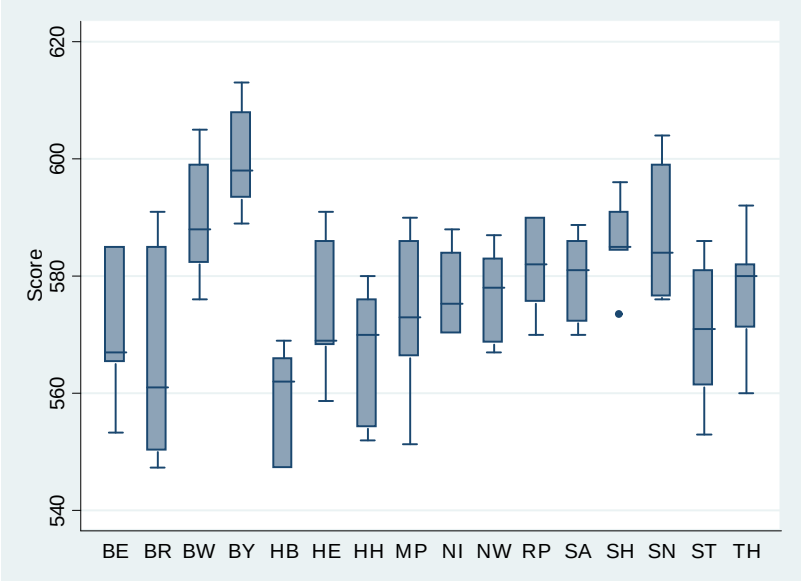

Figure 2. Math and reading scores across states; PISA-E cohorts 2000/03/06/09

As reliable data on CIT, constructed according to the strategy sketched above, cannot be obtained for all test subjects, we restrict our analysis to math and reading. We should also be clear about the point that we do not consider unscheduled shortfall, homework, individual tutoring or private study taken before or after school. In sum, we comprise total curricular hours of the four test cohorts accumulated from year of primary school enrollment (1991/92, 1994/95, 1997/1998, 2000/2001) to test year (2000, 2003, 2006, 2009).

Pooling PISA-E and *Ländervergleich* data for math and reading over states, tested subjects (math, reading), and cohorts allows us to rely on a sample of 112 observations.[2] As can be seen from Figure 1 and Figure 2,[3] German states substantially differ both in test scores and cumulative instruction by subject. For the distributions shown in Figure 1 and Figure 2, there are two central sources of variation: changes over the considered

[2] 16 states, two subjects (math and reading) for three cohorts plus reading for the test cohort of 2009 amounts to $16 \times 2 \times 3 + 16 = 112$. It is frequently claimed that studies relying on data at the level of states or districts suffer from an aggregation bias. Coates (2003) argues that the profession has not yet reached a consensus on whether such bias tends to produce spurious resource effects or not. According to Wössmann (2010) aggregation bias is not an issue in the case of marginal effects estimated using German state-level data.

[3] The following abbreviations are used: Berlin (BE), Brandenburg (BR), Baden-Wurttemberg (BW), Bavaria (BY), Bremen (HB), Hamburg (HH), Hesse (HE), Mecklenburg-West Pomerania (MP), Lower Saxony (NI), North Rhine-Westphalia (NW), Rhineland-Palatinate (RP), Saarland (SA), Schleswig-Holstein (SH), Saxony (SN), Saxony-Anhalt (ST), Thuringia (TH).

four waves and across the two subjects. Aggregated over states the distributions behind these two sources are shown in first and second schedule of Figure 3 and Figure 4, respectively. In order to check, whether it actually makes a difference to consider CIT, we also compared the snap-shot (i.e. only test year concerning) "instruction time" variable used in Wössmann (2010, Table 2, p. 241, Table A.1, p. 266) for PISA-E 2003 and subject math[4] with our corresponding cumulative measure. The correlation is statistically insignificant at all conventional levels.

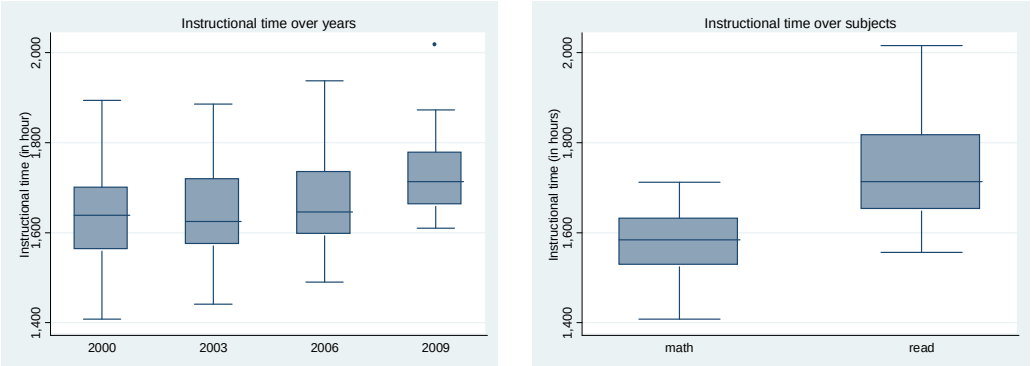

Figure 3. CIT state-means distributions

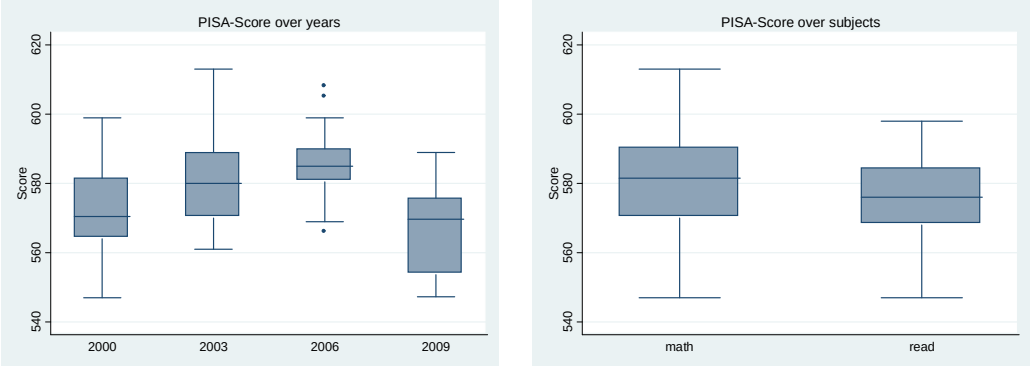

Figure 4. PISA-E test scores state-means distributions

2.2 Baseline estimates

To analyze the impact of CIT on students' test performance we rely on empirical models close to the ones common in the literature on education –or more specifically on cognitive achievement– production functions. See, among others, Hanushek (2002), Todd and

[4]The variable is constructed based on the PISA 2003 student questionnaire, in particular, on Q35b, Section F: Your mathematics classes. It reads as follows: "In the last full week you were in school, how many class periods did you spend in mathematics?" (OECD 2003, p. 24).

Wolpin (2003), Fuchs and Wössmann (2007), and Wössman (2003, 2010). Our baseline specifications are standard in the sense that we consider besides our central regressor (CIT) also sets of control variables that include measures of social environment and institutional features at the state level.

Table 1. Baseline estimates: PISA-E 2000, 2003, 2006

	Without state fixed effects			With state fixed effects		
CIT	0.287*	0.387**	0.450**	0.325*	0.385**	0.380**
	(0.096)	(0.034)	(0.016)	(0.082)	(0.043)	(0.029)
$CIT^2/1000$	-0.075	-0.112**	-0.126**	-0.094*	-0.112**	-0.111**
	(0.141)	(0.036)	(0.022)	(0.084)	(0.044)	(0.030)
Included set of controls						
a) Economic	x	x	x	x	x	x
b) Social		x	x		x	x
c) Educational			x			x
N obs	96	96	96	96	96	96
Adj. R-Squ. (percent)	63.63	74.87	69.87	72.19	72.77	78.61
F statistics	19.47	10.43	8.87	11.28	9.75	9.73

Note: Estimates include subject and year dummies; *, **, *** denotes significance at 10, 5, 1% level, respectively; p-values in parentheses; controls defined in text (and Appendix B).

Table 2. Baseline estimates: PISA-E 2000, 2003, 2006, *Ländervergleich* 2009

	Without state fixed effects			With state fixed effects		
CIT	0.419***	0.327**	0.338**	0.258	0.295*	0.295*
	(0.009)	(0.047)	(0.035)	(0.105)	(0.076)	(0.069)
$CIT^2/1000$	-0.113**	-0.088*	-0.093**	-0.094	-0.086*	-0.087*
	(0.016)	(0.067)	(0.046)	(0.102)	(0.075)	(0.064)
Included set of controls						
a) Economic	x	x	x	x	x	x
b) Social		x	x		x	x
c) Educational			x			x
N obs	112	112	112	112	112	112
Adj. R-Squ. (percent)	62.35	64.98	71.61	73.79	73.00	76.04
F statistics	19.38	13.11	10.65	13.50	11.01	9.81

Note: Estimates include subject and year dummies; *, **, *** denotes significance at 10, 5, 1% level, respectively; p-values in parentheses; controls defined in text (and Appendix B).

Concretely, in order to get a first assessment of the relationship, we estimate the following specifications

$$S_{it} = \alpha_i + \beta f\left(CIT_{it}\right) + \sum\nolimits_{g=1}^{3} \sum\nolimits_{j=1}^{k_g} \gamma_j X_{g,jit} + \varepsilon_{it}, \tag{1}$$

where S_{it} denotes test scores; index i and t refer to state and test period, respectively. We consider up to three sets of control variables, i.e., $X_{g=1}, ..., X_{g=3}$, consisting of $k_1 = 6$ economic and political economy variables, $k_2 = 8$ social environment and socio-demographic variables, and $k_3 = 9$ education policy and institutional variables, respectively. The maximum magnitude of conditioning variables amounts to $k_1+k_2+k_3 = 23$. Set X_1 (economic controls) comprises conservative party shares of governments (*Cons*), per capita (p.c.) public indebtedness (*Debt*), p.c. disposable income (*Disp*), population densities (*Dens*), unemployment rates (*Unemp*), and p.c. GDP (*GDP*) figures. Set X_2 (social controls) consists of data on last and first cohorts experiencing secondary school fees (*Fee*, *Fee2*), female employment rates (*Fem*), shares of foreign population (*For*), segregation measured by the share of 15-year-olds attending high school (*Seg*), shares of students with migration background (*Mig*), and dummies for East Germany (*East*) and city-state (*City*). Finally, set X_3 (education controls) considers secondary school years to final grade, i.e., either 8 or 9 years track (*G9*), average class sizes, student-teacher-ratios, instructional hours per teacher, and shares of part-time teachers in elementary school (*CS1*, *ST1*, *HT1*, *PT1*) and in secondary I (*CS2*, *ST2*, *HT2*, *PT2*), respectively. For further detail and sources of variables see Appendix B. A brief summary on how these variables might affect student test scores is given in Appendix B.

Controlling for state fixed effects α_i addresses the qualification of unobserved heterogeneity across states. In particular, this concerns such unobservables as pedagogical quality, performance, and effectiveness of teachers across states (Hanushek 2006b) as well as differences in the quality of educating teachers. It also implies the quality of text books, instructional methods and materials, and the administration and organization of curricula. For all these dimensions each German state has its own choice and responsibility. As can be seen from the respective first three (common constant model) and last three columns of estimates shown in Table 1 and Table 2, a significant positive effect from CIT on test scores is robust to the inclusion of state fixed effects. In fact, our estimates controlling for fixed effects do not markedly differ from the ones obtained from regressions without considering state effects. In specification (1), we follow the most recent cross-country study by Lavy (2010) in allowing for concavity in the functional re-

lationship $f(\cdot)$ between student performance and CIT. As can be seen from the estimates reported in Table 1 and Table 2, this more flexible specification accords with the data, although less than ten percent of cases lie to the right of the implied upper turning point (Figure 5). For the remaining vast majority of data points in the scatter diagram, the relationship between CIT and test scores is close to linear. As Figure 5 is based on the estimates reported in the last three columns of Table 2, including state fixed effects, it does not show ordinate values. Hence, we interpret it only qualitatively as lending support to a weakly concave, nearly linear relationship. In the estimates reported in Table 1, we abstracted from using data from the *Ländervergleich* 2009, which in contrast to the preceding PISA-E tests did not test math skills of students. Again, the results are qualitatively not sensitive to the inclusion of the 2009 (reading) scores (Table 2). We leave all further quantitative interpretation of estimates for section 3, reporting results from our sensitivity analysis of the bearing CIT has on student test performance. In the following, we outline how we achieve robustness by addressing model uncertainty in an extreme bounds analysis framework.

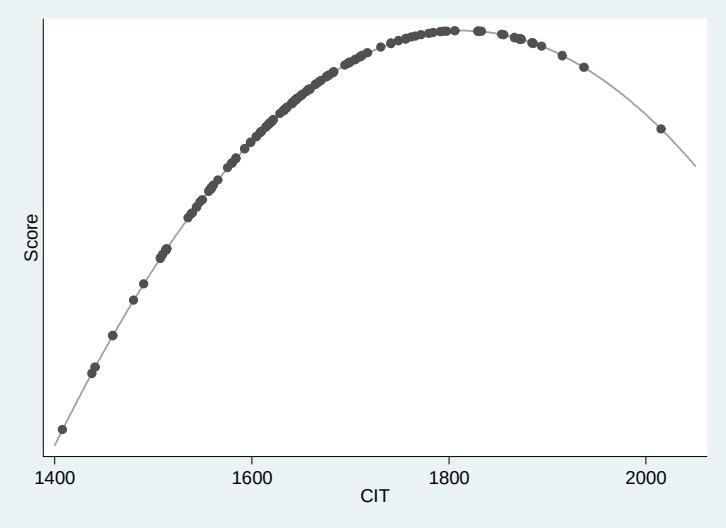

Figure 5. Relationship between CIT and test scores

2.3 Addressing model uncertainty: methodology

In order to address model uncertainty, we subject our empirical model to an extreme bounds analysis (EBA) as originally suggested by Leamer (1983, 1985) and Levine and Renelt (1992) and extended and modified by Granger and Uhlig (1990) and Sala-i-Martin (1997). The use of EBA techniques is fairly popular in the empirics of economic growth

literature. However, its use is not limited to growth regressions. For recent applications in other contexts see Sturm *et al.* (2005) and Mossa (2009). Yet, we are not aware of an EBA application in the area of education production function estimates. In general, EBA does not include the use of state (or country) fixed effects to take account of unobserved heterogeneity. In fact, in the present context the use of state fixed effects implies that different average student PISA test scores between states are not explained but represented by dummy variables. Thus, all that can be explained by these regressions are reactions of test scores over time (see, in a similar context, Kirchgässner 2011, p. 17), which show a comparatively lower variation than the distribution of scores across states. See the distributions shown in Figure 2 as opposed to the ones shown in the first diagram of Figure 4. Following this argumentation and adhering to the EBA practice in the literature, we abstract from the inclusion of state dummies as well as from nonlinear specifications. Both modifications have shown to be not critical for an assessment of the relationship between CIT and student test scores (section 2.2).

Hence, our general EBA specification reads

$$S_{it} = \alpha + \beta CIT_{it} + \sum\nolimits_{j=1}^{n} \gamma_j V_{jit} + \sum\nolimits_{k=1}^{m} \delta_k Z_{kit} + \epsilon_{it}, \tag{2}$$

where V_j represents a set of important variables included in every regression. It contains a dummy for subject math as well as dummy variables identifying the respective year. Z_k is a set of three up to eleven out of 23 possible conditioning variables (section 2.2), where the minimum number of such conditioners ($= 3$) follows the suggestion in Levine and Renelt (1992). To identify differences between the impact of CIT on average PISA-E test scores and of CIT on scores of top and bottom percentile students of each cohort, we also consider $TopX\%$ and $BotX\%$ as dependent variables. To check for robustness, the strategy is to consider all possible $M = n!/(k!(n-k)!)$ regression models that can be estimated by taking combinations of k out of the 23 Z-variables, that is, $1,771$ models for $k = 3$ up to $1,352,078$ models for $k = 11$. For this elaborated sensitivity analysis we also address for every single regression in the procedure possible problems of multicollinearity by dropping models with a variance inflation factor (VIF) for the exogenous at stake exceeding a value of four.[5]

As proposed by Levine and Renelt (1992), so-called "extreme bounds" of estimates

[5] A VIF_j for some exogenous variable x_j is defined as $VIF_j = 1/(1 - R_j^2)$, where goodness-of-fit measure R_j^2 refers to a regression of x_j on all other independent variables in the respective model. For zero collinearity VIF_j takes on a value of one.

can be used to check whether a variable like CIT_{it} in eq. (2) is fragile or robust. They are made of lower and upper bound. The former is defined as the lowest estimated value for β_M minus two standard deviations, the latter as the highest estimated value for β_M plus two standard deviations. If lower and upper extreme bound for estimated β coefficients show the same sign, the explanatory variable at stake is said to be robustly related to the dependent variable.

A critical aspect of EBA-techniques in their original version proposed by Leamer (1983) and Levine and Renelt (1992) is that extreme bounds may be resultant from models that are unreasonable in terms of a corresponding relatively low R^2 statistics. A modified EBA procedure addressing this problem is suggested by Granger and Uhlig (1990). Their idea is to consider only those β_M estimates stemming from models that reach R^2 statistics corresponding to a certain percentage of the $R^2_{\max}$ of all M estimated models, taking into account the goodness-of-fit $R^2_{\min}$ of the basic model (leaving out the control for conditioning variables, i.e., $\sum_{k=1}^{m} \delta_k Z_{ki}$). This approach is referred to as "reasonable extreme bounds analysis" (REBA) in the literature. For model specifications with R^2-values equal to or greater than

$$R^2_\delta = (1-\delta)R^2_{\max} + \delta R^2_{\min}, \tag{3}$$

where $0 < \delta < 1$ and for small δ-values, we consider corresponding specifications as being "reasonable" specifications as they are not too far off from the "best" model –of the M considered ones– in terms of goodness-of-fit as measured by the adjusted R^2.

Sala-i-Martin (1997) argues that a single regression for which the sign of the coefficient β changes or becomes insignificant suffices according to original EBA or REBA standards that a variable is identified to be non-robust. He assesses this procedure as a too hard to pass test for almost any variable at stake: "if the distribution of the parameter of interest has some positive and some negative support, then one is bound to find one regression for which the estimated coefficient changes signs if enough regressions are run" (Sala-i-Martin 1997, p. 179). This insight led Sala-i-Martin to introduce a newly modified approach by moving away from the extreme test and instead assigning some level of confidence by looking at the entire distribution of the estimators of β_M. For each of the M estimated models the likelihood L_M, the point estimates β_M, and the standard deviation σ_M are calculated. They are used to construct the mean estimate of β and the average variance σ^2 as a weighted average of M point estimates and estimated variances, respectively:

$$\beta = \sum_{l=1}^{M} \omega_M \beta_M; \quad \sigma^2 = \sum_{l=1}^{M} \omega_M \sigma^2_M, \tag{4}$$

where weights ω_M are proportional to the likelihoods of the M models according to

$$\omega_M = \frac{L_M}{\sum_{l=1}^{M} L_M}. \tag{5}$$

Once the mean and the variance of the distribution of β, assumed to be normal,[6] are known, the cumulative distribution function (CDF) can be calculated using the standard normal distribution. The level of confidence for the variable of interest is defined as the larger of the two areas under the probability density function (PDF) left and right from zero.[7] In order to be as comprehensive as possible, we apply all three methods, that is, standard EBA, REBA, and EBA in the modified version of Sala-i-Martin (1997), henceforth SiM-EBA. Primarily this is done to check the robustness of the association between CIT and PISA scores, letting k vary between 3 and 11. Going beyond this primary sensitivity analysis, we also scrutinize the impact of the other 23 potential explanatories (see Appendix B for detail) on our measure of cognitive achievement S_{it} relying on the considered portfolio of EBA-techniques.

3 Results

3.1 Cumulative instructional time

Results for all three EBA methods outlined above are reported for three different (maximum) numbers of variables sampled into the conditioning set, i.e., for $k = 3, k = 5$, and $k = 11$, in Table A.3, Table A.4, and Table A.5 of Appendix A, respectively. In the interpretation of these findings, we will follow Sala-i-Martin and focus on the entire distribution (SiM-EBA) and only discuss results from the other two procedures if they deviate from the SiM-EBA based finding. For all used dependent variables, CIT shows a positive significant impact on scores that is robust if we consider different subperiods, even if we apply CDF(0) > 0.95 as more strict criterion of robustness. A first point to note is that variation in k does not qualitatively alter our results as can be seen from Table A.3 to A.5 in Appendix A. Figures 6 to 8 make the point by showing the respective

[6] The normality assumption is justified on the grounds of the central limit theorem as can be seen from Figures 5 to 7.

[7] We follow Sala-i-Martin (1997) by referring to the larger of the two areas as "CDF(0)" irrespective of whether the area lies actually above or below zero.

distribution of estimated β coefficients for different k (the black line drawn through the respective diagram shows the kernel density, while the grey line depicts the normal PDF as a reference case). The distributions virtually have the same mean, while the variance, of course, decreases with k and the number of estimated models M. CDF(0) remains above a value of 0.95 going from $k = 11$ to $k = 3$, that is, narrowing the number of variables contained in the conditioning set. This suggests for the sake of efficiency, that is, for the sake of estimating rather 33,649 ($k = 5$) than 1.35 million ($k = 11$) models for different test waves or subsets of pooled waves (Table 3), to concentrate the further analysis on $k = 5$. Table 3 reports these results for all students' scores as well as for the top-5% and top-10% and the bottom-5% and bottom-10% of students in terms of test scores. Since for *Ländervergleich* 2009 no score data by percentile is available our analysis for bottom-/top-end students is restricted to the PISA-E waves 2000, 2003, and 2006. For the overall test scores as dependent, we consider besides the total pool also a corresponding data set restricted to the year 2000 only and one that leaves out the *Ländervergleich* 2009, when math has not been tested. The year 2K test sub-sample captures the effect of the first year in which the test was conducted. In this sense, it can be seen as relatively free from effects induced by policies that the states started in the aftermath of the first test. This is due to the fact that results from the PISA-E 2000 tests were widely published and extensively discussed in the media and in political debates (Tillmann *et al.* 2008, Pütz 2008). As can be seen from Table 3, we also considered total and sub-set sample separately for math and reading sub-samples. The fourth column of Table 3 displays the unweighted mean of β_M for $M = 33,649$. Multiplying these figures with an average of 360 ($= 40 \times 9$) school weeks over the nine years from first grade to test date, we can calculate the approximate effect of a policy corresponding to one additional hour of instructional time per week over the total learning period. It is shown in the fifth column of Table 3. Finally, the last column in Table 3 reports CDF(0) values from applying the SiM-EBA method. As can be seen from the third line of results displayed in Table 3, the above described stylized policy effect of CIT on scores amounts to sizable 11.59 test-score points or roughly 12 percent of an international standard deviation. Dropping the *Ländervergleich* 2009 data the effect increases to more than 13 percent. The largest average impact of a one hour per week increase policy is calculated, when one restricts the sample to the first year when German states ran the OECD PISA test for the first time, i.e. for PISA-E 2000. It amounts to nearly 17 percent of an international standard deviation. For the sub-samples separating subjects, we find that the

effect is particularly pronounced for math (> 16 percent) but still sizable, that is, above 12 percent of an international standard deviation, for reading. Looking at upper and lower percentiles of test scores, we find that all students would benefit from an increase in CIT. The CIT–score relationship is, however, more pronounced for the bottom-end students in terms of test scores, suggesting that those students would benefit the most.

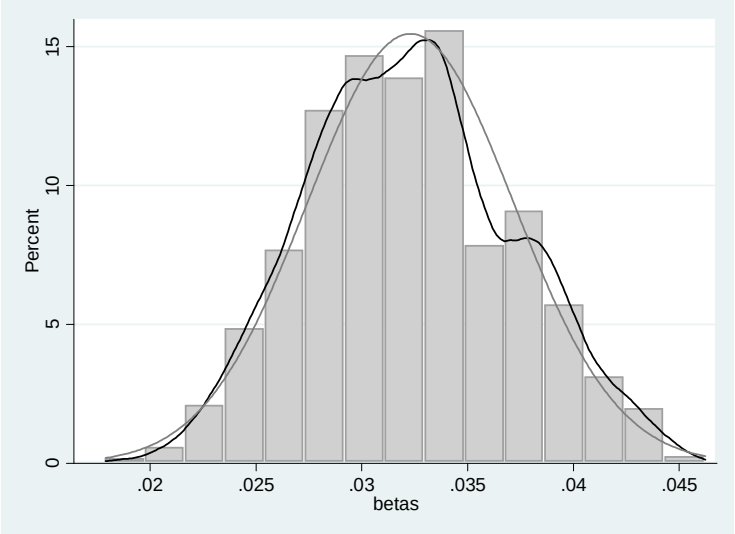

Figure 6. Distribution of estimated betas for
SiM EBA: $k = 3$, pool: 00/03/06/09, math/reading
$M = 1,771$ models, $N = 112$ observations

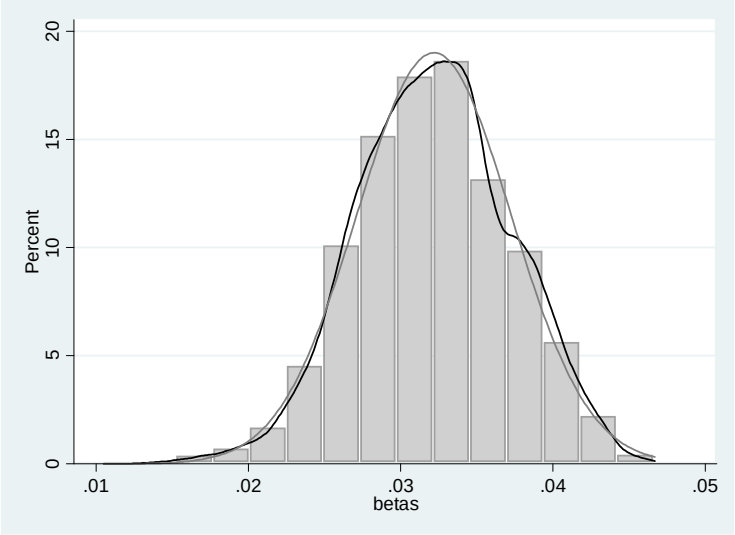

Figure 7. Distribution of estimated betas for
SiM EBA: $k = 5$, pool: 00/03/06/09, math/reading
$M = 33,649$ models, $N = 112$ observations

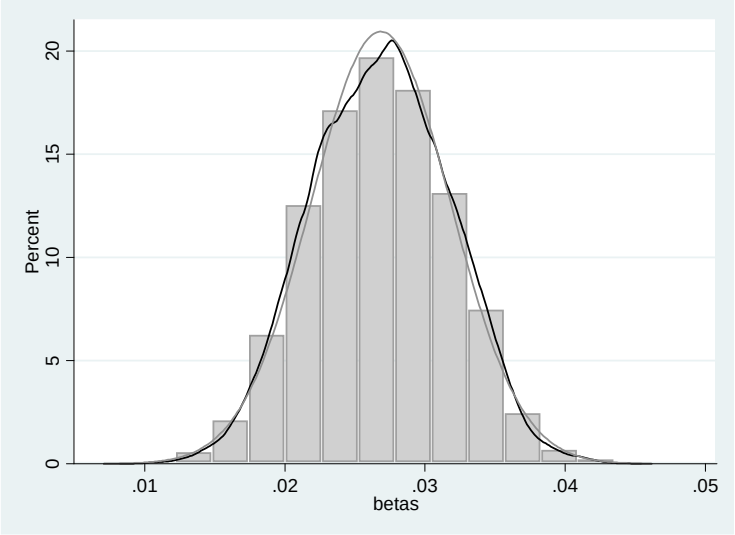

Figure 8. Distribution of estimated betas for
SiM EBA: $k = 11$, pool: 00/03/06/09, math/reading
$M = 1,352,078$ models, $N = 112$ observations

To get an impression of what 16 percent of an international standard deviation is actually to mean, consider the following experiment of thought: Under the assumption that the policy of increasing the instructional time by one additional hour per week over the total learning period has a similar impact on scores for other secondary school types as it has for high schools, we can calculate the consequences in the rankings, interpreting German states "as a microcosm for OECD countries" (Wössmann 2010). Take OECD PISA 2006, in which German students ranked 14th in math compared to the other OECD test participating countries; see Table A.7 in Appendix A. An increase by 16 percent of the standard deviation (normed to 100) of this test would correspond to running up six ranks up to rank 8. In contrast, cutting CIT down by one hour per week would correspond to a drop in the ranking down to rank 24. Similar effects can be calculated for reading.

All results reported in Table 3 are robust in the sense of Sala-i-Martin (1997).[8] CDF(0) values range between 0.9736 to 0.9999. As can be seen from detailed Tables A.3, A.4, and A.5 in Appendix A, virtually all values for lower and upper bounds show positive signs, confirming the highly robust positive effect of CIT on PISA test-scores.

[8]Re-running our estimates relying on a snap-shot measure like the ones discussed in the introduction and used, for example, by Wössmann (2010), throughout generates results that are not robust ("fragile") in the sense of any conventional EBA-criterion (EBA, REBA, SiM-EBA). Results are available on request from the authors.

Table 3. Impact of CIT on PISA-E tets-scores: SiM-EBA ($k = 5$, $Z = 23$)

Dependent	Test year	Test subject	Beta (mean)	Policy effect	CDF(0)
Score	2000	math/reading	0.0468	16.87	0.9736
Score	00/03/06	math/reading	0.0364	13.12	0.9992
Score	00/03/06/09	math/reading	0.0321	11.59	0.9976
Score	00/03/06/09	reading	0.0340	12.27	0.9987
Score	00/03/06	reading	0.0412	14.84	0.9996
Score	00/03/06	math	0.0456	16.44	0.9999
Bot5%	00/03/06	math/reading	0.0444	16.00	0.9960
Bot10%	00/03/06	math/reading	0.0421	15.17	0.9980
Top10%	00/03/06	math/reading	0.0323	11.63	0.9961
Top5%	00/03/06	math/reading	0.0266	9.57	0.9790

Note: Policy effect is one additional hour of CIT per week over total learning period.

3.2 Other robust determinants of cognitive achievement

In order to analyze which of the remaining 23 available explanatories (see Appendix B for detail) have a robust impact on test scores, we rely on the same portfolio of EBA-techniques as for CIT in the preceding paragraphs. Table 4 below summarizes the results for this exercise reported in detail in Table A.6, for which we set $k = 3$ ($M = 1,771$)[9] and consider the 00/03/06/09–math/reading pool with $N = 112$ observations for each of the M estimated models. As central criterion of robustness we again apply the modified SiM-EBA criterion, i.e. CDF(0) > 0.95.[10]

At the federal state level, public indebtedness (*Debt*), disposable income (*Disp*), population density (*Dens*), and the unemployment rate (*Unemp*), each measured for the year corresponding to the respective PISA-E test-year, as well as a dummy for East German states (*East*) are found to be robustly and negatively associated with test scores according to the modified SiM-EBA criterion. Drawn from the respective PISA cohort data, also the share of 15 years old students with migrational background (*Mig*) is identified as

[9] As for CIT, results are qualitatively unaffected by setting $k = 5$ ($M = 33,649$).

[10] Note, applying the less strict CDF(0) > 0.90 instead, as originally proposed in Sala-i-Martin (1997), also identifies variables average class size in elementary schools (*CS1*), average instructional hours per teacher in elementary schools (*HT1*) and segregation (*SEG*) as robust determinants of student achievement.

robust negative correlate with scores. By this standard robust education policy variables are the number of secondary school years to exit exam (*G9*) and average class size in secondary I (*CS2*).

Table 4. Robust and weakly robust determinants of PISA-E test-scores besides CIT
SiM-EBA ($k = 3$, $Z = 23$)

Category	Variable	Beta (mean)	CDF(0)
Economic/Political economy	*Debt*	-2.236	1.000
	Disp	-0.288††	0.955
	Dens	-1.773††	0.977
	Unemp	-1.634†	0.999
Social environment	*Mig*	-0.501†	0.988
	East	-8.931†	0.999
Education policy	*G9*	-5.770†	0.984
	CS2	-0.335†	0.979

Note: †SiM-EBA, REBA: robust, standard EBA: fragile

††SiM-EBA: robust; standard EBA, REBA: fragile

The only variable that measures up to CIT with regard to meeting all robustness criteria of the considered portfolio of EBA-techniques is public indebtedness per inhabitant. It proxies the cost effectiveness of incumbent and former governments of the respective state. In terms of size, the estimated average coefficient (< 0) of the dummy for an East German state (*East*) stands out. This is, in particular, due to the poor performance of students from the East German state of Brandenburg (BR) as well as to the below national average achievement in terms of math and reading test scores of the two East German states Mecklenburg-West Pomerania (MP) and Saxony-Anhalt (ST); see Figure 2. It is also these states of the five East German ones that are known for their notoriously unsound economic and demographic status characterized by a substantial number of movers to the Western states in the decades following German unification. The average negative impact of the institutional grade configuration variable *G9* on test scores also is relatively sizable.[11] The effect is negative and amounts to about six percent of an international standard deviation. It is straightforward to attribute this effect to differences in

[11] Variable *G9* is a dummy that takes on a value of one if the number of secondary school years to final grade, that is, to *Abitur*, the German A-level equivalent, is nine as opposed to eight years. It is up to each state's discretionary education policy to set this length. East German states traditionally practice an eight years system.

the density of curricula: Students might be comparatively more advanced in math and reading skills in a system, where the final exit exam takes place three rather than four years after the PISA test date. State-level population densities, unemployment rates, and shares of students with migration background show the expected sign (see Appendix B), though being less strongly associated with lower test scores in terms of size of estimated coefficients. Another education policy variable that is robust according to the modified SiM-EBA criterion is average class size in secondary I (*CS2*). A decrease of *CS2* by one student over the nine years from enrollment (first grade) to test year (ninth grade) is robustly associated, however, with only a minor increase of 0.335 points or 0.335 percent of an international standard deviation. In sum, the effect of (education) policy variables is either most probably resultant from PISA tests being not adjusted to differences in curricula or is quite small in size compared to the policy effect of increasing CIT. The only counter-intuitive and weakest, in terms of size, effect is the negative average coefficient for disposable income.

A final caveat concerns the above interpretation of results reported in Table 4: Apart from public indebtedness per inhabitant and CIT all other 22 considered determinants of student achievement are either not robust or are fragile, at least, according to one criterion in the used portfolio of EBA-procedures. Hence, they have both some positive and some negative support (Table A.6).

4 Conclusion

Econometric misspecification and failure to account for major determinants of student achievement represent the central problem in correctly identifying the school inputs–student achievement relationship (Todd and Wolpin 2003). By relying on a portfolio of extreme bounds analysis techniques as well as a newly compiled cross-state dataset on cumulative instructional time from first grade to ninth (i.e., OECD PISA-test date) grade for German states, we addressed two fundamental shortcomings in the literature: A serious error-in-variables problem due to using poor proxies of instructional time and the widely ignored issue of model uncertainty. We find that instructional time by subject, measured in cumulative terms, is a highly robust determinant of student cognitive achievement. This finding is insensitive to the inclusion of state fixed effects and to sub-set choices of tests. It is robust according to all conventional EBA-standards.

References

[1] Afonso, A. and M. St. Aubyn, 2006. Cross-country efficiency of secondary education provision: A semi-parametric analysis with non-discretionary inputs, *Economic Modelling* 23, 476-491.

[2] Baker, D.P., Fabrega, R., Galindo, C., and J. Mishook, 2004. Instructional time and national achievement: Cross-national evidence, *Prospects: Quarterly Review of Comparative Education* 34, 311-334.

[3] Baumert, J. (ed.), 2002. *PISA 2000 - die Länder der Bundesrepublik Deutschland im Vergleich*, Wiesbaden: Leske + Budrich.

[4] Büttner, T., Schwager, R., and M. Stegarescu, 2004. Agglomeration, population size and the cost of providing public services: An empirical analysis of German states, *Public Finance and Management* 4, 496-520.

[5] Card, D. and A.B. Krueger, 1992. Does school quality matter? Returns to education and the characteristics of public schools in the United States, *Journal of Political Economy* 100, 1-40.

[6] Card, D. and A.B. Krueger, 1996. Labor market effects of school quality: Theory and evidence, in G. Burtless (ed.), *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, Washington, DC: Brookings Institution.

[7] Coates, D., 2003. Education production function using instructional time as an input, *Education Economics* 11, 273-292.

[8] Dewey, J., Husted, T., and L. Kenny, 2000. The ineffectiveness of school inputs: A product of misspecification?, *Economics of Education Review* 19, 27-5.

[9] Doepke, M. and F. Zilibotti, 2008. Occupational choice and the spirit of capitalism, *Quarterly Journal of Economics* 123, 747-793.

[10] Eren, O. and D.L. Millimet, 2007. Time to learn? The organizational structure of schools and student achievement, *Empirical Economics* 32, 301-332.

[11] Freebody, P., Maton, K., and J. Martin, 2008. Talk, text, and knowledge in cumulative, integrated learning: A response to 'intellectual challenge', *Australian Journal of Language and Literacy* 31, 188-201.

[12] Fuchs, T. and L. Wössmann, 2007. What accounts for international differences in student performance? A re-examination using PISA data, *Empirical Economics* 32, 433-464.

[13] Granger, C. and H. Uhlig, 1990. Reasonable extreme-bounds analysis, *Journal of Econometrics* 44, 159-170.

[14] Hanushek, E.A., 2002. Publicly provided education, in Auerbach, A.J. and M. Feldstein (eds.), *Handbook of Public Economics*, Vol. 4, Amsterdam: Elsevier, 2045-2141.

[15] Hanushek, E.A., 2003. The failure of input-based schooling policies, *Economic Journal* 113, F64-F98.

[16] Hanushek, E.A., 2004. What if there are no 'best practices'?, *Scottish Journal of Political Economy* 51, 156-172.

[17] Hanushek, E.A., 2006a. School resources, in Hanushek, E.A. and F. Welch (eds.), *Handbook of the Economics of Education*, Vol. 2, Amsterdam: Elsevier, 865-908.

[18] Hanushek, E.A., 2006b. Teacher quality, in Hanushek, E.A. and F. Welch (eds.), *Handbook of the Economics of Education*, Vol. 2, Amsterdam: Elsevier, 1051-1076.

[19] Hoxby, C.M., 2000. Does competition among public schools benefit students and taxpayers?, *American Economic Review* 90, 1209-1238.

[20] Kirchgässner, G., 2011, Econometric estimates of deterrence of the death penalty: Facts or ideology?, CESifo Working Paper, No. 3443.

[21] Köller, O., Knigge, M., and B. Tesch (eds.), 2010. *Sprachliche Kompetenzen im Ländervergleich*, Münster: Waxmann

[22] Krueger, A.B., 2003. Economic considerations and class size, *Economic Journal* 113, F34-F63.

[23] Lavy, V., 2010. Do differences in school's instruction time explain international achievement gaps in math, science, and reading? Evidence from developed and developing countries, NBER Working Paper, No. 16227.

[24] Leamer, E.E., 1983. Let's take the con out of econometrics, *American Economic Review* 73, 31-43.

[25] Leamer, E.E., 1985. Sensitivity analysis would help, *American Economic Review* 75, 308-313.

[26] Lee, J.-W. and R. Barro, 2001. Schooling quality in a cross-section of countries, *Economica* 68, 465-488.

[27] Levine, R. and D. Renelt, 1992. A sensitivity analysis of cross-country growth regressions, *American Economic Review* 82, 942-963.

[28] Marcotte, D.E., 2007. Schooling and test scores: A mother-natural experiment, *Economics of Education Review* 26, 629-640.

[29] Marcotte, D.E. and S. Hemelt, 2008. Unscheduled closings and student performance, *Education Finance and Policy* 3, 316-338.

[30] Maton, K., 2009. Cumulative and segmented learning: Exploring the role of curriculum structures in knowledge-building, *British Journal of Sociology of Education* 30, 43-57.

[31] Moser, U. and D. Angelone, 2009. Unterrichtszeit, Unterrichtsorganisation, Leistung und Interesse, in Bundesamt für Statistik (ed.), *PISA 2006: Analysen zum Kompetenzbereich Naturwissenschaften*, Neuchâtel: BFS, 9-40.

[32] Moosa I.A., 2009. The determinants of foreign direct investment in MENA countries: an extreme bounds analysis, *Applied Economics Letters* 16, 1559-1563.

[33] OECD, 2003. *PISA 2003. Student Questionnaire*, Paris: Organisation for Economic Co-operation and Development (OECD)

[34] OECD, 2009. *Take the Test – Sample Questions from OECD's PISA Assessments*, Paris: Organisation for Economic Co-operation and Development (OECD)

[35] Pischke, J.-S., 2007. The impact of length of the school year on student performance and earnings: Evidence from the German short school years, *Economic Journal* 117, 1216-1242.

[36] Prenzel, M., Baumert, J., and W. Blum (eds.), 2005. *PISA 2003 der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche?*, Münster: Waxmann.

[37] Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., and E. Klieme (eds.), 2008. *PISA 2006 in Deutschland die Kompetenzen der Jugendlichen im dritten Ländervergleich*, Münster: Waxmann.

[38] Pütz, M, 2008. *PISA und die Reaktionen der Bildungspolitik*, Munich: Grin.

[39] Riphahn, R., 2011. The effect of secondary school fees on educational attainment, forthcoming in *Scandinavian Journal of Economics*.

[40] Sala-i-Martin, X.X., 1997. I just ran two million regressions, *American Economic Review* 87, 178-183.

[41] Schulte, B., 2004. Teaching subjects and time allocation in the German school system (Berlin), *Prospects: Quarterly Review of Comparative Education* 34, 335-351.

[42] Sturm, J-E., Berger, H., and J. de Haan, 2005.Which variables explain decision on IMF credits? An extreme bounds analysis, *Economics & Politics* 17, 177-213.

[43] Tillmann, K.-J., Dedering, K., Kneuper, D., Kuhlmann, C., and I. Nessel, 2008. *PISA als bildungspolitisches Ereignis: Fallstudien in vier Bundesländern*, Wiesbaden: VS Verlag.

[44] Todd, P.E. and K.I. Wolpin, 2003. On the specification and estimation of the production function for cognitive achievement, *Economic Journal* 113, F3-F33.

[45] Wössmann, L., 2003. Schooling resources, educational institutions and student performance: The international evidence, *Oxford Bulletin of Economics and Statistics* 65, 117-170.

[46] Wössmann, L., 2010. Institutional determinants of school efficiency and equity: German states as a microcosm for OECD countries, *Journal of Economics and Statistics* (*Jahrbücher für Nationalökonomie und Statistik*) 230, 234-270.

[47] Yew, E.H.J., Chng, E., and H.G. Schmidt, 2011. Is learning in problem-based learning cumulative?, forthcoming in *Advances in Health Sciences Education Theory and Practice*.

Appendix A

Table A.1. Summary statistics of PISA test scores S_{it} and cumulated instructional time CIT_{it}

Sample	Math/ reading		Math only		Reading only	
Variable	S_{it}	CIT_{it}	S_{it}	CIT_{it}	S_{it}	CIT_{it}
Mean	577.2	1665.8	580..9	1575.1	574.5	1733.7
Max	613	2015.4	613	1712	598	2015.4
Min	547	1407.8	547	1407.8	547	1556.4
Range	66	607.6	66	304.2	51	459.0
Std. dev.	13.8	120.4	14.9	69.5	12.4	104.9
Median	578	1645.1	581.5	1584.0	576	1713.5
N	112	112	48	48	64	64

Table A.2. Basic and full model: Variants by dependent, test cohorts (Year), and tested subjects (Subjects)

	Dependent	Year(s)	Subjects		Basic				Full			
No.				*N*	Beta *CIT*	p-val (%)	Adj. R^2 (%)	M.E. +1h/week	Beta *CIT*	p-val (%)	Adj. R^2 (%)	M.E. +1h/week
1	*Score*	00-09	math/reading	112	0.02893	2.3	25.99	10.41	0.01836	5.9	70.54	6.61
2	*Score*	00	math/reading	32	0.05182	4.6	7.44	18.66	insufficient degrees of freedom (DF)			
3	*Score*	00-06	math/reading	96	0.03323	1.6	24.43	11.96	0.02360	4.2	67.89	8.50
4	*Score*	00-09	reading	64	0.02937	3.4	20.36	10.57	0.02000	6.5	73.95	7.20
5	*Score*	00-06	reading	48	0.03580	1.4	21.75	12.89	0.04030	1.9	70.32	14.51
6	*Score*	00-06	math	48	0.02606	34.6	24.77	9.38	0.08620	0.5	74.31	31.03
7	*Bot5%*	00-06	math/reading	96	0.04349	3.2	2.18	15.66	0.03038	6.0	62.59	10.94
8	*Bot10%c*	00-06	math/reading	96	0.04151	2.1	6.66	14.94	0.02736	4.6	67.66	9.85
9	*Top10%*	00-06	math/reading	96	0.02692	4.0	40.84	9.69	0.02265	10.0	60.59	8.15
10	*Top5%*	00-06	math/reading	96	0.02137	11.5	42.76	7.69	0.01931	19.4	58.59	6.95

Note: All regressions include dummies for respective test subject (math, reading) and test year (2000, 2003, 2006, 2009).
Basic model includes *CIT* as sole regressor, full model considers all 23 variables listed below.
M.E. +1h/week – effect of one additional hour of instruction per week

Table A.3. EBA with *CIT* as variable of interest, $k = 3$ ($M = 1{,}771$)

No.	Percentile					Leamer EBA			Granger EBA ($\delta = 0.1$)		Sala-i-Martin EBA			
For detail: see Table A.2	5	10	50	90	95	Lower Bound	Upper Bound	% significant at 5%	Lower Bound	Upper Bound	Unweighted mean	Weighted Mean	Weighted Std.error	CDF(0)
1	0.02460	0.02612	0.03224	0.03891	0.04070	0.01789	0.04623	99.0	0.01789	0.03569	0.03230	0.02957	0.0930	0.9993
2	0.03465	0.03814	0.05387	0.06471	0.06691	0.01615	0.07713	77.2	0.03070	0.04662	0.05238	0.03670	0.01511	0.9924
3	0.02726	0.02993	0.03639	0.04486	0.04654	0.01827	0.05214	98.1	0.02073	0.04144	0.03686	0.03423	0.00979	0.9998
4	0.02433	0.02667	0.03407	0.04261	0.04477	0.01538	0.05017	95.4	0.02458	0.04268	0.03423	0.03249	0.00992	0.9995
5	0.02907	0.03284	0.04274	0.05142	0.05326	0.01668	0.06574	97.7	0.03238	0.04776	0.04195	0.03881	0.01010	0.9999
6	0.01553	0.01831	0.03208	0.05809	0.06796	0.00554	0.09675	22.2	0.04650	0.08640	0.03583	0.07006	0.01834	0.9999
7	0.03100	0.03325	0.04425	0.05740	0.06072	0.02090	0.06754	82.8	0.02942	0.05518	0.04488	0.03949	0.01375	0.9980
8	0.03022	0.03230	0.04223	0.05426	0.05685	0.02068	0.06287	89.3	0.02750	0.04649	0.04283	0.03553	0.01159	0.9989
9	0.02425	0.02623	0.03267	0.03815	0.03944	0.01489	0.04492	92.7	0.03037	0.03501	0.03236	0.03209	0.01121	0.9979
10	0.01785	0.02000	0.02688	0.03228	0.03373	0.00862	0.03948	49.2	0.02602	0.02792	0.02648	0.02623	0.01221	0.9841

Table A.4. EBA with *CIT* as variable of interest, $k = 5$ ($M = 33{,}649$)

No.	Percentile					Leamer EBA			Granger EBA ($\delta = 0.1$)		Sala-i-Martin EBA			
For detail: see Table A.2	5	10	50	90	95	Lower Bound	Upper Bound	% significant at 5%	Lower Bound	Upper Bound	Unweighted mean	Weighted Mean	Weighted Std.error	CDF(0)
1	0.02407	0.02588	0.03219	0.03896	0.04057	0.01052	0.04670	97.8	0.01582	0.03658	0.03219	0.02594	0.00918	0.9976
2	0.02774	0.03185	0.04661	0.06191	0.06515	-0.00130	0.09220	71.7	0.01494	0.06122	0.04687	0.02705	0.01396	0.9736
3	0.02636	0.02873	0.03622	0.04481	0.04638	0.01151	0.05283	96.9	0.01892	0.04236	0.03645	0.03127	0.00987	0.9992
4	0.02506	0.02678	0.03374	0.04188	0.04423	0.01158	0.05197	718	0.01520	0.04095	0.03408	0.02806	0.00933	0.9987
5	0.02958	0.03216	0.04150	0.05018	0.05229	0.01426	0.06736	98.2	0.02539	0.04679	0.04123	0.03707	0.01022	0.9996
6	0.01687	0.02119	0.04486	0.07236	0.07902	-0.00230	0.11702	43.8	0.03977	0.11123	0.04566	0.07382	0.01855	0.9999
7	0.02986	0.03312	0.04425	0.05661	0.05980	0.01316	0.06856	87.9	0.02232	0.05372	0.04444	0.03655	0.01377	0.9960
8	0.02874	0.03164	0.04174	0.05354	0.05624	0.01254	0.06377	91.8	0.02152	0.05011	0.04215	0.03335	0.01161	0.9980
9	0.02261	0.02503	0.03275	0.03899	0.04034	0.01170	0.04684	89.4	0.02311	0.03830	0.03231	0.03002	0.01129	0.9961
10	0.01618	0.01878	0.02727	0.03343	0.03505	0.00631	0.04304	53.7	0.01683	0.03886	0.02660	0.02614	0.01236	0.9790

Table A.5. EBA with *CIT* as variable of interest, $k = 11$ ($M = 1{,}352{,}078$)

No.	Percentile					Leamer EBA			Granger EBA ($\delta = 0.1$)		Sala-i-Martin EBA			
For detail: see Table A.2	5	10	50	90	95	Lower Bound	Upper Bound	% significant at 5%	Lower Bound	Upper Bound	Unweighted mean	Weighted Mean	Weighted Std.error	CDF(0)
1	0.01876	0.02041	0.02688	0.03331	0.03490	0.00707	0.04617	92.42	0.00867	0.03372	0.02684	0.02065	0.00882	0.9904

Table A.6. EBA with variables of interest different from *CIT*, $k = 3$ ($M = 1{,}771$). Reported only if CDF(0) > 0.95

Variable of interest	EBA		REBA ($\delta = 0.1$)		SiM-EBA			
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Unweighted mean	Weighted Mean	Weighted Std.error	CDF(0)
Debt	-3.386	-1.050	-2.382	-1.438	-2.236	-2.177	0.272	1.000
Dens	-4.208	8.247	-0.299	7.084	-1.773	2.771	1.379	0.9777
Disp	-4.208	8.247	-3.007	0.130	-0.288	-1.080	0.637	0.9550
Unemp	-3.563	0.451	-2.370	-0.337	-1.634	-0.960	0.283	0.9997
G9	-12.263	6.595	-7.193	-4.118	-5.770	-5.736	2.652	0.9847
East	-32.878	15.607	-17.570	-3.073	-8.931	-9.346	2.615	0.9998
CS2	-2.605	1.817	-1.661	-0.666	-0.335	-1.223	0.596	0.9799
Mig	-1.558	0.426	-0.822	-0.037	-0.501	-0.341	0.151	0.9880

Table A.7. OECD PISA test scores in math (06) and reading (06/09): [international standard deviation units]

Reading 2006	**Math 2006**	**Reading 2009**
Korea 556	Finland 548	Korea 539
Finland 547	Korea 547	Finland 536
Canada 527	Netherlands 531	Canada 524
New Zealand 521	Switzerland 530	New Zealand
Irland 517	Canada 527	Japan 520
Australia 513	Japan 523	Australia 515
Poland 508	New Zealand 522	Netherlands 508
Sweden 507	Belgium 520	Belgium 506
Netherlands 507	Australia 520	Norway 503
Belgium 501	Denmark 513	Estonia 501
Switzerland 499	Czech Republic 510	Switzerland 501
Japan 498	Iceland 506	Poland 500
United Kingdom 495	Austria 505	Iceland 500
Germany495	**Germany** 504	United States 500
Denmark 494	Sweden 502	Sweden 497
Austria 490	Irland 501	**Germany 497**
France 488	France 496	Ireland 496
Iceland 484	United Kingdom 495	France 496
Norway 484	Poland495	Denmark 495
Czech Republic 483	Slovak Republic 492	United Kingdom 494
Hungary 482	Hungary 491	Hungary 494
Luxembourg 479	Luxembourg 490	Portugal 489
Portugal 472	Norway 490	Italy 486
Italy 469	Spain 480	Slovenia 483
Slovak Republic 466	USA 474	Greece 483
Spain 461	Portugal 466	Spain 481
Greece 460	Italy 462	Czech Republic 478
Turkey 447	Greece 459	Slovak Republic 477
Mexico 410	Turkey 424	Israel 474
OECD Mean 492	Mexico 406	Luxembourg 472
	OECD Mean 498	Austria 470
		Turkey 464
		Chile 449
		Mexico 425
		OECD average 493

Appendix B

Dependent Variables

Score (*S*)	Federal state-mean score of test subject mathematics and reading; high school ninth graders: OECD PISA-E 00/03/06, 2009: *Ländervergleich*, OECD PISA (reading only)
BotX%	State-mean score of bottom-X% of students; subjects: math and reading; high school ninth graders: PISA-E 00/03/06, 2009: *Ländervergleich*, OECD PISA (reading only)
TopX%	State-mean score of top-X% of students; subjects: math and reading; high school ninth graders: PISA-E 00/03/06, 2009: *Ländervergleich*, OECD PISA (reading only)

Explanatory Variables (in alphabetical order)

CIT	Cumulative instructional time of PISA-E cohorts 00/03/06 for math and reading, respectively; *Ländervergleich* 2009 for reading; aggregated curricular hours (see Section 2.1)
City	Dummy = 1, if city-state (Berlin, Bremen, Hamburg)
Cons	Election result of CDU/CSU (conservative parties), federal election preceding respective test, *Source*: State parliaments
CS1	State-mean of class size in elementary schools, respective test year, *Source*: Statistisches Bundesamt
CS2	State-mean of class size in secondary I (Gymnasium), respective test year, *Source*: Statistisches Bundesamt
Debt	State-public-indebtedness per inhabitant in respective test year, 1,000 Euros, *Source*: Statistisches Bundesamt
Dens	State-population per square-kilometre (km^2) in respective test year, *Source*: Statistisches Bundesamt
Disp	State-disposable-income per inhabitant in respective test year, 1,000 Euros, *Source*: Statistisches Bundesamt
East	Dummy = 1, if East-German state ($\in$ Neue Bundesländer)
Fee	Last cohort of state experiencing secondary school fees, birth year, *Source*: Riphahn (2011)
Fee2	First cohort of state after abolishment of secondary school fees, birth year, *Source*: Riphahn (2011)
Fem	State-employment-rate of females in respective test year, percent, *Source*: Statistisches Bundesamt
For	Foreigner share of state-population in respective test year, percent, *Source*: Statistisches Bundesamt
G9	Dummy = 1, if no. secondary school years to final grade = 9 in respective test year, *Source*: Kultusministerkonferenz, www.kmk.org (see Section 3.2)

GDP	Gross domestic state-product per inhabitant in respective test year, 1,000 Euros, *Source*: Statistisches Bundesamt
HT1	State-mean of hours per teacher in elementary schools, respective test year, *Source*: Statistisches Bundesamt
HT2	State-mean of hours per teacher in secondary I in respective test year, *Source*: Statistisches Bundesamt
Mig	State-share of 15 years old students with migration background in respective test year, *Source*: Baumert (2002), Prenzel *et al.* (2005, 2008), Köller *et al.* (2010)
PT1	State-share of part-time teachers in elementary schools in respective test year, *Source*: Statistisches Bundesamt
PT2	State-share of part-time teachers in secondary schools in respective test year, *Source*: Statistisches Bundesamt
Seg	State-share of 15 years old students attending high school (Gymnasium) in respective test year, *Source*: Baumert (2002), Prenzel *et al.* (2005, 2008), Köller *et al.* (2010)
ST1	State-mean of student-teacher-ratio in elementary schools in respective test year, *Source*: Statistisches Bundesamt
ST2	State-mean of student-teacher-ratio in secondary I (Gymnasium), respective test year, *Source*: Statistisches Bundesamt
Unemp	State-unemployment-rate in respective test year, percent, *Source*: Statistisches Bundesamt

Table B.1. Summary Statistics of (Non-qualitative) Explanatory Variables

Variable	*Cons*	*Debt*	*Dens*	*Disp*	*Emp*	*Fem*	*For*	*G9*	*GDP*	*Seg*	*Mig*	*CS1*	*CS2*	*ST1*	*ST2*	*HT1*	*HT2*	*PT1*	*PT2*
Mean	39.1	8.5	0.7	16.8	12.8	68.1	6.5	0.8	24.9	30.5	22.3	21.3	26.3	18.3	15.4	20.5	19.5	1.4	2.6
Max	60.7	24.5	3.9	23.5	21.4	77.4	15.4	1.0	44.0	41.2	51.7	25.1	31.2	21.7	19.1	24.0	23.3	12.1	22.6
Min	19.4	2.9	0.1	12.8	5.5	57.7	0.1	0.0	16.5	24.8	2.9	17.0	22.5	11.8	10.7	15.5	14.3	0.1	0.2
Range	41.3	21.6	3.8	10.7	15.9	19.7	15.3	1.0	27.5	16.4	48.8	8.1	8.7	9.9	8.4	8.5	9.0	12.0	22.5
Std. dev.	9.8	4.0	1.0	2.5	4.8	5.0	4.9	0.4	6.9	3.8	12.5	1.9	1.9	3.7	2.0	3.3	1.5	2.0	3.1
Median	40.2	7.9	0.2	16.5	11.3	67.9	6.7	1.0	23.2	30.0	23.5	21.5	21.5	18.1	15.8	20.7	19.5	0.8	1.9

(a) Path-dependent and institutional variables

Segregation. It is straightforward to assume a negative relationship between the relative share of a cohort of ninth graders attending high school (in Germany *Gymnasium*) and the average PISA-test score of this group of students (Baumert 2002, p. 92, 124, 141). As a smaller proportion might reach better learning outcomes, the selection of these students (those attending *Gymnasium*) might matter. Ultimately, a negative relationship might indicate future academics being educated and promoted better in smaller groups. Undesirable side effects are social disparities and inequity (Hoxby 2000).

Family background and path dependency. The historical time of abolishment of secondary school fees at the federal state level in Germany can be seen as a path-dependent determinant of student achievement. It is immanent to the respective schooling system. For example, the state of Rhineland-Palatinate (RP) continued to raise tuition fees for secondary education up to two decades after world war II. According to the estimates of Riphahn (2011), the abolishment of these fees has increased secondary school attendance by about six percent. The positive enrollment effect is found to have been particularly pronounced for female students. This finding suggests two lines of reasoning. First, families with a lower social status were able to send their children to secondary school after the abolishment of fees. Ninth-graders of the PISA-test cohorts 2000, 2003, 2006, and 2009 may have parents or grandparents who were able to attain a high school degree after abolishment of fees. A corresponding generation of parents or grandparents from another state, however, may not have had this chance due to fees and hence may not have started a tradition of higher education (*first-cohort-without effect*). Secondly, the awareness of costs related to secondary education witnessed by the last birth cohort who paid fees might matter for today's students' work ethic as this awareness might have been passed on to next generations (*last-cohort-with effect*). For a recent theoretical rationalization of both arguments see Doepke and Zilibotti (2008).

(b) Political economy factors

Conservative party effects. Post-war Germany witnessed a four-party and as of German unification a five-party representative democracy with two dominating parties: the conservative Christian Democratic Union (in the state of Bavaria, BY: the Christian Social Union) and the left-of-center social democrats (SPD). Party platforms differ in their education policy programs at the state-level.

(c) Socio-demographic framework

Socio-demographic conditions. Some authors find for German states a significant positive impact of population density on the general support of education in a federal state (Büttner *et al.* 2004). Thus, density might proxy a pro-education environment with regard to public spending. On the other hand, particularly urbanized, densely populated regions typically attract immigrants bearing potential adverse effects on learning outcomes. This might be due to non-native speakers requiring a higher teaching intensity.