High Stakes Testing and Student Achievement

Giorgio Brunello*, David Kiss†

Preliminary version. Please do not cite or circulate without the authors' permission.

Abstract

We investigate the effect of high stakes exam periods on math test scores of German primary and secondary students. In our setting, stakes are high during the final grades in primary and secondary education because student performance at these grades determines subsequent educational or labor market success. Our difference-in-differences estimates reveal substantial high stakes effects: on average, students nearing the final grades of their primary or secondary education experience a 0.20 s.d. gain in math scores. As high stakes effects only develop towards the end of a schooling level, there are potentially large gains from better motivating students right upon enrollment into new schools.

JEL: I26, J24, D91

Keywords: high stakes testing, student motivation, achievement, (perceived) returns to education

*Department of Economics and Management "Marco Fanno", University of Padova, Via del Santo 33, 35123, Padova, Italy, Email: giorgio.brunello@unipd.it

†Institute of Labor Economics, Leibniz University Hannover, Koenigsworther Platz 1, 30167 Hannover, Germany, Email: kiss@aoek.uni-hannover.de.

This project is funded by the German Research Foundation (DFG), Priority Programme SPP1646, grant number KI 1948/1-1, which is gratefully acknowledged. It employs data from the National Educational Panel Study (10.5157/NEPS:SC2:6.0.1 and 10.5157/NEPS:SC3:6.0.1). They were collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). Since 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

1 Introduction

This study investigates the effect of high stakes, i.e. periods with high returns to test preparation, on math achievement. When it comes to the motivation to learn, a rich body of research has shown that student outcomes are mainly determined by the efforts exerted by the student, her parents, peers, and teachers.[1]

To some extent, these agents can be incentivized – well-known examples include school vouchers (Rouse 1998, Behrman et al. 2016), performance-based teacher compensation (Glewwe et al. 2010, Duflo et al. 2012), school accountability systems (Figlio and Loeb 2011), or central examinations (Jürges et al. 2005). However, less is known about incentives targeted at students: though it has been shown that cash rewards can increase student performance (Angrist and Lavy 2009, Dearden et al. 2009), student effort is also shaped by the prevailing social norms in the class (Bursztyn and Jensen 2015).

In this paper, we focus on another mechanism that presumably raises both, student effort and parental supervision. The central assumption of this paper is that "stakes" are "high" during the final grades in primary and secondary school. In several countries, students are tracked by ability into various types of secondary schools as they make the transition from primary to secondary education.[2] Enrollment into these secondary schools is mainly determined by a student's final marks from primary school. Ability-tracked systems therefore generate strong learning incentives, particularly towards the end of primary education. A similar argument applies to the final grade of secondary school, after which students make the transition either towards the labor market, vocational training, or tertiary education. Again, one would expect student effort to steadily increase as the end of secondary school is nearing.

Our empirical strategy exploits the fact that the final grades of primary and secondary education are varying among students of the same age in Germany. In most German federal states, primary school ends after the fourth grade. In Brandenburg,

[1] The extensive empirical evidence on the agents involved in the production of student skills is summarized by Cunha and Heckman's (2007) well-known theoretical model on skill formation.

[2] See Hanushek and Wößmann's (2006) assessment of ability-tracked school systems for more details.

however, primary education lasts six years. Stakes are therefore not the same among fourth-graders: they are low for fourth-graders in Brandenburg but high in the remaining federal states.[3] The same reasoning applies to ninth-graders in secondary school. The ninth grade is the final grade in lower-secondary school (Hauptschule). Compared to ninth-graders in middle-secondary school (Realschule), one would therefore expect lower-secondary students to be more motivated because middle-secondary school ends one year later after the tenth grade. As math achievement was assessed three times in our panel data, we can employ a difference-in-differences (DiD) framework and test for parallel trends.

Our DiD estimates reveal substantial high stakes effects. Compared to earlier grades, math scores measured at the final grades in primary and secondary education are increasing by 0.24 and 0.15 standard deviations, respectively.[4] We attribute this effect to changes in stakes as the math score differential between control and treatment students remains fairly constant during earlier math skill assessments.

There exist a a few papers that find a positive association between student outcomes and final exam periods (Winfield 1990, Frederiksen 1994, Bishop 1998, Neill and Gayler 2001). However, we are aware of only two studies (Jacob 2005, Federivcova and Munich 2017) that intend to estimate average treatment effects by employing a difference-in-differences framework similar to ours. Both papers find that high stakes testing boosts student math achievement. Using data on Slovak primary students, Federivcova and Munich (2017) report that high stakes admission exams towards the end of primary education raise student math scores by 0.2 standard deviations, which is similar to our findings. In their setting, Czech students (who were not affected by the policy) constitute the control group. We complement these findings by additionally testing for parallel trends during low stakes periods. In a similar manner, Jacob (2005) compares mean math achievement levels of students living in regions where high stakes testing was in

[3] We will later discuss why we excluded Berlin, where primary education also lasts six years.
[4] As a rule of thumb, students learn between 0.25 and 0.33 standard deviations during a school year (Wößmann 2016).

effect to regions where high stakes testing was absent. Their high stakes estimates are more sizable than ours (around 0.3 s.d.). We provide a robustness check for their approach by choosing a school type (rather than a district) not affected by high stakes testing as a control group in our sample of secondary students.

The paper proceeds as follows. The next section covers our data. Section 3 develops the empirical strategy in three steps: we first discuss the relationship between stakes and timing of assessments, then define our treatment and control groups before turning to our DiD estimation framework. Results and potential mechanisms are discussed in Section 4. Section 5 concludes.

2 Data

This study employs the National Educational Panel Study (NEPS), Starting Cohorts 2 and 3, a representative panel of German students. Students were interviewed and tested the first time in fall 2010, upon enrollment in primary education (Starting Cohort 2) or at the the beginning of secondary school (Starting Cohort 3). In either case, a student's educational progress was tracked in yearly follow-up assessments.[5]

All variables used in this study are summarized in Tables 1a (primary students) and 1b (secondary students). One can infer from Table 1a that students were first assessed during grade 1, and re-assessed during grades 2 and 4. Average math scores lie between .13 and .19 standard deviations. The reason for this as follows: we first standardized math scores by grade for all students in the original data before constraining them to our analysis sample. By doing so, we implicitly set the math scores of all German students as the reference. In our study, primary students are positively selected as their mean math scores are larger than zero.

On average, students were 7.86 years old when they enrolled in primary school. A student's initial math achievement denotes her standardized math score from the first

[5] See Bela et al. (2012) for details and Blossfeld et al. (2011) for a general introduction to the NEPS.

assessment.[6] As one may expect, initial math skills are closely related to current math skills.

In most cases, the highest educational background of the student's parents is tertiary education (around 67%), followed by vocational training (21%). Whereas observations with missing values in math achievement, age or gender were deleted, we allow for missing values in parental education. The first assessment contains data on 1457 students. Attrition rates are low in the primary sample, which can be inferred from the modest drops in observation numbers. Finally, as will be explained in detail in Section 3, the vast majority of primary students (94%) are classified as treatment students.

As one can observe in Table 1b, students in our secondary school sample have below-average math skills. The reason for this is the following: in Germany, students are tracked by ability into various types of secondary schools once they complete their primary education. The most common types of secondary schools are lower-secondary (Hauptschule), middle-secondary (Realschule), and upper-secondary (Gymnasium). As we will explain in Section 3, our secondary sample only includes lower- and middle-secondary students, leading to average math scores smaller than zero.

The lower math achievement levels of our secondary students is further reflected by the educational background of their parents: in Table 1a, the majority of parents held a tertiary degree. In our sample of secondary students (Table 1b), however, most parents only completed vocational training. The observation numbers also warrant some explanation: initially, 2098 fifth graders were assessed. This number grew to 2591 during the second assessment (grade 7) because new students were added to the panel. One can further observe that a significant number of students left the panel between grades 7 and 9. Because initial math skills are increasing between the second and third assessments, attrition is primary caused by low-ability students who left the panel. Finally, roughly 30% of our secondary students belong to the treatment group, which will be further

[6] The few missing values in this variable were imputed by the math score from the second assessment.

clarified in the following section.

3 Empirical Strategy

3.1 Timing and Stakes

Even though panel participants were interviewed each year, math scores were assessed irregularly. The precise timing of the math score assessments is summarized in Table 2: As one can see, primary students were tested during grades 1, 2, and 4. Therefore, in our sample of primary students, the first period $t = 1$ corresponds with the first grade, $t = 2$ with the second grade, and $t = 3$ with the fourth grade. Note that stakes are low during periods 1 and 2 because primary education usually lasts till the fourth grade ($t = 3$). Consequently, we assume that stakes are not changing between periods 1 and 2 (i.e., grades 1 and 2) but may increase between periods 2 and 3 (i.e., between grades 2 and 4).

The math skills of secondary students were also assessed three times: during grades 5, 7, and 9. Again we associate each grade with a period $t \in \{1, 2, 3\}$. Because lower-secondary education ends after the ninth grade, stakes are assumed to be low between periods 1 and 2, but to increase between periods 2 and 3.

3.2 Treatment and Control Groups

Table 3 summarizes how we define our treatment and control groups. In the sample of primary students, the treatment indicator D equals one if students live in Bavaria or Saxony, and $D = 0$ for students from Brandenburg. One may wonder why we focus on three out of a total of 16 German federal states: primary teachers have to propose the secondary school track they consider most suitable for a child. These so-called teacher recommendations are in written form. However, most federal states allow parents to ultimately decide in which type of secondary school (lower-, middle-, or upper-secondary) to enroll their child. The only exceptions are Bavaria, Brandenburg, and Saxony where

either the teacher's recommendation is binding[7], or where secondary schools are allowed to administer placement tests, see Kultusministerkonferenz (2009). Put differently, enrollment into upper-secondary school, the most academic secondary school track, is most restrictive in these three German federal states. As the educational systems of these three states are similar with respect to this highly relevant feature (and, presumably, additional unobserved characteristics), we chose to restrict our primary sample in this way.

As one can see in Table 3, primary students from Bavaria and Saxony constitute our treatment group $D = 1$ because stakes are high at the fourth grade, i.e., at $t = 3$. In Brandenburg, however, primary education lasts 6 years. We therefore assume that stakes are low in Brandenburg at $t = 3$ (i.e., the fourth grade) because primary school still lasts for two more years.[8] To summarize, the three states have highly selective secondary schooling systems but differ in the length of primary education – the classification into treatment and control solely depends on whether stakes are high or low during period $t = 3$.

A similar logic applies to our sample of secondary students in Table 3. However, this time students are grouped by the type of secondary school they are enrolled in, not by regions. In NEPS SC3, the third assessment of student math skills occurred during the ninth grade ($t = 3$), which is the final grade in lower-secondary school. Because middle-secondary school ends after the tenth grade, i.e. one year later, we assume that stakes are low (or at least lower) for middle-secondary students during grade nine. As before, the classification into treatment or control depends on whether stakes are high or low during period $t = 3$: lower-secondary students constitute the treatment group ($D = 1$) because stakes are high for them during the ninth grade whereas stakes are still low in middle-secondary school during grade nine ($D = 0$).

[7] That is, parents cannot enroll their child to school types that are not explicitly listed in the teacher's recommendation.

[8] As already mentioned in the introduction, primary education also lasts six years in Berlin. However, above-average primary students have the possibility to make the transition to upper-secondary school already after the fourth grade. As this may affect the composition of classes already during the third and fourth grades, we decided to omit Berlin from the analysis. In addition, Berlin is not comparable to the selected three federal states because secondary school track choice is very lenient in Berlin.

3.3 Difference-in-Differences Model

Throughout, we estimate two specifications of the following baseline model:

$$\begin{aligned} mathscore_{it} &= \alpha + \lambda_2 \cdot t_2 + \lambda_3 \cdot t_3 + \\ &\quad \beta \cdot D_i + \delta_2 \cdot (t_2 \times D_i) + \delta_3 \cdot (t_3 \times D_i) + \varepsilon_{it} \end{aligned} \tag{1}$$

The standardized math score of student i measured in period $t \in \{1, 2, 3\}$ is regressed on two time fixed effects t_2 and t_3.[9] The second row in (1) simply interacts the first row with our treatment indicator D_i as defined in Table 3. Errors ε_{it} are clustered at the school level.

The first specification of (1) further contains the control variables listed in Tables 1a and 1b, including a student's initial math achievement. The second specification instead adds student fixed effects to (1). This more restrictive specification requires a student to have participated in all three assessments, and mainly serves as a robustness check for the results obtained from the first specification of our baseline model.

Estimates of δ_2 can be interpreted as a parallel trends test for the following reason: during the first and second assessments, stakes are low for both treatment and control students. Therefore, as one moves from $t = 1$ to $t = 2$, one would expect no significant change in the math score differential between treatment and control students.[10] As one moves further to $t = 3$, however, stakes become high for students in the treatment group but remain low for control students. We therefore interpret δ_3 as the average treatment effect of high stakes on math achievement.

Our DiD framework is illustrated in Figures 1 and 2. Figure 1 simply depicts the raw

[9] The two time dummy variables are defined as follows: $t_2 = \begin{cases} 1 & \text{if } t = 2 \text{ (second assessment)} \\ 0 & \text{else} \end{cases}$ and $t_3 = \begin{cases} 1 & \text{if } t = 3 \text{ (third assessment)} \\ 0 & \text{else} \end{cases}$.

[10] Mathematically, we have the following. Let $\Delta_t \coloneqq E(mathscore \mid D = 1, t) - E(mathscore \mid D = 0, t)$ denote the math score differential between treatment and control students in period t. This implies $\Delta_1 = \beta$, $\Delta_2 = \beta + \delta_2$, and $\Delta_3 = \beta + \delta_3$. Hence, the difference in differences $\delta_2 = \Delta_2 - \Delta_1$ corresponds to the change in the math score differential between the first and second assessments, and the difference in differences $\delta_3 = \Delta_3 - \Delta_1$ reflects the change in the differential between the first and third assessments.

math score differential between treatment and control students for each period $t = 1, 2, 3$. As one can see, differentials during low stakes periods (i.e., $t = 1$ and $t = 2$) are quite similar. However, the differential between treatment and control students grows as stakes become high for treatment students in $t = 3$.

This development is depicted in Figure 2. The first difference in differences, $\delta_2 = \Delta_2 - \Delta_1$ reflects the change in the math score differential between periods $t = 1$ and $t = 2$. Because stakes remain low for all students during these two periods, $\hat{\delta}_2 \approx 0$ can be interpreted as a successful test for parallel trends. The second difference in differences, $\delta_3 = \Delta_3 - \Delta_1$, reflects the effect of high stakes testing on math scores as stakes switch from "low" to "high" for treatment students, but remain "low" for control students.

4 Results

4.1 Main Findings

Table 4 presents our main findings. The first two columns refer to primary school sample, the remaining two columns cover secondary students. We estimate two specifications of the baseline model (1): the first specification adds the control variables listed in Tables 1a and 1b, including a student's initial math score. To check the robustness of our results, the second specification instead adds student fixed effects and age to (1) which requires a student to have participated in all three assessments.

Turning to the estimates, one can see that $\hat{\delta}_2$ is both insignificant and of relatively small magnitude for both school types and in both specifications. We interpret this finding as an empirical support for the parallel trends assumption. Estimates of $\hat{\delta}_3$, one the other hand, turn out to be positive and statistically significant: according to a rule of thumb, students learn the equivalent of one quarter to one third of a standard deviation during a school year (Wößmann 2016). Our estimated high stakes effects, ranging between 0.15 and 0.24 standard deviations, are therefore sizable.

4.2 Subgroup analyzes

Tables 5a and 5b provide subgroup analyses based on the control variable specification of (1). Table 5a reports high stakes effects for models that were separately estimated for girls and boys. Note that this procedure affects how treatment and control groups are specified: treatment girls are now compared to control girls, and the same rationale applies to the male subsample.

Turning to the findings reported in Table 5a, one can see that high stakes effects are larger for girls than for boys. This pattern is also confirmed by the fixed effects specifications (results not reported), though at lower significance levels. There are two plausible explanations why girls are more affected by high stakes than boys: first, girls may take high stakes exams more seriously than boys. As shown by Xu (2006) and Wagner et al. (2008), for example, girls spend more time on homework than boys. In addition, because girls have lower average math scores than boys, there might be simply more room "to catch up" for girls during high stakes exam periods.

Table 5b reports high stakes effects for students whose parents have at most completed vocational training ("≤voc."), and students whose parents have completed either upper-secondary school or tertiary education (">voc."). As expected, a higher educational background of the parents is associated with higher mean math scores. Interestingly, estimated high stakes effects are larger for students with less-educated parents. We interpret this finding as follows: if parents are well-educated, their supervision and monitoring might be high in general, regardless whether stakes are low or high. In contrast, less-educated parents may respond to changes in stakes and intensify their monitoring efforts as stakes become high. An alternative explanation could be centered around adjustments in student effort: children of less-educated parents may receive less help in general, and decide to take responsibility for their own educational success.

4.3 Mechanisms

Once the existence of high stakes effects has been empirically established, it is natural to ask for the mechanisms behind them. In NEPS, students, parents, and educators were asked several questions related to student effort. Example questions include: *To which extent do you agree that...*

- *your child quickly gives up if challenged?* (respondents: parents)
- *your class takes things seriously?* (respondents: teachers)

Table 6a provides estimates of our DiD-model (1) with various effort measures as the dependent variable. This allows us to test for both parallel trends and adjustments in effort as stakes increase. In the first column of Table 6a, the dependent variable denotes parental agreement with the statement *"my child quickly gives up if challenged"*. Parents could reply on a 5-point Likert scale ranging from 1 ("fully disagree") to 5 ("fully agree").

Results in Table 6a, Column 1, reveal the following: as stakes increase, parents are less satisfied with the learning efforts of their children. This result is difficult to interpret: the negative estimate of δ_3 might suggest that parents increase their supervision during high stakes periods and are not satisfied with their child's progress. In that case, it is still possible that student effort actually increases during high stakes periods. This interpretation is confirmed by the second column in Table 6a: even though the effect of high stakes on homework completion is insignificant, its positive sign and magnitude suggest that students take their studies more seriously during high stakes periods. The third column somewhat resembles the first: here, teachers were asked how well their students pay attention in class. Similar to Column 1, the negative high stakes effect on teacher-reported student attention might primarily reflect that teachers are unsatisfied with the (actually elevated) effort levels of their students.

This interpretation is confirmed by Table 6b, which refers to our sample of secondary students. One can see that teachers observe an increase in student effort (Column 1) but are still dissatisfied with their students' attention levels (Column 2). Taken together,

these findings are difficult to interpret as the respondents – particularly students – might respond (strategically) to changes in stakes. We further want to note that we investigated several measures of student effort, but only the ones reported in Tables 6a and 6b turned out to be sizable.

5 Summary and Conclusions

Using a difference-in-differences framework, we find large increases in math achievement as stakes switch from "low" to "high". On average, high stakes boost math achievement by .24 and .15 standard deviations in primary and secondary school, respectively. Further analyses reveal that high stakes effects are more pronounced among girls and students with less-educated parents.

Because high stakes effects are prevalent only during the final grades of primary and secondary school, our results suggest that students do not realize their full potential right from the beginning as they enroll into a primary or secondary school. Because learning begets learning (Cunha and Heckman 2007, Cunha et al. 2010), possible losses form unrealized student potential might be large – if students catch up by roughly one fifth of a standard deviation towards the end of primary and secondary school, how much could they have learned if they were properly motivated right from the beginning?

Of course, this interpretation of our findings crucially depends on whether high stakes effects are persistent our rather short-lived. As our data do not permit to conduct such an analysis, this important question must be left for further investigations. In addition, further research may attempt to identify the mechanisms behind high stakes effects.

References

Angrist, Joshua, and Victor Lavy. 2009. "The effects of high stakes high school achievement awards: Evidence from a randomized trial." *American economic review*, 99(4): 1384–1414.

Behrman, Jere, Michela Tincani, Petra Todd, and Kenneth Wolpin. 2016. "Teacher quality in public and private schools under a voucher system: The case of Chile." *Journal of Labor Economics*, 34(2): 319–362.

Bela, Daniel, Sebastian Pink, and Jan Skopek. 2012. "Data manual. Starting Cohort 3 – From lower to upper secondary school. NEPS SC3 1.0.0." NEPS Research Data Paper, University of Bamberg.

Bishop, John. 1998. "Do curriculum-based external exit exam systems enhance student achievement?" Working Paper.

Blossfeld, Hans-Peter, Hans-Günther Rossbach, and Jutta von Maurice. 2011. *Education as a lifelong process – The German National Educational Panel Study (NEPS).* Zeitschrift für Erziehungswissenschaft - Sonderheft 14. Wiesbaden: Springer.

Bursztyn, Leonardo, and Robert Jensen. 2015. "How does peer pressure affect educational investments?" *Quarterly Journal of Economics*, 130(3): 1329–1367.

Cunha, Flavio, and James Heckman. 2007. "The technology of skill formation." *American Economic Review*, 97(2): 31–47.

Cunha, Flavio, James Heckman, and Susanne Schennach. 2010. "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica*, 78(3): 883–931.

Dearden, Lorraine, Carl Emmerson, Christine Frayne, and Costas Meghir. 2009. "Conditional cash transfers and school dropout rates." *Journal of Human Resources*, 44(4): 827–857.

Duflo, Esther, Rema Hanna, and Stephen Ryan. 2012. "Incentives work: Getting teachers to come to school." *American Economic Review*, 102(4): 1241–78.

Federivcova, Miroslava, and Daniel Munich. 2017. "The impact of high-stakes school admission exams on study achievements: quasi-experimental evidence from Slovakia." *Journal of Population Economics*, 30(4): 1069–1092.

Figlio, David, and Susanna Loeb. 2011. "School accountability." Vol. 3 of *The Handbook of the Economics of Education*, ed. by Eric Hanushek and Finis Welch, Chapter 8, 383–421. Elsevier.

Frederiksen, Norman. 1994. "The influence of minimum competency tests on teaching and learning." Working Paper.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher incentives." *American Economic Journal: Applied Economics*, 2(3): 205–27.

Hanushek, Eric, and Ludger Wößmann. 2006. "Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries." *Economic Journal*, 116(510): 63–76.

Jacob, Brian. 2005. "Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools." *Journal of Public Economics*, 89(5-6): 761–796.

Jürges, Hendrik, Kerstin Schneider, and Felix Büchel. 2005. "The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany." *Journal of the European Economic Association*, 3(5): 1134–1155.

Kultusministerkonferenz (Ed.). 2009. *Das Bildungswesen in der Bundesrepublik Deutschland 2008. Darstellung der Kompetenzen, Strukturen und bildungspolitischen Entwicklungen für den Informationsaustausch in Europa.* Bonn: Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

Neill, Monty, and Keith Gayler. 2001. "Do high-stakes graduation tests improve learning outcomes? Using state-level NAEP data to evaluate the effects of mandatory graduation tests." *Raising standards or raising barriers*, 107–126.

OECD (Ed.). 1999. *Classifying educational programmes: Manual for ISCED-97 implementation in OECD Countries.* Paris: OECD Publishing.

Rouse, Cecilia Elena. 1998. "Private school vouchers and student achievement: An evaluation of the Milwaukee Parental Choice Program." *Quarterly Journal of Economics*, 113(2): 553–602.

Wagner, Petra, Barbara Schober, and Christiane Spiel. 2008. "Time students spend working at home for school." *Learning and Instruction*, 18(4): 309–320.

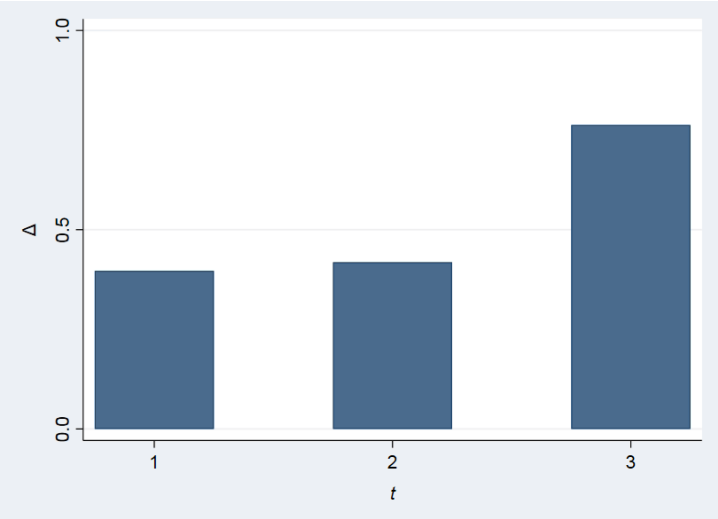
Winfield, Linda. 1990. "School competency testing reforms and student achievement: Exploring a national perspective." *Educational Evaluation and Policy Analysis*, 12(2): 157–173.

Wößmann, Ludger. 2016. "The importance of school systems: Evidence from international differences in student achievement." *Journal of Economic Perspectives*, 30(3): 3–32.

Xu, Jianzhong. 2006. "Gender and homework management reported by high school students." *Educational Psychology*, 26(1): 73–91.

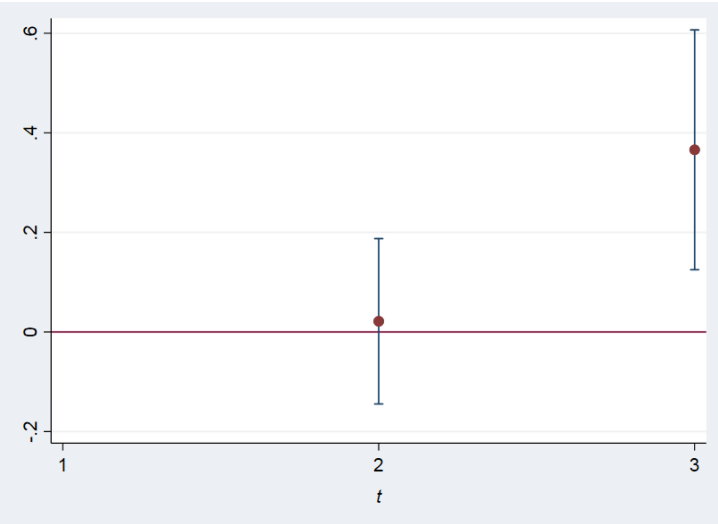
Figures

Figure 1: Math score differentials between treatment and control students for various periods



This figure depicts Δ_t, the difference in (raw) math scores between treatment and control students during periods $t = 1, 2, 3$. As one can see, the math score gap between treatment and control students is larger in high stakes periods ($t = 3$) than in low stakes periods ($t = 1, 2$). For further details, see Tables 2 and 3 as well as Section 3.3 and Footnote 9.

Figure 2: Changes in the math score differential over time



This figure depicts the change in Δ, the math score differential between treatment and control students, over time. The first difference in differences, $\delta_2 = \Delta_2 - \Delta_1$, denotes the change in the math score differential between periods 1 and 2. Similarly, $\delta_3 = \Delta_3 - \Delta_1$ reflects the change in Δ between periods 1 and 3. As stakes remain low during $t = 1$ and $t = 2$, we observe no change in the math score differential Δ. However, between $t = 1$ and $t = 3$, stakes switch from "low" to "high", resulting in an increase in the math score differential. We interpret $\hat{\delta}_2 \approx 0$ as a successful empirical test for parallel trends and $\hat{\delta}_3$ as the average treatment effect of high stakes periods on math achievement.

Tables

Table 1a: Descriptive statistics (panel of German primary students)

Grade	1st		2nd		4th	
	mean	s.d.	mean	s.d.	mean	s.d.
Math achievement	.13	.92	.18	.96	.19	.92
Age	7.86	.70	8.87	.70	10.89	.70
Female	.51		.51		.51	
Initial math achievement	.12	.92	.13	.92	.15	.89
Highest parental education:						
lower- or middle-secondary	.03		.03		.03	
upper-secondary	.09		.08		.09	
vocational training	.21		.21		.20	
tertiary	.67		.68		.68	
% missing parental educ.	7%		6%		6%	
N(students)	1457		1418		1298	
% in treatment group	94%		94%		94%	

Data source: NEPS Starting Cohort 2. Unit of analysis: primary students who were assessed during the first, second, and fourth grades. Standard deviations are not reported for binary random variables. Prior to restricting the original data to our analysis sample, math achievement was standardized to mean 0 and standard deviation 1 within each grade. The highest educational degree of the student's parents is classified according to the International Standard Classification of Education (ISCED) guidelines, see OECD (1999) for details.

Table 1b: Descriptive statistics (panel of German secondary students)

Grade	5th		7th		9th	
	mean	s.d.	mean	s.d.	mean	s.d.
Math achievement	−.49	.85	−.49	.89	−.49	.81
Age	11.18	.82	13.17	.83	15.17	.83
Female	.47		.46		.46	
Initial math achievement	−.49	.85	−.43	.85	−.40	.86
Highest parental education:						
lower- or middle-secondary	.07		.05		.06	
upper-secondary	.06		.06		.07	
vocational training	.60		.63		.60	
tertiary	.27		.26		.27	
% missing parental educ.	30%		39%		38%	
N(students)	2098		2591		1995	
% in treatment group	33%		27%		25%	

Data source: NEPS Starting Cohort 3. Unit of analysis: lower- and middle-secondary students who were assessed during the fifth, seventh, and ninth grades. Standard deviations are not reported for binary random variables. Prior to restricting the original data to our analysis sample, math achievement was standardized to mean 0 and standard deviation 1 within each grade.

Table 2: Timing of math skill assessments

t	$= 1$	$= 2$	$= 3$
Primary sample	1st grade	2nd grade	4th grade
Secondary sample	5th grade	7th grade	9th grade

Table 3: Definition of treatment and control groups (by schooling level)

Sample	D	Students are...	$t = 3$
Primary	Treated $D = 1$	...from Bavaria and Saxony	Final grade in primary school
	Control $D = 0$	...from Brandenburg	Primary school lasts two more years
Secondary	Treated $D = 1$	...enrolled in lower-secondary school	Final grade in lower-sec. school
	Control $D = 0$	...enrolled in middle-secondary school	Middle-sec. school lasts one more year

Table 4: High stakes effects on math scores in primary and secondary education

	Primary school		Secondary school	
Specification:	cont.	FE	cont.	FE
$\hat{\delta}_2$	0.05	0.01	-0.03	-0.05
	(0.089)	(0.094)	(0.032)	(0.054)
$\hat{\delta}_3$	0.24^{***}	0.21^{***}	0.15^{***}	0.17^{***}
	(0.077)	(0.085)	(0.048)	(0.064)
R^2	0.58	0.02	0.69	0.02
N	4173	3579	6684	3573

This table provides estimates of δ_2 and δ_3 for two specifications of the baseline model (1). The first specification ("cont.") includes the control variables listed in Tables 1a and 1b plus a dummy variable for missing parental education. As a robustness check, the second specification ("FE") instead adds student fixed effects and age to the baseline model (1). Sample sizes are smaller in the fixed effect specifications because students must have participated in all three assessments.

Table 5a: High stakes effects by gender

	Primary school		Secondary school	
Subgroup:	girls	boys	girls	boys
mean(*mathscore*)	.07	.27	$-.65$	$-.35$
$\hat{\delta}_2$	0.02	0.07	-0.00	-0.05
	(0.099)	(0.107)	(0.041)	(0.041)
$\hat{\delta}_3$	0.29^{***}	0.20	0.18^{***}	0.13^{**}
	(0.089)	(0.154)	(0.056)	(0.062)
R^2	0.58	0.56	0.67	0.68
N	2113	2060	3147	3537

This table provides estimates of δ_2 and δ_3 that were separately obtained for girls and boys. All models include the control variables listed in Tables 1a and 1b plus a dummy variable for missing parental education.

Table 5b: High stakes effects by parental education

	Primary school		Secondary school	
Subgroup:	$\leq$voc.	$>$ voc.	$\leq$voc.	$>$ voc.
mean(*mathscore*)	$-.23$	.31	$-.65$	$-.16$
$\hat{\delta}_2$	-0.01	0.08	-0.01	-0.09
	(0.212)	(0.090)	(0.034)	(0.065)
$\hat{\delta}_3$	0.36^{**}	0.19^{**}	0.16^{***}	0.12
	(0.176)	(0.083)	(0.046)	(0.098)
R^2	0.60	0.53	0.67	0.64
N	1094	3079	4532	2152

This table provides estimates of δ_2 and δ_3 that were separately obtained for students whose parents have at most completed vocational training ("$\leq$voc."), and students whose parents have completed either upper-secondary school or tertiary education ("$>$voc."). All models include the control variables listed in Tables 1a and 1b plus a dummy variable for missing parental education.

Table 6a: Potential transmission channels (primary school)

Dep. variable:	*"my child quickly gives up if challenged"*	*"my child does her homework with care"*	*"my students pay good attention"*
Respondent:	parents	parents	teachers
$\hat{\delta}_2$	0.02	0.07	-0.07
	(0.158)	(0.053)	(0.069)
$\hat{\delta}_3$	-0.21**	0.19	-0.16
	(0.084)	(0.163)	(0.188)
N	3562	3567	3552

The dependent variables range from 1 ("fully disagree") to 5 ("fully agree").

Table 6b: Potential transmission channels (secondary school)

Dep. variable:	*"my students show high effort levels"*	*"my students pay good attention"*
Respondent:	teachers	teachers
$\hat{\delta}_2$	0.01	0.03
	(0.217)	(0.180)
$\hat{\delta}_3$	0.16	-0.16
	(0.253)	(0.199)
N	5328	5328

The dependent variables range from 1 ("fully disagree") to 5 ("fully agree").