Education Curriculum and Student Achievement: Theory and Evidence*

Vincenzo Andrietti†

Università "G. D'Annunzio" di Chieti e Pescara, Italy

Xuejuan Su‡

University of Alberta, Canada

September 8, 2018

Abstract

We propose a theory of education curricula as horizontally differentiated by their paces. The pace of a curriculum and the preparedness of a student jointly determine the match quality of the curriculum for this student, so different students derive different benefits from learning under the same curriculum. Furthermore, a change in the curricular pace has distributional effects across students, benefiting some while hurting others. We test the model prediction using a quasi-natural experiment we call the G8 reform in Germany, which introduced a faster-paced curriculum for academic-track students. We find evidence consistent with our theory: While the reform improves students' test scores on average, such benefits are more pronounced for well-prepared students. In contrast, less-prepared students do not seem to benefit from the reform.

- **JEL classification**: I21, D20, I28
- **Keywords**: Education curriculum, horizontal differentiation, distributional effects

---

*We are grateful to Nicole Fortin for sharing her STATA codes with us. We thank seminar participants at University of Alberta, Canadian Economic Association conference, European Economic Association conference, and International Workshop of Applied Economics of Education for helpful comments. All remaining errors are our own.

†Vincenzo Andrietti, Università "G. d'Annunzio"di Chieti e Pescara, Dipartimento di Scienze Filosofiche, Pedagogiche ed Economico-Quantitative (DiSFPEQ), Viale Pindaro 42, 65127 Pescara, Italy. E-mail: vincenzo.andrietti@unich.it.

‡Xuejuan Su, University of Alberta, Department of Economics, 8-14 Tory Building, Edmonton, Alberta Canada T6G 2H4. E-mail: xuejuan1@ualberta.ca.

1 Introduction

Students of different levels of preparedness (or prior knowledge) have different learning needs. Given the hierarchical nature of learning, students need to comprehend, apply, and synthesize basic materials before they can effectively learn more advanced ones. In other words, their human capital output from an earlier stage becomes the input for learning at a subsequent stage. As a result, well-prepared students, i.e. those with high human capital output from previous learning, can learn new topics more quickly, while less-prepared students learn more slowly. In this sense, when the pace of learning is ideally matched to a student's preparation, she will have better learning outcomes.

To capture this concept of a match between a student's preparedness and the ideal pace of learning, we propose a model of education curricula as horizontally differentiated products. In particular, a curriculum is characterized by two parameters: its pace of learning, and a corresponding minimum requirement on student preparation. Intuitively, faster-paced curricula impose higher requirements on students' preparation to keep up, while slower-paced curricula lower the requirements by allowing extra time for remedial work and repeated exposure to the same topics. Accordingly, well-prepared students prefer the faster-paced ones, and less-prepared students prefer the slower-paced ones.

This horizontal feature of curriculum differs from many other measures of schooling quality such as class size, teacher quality, and peer effects, which are better viewed as vertical differentiation.[1] In the case of vertical differentiation, all students would prefer to have smaller classes, better teachers, and better peers, for instance. In contrast, not all students prefer a fast-paced curriculum; instead their preferences differ according to their preparation levels.

An immediate implication of the horizontal feature of curriculum is its distributional impact across students. When many students study under the same curriculum, the learning pace may be just right for some, but too fast or too slow for others. So they experience

[1] We have borrowed the terminology of horizontal versus vertical differentiation from the industrial organization literature, see Tirole (1988) for example.

different match qualities. Now consider a change in the curriculum, for example, it becomes faster paced than the old one. In this case, well-prepared students will benefit from such a change, but less-prepared students will suffer. So a change in the curriculum generates heterogeneous effects on student learning outcomes, depending on their preparation levels.

We empirically test this theoretical prediction to see whether such pattern of distributional impact exists in the data. There are two major challenges for our analysis. The first is that it is difficult to measure the pace of a curriculum. While most teachers intuitively adjust the pace of teaching to better serve their students, one can hardly assign a numerical value to reflect such pace adjustment in practice. We address this measurement problem by focusing on ordinal comparison between two curricula, namely, which curriculum is faster-paced than the other. The second challenge is that any curriculum implemented in practice is by definition an endogenous choice. This choice depends on the student composition as well as the objective of the decision maker, both of which may not be observed in the data. These unobserved factors may confound curricular variation in observational data, making it unsuitable for identification purpose. We deal with this endogeneity problem by relying on a quasi-natural experiment, which leads to an arguably exogenous change in the pace of curriculum.

More specifically, we take advantage of the G8 reform in Germany. This reform compresses secondary schooling duration for academic-track students from nine to eight years, but keeps the academic content required for graduation fixed. Thus, the reform requires more content to be learned on an annual basis, implying a faster-paced curriculum. Furthermore, the reform was adopted by states mainly based on considerations of labor market conditions and demographic changes, so the implied curriculum change is exogenous to the learning process itself. This enables us to identify its distributional effects on student learning outcomes.

We use five waves of PISA data to measure student achievement as their test scores at the end of grade 9. Since the pooled PISA data are repeated cross-sections rather than panel data, we have very little information on students' preparedness when they entered secondary schooling. Thus we rely on two approaches to handle the unobserved nature of student preparedness, namely the quantile difference-in-difference (Q-DiD) and the recentered influence function difference-in-difference (RIF-DiD) methods. The results reflect the reform effects at different quantiles of the test score distribution, the former method estimating quantile effects on conditional distributions given the observed controls, while the latter estimating quantile effects on the single unconditional distribution. What we find is broadly consistent with our theory. While the faster-paced curriculum after the G8 reform improves student test scores on average, the benefits are more pronounced for well-prepared students. In contrast, there is little evidence that less-prepared students benefit at all.

The rest of the paper is structured as follows. Section 2 reviews the related literature. Section 3 introduces the theoretical curriculum model and derives its implications. Section 4 presents the regression models within the context of the curriculum theory. Section 5 describes the data used for empirical analysis and the estimation results. Section 6 offers concluding remarks.

2 Related literature

This paper is linked to several strands of the existing literature. On the theoretical side, there is a growing literature that focuses on the hierarchical nature of the education process, namely the human capital output from an earlier stage is an input for human capital accumulation and improves the learning effectiveness at a subsequent stage of education (see, for example, Ben-Porath, 1967; Lucas, 1988; Su, 2004, 2006; Blankenau, 2005; Blankenau et al., 2007; Cunha and Heckman, 2007; Gilpin and Kaganovich, 2012). In these studies, the hierarchical education production function is taken as exogenously given. Kaganovich and Su (forthcoming) are the first to examine the strategic choices of education production functions by schools (colleges). When colleges compete for students, they optimally choose to differentiate their curricular offerings, with those caring more about student quality than quantity choosing more challenging curricula, while those caring more about student quantity relative to quality choosing less challenging curricula. This paper focuses on the distributional impact of an exogenous change to a faster-paced (more challenging) curriculum, and offers empirical evidence that is broadly consistent with the theoretical predictions.[2]

On the empirical side, there is a large literature estimating the impact of various education inputs on students' learning outcomes, such as school quality and school resources (Card and Krueger, 1992; Currie and Dee, 2000; Hanushek, 2006), class size (Angrist and Lavy, 1999; Hoxby, 2000; Krueger, 2003; Ding and Lehrer, 2010), teacher quality (Clotfelter et al., 2006; Aaronson et al., 2007; Rothstein, 2010; Mueller, 2013), and peer effects (Sacerdote, 2001; Zimmerman, 2003; Arcidiacono and Nicholson, 2005; Carrell et al., 2009). As discussed in the Introduction, these measures of education inputs represent vertical differentiation, and economic theory is unambiguous about the qualitative impact (the direction) of a change in any of the vertical measure on students' learning outcomes. Accordingly, this literature tends to focus on quantifying the impact, typically in the form of the average treatment effect.

There is a small but fast growing empirical literature that focuses on the distributional effect of the match quality between students and schools. For example, Light and Strayer (2000) examine whether the match between student ability and college quality (measured as the average ability of its student body) affects the student's college graduation rate. They find that students are more likely to graduate if they attend colleges with quality level matching their ability level. In other words, high-ability students are more likely to graduate when attending high-quality colleges, while low-ability students are more likely to graduate if they attend low-quality colleges. More recently, Arcidiacono et al. (2016) examine the difference in the graduation rates for minority science students across University of California campuses under affirmative action policies. They find that less-prepared minority students at higher-ranked campuses had lower persistence rates in science and took longer to graduate. In

[2] The paper is also related to the "peer effects" literature pioneered by Epple and Romano (1998), where any student would benefit from having better-ability peers. In contrast, our curriculum model suggests that a less-prepared (low-ability) student may actually suffer from having well-prepared (high-ability) peers, if the curriculum is geared toward her peers and consequently too fast-paced for her own learning.

other words, had these minority students attended lower-ranked campuses according to their preparedness instead of affirmative action, they would have reached higher graduation rates in STEM fields. This line of evidence contradicts the peer effects prediction, but is consistent with our curriculum model when the curricula at low-quality colleges are better suited for the learning needs of low-ability students. Arcidiacono and Lovenheim (2016) provide an excellent survey of this empirical literature.

Another paper that is closely related to our paper is Duflo et al. (2011), who examine whether academic tracking helps or hurts low-ability students. Using randomized experimental data from Kenya, they find that tracking students by prior-achievement raises scores for all students, even those assigned to lower achieving peers. To interpret these results, they argue that tracking allows teachers to better tailor their instruction level, and lower-achieving pupils are particularly likely to benefit from tracking because teachers have incentives to teach to the top students. Unlike our model of curriculum, they model the pace of learning as a result of teacher effort, which can be changed independently of the target level of instruction. In a sense, our paper can be viewed as moving along a given efficiency frontier of the education technology consisting of different curricula, while their paper can be viewed as improving the efficiency frontier when changes in the teaching environment (tracking) and stronger incentives (contract teachers) induce higher levels of teacher effort, again a vertical measure because all students prefer more teacher effort regardless of their ability levels. Furthermore, we estimate the distributional effect of an arguably exogenous curricular change mandated by the G8 reform, holding the composition of the student body fixed. In comparison, they allow the composition of the student body to change according to the random experiment (tracking versus pooling), and estimate the impact of the endogenous curricular change made by the teachers facing different classes.

Finally, there is an empirical literature evaluating the G8 reform effects on student outcomes such as personality (Thiel et al., 2014), cognitive skills at high school graduation (Büttner and Thomsen, 2015; Dahmann, 2017), high school graduation rate and graduation age (Huebener and Marcus, 2017), post-secondary enrollment (Meyer and Thomsen, 2016; Meyer et al., forthcoming), and motivation, abilities and achievements at university (Meyer and Thomsen, 2017). Andrietti and Su (forthcoming) examine the G8 reform effects on student performance in high school, and find that while the reform effect is positive on average, it can be different across subgroups of students.[3] However, after exploring various channels, Andrietti and Su (forthcoming) conclude that differences in observed characteristics play only a limited role in understanding the *mechanism* that gives rise to the heterogeneous reform effects. This suggests that unobserved factors (in particular, differences in student preparation) are important. In this paper, we develop a theory and find supporting evidence that student preparation, an element of unobserved heterogeneity, plays a central role in shaping the pattern of heterogeneous reform effects, as observed in the data.[4]

3 The curriculum model

Consider an economy with heterogeneous students. Students differ by their preparedness $p_i \in [\underline{p}, \overline{p}]$, the prior knowledge they possess before the current learning process. Note that it is tempting to call student preparedness "ability", a term we intentionally avoid for the following reason. To highlight the hierarchical structure of the learning process, namely human capital output from a previous stage becomes the input at the subsequent stage, we reserve the term "ability" to mean innate ability, which affects only the initial learning stage (i.e., early childhood development). In contrast, innate ability and the learning experience at an earlier stage (e.g., primary schooling) jointly determine a student's preparedness for

[3] Two recent studies, Homuth (2017) and Huebener et al. (2017), perform empirical analysis similar to that in Andrietti and Su (forthcoming) (whose first version appeared as Andrietti (2015)), and find similar results.

[4] In a related context, Morin (2013) and Krashinsky (2014) examine the impact of an Ontarian high school reform on students' university outcomes. Similar to the G8 reform, the Ontario reform reduces the high school duration for most students from five to four years. Unlike the G8 reform, instead of keeping the total amount of academic content required for high school graduation fixed, the Ontario reform also reduces the number of course credits and the academic content typically associated with the fifth and last year. As a result, the Ontario reform has no impact on the pace of the curriculum *per se*, but rather delivers less content in a shorter period. Both studies find that, on average, the four-year graduates perform significantly worse than their five-year counterparts.

learning at a subsequent stage (e.g., secondary schooling). We will discuss the distribution of student preparation in detail later.

3.1 Education curriculum

For simplicity, we assume that human capital is measured on a single dimension. A *curriculum* is represented by two parameters: the pace of learning A, and the corresponding minimum requirement on student preparedness $c(A)$. For a student with preparedness p (we omit the subscript in p_i when there is no risk of confusion), if she studies under the curriculum $(A, c(A))$, her human capital output h from a period of study is

$$h = \begin{cases} (1-\lambda)p & \text{if} \quad p \leq c(A), \\ (1-\lambda)p \quad +A(p - c(A)) & \text{if} \quad p > c(A). \end{cases} \tag{1}$$

The first term allows for potential depreciation (at a rate $0 \leq \lambda \leq 1$) of her existing human capital p. If she is not sufficiently prepared (i.e., when $p \leq c(A)$), the student does not benefit from the learning process at all. If she is adequately prepared, her value-added human capital from the learning process is $A(p - c(A))$.[5]

It is obvious that if there were a curriculum $(A, c(A))$ with a very large value for A and a very small value for $c(A)$, it would give large benefit to students of almost any level of preparation. In a world of trade-offs, such a curriculum is unlikely to be feasible. At the efficiency frontier, different curricula inevitably involve a tradeoff. That is, larger values for A (faster pace) requires larger values for $c(A)$ (higher requirement on student preparedness).

Assumption 1 *Let $c(A)$ be a differentiable function of A, and $c'(A) > 0$.*

We maintain Assumption 1 hereinafter.

[5] The current one-stage curriculum model can be easily extended to a multi-stage framework, where the human capital output h^s from a given stage s determines the student preparation for a subsequent stage $s+1$, with the initial level h^0 being the innate ability. This is in line with the literature on human capital accumulation pioneered by Ben-Porath (1967) and more recently Cunha and Heckman (2007).

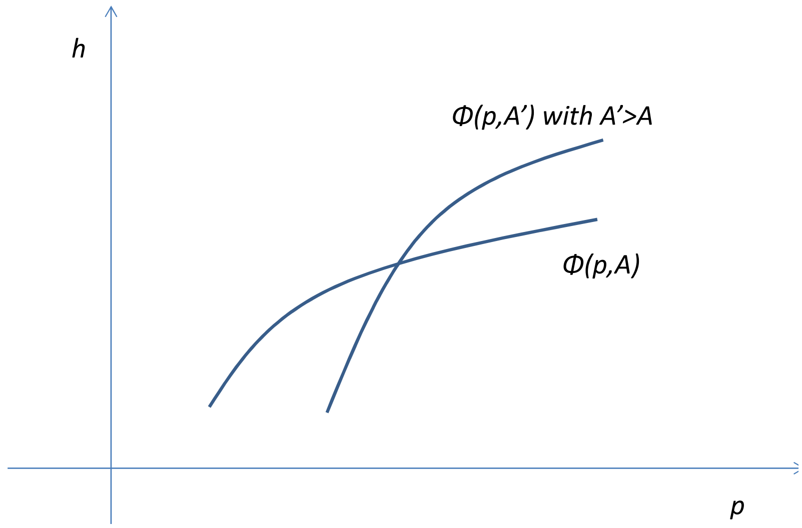

Fig. 1. More general formulation of the curriculum

It is worth pointing out that even though our formulation of the curriculum may appear restrictive at first glance, it actually is quite general. More specifically, (1) can be viewed as a local linear approximation of a general nonlinear specification $h = \max\{0, \Phi(p, A)\}$ that exhibits the single-crossing property illustrated in Figure 1, namely a faster-paced curriculum crosses a slower-paced curriculum from below. Assuming $\Phi(p, A)$ is twice continuously differentiable, we can impose the following three conditions: (1) $\Phi_1(p, A) > 0$, i.e., in a given curriculum, better-prepared students do better than less-prepared students; (2) $\Phi_{22}(p, A) < 0$ with $\lim_{A\to 0} \Phi_2(p, A) > 0$ and $\lim_{A\to\infty} \Phi_2(p, A) < 0$, i.e., for a given student, there is a unique, interior optimal curricular pace; and (3) $\Phi_{12}(p, A) > 0$ with $\lim_{p\to 0} \Phi_2(p, A) < 0$ and $\lim_{p\to\infty} \Phi_2(p, A) > 0$, i.e., there is complementarity between student preparedness and curricular pace.

It is easy to check that (1) satisfies the above three conditions, and all our results hold true under the more general formulation.[6] For the purpose of exposition, we stick with the simple linear formulation hereinafter. A direct implication of (1) is the result below:

[6]Note that since (1) represents only a local approximation, the results may not hold globally, in particular at the far end of the preparedness distribution.

Proposition 1 *Given $p' > p > c(A)$, we have $A(p' - c(A)) > A(p - c(A))$.*

Namely, well-prepared students have "absolute advantage" over less-prepared students in a given curriculum. This is a distinguishing feature of our model from Duflo et al. (2011), and follows the convention of the theoretical literature on human capital accumulation discussed earlier. Furthermore, this monotonic relationship between student preparation and her human capital output is necessary for interpreting our quantile analysis results, where there is a one-to-one mapping from the unobserved heterogeneity (student preparation) to the observed outcomes (test scores).

3.2 Ideal curriculum for a student

If an educator were able to customize the education curriculum to serve the individual learning needs of a given student with preparation p, the educator would have chosen a curriculum that maximizes the student's human capital output h according to (1). This optimal choice would be the ideal curriculum for this given student. For example, if we make the assumption that $c(A) = CA^r$ with $r > 0$, the ideal curriculum for a student with preparation p is $\hat{A}(p) = \operatorname{argmax} A(p - CA^r) = (\frac{p}{C(r+1)})^{1/r}$, which is strictly increasing in p. More generally, without a specific functional form for $c(A)$, we may not explicitly solve for the ideal curriculum $\hat{A}(p)$. Nonetheless, it is implicitly defined as the solution to the first-order equation:

$$p - c(A) - Ac'(A) = 0 \tag{2}$$

assuming that the second-order sufficient condition is also satisfied, namely $-2c'(A) - Ac''(A) < 0$. Applying the Implicit Function Theorem to (2), and then invoking the second-order sufficient condition, we have the following result:

Proposition 2 *Let $\hat{A}(p)$ be the ideal curriculum for a student with preparedness p, we have $\hat{A}'(p) > 0$.*

Here, as the ideal curricular pace for a student is strictly increasing in her preparedness, well-prepared students enjoy comparative advantage in faster-paced curricula compared to slower-paced ones, and less-prepared students enjoy comparative advantage in slower-paced curricula compared to faster-paced ones. This is the main implication of our model, that students prefer different paced curricula to suit their learning needs according to their levels of preparation.

3.3 Implemented curriculum

In practice, it is typically infeasible for an educator to customize the curriculum to serve an individual student. Instead, many students study under the same curriculum despite their different levels of preparation. When this happens, the given curriculum may not be ideal for all but a few students.

In this paper, we do not explicitly model how a curriculum gets chosen. In principle, the optimal choice will be derived from maximizing the objective function of a decision maker such as a teacher or a school principal. For example, consider N students with different preparation p_i, where $p_1 \leq p_2 \leq ... \leq p_N$. Assuming that the objective function is linear in each student's human capital output from the learning process, the optimal curriculum can be expressed as

$$A^* = \operatorname{argmax} \Sigma_{i=1}^{N} \gamma_i h_i \quad s.t.(1), \quad \Sigma_{i=1}^{N} \gamma_i = 1, \tag{3}$$

where γ_i is the relative weight the decision maker assigns to student i. So, similar to the interpretation of a social welfare function, when $\gamma_i = \gamma$ for all i, the decision maker is "utilitarian" and treats all students equally; when $\gamma_1 \geq \gamma_2 \geq ... \geq \gamma_N$, the decision maker is more concerned about less-prepared students (e.g., "no child left behind"); and when $\gamma_1 \leq \gamma_2 \leq ... \leq \gamma_N$, the decision maker is more concerned about better-prepared students.

From this setup, it is obvious that the optimal curriculum depends critically on two factors: the distribution of student preparedness, and the relative weights assigned to these students. We can characterize comparative statics of the optimal curriculum in some special cases. For instance, holding the set of relative weights fixed, the optimal curriculum would be faster paced when the distribution of student preparedness shifts to the right, i.e., students are better prepared. Alternatively, holding the distribution of student preparedness fixed, the optimal curriculum would be faster paced when the weights are shifted from less-prepared to better-prepared students. However, with more complex changes, the comparative statics cannot be characterized qualitatively, but will be sensitive to quantitative comparisons instead. An immediate implication is that, to identify and estimate the impact of curriculum on learning outcomes, cross-sectional variation of actual curricula in the observational data has limited value. In particular, unless we have perfection information on student preparedness, self-selection bias poses a major challenge. In our empirical analysis, we overcome this hurdle by relying on an arguably exogenous change in the curriculum as a result of a quasi-natural policy experiment, where the distribution of student preparedness remains stable overtime.

It is reasonable to assume that regardless of the specific objective function, an implemented curriculum has to fall within two extremes: the ideal curricula for the least-prepared ($p_i = \underline{p}$) and the most-prepared ($p_i = \overline{p}$) students. Otherwise, any curriculum with $A < \hat{A}(\underline{p})$ would be Pareto dominated by $\hat{A}(\underline{p})$, and any curriculum with $A > \hat{A}(\overline{p})$ would be Pareto dominated by $\hat{A}(\overline{p})$. Comparing two curricula within this interval, we have the following stratification result:

Proposition 3 *Consider two different paced curricula A and A', where $\hat{A}(\underline{p}) < A < A' < \hat{A}(\overline{p})$. There exists a cutoff level $\hat{p} \equiv \frac{A'c(A') - Ac(A)}{A' - A} \in (\underline{p}, \overline{p})$ such that students with $p_i = \hat{p}$ accumulate the same human capital under both curricula; students with $\underline{p} < p_i < \hat{p}$ accumulate more human capital under the slower-paced curriculum A; and students with $\hat{p} < p_i < \overline{p}$ accumulate more human capital under the faster-paced curriculum A'.*

4 Regression models

Our empirical analysis focuses on the interaction between student preparedness p and curricular pace A, neither of which is directly measured. As will be described in more detail later, the G8 reform leads to an exogenous change to a faster-paced curriculum in reformed states, i.e., an increase in A. On the other hand, we treat student preparedness p as an unobserved variable, and use two methods to estimate the reform effects at different quantiles of the student distributions.

First we consider the following quantile difference-in-difference model which we call QDiD:

$$y_{ist} = \alpha_\tau X_{ist} + \beta_\tau G8_{st} + \gamma_\tau D_s + \delta_\tau P_t + \epsilon_{ist}, \tag{4}$$

where y_{ist} is the test score for student i in state s and year t. On the right hand side, the vector X_{ist} represents a set of control variables that may affect a student's learning outcome, including demographics, family background, and school characteristics. The policy variable $G8_{st}$ is an indicator that equals 1 if the grade-9 student cohort in state s is affected by the G8 reform in year t, and 0 otherwise. State dummies are captured by D_s, and year dummies are captured by P_t. The error term is ϵ_{ist}. We estimate this model at various quantiles $\tau \in (0, 1)$, thus the parameters are quantile-specific and indexed by the subscript τ.

To link (4) to our theory, we interpret the error term ϵ_{ist} as capturing the (de-meaned) unobserved student preparation. Thus, (4) implies that after controlling for observed variables, better prepared students have higher test scores, a strictly increasing relationship linking student preparation to their test score outcomes. This is precisely what our theory predicts (Proposition 1).[7] Furthermore, we need the "common distribution" assumption for the error term ϵ_{ist}, namely at a given quantile τ, students have the same level of preparation across states and years. With this assumption, after controlling for observed variables, test

[7] In contrast, such an interpretation would be invalid in an alternative framework, e.g., Duflo et al. (2011), where education benefits are hump-shaped (instead of strictly increasing) across student ability levels. In that framework, the same test score may be obtained by students on either side of the hump shape, i.e., different ability levels.

score differences between treated and control states before and after the reform reflects how the reform affects students who have the same level of preparation at the given quantile τ. Thus, the distributional impact of a change in the curricular pace on students across the preparation distribution, as our theory predicts (Proposition 3), can be seen by tracing the parameter β_τ across different quantiles $\tau \in (0, 1)$.

However, the common distribution assumption is a well-known limitation, which cannot be expected to hold in general. To relax this assumption, we also use the Recentered Influence Function (RIF) method recently developed by Firpo et al. (2009).[8] More specifically, when the observed outcome (i.e. test score) varies monotonically with the unobserved variable (i.e. student preparedness), RIF for the τth quantile is given by

$$RIF(Y; q_\tau, F_Y) = q_\tau + \frac{\tau - \mathbb{1}\{Y \leq q_\tau\}}{f_Y(q_\tau)}, \tag{5}$$

where q_τ is test score at the τth quantile of the marginal (unconditional) distribution, $F_Y(y)$, and $f_Y(q_\tau)$ is the density at that point. It has been shown that a RIF regression, defined as $E[RIF(Y; q_\tau, F_Y)|X] = m_\tau(X) \approx X'\beta_\tau$, leads to a consistent estimate of the unconditional quantile treatment effect.[9]

As illustrated in Figure 2 below, when applied to our analysis, QDiD compares students at the same quantile across states and years, and the RIF method compares students with the same test score and hence located at potentially different quantiles of the distributions across states and years. More specifically, let $F_t(y)$ ($G_t(y)$) represent the outcome distributions for the treated (control) states before ($t = 0$) and after ($t = 1$) the G8 reform. The treatment effect under the QDiD method is given by $(a - e) - (f - g)$, for a given quantile $\tau \in (0, 1)$. In comparison, using the RIF method, for a given test score $q_\tau = y'$, we can measure the changes in *population shares* between the treated and the control states as $-[(a - b) - (c - d)]$. This difference is then divided by a kernel estimation of the density at this point, $f_Y(q_\tau)$, to arrive

[8] Given its flexibility, the RIF method has recently been applied to analyze a range of issues such as cigarette taxes (Maclean et al., 2014) and child care (Havnes and Mogstad, 2015).

[9] See Firpo et al. (2009); Borah and Basu (2013).

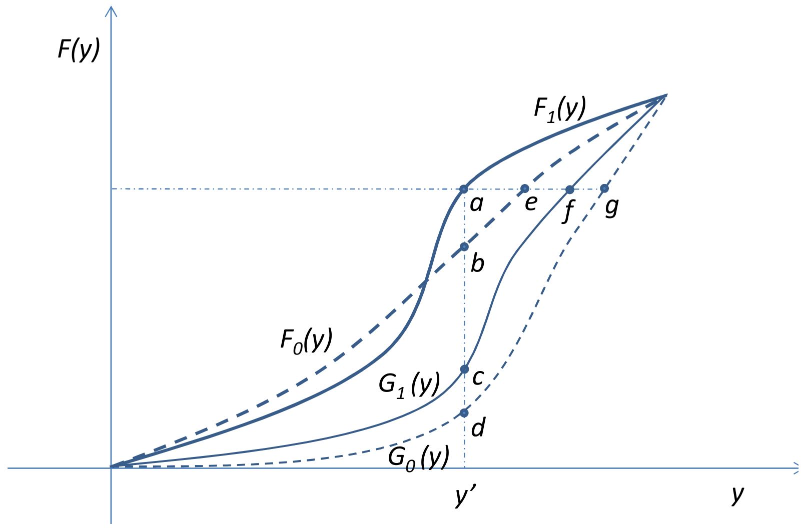

Fig. 2. Illustration of QDiD and RIF-DiD

at the associated quantile treatment effect in test score. We call this the RIF-DiD method, which relies on the assumption that but for the reform, the changes in the population shares would have been the same for treated and control states. This is less restrictive compared to the common distribution assumption necessary for the QDiD method.

The QDiD and RIF-DiD estimates have different interpretations. The QDiD estimates are the conditional quantile treatment effect, namely the reform effect conditioning on the observed variables. This can be easily linked to our theoretical framework, representing the treatment effect of a faster-paced curriculum under the G8 reform, holding everything else constant. However, the conditional quantile treatment effect can be quite sensitive to the variables that it conditions on (Borah and Basu, 2013; Maclean et al., 2014), and heterogeneity in the observed variables implies potentially many different distributions. On the other hand, the RIF-DiD estimates are the unconditional quantile treatment effect, namely the reform effect on a single unconditional test score distribution, even though students do differ along their observed dimensions.

A limitation of both the QDiD and the RIF-DiD method is that, despite the importance of clustering standard errors at the treatment (state) level to avoid overstating precision

(Bertrand et al., 2004) is widely recognized, a statistically valid method to cluster standard errors has not been developed yet. This is further complicated by the sampling weights associated with the observations in the complex survey design. As a result, we can only report the standard error for QDiD assuming i.i.d. residues, while that for RIF-DiD is bootstrapped using 200 repetitions.[10]

5 Data and estimation results

5.1 PISA data

Our empirical analysis is based on a dataset that pools the first five waves of PISA assessment (2000, 2003, 2006, 2009, and 2012) for Germany.[11] PISA tests cover three different subjects (reading, mathematics, and science). Each subject is tested using a broad sample of tasks with differing levels of difficulty.[12] The tests for all three subjects are periodically reviewed and revised to ensure comparability across PISA cycles.

While PISA is conducted by the OECD in a number of countries sampling 15-year-old students regardless of their grades, national grade- and/or age-based extensions of the study were conducted in Germany for all PISA cycles, with the purpose of providing a sample large enough to allow comparisons across the sixteen federal states. Given that the age-based PISA 2009 sample has not been released with state identifiers, our empirical analysis is based on the grade-9 sample of academic-track students.

Besides test scores, we also have student and school characteristics as controls. At the student level, there are demographic and socio-economic variables such as gender, age, immi-

[10] We implement the RIF-DiD estimation procedure using the STATA ado file `rifreg` – downloaded from http://faculty.arts.ubc.ca/nfortin/datahead.html (last accessed December, 2015). The RIF is computed using a Gaussian kernel with an optimal bandwidth.

[11] Baumert (2009); Prenzel (2007, 2010); Klieme (2013); Prenzel (2015)

[12] Using item response theory, PISA maps student performance in each subject on a scale with an international mean of 500 and a standard deviation of 100 across the OECD countries. The scores are averages of plausible values, which are drawn from a distribution of values that a student with the given amount of correct answers could achieve as a test score (OECD, 2012).

gration status, parental highest educational level (ISCED), parental socio-economic status (ISEI), etc. At the school level, we have variables on the total number of enrolled students, the percentage of female students, student-teacher ratio, student-computer ratio, etc. The full set of variables and their summary statistics are reported in Table 1.

5.2 The G8 reform

In Germany, education policies fall within the jurisdiction of the sixteen federal states. Children typically enroll in primary school for four years, and at the beginning of grade 5, they are tracked into three types of secondary schools.[13] The basic track and the middle track both provide vocational oriented schooling through grade 9 or 10. The academic track (*Gymnasium*) leads to university entrance qualification called "*Abitur*".

Before the German reunification, West German states required nine years of schooling in the academic track, while East German states required eight years. After the reunification in 1990, eastern states switched to the nine-year system except for two states Saxony and Thuringia, which held on to the eight-year system.[14] From 2001 to 2008, the fourteen nine-year states began to implement what we call the G8 reform. This reform shortens the length of Gymnasium by one year, but holds the total amount of academic content for graduation fixed. Thus, academic content and the corresponding lecture hours are reallocated from nine to eight grades, introducing a faster-paced curriculum for academic-track students. For our analysis, this curriculum change is arguably exogenous because the G8 reform was mostly driven by considerations of labor market conditions and demographic changes, instead of concerns for student learning outcomes.[15]

[13] Two states, Berlin and Brandenburg, have tracking start in grade 7.

[14] For our sample period (2000–2012), Saxony and Thuringia are thus always "treated" with the G8 status.

[15] For example, in earlier policy discussions, then-federal secretary of education Jürgen Mölleman strongly argued for the reform because "[German] graduates are two to three years older than their peers against whom they compete for jobs in the European labor market. ... German pension systems and demographics (characterized by a significant fraction of senior, retired citizens) cannot support such a late start of employment by young adults. ... Students reach the age of majority at 18 and should have completed secondary schooling by then. (Translation by author)" (Wiater, 1996). When the reform was actually implemented, its was implemented for similar reasons: "As mentioned earlier, reducing the number of years of education is one

Except for a few states, the G8 reform is typically implemented on the cohort of students just entering the academic track. Since in most states students make their tracking choices at the beginning of grade 5, the first treated cohort reaches the end of grade 9 (if the PISA test is administered in that year) after a five-year lag. For the two states Berlin and Brandenburg, students make their tracking choices at the beginning of grade 7, so the first treated cohort reaches the end of grade 9 after a three-year lag. The reform status is slightly more complicated for two states Bavaria and Lower Saxony, which implemented the G8 reform in 2004 on their grade-5 and grade-6 students. However, given the triennial nature of the PISA data, the first treated cohorts in both states (who were in grade 5 in 2004) reach grade 9 in 2009, so these two states are treated in the PISA data from 2009 onward. Finally, Saxony-Anhalt implemented the G8 reform in 2003 on its grades 5–9 students, Mecklenburg-Vorpommern implemented the G8 reform in 2004 on its grades 5–9 students, and Hessen implemented the G8 reform on its grade-5 students for 10% of the schools in 2004, 60% of the schools in 2005, and the remaining 30% of the schools in 2006. Thus Saxony-Anhalt and Mecklenburg-Vorpommern are designated as treated (albeit partially) in the PISA data from 2006 onward, and Hessen is designated as treated in the PISA data from 2012 onward (since only 10% of the schools have been treated in 2009). Table 2 summarizes the timing of the G8 reform as well as the treatment status of student cohorts in the PISA data.

5.3 Estimation results

Table 3 reports the Q-DiD results at all deciles of the distribution. Panel A reports the reform effects on reading test scores for both the baseline and the main specifications. Here we see that the effects are not uniform across students located at different deciles. Conditional

of several measures aimed at lowering the age at which academically qualified workers enter the labor force, which is regarded as too high when compared internationally and, in light of the rising demand for highly educated workers in a globalizing world, is expected to result in a competitive disadvantage for German university graduates, and hence for Germany itself. ... In order to protect social insurance systems, the palpable aging of the population, coupled with the simultaneous decline in births and population, necessitates an earlier entry of young adults into a longer phase of gainful employment. (Translation by author)" See Kühn et al. (2013) for more discussions.

on state and year fixed effects, we find that the G8 reform is insignificant at the first two deciles, and becomes significant at the 5% level or above from the third decile upward. Furthermore, the point estimates range from 0.045 to 0.115, exhibiting an overall increasing pattern. This pattern is broadly consistent with our theory that better-prepared students benefit more from a faster-paced curriculum, in that the reform effect increases as we move up the deciles of the distribution. Adding student and school controls does not change the results qualitatively. Similar patterns also show up in mathematics (panel B) and science (panel C). In these subjects, the reform effects seem to flatten out above the sixth decile, but they are nonetheless larger than those at the lower deciles.

Table 4 reports the RIF-DiD results at all deciles of the distribution. Again panel A uses reading test scores as the outcome variable. Here the RIF-DiD estimates exhibit a pattern similar to those from QDiD estimation, and adding student and school controls does not affect the results qualitatively.

However, when mathematics test scores are considered (panel B), the pattern changes. The RIF-DiD estimates are statistically insignificant at both the low end and the high end of the distribution, but significant in the middle. Furthermore, the point estimates appear to exhibit a hump shape, first increasing and then decreasing. As discussed before, RIF-DiD gives us the unconditional treatment effect, which captures both the within-group difference and between-group difference, where groups are defined by their heterogeneity in the observed control variables. Thus, similar to Firpo et al. (2009),[16] what we find is that while the conditional treatment effect (given by QDiD estimates) in mathematics is generally increasing as we move up the deciles, the unconditional treatment effect (given by RIF-DiD estimates) exhibits a non-monotonic relationship. In our case, the faster-paced curriculum after the G8 reform widens the within-group performance gap across students depending on their levels

[16] As a comparison between the quantile regression and RIF-regression, Firpo et al. (2009) examine the union status on male wages. What they find is that while the union status compresses within-group wage inequality among unionized workers, it increases between-group wage inequality of unionized workers relative to non-unionized workers. As a result, the conditional treatment effect (given by the quantile regression) is monotonically decreasing, while the unconditional treatment effect (given by RIF regression) exhibits an inverted-U shape.

of preparedness, holding control variables constant. At the same time, it also affects the between-group performance gap for students with observed heterogeneity in treated states relative to those in control states. The composition of within-group and between-group effects then give rise to the non-monotonic pattern, as seen here in mathematics.

In panel C, the RIF-DiD result using science test scores is insignificant at the first decile and becomes significant from the second decile onward. The point estimates exhibit a mild increase in the first three deciles, and then remain essentially flat. Across the three subject, it appears that while the within-group effect of the faster-paced curriculum leads to similar increases in the performance gap from lower to upper deciles, the between-group effect is most notable in mathematics, less so in science, and minimal in reading.

6 Conclusion

The horizontal feature of the curricular pace is an important determinant of student learning outcome, yet so far it has been largely overlooked in the literature. This paper is our first step toward understanding this issue. We develop a theory of education curricula as horizontally differentiated by their paces, and empirically test the model prediction, using the quasi-natural experiment of the G8 reform for identification. The evidence we find, namely heterogeneous reform effects depending on students initial preparation, is broadly consistent with our theory. While the average effect of the faster-paced curriculum after the G8 reform is an increase in student test scores, such a benefit is much more pronounced for well-prepared students. In contrast, less-prepared students do not seem to benefit at all, resulting in a widening performance gap.

Such potential mean-variance trade-off in the reform effects may have important policy implications. When policy makers evaluate the reform impact on students' academic achievement, a dimension outside the original policy target, whether they find the results satisfactory or not will depend on the relative weights they assign to different groups of students. Some may decide that the average improvement in test scores outweighs the harm of some students falling further behind, but others may hold opposite opinions. Our model provides a context where such mean-variance trade-off can arise due to a change in the curricular pace, a feature of horizontal differentiation.

Our model can be extended in different directions. For example, currently for tractability, we assume that the pace of curriculum (a horizontal feature) and other measures of schooling quality (vertical features) are additively separable, i.e., there is no interaction between them. In practice, such interaction can play an important role in determining student outcomes, and the joint impact may be quite different from the individual impact for each of the measures respectively. Next, our current analysis assumes a constant level of student effort, which again can change depending on a student's objective. For example, when a well-prepared student faces a faster-paced curriculum and hence improved match quality, she may increase her study effort if effort and match quality are complements, or decrease her study effort if they are substitutes. Endogenizing students' effort choices may either magnify or mitigate the distributional effects of a curriculum change, compared to what we have characterized here. Finally, our empirical analysis relies on an exogenous change in the curriculum for identification, and the change occurs at an aggregate level. With better information on student preparedness, e.g., panel data that track student performance in a series of education stages, we can extend our framework to endogenous curriculum choices, where the choices occur at a disaggregate level such as schools or even classes. This will enable us to recover educators' preference parameters, i.e., the weights they assign to different students, when they adjust the curriculum to meet students' learning needs. Such preference parameters will be useful to help design mechanisms to align educators' incentives with policy makers' targets.

References

Aaronson, D., L. Barrow, and W. Sander (2007) 'Teachers and student achievement in the chicago public high schools.' *Journal of Labor Economics* 25(1), 95–135

Andrietti, Vincenzo (2015) 'The causal effects of increased learning intensity on student achievement: Evidence from a natural experiment.' UC3M WP Economic Series 15-06, Universidad Carlos III de Madrid, June

Andrietti, Vincenzo, and Xuejuan Su (forthcoming) 'The impact of schooling intensity on student learning: evidence from a quasi-experiment.' *Education Finance and Policy*

Angrist, Joshua D., and Victor Lavy (1999) 'Using maimonides' rule to estimate the effect of class size on children's academic achievement.' *The Quarterly Journal of Economics* 114(2), 533–575

Arcidiacono, P., and M. Lovenheim (2016) 'Affirmative action and the quality-fit trade-off.' *Journal of Economic Literature* 54, 3–51

Arcidiacono, P., Esteban M. Aucejo, and V. Joseph Hotz (2016) 'University differences in the graduation of minorities in stem fields: Evidence from california.' *The American Economic Review* 3(106), 525–562

Arcidiacono, Peter, and Sean Nicholson (2005) 'Peer effects in medical school.' *Journal of Public Economics* 89(2-3), 327–350

Baumert, J. (2009) *Programme for International Sudent Assessment 2000 (PISA 2000). Version: 1* (IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2000_v1)

Ben-Porath, Yoram (1967) 'The production of human capital and the life cycle of earnings.' *Journal of Political Economy* 75(1), 352–365

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) 'How much should we trust differences-in-differences estimates?' *The Quarterly Journal of Economics* 119(1), 249–275

Blankenau, William (2005) 'Public schooling, college subsidies and growth.' *Journal of Economic Dynamics and Control* 29(3), 487–507

Blankenau, William, Steven P. Cassou, and Beth F. Ingram (2007) 'Allocating government education expenditures across k-12 and college education.' *Economic Theory* 31(1), 85–112

Borah, Bijan J., and Anirban Basu (2013) 'Highlighting differences between conditional and unconditional quantile regression approaches through an application to assess medical adherence.' *Health Economics* 22(9), 1052–1070

Büttner, Bettina, and Stephan L. Thomsen (2015) 'Are we spending too many years in school? causal evidence of the impact of shortening secondary school duration.' *German Economic Review* 16(1), 65–86

Card, David, and Alan B. Krueger (1992) ‘Does school quality matter? returns to education and the characteristics of public schools in the united states.’ *Journal of Political Economy* 107(1), 151–200

Carrell, Scott E., Richard L. Fullerton, and James E. West (2009) ‘Does your cohort matter? measuring peer effects in college achievement.’ *Journal of Labor Economics* 27(3), 439–464

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor (2006) ‘Teacher-student matching and the assessment of teacher effectiveness.’ *Journal of Human Resources* 41(4), 778–820

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

Currie, Janet, and Thomas Dee (2000) ‘School quality and the longer-term effects of head start.’ *Journal of Human Resources* 35(4), 755–774

Dahmann, Sarah (2017) ‘How does education improve cognitive skills? instructional time versus timing of instruction.’ *Labour Economics* 47, 35–47. EALE conference issue 2016

Ding, Weili, and Steven F. Lehrer (2010) ‘Estimating treatment effects from contaminated multiperiod education experiments: the dynamic impacts of class size reductions.’ *The Review of Economics and Statistics* 92(1), 31–42

Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011) ‘Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya.’ *The American Economic Review* 101(5), 1739–1774

Epple, Dennis, and Richard E. Romano (1998) ‘Competition between private and public schools, vouchers, and peer-group effects.’ *The American Economic Review* 88(1), 33–62

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux (2009) ‘Unconditional quantile regressions.’ *Econometrica* 77(3), 953–973

Gilpin, Gregory, and Michael Kaganovich (2012) ‘The quantity and quality of teachers: Dynamics of the trade-off.’ *Journal of Public Economics* 96(3-4), 417–429

Hanushek, Eric A. (2006) ‘School resources.’ In *Handbookof the Economics of Education (Vol. 2),* ed. E. A. Hanushek and F. Welch (Amsterdam: Elsevier)

Havnes, Tarjei, and Magne Mogstad (2015) ‘Is universal child care leveling the playing field?’ *Journal of Public Economics* 127, 100–114

Homuth, Christoph (2017) *Die G8-Reform in Deutschland: Auswirkungen auf Schülerleistungen und Bildungsungleichheit.* (Springer VS)

Hoxby, Caroline (2000) ‘The effects of class size on student achievement: New evidence from population variation.’ *The Quarterly Journal of Economics* 115(4), 1239–1285

Huebener, Mathias, and Jan Marcus (2017) ‘Moving up a gear: the impact of compressing instructional time into fewer years of schooling.’ *Economics of Education Review* 58, 1–14

Huebener, Mathias, Susanne Kuger, and Jan Marcus (2017) ‘Increasing instruction hours and the widening gap in student performance.’ *Labour Economics* 47, 15–34. EALE conference issue 2016

Kaganovich, Michael, and Xuejuan Su (forthcoming) ‘College curriculum, diverging selectivity, and enrollment expansion.’ *Economic Theory*

Klieme, E. (2013) *Programme for International Sudent Assessment 2009 (PISA 2009). Version: 1* (IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2009_v1)

Krashinsky, Harry (2014) ‘How would one extra year of high school affect academic performance in university? evidence from an eudcational policy change.’ *Canadian Journal of Economics* 47(70-97), 1

Krueger, Alan B. (2003) ‘Economic considerations and class size.’ *Economic Journal* 113, F34–F63

Kühn, Svenja M., Isabell van Ackeren, Gabriele Bellenberg, Christian Reintjes, and Grit im Brahm (2013) ‘Wie viele schuljahre bis zum abitur? eine multiperspektivische standortbestimmung im kontext der aktuellen schulzeitdebatte.’ *Zeitschrift für Erziehungswissenschaft* 16, 115–136

Light, Audrey, and Wayne Strayer (2000) ‘Determinants of college completion: School quality or student ability?’ *Journal of Human Resources* 35(2), 299–332

Lucas, Robert (1988) ‘On the mechanics of economic development.’ *Journal of Monetary Economics* 22(1), 3–42

Maclean, Johanna Catherine, Douglas A. Webber, and Joachim Marti (2014) ‘An application of unconditional quantile regression to cigarette taxes.’ *Journal of Policy Analysis and Management* 33(1), 188–210

Meyer, Tobias, and Stephan L. Thomsen (2016) ‘How important is secondary school duration for post-school education decisions? evidence from a natural experiment.’ *Journal of Human Capital* 10(1), 67–108

Meyer, Tobias, and Stephan L. Thomsen (2017) ‘The role of high-school duration for university students’ motivation, abilities and achievements.’ *Education Economics* pp. 1–22

Meyer, Tobias, Stephan L. Thomsen, and Heidrun Schneider (forthcoming) ‘New evidence on the effects of the shortened school duration in the german states: An evaluation of post-secondary education decisions.’ *German Economic Review*

Morin, Louis-Philippe (2013) ‘Estimating the benefit of high school for university-bound students: evidence of subject-specific human capital accumulation.’ *Canadian Journal of Economics* 46(441-468), 2

Mueller, Steffen (2013) 'Teacher experience and the class size effect: Experimental evidence.' *Journal of Public Economics* 98, 44–52

OECD (2012) *PISA 2009 technical report* (OECD Publishing)

Prenzel, M. (2007) *Programme for International Sudent Assessment 2003 (PISA 2003). Version: 1* (IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2003_v1)

— (2010) *Programme for International Sudent Assessment 2006 (PISA 2006). Version: 1* (IQB - Institut zur Qualitätentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2006_v1)

— (2015) *Programme for International Sudent Assessment 2012 (PISA 2012). Version: 1* (IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2012_v1)

Rothstein, Jesse (2010) 'Teacher quality in educational production: tracking, decay, and student achievement.' *The Quarterly Journal of Economics* 125(1), 175–214

Sacerdote, Bruce (2001) 'Peer effects with random assignment: Results for dartmouth roommates.' *The Quarterly Journal of Economics* 116(2), 681–704

Su, Xuejuan (2004) 'The allocation of public funds in a hierarchical educational system.' *Journal of Economic Dynamics and Control* 28(12), 2485–2510

— (2006) 'Endogenous determination of public budget allocation across education stages.' *Journal of Development Economics* 81(2), 438–456

Thiel, Hendrik, Stephan L. Thomsen, and Bettina Büttner (2014) 'Variation of learning intensity in late adolescence and the effect on personality traits.' *Journal of the Royal Statistical Society: Series A* 177(4), 861–892

Tirole, Jean (1988) *The Theory of Industrial Organization* (Caambridge: MIT Press)

Wiater, W. (1996) 'Zwölf jahre bis zum abitur? positionen im streit um die verkï£¡rzung der gymnasialen schulzeit.' In *Schulreform in der Mitte der 90er Jahre: Strukturwandel und Debatten um die Entwicklung des Schulsystems in Ost- und Westdeutschland,* ed. W. Melzer and K.-J. Tillmann (Opladen: Leske + Budrich) pp. 121–139

Zimmerman, David J. (2003) 'Peer effects in academic outcomes: Evidence from a natural experiment.' *The Review of Economics and Statistics* 85(1), 9–23

Table 1. PISA Summary statistics

Variable	Mean	SD
PISA scores		
Reading	572.13	55.51
Mathematics	578.39	58.26
Science	587.05	61.10
Student controls:		
Female	0.53	0.50
Age (in months)	185.22	5.54
Parents' education: Upper secondary	0.18	0.39
Parents' education: Tertiary	0.62	0.49
Parents' ISEI	59.25	17.34
Books in house: > 100	0.58	0.49
Only child	0.29	0.45
Kid born in foreign country	0.04	0.20
Parents born abroad	0.13	0.34
Foreign language spoken at home	0.04	0.20
School controls:		
School enrollment	793.93	352.15
% of girls enrolled	49.42	15.07
Student-teacher ratio	14.66	5.93
Student-computer ratio	26.78	62.84
Urban school	0.26	0.44
Private school	0.08	0.26
Observations	33,996	

Notes: The sample includes academic-track ninth graders in PISA 2000–2012 with valid reading scores.

Table 2. G8 reform timing and treatment status in PISA data

State	Tracking begins in grade	G8 reform since	Grade(s) affected	PISA cohorts 2000	2003	2006	2009	2012
Saxony (SN)	5	before 1990	5	T	T	T	T	T
Thuringia (TH)	5	before 1990	5	T	T	T	T	T
Saarland (SL)	5	2001	5	C	C	T	T	T
Hamburg (HH)	5	2002	5	C	C	C	T	T
Saxony-Anhalt (ST)	5	2003	5–9	C	C	T*	T	T
Baden-Württemberg (BW)	5	2004	5	C	C	C	T	T
Bremen (HB)	5	2004	5	C	C	C	T	T
Bavaria (BY)	5	2004	5–6	C	C	C	T	T
Lower Saxony (NI)	5	2004	5–6	C	C	C	T	T
Mecklenburg-Vorpommern (MV)	5	2004	5–9	C	C	T**	T	T
Hessen (HE)	5	2004–2006	5	C	C	C	C***	T
Berlin (BE)	7	2006	7	C	C	C	T	T
Brandenburg (BB)	7	2006	7	C	C	C	T	T
North Rhine-Westfalia (NW)	5	2005	5	C	C	C	C	T
Rhineland-Palatinate (RP)	5	2008	5	C	C	C	C	C
Schleswig-Holstein (SH)	5	2008	5	C	C	C	C	C

Notes: Treatment status: C for control and T for treated. * In Saxony-Anhalt the G8 reform was implemented on grades 5–9 students in 2003. The 2006 PISA cohort, designated as treated, receives partial treatment since grade 7. ** In Mecklenburg-Vorpommern the G8 reform was implemented on grades 5–9 students in 2004. The 2006 PISA cohort, designated as treated, receives partial treatment since grade 8. *** In Hessen the G8 reform was introduced in three waves on 10%, 60%, and the remaining 30% of schools in 2004, 2005, and 2006 respectively. The 2009 PISA cohort, designated as control, actually includes 10% of treated schools. *Source*: Kulturministerkonferenz (KMK).

Table 3. G8 distributional effects: QDiD

	Quantiles								
	0.10	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
Panel A: Reading ($N = 33,996$)									
Baseline									
G8	0.049	0.064*	0.071**	0.066**	0.078***	0.072**	0.080**	0.096***	0.105***
	(0.049)	(0.038)	(0.033)	(0.033)	(0.029)	(0.028)	(0.032)	(0.030)	(0.039)
Main									
G8	0.044	0.055*	0.064**	0.089***	0.098***	0.095***	0.103***	0.111***	0.108**
	(0.039)	(0.031)	(0.029)	(0.029)	(0.023)	(0.023)	(0.027)	(0.033)	(0.044)
Panel B: Math ($N = 30,205$)									
Baseline									
G8	0.064	0.049	0.059*	0.074***	0.098***	0.104***	0.104***	0.104***	0.090*
	(0.052)	(0.046)	(0.035)	(0.027)	(0.034)	(0.029)	(0.033)	(0.032)	(0.049)
Main									
G8	0.059	0.049	0.056*	0.076***	0.093***	0.113***	0.108***	0.121***	0.101**
	(0.045)	(0.040)	(0.033)	(0.026)	(0.028)	(0.028)	(0.031)	(0.030)	(0.050)
Panel C: Science ($N = 30,202$)									
Baseline									
G8	0.088	0.082**	0.096***	0.078***	0.082***	0.107***	0.092***	0.098***	0.103**
	(0.054)	(0.041)	(0.033)	(0.030)	(0.029)	(0.027)	(0.032)	(0.035)	(0.043)
Main									
G8	0.070	0.078*	0.073***	0.087***	0.096***	0.107***	0.118***	0.115***	0.110**
	(0.059)	(0.041)	(0.026)	(0.026)	(0.030)	(0.024)	(0.025)	(0.036)	(0.052)

Notes: Final student weights are used in estimation. Conventional standard errors are reported in parentheses. $***$, $**$, and $*$ indicate significance at 1, 5, and 10 percent levels respectively. The samples in panel A, B, and C include academic-track ninth graders in PISA 2000–2012 with valid reading, math, and science scores, respectively.

Table 4. G8 distributional effects: RIF-DiD

	Quantiles								
	0.10	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
Panel A: Reading ($N = 33,996$)									
Baseline									
G8	0.063	0.069*	0.072*	0.073**	0.077***	0.075***	0.090***	0.089***	0.094**
	(0.056)	(0.041)	(0.039)	(0.031)	(0.030)	(0.026)	(0.028)	(0.029)	(0.039)
Main									
G8	0.071	0.079**	0.080**	0.081***	0.084***	0.084***	0.101***	0.098***	0.107***
	(0.059)	(0.039)	(0.038)	(0.031)	(0.029)	(0.025)	(0.029)	(0.030)	(0.041)
Panel B: Math ($N = 30,205$)									
Baseline									
G8	0.074	0.067*	0.086***	0.096***	0.106***	0.104***	0.092***	0.069*	0.056
	(0.048)	(0.039)	(0.033)	(0.030)	(0.031)	(0.033)	(0.031)	(0.041)	(0.048)
Main									
G8	0.073	0.071*	0.094***	0.104***	0.114***	0.113***	0.101***	0.077*	0.065
	(0.044)	(0.037)	(0.033)	(0.029)	(0.031)	(0.032)	(0.030)	(0.042)	(0.049)
Panel C: Science ($N = 30,202$)									
Baseline									
G8	0.088	0.082*	0.103***	0.091***	0.101***	0.096***	0.094***	0.086***	0.101**
	(0.067)	(0.047)	(0.032)	(0.031)	(0.029)	(0.031)	(0.029)	(0.032)	(0.046)
Main									
G8	0.091	0.087*	0.111***	0.100***	0.111***	0.107***	0.104***	0.095***	0.111**
	(0.063)	(0.046)	(0.032)	(0.030)	(0.028)	(0.029)	(0.029)	(0.033)	(0.046)

Notes: Final student weights are used in estimation. Standard errors reported in parentheses are based on 200 bootstrap replications. $***$, $**$, and $*$ indicate significance at 1, 5, and 10 percent levels respectively. The samples in panel A, B, and C include academic-track ninth graders in PISA 2000–2012 with valid reading, math, and science scores, respectively.