Public Disclosure Authorized

Public Disclosure Authorized

WPS5465

Policy Research Working Paper 5465

Short-run Learning Dynamics under a Test-based Accountability System

Evidence from Pakistan

Felipe Barrera-Osorio
Dhushyanth Raju

The World Bank
South Asia Region
Education Unit
&
Human Development Network
Education Unit
November 2010

Policy Research Working Paper 5465

Abstract

Low student learning is a common finding in much of the developing world. This paper uses a relatively unique dataset of five semiannual rounds of standardized test data to characterize and explain the short-term changes in student learning. The data are collected as part of the quality assurance system for a public-private partnership program that offers public subsidies conditional on minimum learning levels to low-cost private schools in Pakistan. Apart from a large positive distributional shift in learning between the first two test rounds, the learning distributions over test rounds show little progress. Schools are ejected from the program if they fail to achieve a minimum pass rate in the test in two consecutive attempts, making the test high stakes. Sharp regression discontinuity estimates show that the threat of program exit on schools that barely failed the test for the first time induces large learning gains. The large change in learning between the first two test rounds is likely attributable to this accountability pressure given that a large share of new program entrants failed in the first test round. Schools also qualify for substantial annual teacher bonuses if they achieve a minimum score in a composite measure of student test participation and mean test score. Sharp regression discontinuity estimates do not show that the prospect of future teacher bonus rewards induces learning gains for schools that barely did not qualify for the bonus.

This paper—a product of the Education Unit, South Asia Region; and the Education Unit, Human Development Network—is part of a larger effort of the departments to rigorously evaluate innovative government programs supported by World Bank lending operations. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at fbarrera@worldbank.org and draju2@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Produced by the Research Support Team

Short-run learning dynamics under a test-based accountability system: Evidence from Pakistan*

Felipe Barrera-Osorio†
Dhushyanth Raju†

JEL classification codes: I21; I28
Keywords: education; Pakistan; private schools; subsidies; learning; test accountability; teacher incentives; regression-discontinuity design.

* We thank the Punjab Education Foundation, in particular, Mian Kashif Ijaz and Huma Rizvi, for extensive discussions on program and test design and implementation and assistance with the program administrative data. We also thank Amit Dar, Sofia Shakil, Huma Waheed, and the Government of Punjab School Education Department/Project Management and Implementation Unit for their encouragement and support of the research project. The findings, interpretations, and conclusions expressed herein are our own and do not necessarily represent the views of the World Bank, its Board of Directors, or any of its member countries. All remaining errors are our own.
† World Bank, Washington, DC. Email addresses: fbarrera@worldbank.org; draju2@worldbank.org.

1. Introduction

Low student learning is an important feature in Pakistan, as it is in much of the developing world (Andrabi et al. 2007; Glewwe and Kremer 2006; Hanushek and Wößmann 2007). The shortfalls relative to national curricular standards are exacerbated by the fact that a large share of children, particularly rural, female, and poor children, obtains little or no formal schooling in the country, likely seriously impairing their long-term socioeconomic prospects. In addition, in general, the factors and processes that produce real and sustained gains in student learning are not as well understood as the factors and processes that produce real and sustained gains in school participation and attainment. Evidence also shows that the factors and processes behind learning can differ from the factors and processes behind participation (Andrabi et al. 2007). This background makes the search for effective policy solutions to improve learning outcomes pressing yet tenuous.

In this study, we use five semiannual rounds of standardized test data to characterize and explain the evolution of student learning over a period which spans three academic years (2007/08-2009/10). The tests were conducted in low-cost private schools supported by public cash subsidies under an innovative test-based accountability program in the province of Punjab, Pakistan administered by the Punjab Education Foundation (PEF), called the Foundation Assisted Schools (FAS) program. The data are repeated cross-sections at the student level and longitudinal at the school level. Such data are relatively rare for a developing country. Additionally, the data are unusual for a low-income setting given that they are generated from what appears to be a well-designed and robust quality assurance testing system integrated into the administration of the program, providing a high degree of confidence that the test scores can be interpreted as sound estimates of underlying learning.

Initiated in 2005, the FAS program provides conditional cash subsidies to low-cost private schools with the objective of offering private schooling opportunities for children from low-income households and raising the level of learning in low-cost private schools. The cash subsidies are provided monthly and on a per-student basis, with essentially no conditions on how the subsidy is to be used by the program school. The subsidy amount is purposely set low to

ensure that only low-cost private schools self-select to participate in the program; the original amount was also purposely set at half the estimated per-student cost in the public school system.[1]

In return for receiving the subsidy benefit, the program school has to, among other things, waive tuition and fees for all students and ensure that the school achieves a minimum student pass rate in the Quality Assurance Test (QAT). Program schools that satisfy the above conditions are also eligible for other substantial cash benefits offered on an annual basis: group-based bonuses for teachers in schools that achieve high QAT pass rates/mean scores and competitive bonuses for schools that rank highest in the QAT in each main program district.

Each of the above incentive features (the structure of the benefits as well as the benefit eligibility rules) represents key program innovations in the education field that may induce participation, equity, and learning gains.[2] We are not aware of any other program in a developing country which combines these innovations into a single intervention. As of June 2010, the FAS program has proceeded through six phases of expansion and supports 798,000 students in 1,779 schools in 29 of the 36 districts in the province, making it one of the largest PPP initiatives in education in the developing world.

The learning data used in the study come from the QAT. The QAT is a curriculum-based, multi-subject, written test offered biannually in program schools. It is designed by professional subject specialists in PEF, and, starting with QAT 4, the administration of the test was competitively outsourced to independent testing agencies. The testing agencies are expected to adhere to the testing protocols set by PEF, with testing agency compliance monitoring by PEF at all program schools. Importantly, the testing agencies are instructed by PEF to follow pre-established procedures to control for certain types of potential strategic responses of schools (i.e., cheating in a broad sense) that may artificially raise test scores.

To date, nine rounds of the QAT (QAT 1-QAT 9) have been administered, of which data from five rounds, QAT 4-QAT 8, administered between November 2007 and November 2009, are used in this study. QAT 4 is the first round of test data used in the study due to sample size considerations: it is the first QAT administered after the first major expansion of the FAS program.

[1] Setting the subsidy level in such a way was considered important by PEF for securing political buy-in for the use of public funds for the initiative.

[2] The typical subsidy program is designed to finance stipulated education inputs (Gauri and Vawda 2003).

We examine four questions. The first three questions are descriptive in nature. First, we explore how student learning at the school level has evolved in program schools over QAT rounds by examining summary statistics of the mean QAT score distributions and stochastic dominance relations between pairs of successive mean QAT score distributions. Second, we estimate the share of student-level QAT score variation attributable to different levels (district, school, grade, and student) and how the shares and levels of QAT score variation have evolved over QAT rounds. Third, we estimate the conditional mean relationship at the school level between basic pre-program school characteristics and the evolution of mean QAT scores over QAT rounds, controlling for school-level heterogeneity using alternative panel regression methods.

The fourth question is causal in nature. We estimate the causal effects at the school level of specific program benefit rules on student learning in program schools. The first subquestion we ask is whether high-powered stick incentives can induce learning gains in program schools. Program schools that fail to achieve the minimum pass rate in a given QAT are offered a second opportunity in the following QAT. If the program school fails to achieve the minimum pass rate in two consecutive attempts, the school is permanently disqualified. This creates a high-stakes situation for first-time failers to rapidly boost their learning performance in order to avoid program disqualification. Applying a sharp regression discontinuity (RD) design to the data, we examine whether accountability pressure induces gains in mean scores in the ultimatum QAT among marginal first-time failers. We also examine whether any gains found persist over post-ultimatum QAT rounds. Such persistence is at times interpreted in the literature as indicating that gaming behavior is unlikely to be an explanatory factor though it cannot be ruled out.

Several studies have examined the effects of (the threat of) sanctions and assistance on low-performing schools on performance in subsequent test rounds (see, e.g., Figlio and Rouse 2006; Chiang 2009; Rockoff and Turner 2008; West and Peterson 2006; and Rouse et al. 2007). Much of this research has focused on state public education systems in the United States where schools are rated on a letter grade scale from A to F based on summarizing multiple performance measures, including student performance on standardized tests, and where F-rated schools are

subjected to (the threat of) sanctions.[3] Typically applying a sharp regression discontinuity design to the data, these studies consistently find short-run gains in test scores among (students in) marginal F-rated schools in the test which is included in the determination of the rating, with the estimated impacts ranging between 0.05−0.15 standard deviations. Some of the studies additionally find that (i) test score gains persist in the medium run among (students from) marginal F-rated schools, (ii) the estimated effects are unlikely to be driven by test-taking pool selection; and (iii) "teaching to the test" is likely to be an important driver as inferred from the small or lack of effects in standardized tests that are not used in the determination of the rating.

The second subquestion we ask is whether carrot incentives can induce learning gains in program schools. Program schools qualify for substantial group-based teacher bonuses if they de facto achieve a minimum score of 100 in a composite measure of the number of students that took the QAT and the mean percentage score in the QAT. Also applying a sharp RD design to the data, we examine whether bonus nonqualification induces gains in mean scores among marginal nonqualifiers in the QAT round following the distribution of the teacher bonuses. We also examine the persistence of the carrot effect over subsequent QAT rounds. Though the carrot incentive question posed here is not perfectly correspondent (as the focus here is on the effects on nonbeneficiaries as opposed to beneficiaries), it broadly fits into the literature on the causal effects of teacher incentive schemes which directly tie teacher compensation to student learning levels and changes (Glewwe et al. 2009). To date, the limited but growing evidence is inconclusive with, for example, some studies finding positive effects on student learning (Lavy 2002, 2009; Muralidharan and Sundararaman 2009) and others finding positive effects on student learning which disappear post-intervention, suggesting temporary artificial gains in learning (Glewwe et al. 2003).

The remainder of the paper is organized as follows. Section 2 describes the education context in Pakistan at the time that the FAS program was introduced. It also describes in detail the main design and implementation features of the program. Section 3 describes in detail the design and administration features of the Quality Assurance Test as well as the data used in the study. Section 4 discusses the evolution of the school-level mean QAT scores in core subjects over QAT rounds. Section 5 discusses the student-level QAT score variance decomposition

[3] Sanctions for F-rated schools include a combination of stigmatization from publicly disclosing ratings, higher scrutiny and oversight by school system administrations, permission/private school vouchers for students to shift out of F-rated schools, changing school management and staff, and school closure.

analysis and findings. Section 6 discusses the school-level QAT score regression analysis and findings. Section 7 discusses the regression-discontinuity analysis and findings on the threat effect on mean QAT scores arising from QAT failure on marginal first-time failers and the carrot effect on mean QAT scores and test participation arising from missing teacher bonus qualification on marginal bonus nonqualifiers. Section 8 summarizes the main findings and provides some concluding remarks.

2. Context and program features

The Foundation Assisted Schools (FAS) program was introduced into an education landscape characterized by three defining features. One: equitable access to schooling and attainment and achievement are acute, persistent challenges. Estimates using household sample survey data for Punjab from 2004/05, the year prior to the launch of the FAS program, indicate that the school participation rate (grade 1+) of children ages 6–15 years was 65.7%—this share drops to 61.2%, 60.9%, and 48.7% when the sample is restricted to girls, children from rural households, and children from the poorest (bottom expenditure quintile) households, respectively. Conditional on any schooling (grade 1+), while 91.8% of individuals ages 17–21 years completed primary school (grade 5), only 40.7% of them completed secondary school (grade 10), with significantly lower shares for rural children and, in particular, poor children. The National Education Assessment System (NEAS) for Pakistan finds that scaled mean scores in mathematics and language assessments for a sample of grade-4 students in public schools in Punjab in 2005 were significantly less than 500, suggesting that, on average, students, obtained less than 50% of the points in the assessments (Government of Pakistan 2006). Das et al. (2006) and Andrabi et al. (2007) also find that the learning levels of grade-3 students in a selected sample of villages in Punjab measured in 2004 were far below curriculum standards in English, Urdu, and mathematics.

Two: the public sector, which is the dominant provider of education, suffers from chronic weaknesses which impair its ability to effectively address the challenges in education access, equity, and quality. To a large extent, public sector performance is hampered by weak accountability and incentive systems (Government of Pakistan 2009; Social Policy and Development Center 2003). This state-of-affairs is exemplified by recent evidence on relative teacher performance between private and public schools in rural Punjab. Andrabi et al. (2007)

find that, while public school teachers tend to be better paid and have higher levels of education, training, and teaching experience than their private school counterparts, the teacher absenteeism rate in public schools is almost double that in private schools (15% vs. 8%). They also find that, while private school teacher salaries are increasing in teacher *and* student test scores and decreasing in teacher absenteeism, public school teacher salaries are largely unresponsive to these variables and are mainly a function of teacher credentials (education, training, and experience). The greater accountability in the private sector is likely due to market competitive forces which the public sector is not subject to, at least directly.

Three: in the wake of the public sector failure to adequately address these issues, the private sector has emerged as a major alternate provider of education, growing dramatically in size and reach. In particular, the rapid growth of a private sector that offers schooling opportunities to low-income households has created a major policy opportunity. Using data from 2000, Andrabi et al. (2006) find that there was an exponential increase in the number of private schools over the 1990s, with over 50% of existing private schools established on or after 1996. They also find that the birth rate of private schools in the recent period is higher in rural areas. These changes in the patterns of growth in private schools are also reflected in changes in the patterns of participation growth among students. Data from 2004/05 show that 18.7% of children ages 6–15 years were enrolled in private schools in Punjab, a 36% increase from 1998/99, with this increase largely attributable to private school participation increases among children from rural and poor households. The demand for private schooling is likely driven by the fact that fees in private schools tend to be low, accounting for a small percentage of mean annual household expenditure (Andrabi et al. 2008). It is also likely driven by the perceived higher quality of private schooling—evidence from rural Punjab shows that the learning levels in private schools are significantly higher than in public schools, even after controlling for a range of village, household, and school characteristics (Andrabi et al. 2007).

The FAS program and PEF are direct outcomes of the government's recognition of the importance of leveraging the potential of the growing low-cost private sector in addressing access, equity, and quality issues in primary and secondary education. PEF is a publicly-funded semi-autonomous statutory organization established in 1991 by the government. It serves as the main institutional conduit for PPP programs in education in Punjab. The organization's primary aims are to provide affordable private school opportunities to socioeconomically-disadvantaged

households and raise the quality of education in low-cost private schools. To these ends, it employs a variety of instruments such as providing vouchers to poor households in disadvantaged urban neighborhoods to attend selected low-cost private schools and supporting low-cost private schools with monthly per-student subsidies conditional on school quality standards, teacher training courses on subject content and pedagogical and classroom management techniques, and the insertion of competitively-hired subject specialist trainers in secondary schools for fixed terms.[4]

The FAS program is PEF's largest program. In the fiscal year 2009−10, PEF spent roughly 2.4 billion rupees (US$28.7 million, in current dollars) on the program—this amount accounts for 93% of total expenditures by the organization in that year. It was initiated in November 2005 on a pilot basis in 54 schools in seven districts in Punjab. Since then, PEF has rapidly expanded program coverage in phases in terms of additional districts as well as more schools within districts. The program has proceeded through six phases (the pilot phase is considered by PEF to be phase 1). Phase-6 program schools entered the program in April 2010. As of June 2010, the program covers 798 thousand students in 1,779 private schools in 29 out of the 36 districts in Punjab. As a result of explicitly targeting the program at high illiteracy districts over phases 3−5, the majority of program students and schools (70% and 68%, respectively) are in seven districts located in the southern part of the province (see Figure 1).

Table 1 provides summary statistics based on current administrative data on program schools separately by phase of entry. The sample consists of 1,776 primary (classes 1–5), middle (classes 1–8), and secondary schools (classes 1–10, 6–10).[5] Looking at the aggregate sample (pooled across phases), the mean program school size is 455 students; in schools with both girls and boys enrollment, the mean ratio is 1 (gender parity). Program schools have on average 17 teachers and 16 classrooms. The mean student-teacher and student classroom ratios are 27:1 and 29:1, respectively. The majority of program schools is middle level (64%), coeducational (97%), rural (59%), and registered (100%). The phase-wise statistics indicate that mean enrollment is positively associated with length of program participation, this relationship also holds for mean boys and girls enrollment. The number of teachers and classrooms are also positively associated

[4] The organization is continuing to expand its set of instruments: for example, it is now considering supporting private organizations/individuals that adopt and manage non-functional public schools handed over by the government for this initiative.

[5] The three higher secondary schools in the program are excluded.

with length of program participation, increasing commensurately with enrollment increases as inferred from the stable student-teacher and student-classroom ratios across phases.

The FAS program offers three types of cash benefits to schools. These benefits were introduced at different points in time. First, the program offers a per-student subsidy, which was introduced at program inception. The monthly per-student subsidy amount was fixed at Rs. 300 (US$3.5) for all students, which was roughly half of the estimated per-student cost in the public primary and secondary education system at the time the program was initiated. Starting in September 2008, the monthly per-student subsidy was raised to Rs. 350 (US$4.1) for students in grades 1–8 and Rs. 400 (US$4.7) for students in grades 9–12. The subsidy benefit is provided to schools on a monthly basis for all twelve months in the year. To facilitate timely and regular payments, starting in August 2007, the subsidy benefit amounts are electronically transferred to the bank accounts of program schools.

Second, first announced in December 2006, every academic year, the program offers cash bonuses to teachers in high-performing schools. Details on the eligibility criteria for the bonuses are provided in Section 7 but a maximum of five teachers in program schools that achieve (in part) a minimum student pass rate of 90% (i.e., at least 90% of tested students obtain a minimum percentage of 40%) in the QAT receive a bonus award of Rs. 10,000 (US$118) each. This is a substantial bonus amount for teachers in program schools: using available data on maximum and minimum monthly teacher salaries from applications to the program by phase-3 and phase-4 program schools, we estimate that the bonus amount represents 29% and 59% of pre-program mean annualized maximum and minimum salaries. This bonus size is approximately an order of magnitude larger than the range of sizes observed in most teacher incentive programs (see Glewwe et al. 2003). In practice, the bonuses are offered to teachers who taught the classes and/or subjects that were tested in the QAT. Similar to the subsidy benefits, the bonuses are transferred electronically to the personal bank accounts of the teacher awardees. To date, four rounds of teacher bonuses have been awarded.

Third, first announced in February 2007, every academic year, the program offers a competitive school bonus to the top-performing school in each major program district (there are seven major program districts). The program school with the highest student pass rate in the QAT is awarded Rs. 50,000 (US$588).[6] This bonus size is relatively modest: given a pre-

[6] Ties between schools in pass rates are broken by looking at mean percentage scores in the QAT.

program mean school size of 252 students in phase-3 and phase-4 schools based on data from program applications and a per-student subsidy of Rs. 300, the bonus amount represents 6% of the expected mean annual subsidy payment to program schools. Again, the bonuses are transferred electronically to the bank accounts of the school awardees. To date, three rounds of school bonuses have been awarded.

Once schools qualify and join the FAS program, the partnership contract stipulates several conditions for maintaining benefit eligibility. The conditions that are stringently applied by PEF are (1) schooling is offered to students without charging them any tuition or fees (and displaying this status prominently on a PEF-issued signboard outside the school gate) and (2) participation of the program school in the QAT and that at least 66.67% of the tested students score 40% or higher on the QAT. A one-time violation of these conditions typically results in a warning and the capping of enrollment figures for the subsidy payment until the next QAT round. A second violation results in the permanent disqualification of the school with immediate effect. With the QAT-related condition, the school is permanently disqualified if it fails to achieve the minimum pass rate in the QAT in two consecutive attempts. Since the start of the program, there have been 51 schools that have been disqualified from the program due to double failure on the QAT.

There are also other conditions for maintaining benefit eligibility. These include (1) registering the school with the District Registration Authority within one year of joining the program; (2) conducting only one class in a classroom in any period; (3) maintaining or upgrading the quality of the school's physical infrastructure (e.g., adequate classroom space, properly-constructed rooms and buildings, sufficient ventilation, and sufficient artificial and natural light); (4) adequate furniture and teaching tools (e.g., benches, desks, and blackboards); (5) monthly reporting on enrollment figures; (6) maximum student-teacher and student-classroom ratios of 35:1; (7) a minimum enrollment of 100 students; and (8) no after-hours classes or tutoring services at the school. These additional conditions are applied more leniently; typically, when PEF detects a violation among this subset of conditions, schools are provided with a warning and a grace period within which to comply. To date, no program schools have been disqualified for repeated violations of these conditions.

3. The Quality Assurance Test

The Quality Assurance Test (QAT) is the backbone of the FAS program, as eligibility for program benefits are directly tied to the performance of the school in this test. The test has several design and administration strengths. In terms of salient design features, the QAT is designed by professional subject specialists at the Academic Development Unit (ADU) in PEF. It is a criterion-referenced test based on learning standards in the national curriculum.[7] It tests five levels of cognitive learning under Bloom's (1956) classification: knowledge, comprehension, application, analysis, and synthesis. The test duration varies between 45–65 minutes, with shorter test durations for lower grades. The QATs designed for students in primary and elementary grades have sections in English, Urdu (the vernacular language in Punjab), mathematics, and general science; those designed for students in secondary grades have sections in English, Urdu, mathematics, and a combination of three of the following five subjects: biology, chemistry, and physics, electrical application, and computer science (the combination of the physical science subjects is the norm).[8] Except in the English and Urdu sections, where the questions are matching, short-answer, or essay type (Urdu section only), questions in the rest of the subjects follow a four-option multiple-choice format. Instructions for the QAT are provided in Urdu; the questions and multiple answer choices in the non-language sections are in both Urdu and English. The QAT is offered twice every academic year, first in November and second in March. The tests are structured identically and are of equal difficulty but more material is covered in the March test, in line with the syllabi in the program schools.[9] To help schools familiarize themselves with the test content and format, PEF periodically shares sample test papers with program schools. PEF has also developed and periodically shares content lists which delineate expected content knowledge, separately by grade and subject.[10]

---

[7] The majority of program schools use textbooks from the Punjab Textbook Board as they are provided free of charge through the program—these textbooks are based on the national curriculum.

[8] Each program school is categorized into a subject group based on information provided prior to the QAT. The subject group determines the subjects tested in classes 9 and 10. The groups are Biology (physics, chemistry, biology, math, English, and Urdu), Electrical Application (physics, chemistry, electrical application, math, English, Urdu), and Computer Science (physics, chemistry, computer science, math, English, Urdu).

[9] The November QAT tests material from the first six months of the academic year. The March QAT tests material from the full academic year.

[10] Sample QAT papers and content lists are available on PEF's website at http://www.pef.edu.pk/downloads-model-papers.html and http://www.pef.edu.pk/downloads-content-list.html, respectively (Last accessed: June 15, 2010). Content lists are also shared with program schools in booklet form. The first time the contents lists were shared was in June 2007, prior to QAT 4.

In terms of salient administration features, the testing is school-based, with students tested in their school during normal school hours. The test papers are written, formatted, sorted, counted, and sealed in envelopes by ADU. PEF also administered the tests at program schools in QAT 1−QAT 3. Starting from QAT 4, the administration of the test was competitively outsourced to independent testing agencies, namely the Board of Intermediate and Secondary Education (BISE), Bahawalpur and the Punjab University Education Testing Service, Lahore.[11] The testing agencies are expected to strictly follow the test administration guidelines prepared by PEF and are responsible for the administration steps starting from securely transporting the test materials to the schools and ending with hand-scoring the test following the provided scoring guides and the delivery of the student- and subject-wise test score database and filled-in and unused question papers, as well as brief test administration reports and test administration checklists for each school, to ADU. PEF is not completely divorced from test administration however—it sends a PEF staff member to be present at each test to monitor test agency compliance with the guidelines; if any problems arise, the staff member is authorized to take corrective steps. The staff member is expected to fill in and submit a monitoring report (accompanied by photographs and video footage [see Figure 2 for examples]) for each program school covered.

Although the QAT is a pre-announced test, it attempts to control for test pool selection by the school at the grade level. Prior to administering the QAT, PEF sends an intimation letter to the school some two-three weeks in advance of the planned test date, with key test planning information. The testing agency also communicates the same information to the school a few days before the planned test date, either by letter or personal visit. Importantly, schools are notified that 100% attendance is required on the test date, with attendance numbers checked against the grade-wise enrollment statement submitted to PEF for the latest subsidy benefit payment. Consistent with the test administration guidelines, in practice, the QAT is administered at the program school if the student attendance rate is at least 80% (PEF reports that test cancellation due to lower-than-required student attendance is a rare event). The number of test papers in each grade-specific packet is determined by ADU based on the latest submitted monthly enrollment statement. If the number of students in the grade selected for the QAT

[11] PEF reportedly made the decision to outsource the testing in order to (i) increase program school confidence in the integrity of the testing process and the test results as well as (ii) release the organization from the administrative burden of performing these critical activities during a period of significant program scale-up.

exceeds the number of enclosed test papers, the testing agency is directed to select the longer-tenure students by examining the school enrollment records for the last three months (again, ADU reports that this is a rare event, as changes in enrollment usually occur at the start of the academic year and the QATs are administered well into the academic year). The testing agency also checks the enrollment and attendance records over the last three months for fictitious students and impersonation cases.

QAT procedures have been developed to discourage schools from strategically concentrating on some grades for test preparation and to limit the risk of test leakage and its consequences. In primary schools, the QAT tests two grades; in middle and secondary schools, it tests three grades. Which specific grades are selected to be tested in a given school are kept strictly confidential by the ADU team—both the testing agency and the program school learn which grades will be tested only when the sealed test packets are opened by the testing agency at the school in the presence of the school administrator(s) just before testing time. The testing agency also only brings in the sealed test packets to be used at that school on the day of the test. In addition, for each grade, eight different question papers are prepared by the ADU. These questions papers are randomized for each grade. Table 2 shows the practical results of implementing these procedures: it presents the top-four combinations of tested grades and the associated shares of program schools by QAT round.

A number of additional, more standard procedures are applied to prevent cheating. Teachers and school administers are not permitted to enter the testing area during the test. Test invigilation is typically carried out by a team of at least three persons. Children are not allowed to bring any materials or stationery (other than pens) into the testing area. The grades selected for the QAT are tested simultaneously in the same testing area and the seating for the students is mixed. All test materials provided to the test takers are collected back. All unused test materials are retained by the testing agency. ADU also reports that test item recycling across QAT rounds is presently rare given the size of its test item bank.

The QAT serves as the main source of data for the analysis of learning dynamics in program schools. While PEF has conducted nine QATs since the inception of the FAS program (QAT 9, the latest QAT, was administered in March 2010), this study uses data from QAT 4−QAT 8 which spans November 2007–November 2009, a two-year period, covering, in whole or in part, three academic years. QAT 4 is the first QAT round in the analysis as it represents the

first QAT after the first major expansion of the program: phase-3, which increased the number of program schools from 194 to 676 schools or by 248%.

The sample for the learning dynamics analysis is restricted to phase-3 and phase-4 program schools. This is for two reasons. First, large numbers of schools entered into the program in these phases, and these schools have been in the program for a sufficiently long period to examine short-run dynamics (phase-3 schools have been subjected to five QATs and phase-4 schools four). Second, by phase-3, PEF had formalized the program entry qualification process, introducing standardized application forms, unannounced school inspections by PEF staff with standardized inspection forms, and an entry test, called the Short-Listing Quality Assurance Test (SLQAT), which is a pared-down version of the QAT. As a result, pre-program information on schools from the entry qualification stages is available for phase-3 and phase-4 schools, allowing the examination of the correlations between pre-program characteristics and the learning levels and changes, as well as their use as a form of specification testing in the causal analysis of the effects on learning from program design features that introduce stick and carrot incentives.

The data were provided by PEF in QAT-specific databases. Each QAT database is at the student level, with the number of points scored and/or the percent score in each relevant subject. Information on the school and grade of the tested student is also provided. Schools in the databases for QAT 5−QAT 8 were identified by both their names and unique identification numbers provided by PEF. School-level summary statistics on QAT performance—such as the number of test takers, test score means, test score standard deviations, and test pass rates—derived from the student-level data were linked across QAT rounds using the unique school identifiers. There were no issues in the linking process: when a school in a given QAT was expected to be also found in another QAT, it was. Program schools in the QAT 4 database were however only identified by school name (at times augmented by location markers inserted into the school name field). Thus, schools in QAT 4 were linked to the schools in QAT 5 by visually matching on school names. There were 674 program schools that took QAT 4; 658 of them were expected to be found in QAT 5. Matching on school names yielded 625 linked schools (a success rate of 95%).

Table 3 presents the number of phase-3 and phase-4 program schools that took each of the QATs. It also presents the number of phase-3 and phase-4 schools that qualified for entry and

joined the program. The first QAT that phase-3 program schools took was QAT 4 in November 2007. The first QAT that phase-4 program schools took was QAT 5 in March 2008. It appears that until QAT 8, there was virtually no school attrition out of the program. Even in QAT 8, the attrition was minor: only 3% of phase-3 and phase-4 program schools exited the program. The majority of schools that exited did so due to disqualification for consecutive failures on the QAT. This finding of low attrition from the program indicates that sample selection of this form is unlikely to be a source of bias in the learning dynamics analysis.

Table 4 presents the number of students tested in the various QAT rounds, separately by grade. The sample comprises of all program schools. We discern two patterns that signal test design and administration procedures at work. First, the number of grades that are potential targets for testing has expanded over QAT rounds: in QAT 4, the tested grades were drawn from a universe of four grades. By QAT 8, it was nine. Second, the changing shape of the frequency distribution of test takers by grade between QAT rounds is consistent with the procedure of changing the combination of grades tested in a given program school over QAT rounds.

4. Evolution in learning

In this section, we examine how learning has evolved over QAT rounds among phase-3 and phase-4 program schools, where learning is measured in terms of mean QAT scores in the core subjects of English, mathematics, and Urdu. Table 5 presents means and standard deviations for school-level mean QAT scores by QAT round. These statistics are presented for total mean QAT scores as well as the subject-specific mean QAT scores. Figures 3 and 4 depict kernel estimates of the probability density functions (PDFs) for total mean scores by QAT round for phase-3 and phase-4 program schools, respectively. Mean QAT scores for both phase-3 and phase-4 programs are normalized using the mean QAT 4 score distribution for phase-3 program schools as the base.

Looking first at phase-3 program schools, we find that mean scores jumped 2.90 standard deviations between QAT 4 and QAT 5. Figure 3 clearly displays this distributional shift. This is a massive change that occurred over a short period of four months. In Section 7, we examine whether this substantial change in learning is attributable to some extent to accountability pressure on a large number of phase-3 program schools that failed to achieve the minimum pass rate for program benefit maintenance in QAT 4. Changes in learning after QAT 5 are relatively

more modest, with mean scores backtracking somewhat in QAT 7 and QAT 8 relative to those of QAT 6 by 0.42 and 0.36 standard deviations, respectively. Decomposing the mean total score into its constituent subjects, we find that the mean scores in all subjects depict the same major increase in learning between QAT 4 and QAT 5 as the mean total score. Mean English and mathematics scores continue to show similar trends over QAT rounds as the mean total scores; mean Urdu scores however appear to follow a counter-trend to the mean total scores in later QAT rounds. Similar to the trend for the level of mean scores, the *variation* in mean scores as measured by standard deviations shows a pattern of increase (peaking in QAT 6) and then decline over QAT rounds. Variations in mean English and mathematics scores depict similar trends to the variation in mean total scores, while variation in mean Urdu scores follows a counter-trend to the variation in mean total scores.

Phase-4 program schools entered in the program prior to QAT 5. Mean scores in QAT 5 were 2.51 standard deviations higher than the mean scores for phase-3 program schools in their first QAT, QAT 4. However, the mean scores were 0.39 standard deviations lower than how phase-3 program schools performed in QAT 5. Over QAT 5−8, mean scores for phase-4 program schools do not show any discernible trend, with mean scores fluctuating between 2.5 and 2.9 standard deviations. Similarly, the subject-specific mean scores show no discernible trend, displaying stronger fluctuations over QAT rounds. The effects of these fluctuations on mean total scores are however dampened by the counter-fluctuation of mean Urdu scores vis-à-vis mean English and mathematics scores. Variation as measured by standard deviations in mean total scores shows an increase from QAT 5 to QAT 6 and then a decline over subsequent QAT rounds; similar trends are exhibited by the subject-specific mean scores.

In addition to examining summary statistics of the mean QAT score distributions, we test for stochastic dominance relations between successive pairs of mean QAT score distributions for phase-3 and phase-4 program schools separately. Suppose we have two distributions *A* and *B*, characterized by their cumulative distribution functions (CDFs) F_A and F_B, respectively. The distribution *B* stochastically dominates distribution *A* at first order, if for any argument *z*, $F_A(z) \geq F_B(z)$. If *Z* denotes learning and the observations are schools, this implies that the incidence of schools with learning values equal to or less than a 'low' learning level *z* is higher under distribution *A* than under distribution *B*. Higher orders of stochastic dominance can be

defined analogously. Denote D_k^s to be dominance function of order s for distribution $K = A, B$, defined recursively by the relations

$$D_k^1(z) = F(z), \;\; D_k^{s+1}(z) = \int_{\underline{z}}^{\overline{z}} D_k^s(x)\,dx, \;\; s = 1,2,3,...$$

Distribution *B* stochastically dominates distribution A at order *s* if $D_A^s(z) > D_B^s$ for all values of *z* in the joint support of the two distributions. Stochastic dominance of the second order implies that the low learning gap is higher under distribution *A* than under distribution *B*. Stochastic dominance of the third order implies that the squared low learning gap is higher under distribution *A* than under distribution *B*. We test for stochastic dominance based on an empirical likelihood ratio statistic proposed by Davidson and Duclos (2007) and extended by Davidson (2009).[12] This test tests for the rejection of the null hypothesis of nondominance between distributions *A* and *B*. The test tests first order as well as higher order stochastic dominance relations and permits the distributions under examination to be correlated, which is likely given serial correlation in the mean QAT score distributions. The test is restricted to the mean QAT score values between the 5th and 95th percentiles of the joint support of the two distributions.

Figures 5 and 6 present CDFs of the mean QAT scores by QAT round for phase-3 and phase-4 program schools, respectively. For phase-3 program schools, pair-wise tests suggest that the distribution of mean QAT scores for QAT 5 and subsequent ones first-order dominate the corresponding distribution for QAT 4. There is no evidence of stochastic dominance up to the third order between successive mean score distributions among QAT 5−QAT 8. Pair-wise tests of the successive mean score distributions among QAT 5−QAT 8 for phase-4 program schools also suggest that the null hypothesis of nondominance up to the third order cannot be rejected under standard significance levels.

Thus the evidence collectively points to two main findings. First, there was a major jump in learning as measured by mean scores between QAT 4 and QAT 5 for phase-3 program schools. Second, between QAT 5−QAT 8 learning appears to be in an oscillating steady state over QAT rounds for both phase-3 and phase-4 program schools. This finding remains valid when mean total scores are decomposed into mean scores for English, mathematics, and Urdu.

[12] The test is automated in `Stata` by Araar and Duclos (2009).

5. Learning variance decomposition

In this section, we examine how much of the total variability in QAT scores is attributable to the variability in QAT scores *between* program schools (i.e., due to school heterogeneity) versus the variability in QAT scores between students *within* program schools (i.e., due to student heterogeneity). Decomposing further, we examine how much of the total variability in QAT scores is attributable to the variability in QAT scores (1) between program districts, (2) between schools within district, (3) between grades within school, and (4) between students within grade. These are important questions as education policymakers and practitioners may have at their disposal the means to effectively reduce the extent of heterogeneity at the level of the school and grade (if they are found to be major contributing factors), thereby reducing total variability in student achievement.

To answer the above questions, we fit two-level and four-level variance-components models (nested random effects models) to the student-level QAT score data. This exercise is performed separately for each round of QAT score data.[13] It is also performed separately for phase-3 and phase-4 program schools.

The two-level variance-components model is specified as

$$y_{is} = x_g\beta + \alpha_s + \varepsilon_{is}, \;\; i = 1,...,I, \; s = 1,...,S, \tag{1}$$

where *i* indexes the student and *s* the school, and y_{is} denotes the mean QAT score in the core subjects of English, mathematics, and Urdu for student *i* in school *s*. Students are nested within schools—thus student QAT scores within schools are likely to be correlated. The QAT sometimes tested students in the same grades across schools. As a result, the grade *g* is viewed as a cross factor with the school *s* and is treated in the model as a fixed vector of grade dummies x_g with a vector of fixed parameters β to be estimated. The error components α_s and ε_{is} are assumed to be independently normally distributed with means zero and standard deviations σ_{α_s} and σ_{ε} , respectively, and are estimated correcting for potential correlation among students within schools and heteroskedasticity of arbitrary form. Since the error components (α_s and ε_{is})

[13] We are unable to pool the student-level QAT score data over QAT rounds in order to examine how much the variation in student QAT scores is explained by inter-round variation, as a panel only exists at the school level; at the student level, we have repeated cross-sections.

are assumed to be independently distributed, the total variance in QAT scores $Var\left(y_{is}\right)$ is equal to the sum of the between-school variance $\left(\sigma^2_{\alpha_s}\right)$ and the within school, between-student variance $\left(\sigma^2_{\varepsilon}\right)$.

The four-level variance components model is specified as

$$y_{igsd} = x'_g\beta + \alpha_d + \alpha_{sd} + \alpha_{gsd} + \varepsilon_{igsd}, \tag{2}$$

where *i* indexes the student, *g* the grade, *s* the school, and *d* the district, and y_{igsd} denotes the mean QAT score in core subjects for student *i* in grade *g* in school *s* in district *d*. There are four factors in this model: district, school, grade, and students. Students are nested within grades, grades within schools, and, in turn, schools within districts. These factors are treated as additive random effects in the model, where α_d, α_{sd}, and α_{gsd} denote the random effects varying over district *d*, school *s*, and grade *g*, respectively. These random effects are assumed to be independently and normally distributed with means zero and standard deviations σ_{α_d}, $\sigma_{\alpha_{sd}}$, and $\sigma_{\alpha_{gsd}}$, respectively, and are estimated correcting for potential correlation among students within schools and heteroskedasticity of arbitrary form. As in (1), the grade *g* is additionally treated in the model as a fixed vector of grade dummies x_g with a vector of fixed parameters β to be estimated. Given that the error components are assumed independent, total QAT score variation is equal to the sum of the variances of the error components. Both variance-components models are estimated via maximum likelihood.

The results of the QAT score variance decomposition are presented in Table 6. Panel 1 presents the estimated shares of total variation at the selected levels for phase-3 FAS program schools, separately by QAT round. Panel 2 presents analogous results for phase-4 FAS program schools. We highlight four main results. First, the results in general suggest that most of the variation in QAT scores is at the level of the student—for example, the two-level variance components estimation suggests that roughly 60–70% of total QAT score variation is between students within schools, while the remaining 30–40% is between schools. This finding does not change qualitatively when we look at the results from the four-level variance-components estimation: though the share falls to 50–60%, the variation in QAT scores between students remains the majority contributor to total QAT score variation.

Second, the four-level variance-components estimation suggests that the variation in QAT scores between program districts explains a small share of the total variation in QAT scores: between roughly 1–5%. This finding implies that the QAT score distributions are essentially identically centered across districts**.** Third, the estimation also suggests that within schools, the variation in QAT scores between tested grades explains a nontrivial share of total QAT score variation (around 15–18%). In addition, the share of total variation due to between-grade variation appears to be drawn from the shares due to between-school variation and between-student variation—this absorption from both directions likely arises from the fact that tested grades are not only nested within schools but are also crossed with schools, given that sometimes students in the same grades are tested across schools.

Fourth, the estimated total variation in student scores shows substantive movements over QAT rounds. Figure 7 depicts the evolution of the *levels* of total, between-school, and within-school variation in QAT scores, separately for students in phase-3 and phase-4 schools. In phase-3 program schools, the level of total student score variation declines from QAT 4 to QAT 5 before rising sharply with QAT 6, and then declines monotonically over QAT 7 and QAT 8. The level of total student score variation for phase-4 program schools shows an identical trend, with a sharp increase between QAT 5 and QAT 6 before declining over the subsequent two rounds. It appears that the evolution in the level of total student score variation between QAT 4–QAT 6 in phase 3 schools is largely explained by the evolution in the level of within-school student score variation; between-school student score variation shows little movement over these QAT rounds.

Notwithstanding, what explains the uptick in total score variation between QAT 5 and QAT 6? There were no major changes in program design or implementation between those two rounds. The one change that did occur was the reduction in the number of QAT questions and points in the English, mathematics, and science sections from 15 to 10, while the points for the single question in the Urdu section remained fixed at 10. This test structure has been maintained through the subsequent QAT rounds. This change implies that the Urdu section now receives a higher weight in the QAT—if the student only answered the Urdu question correctly in QAT 5, she would receive 18.2% on the test; if the same event occurred in QAT 6, she would receive 25% on the test.

Given the reweighting of the QAT across subjects, the increase in total QAT score variation between QAT 5 and QAT 6 suggests that the variation in Urdu scores is higher than in

the other core subjects of English and mathematics. Figure 8 depicts the evolution of the level of QAT score variation by subject. While the level of QAT score variation in Urdu is not always higher than those for the other subjects, there was a major spike in the level of its variation between QAT 5 and QAT 6: Urdu score variation increased by 61% and 57% for phase-3 and phase-4 program schools, respectively. Thus, it appears that the score reweighting amplified the effect of the spike in Urdu score variation on total QAT score variation. We however cannot provide an explanation for the observed spike in Urdu score variation.

How do these test score variance decomposition estimates compare to estimates from other recent studies from the South Asia region? Our estimates of inter-school differences appear to lie on the lower end of the range of available estimates. For example, similar to our findings, on the lower end, Goyal and Saiens (2009), using 2007 data from a nationally-representative sample of grade-2 and grade-4 students from Bhutan, find that 26–41% of the total variation in test scores across tested grades and subjects is explained by variation between schools. Andrabi et al. (2007) find that, in a sample of grade-3 students in private and public schools in rural Punjab, Pakistan, roughly 50% of total variation is explained by variation between schools. On the higher end, although the statistic is not reported, Saiens (2009) finds that, in a sample of public schools from rural Sindh, Pakistan, the majority of the variation in test scores for grade-3 and grade-5 students is explained by variation between schools.

6. Pre-program correlates of learning dynamics

In this section, we examine what factors are associated with learning levels and changes in program schools by essentially estimating school effectiveness regressions. The factors that we examine are basic school-level characteristics captured at the program application stage (i.e., pre-program) for phase-3 and phase-4 program schools such as school size, number of teachers, maximum and minimum teacher salaries, number of classrooms, level, gender type, location, registration status, and mean SLQAT score. Learning is measured in terms of the school-level mean QAT score in the core subjects of mathematics, English, and Urdu. The conditional relationships between these factors and learning (changes) are investigated via simple pooled ordinary least squares (OLS) regression as well as via fixed-effects (FE) and random-effects (RE) regressions which take advantage of the panel structure of the data (multiple test rounds over schools) to control for heterogeneity at the school level.

The pooled OLS regression model is formulated as

$$y_{sq} = \alpha + QAT_q\lambda + QAT_q \times x_s\gamma + x_s\beta + \varepsilon_{sq}, \;\; s = 1,...,S, \; q = 1,...,Q, \tag{3}$$

where s indexes the school and q the QAT round, and y_{sq} is the normalized mean QAT score in the core subjects for school s in QAT q, QAT_q and x_s are vectors of the QAT rounds and the time-invariant pre-program covariates, respectively; the associated parameter vectors λ and β to be estimated capture the main effects of the covariates on mean QAT scores. The vector $QAT_q \times x_s$ captures interactions between the pre-program covariates and the QAT rounds; the associated parameter vector γ to be estimated captures the differential effects of the pre-program covariates over QAT rounds. The stochastic error term ε_{sq} is assumed to be uncorrelated with the covariates and distributed normally. The classical assumption of identically and independently distributed errors is however relaxed. The error term is likely to be correlated over QAT rounds for a given school; hence, we estimate standard errors corrected for potential serial correlation as well as cross-sectional heteroskadasticity of arbitrary form. If the model is correctly specified and the error term assumptions are valid, the pooled OLS estimator is consistent.

The pooled OLS estimator is inconsistent if the true model is the fixed effects model. The fixed-effects (FE) regression model is formulated as

$$y_{sq} = \alpha_s + QAT_q\lambda + QAT_q \times x_s\gamma + \varepsilon_{sq}, \;\; s = 1,...,S, \; q = 1,...,Q, \tag{4}$$

where α_s is a vector of random school-specific variables treated as unknown parameters to be estimated that are potentially correlated with the other regressors QAT_q and x_s. The stochastic error term ε_{sq} is assumed to be uncorrelated with the covariates. The estimates of the parameter vectors γ and λ reflect *within*-school effects of the QAT rounds and the pre-program covariates over QAT rounds on mean QAT scores (relative to the base QAT, QAT 4), respectively. As with the OLS standard errors, the FE standard errors are adjusted for serial correlation and cross-sectional heteroskedasticity of arbitrary form.

If the true model is a random effects model, the pooled OLS estimator and the FE estimator are consistent but inefficient. The random-effects (RE) regression model is formulated as

$$y_{sq} = QAT_q\lambda + QAT_q \times x_s\gamma + x_s\beta + \alpha_s + \varepsilon_{sq}, \;\; s = 1,...,S, \; q = 1,...Q, \tag{5}$$

where α_s are assumed to be random intercepts that are distributed independently of the covariates. The random intercepts α_s and the stochastic error term ε_{sq} are assumed to be distributed normally; they are however not assumed to be independently and identically distributed.

The FE estimator is likely to be less biased than the pooled OLS estimator as the school fixed effects control for time-invariant unobserved factors that vary at the school level or higher. The model also allows for the school fixed effects to be correlated with the other covariates. However, as the estimator ignores the significant between-school variation present, the parameter estimates are likely to be imprecise. In addition, the independent effects of the time-invariant pre-program covariates cannot be estimated and are excluded from the model as they are perfectly collinear with the school fixed effects. The RE estimator allows the independent effects of the time-invariant pre-program covariates to be estimated. However, the RE model assumes that the school random effects are uncorrelated with the covariates, which is usually unrealistic in most applications. We use the artificial regression approach to test the null hypothesis that the more efficient RE estimator yields similar parameter estimates to the consistent FE estimator (Wooldridge 2002; Baltagi 2005). This test is robust version of the standard Hausman test, accounting for cross-sectional conditional heteroskedasticity and serial correlation of arbitrary form.[14]

As discussed in Section 4, given that significant changes in learning were only observed with phase-3 program schools, we only present results from estimating school-level learning growth-curve regressions for this sample over QAT 4–8. Table 7 presents the learning growth-curve regression results. The dependent variable in the models is the mean QAT score in the core subjects normalized using the QAT 4 score distribution. All regressions include district dummies although their parameter estimates are not reported. We find that the parameter estimates for pre-program characteristics appear to be of comparable sizes across the three estimation models. Estimates of the standard errors also appear to be similar across models; this implies that

[14] The test is automated into `Stata` by Schaffer and Stillman (2010).

statistical inference is largely identical across models. Hence, we discuss the estimation results without differentiating by model.

Learning in QAT 5–QAT 8 was higher than in QAT 4. The effects largely vary between 4–5 standard deviations at the conditional mean. This evidence is consistent with the unconditional evidence on changes in learning presented in Section 4. While the main effect lacks significance, middle schools have lower learning levels than primary schools in QAT 7 and QAT 8 relative to QAT 4. Similarly, while the main effect lacks significance, secondary schools have lower learning levels than primary schools in all rounds relative to QAT 4. The main and interactions effects for both middle and secondary schools are jointly significant. Registration status of the school has a positive effect on learning but this effect disappears in QAT 6 and QAT 8. While the main effect lacks significance, program school size appears to have a small positive effect on learning in QAT 5. The number of teachers also has a small positive effect on learning. Given that we control for school size, this effect can be interpreted as the effect of lower student-teacher ratios on learning. However, the effect disappears in QAT 5–8. Neither the location of the program school in terms of rural versus urban nor the gender status of the school in terms of coeducational versus single-sex appear to matter for learning levels or changes. Pass rates in the program entry test is positively associated with learning; the effect however does not vary with QAT round. Finally, maximum and minimum teacher salaries, which taken together may be interpreted as reflecting the level of teacher quality in the program school, are not associated with learning levels or changes.

Explaining changes in learning is effectively tantamount to explaining the jump in learning between QAT 4 and QAT 5, as post-QAT 5, the learning distributions of phase-3 program schools appear to show some relatively small oscillating movement. The same situation applies to phase-4 program schools: mean scores appear to show only slight oscillating movement over QAT rounds. Tables A1 and A2 present learning growth-curve regression results over QAT 5–QAT 8 for phase-3 and phase-4 program schools, respectively. The explanatory power of the models estimated over these samples drops precipitously.

All the estimations presented above do not attempt to correct for potential measurement error in the regression covariates. Measurement error in covariates such as location, level, gender type, and registration status is unlikely but schools may overreport such characteristics as enrollment, teachers, and classrooms in the program application forms if they feel that it may

increase the likelihood of program entry. However, the fact that PEF visits all applicants and checks all reported figures is likely to help arrest this tendency. Notwithstanding, if present, the extent of overreporting is likely decreasing with true enrollment, teachers, and classrooms. For example, given that only private schools with at least 100 students can apply to the program, small schools may inflate the enrollment to exceed and distance themselves from this floor. If the observed values of a single covariate are inflated with measurement error negatively correlated with true values, the associated parameter estimate is upwardly biased. In a multiple regression framework, systematic measurement error of this form among multiple covariates is likely to bias regression parameters in unknown ways.

7. Causal effects of stick and carrot incentives on learning

Exploiting the design feature that program assignment was ultimately strictly determined on the basis of SLQAT pass rates relative to a distinct, known cutoff and applying a sharp regression-discontinuity design to the data for phase-4 program schools, Barrera-Osorio and Raju (2009) find that the FAS program produced sizeable, positive effects on school size, school inputs such as teachers, classrooms, and blackboards, and student achievement in program schools with SLQAT pass rates near the cutoff. The documented learning gains induced by the program are theoretically attributable to several drivers. For example, a first potential mechanism is the QAT pass rate floor for continued program eligibility which is likely to create a strong push from below for the program school to invest resources, organize itself, and set and activate its own internal incentives to ensure that the majority of students learn enough to regularly pass the QAT. The effort exerted may also be dynamically continuous: more risk-averse program schools are likely to continue to exert an effort to raise the QAT performance of their students in order to extend the distance between their pass rate positions and the QAT pass rate floor, creating further "breathing room".

A second potential mechanism is competition-driven pressure exerted by parents and other local stakeholders on the program school to raise student learning created by the FAS program requirement initiated with QAT 6 in November 2008 that the program school prominently publicly display the QAT pass rate performance rankings of all the program schools in the school's district. The first and second mechanisms are likely to produce learning gains by compressing the QAT distribution from the left.

A third potential mechanism is the offer of annual cash bonuses to teachers in program schools that (first) achieve a QAT pass rate of at least 90% (and, second, pass another threshold based on a simple addition of the number of testtakers and mean QAT score), as well as the competitive annual cash bonus to the program school with the highest QAT pass rate[15] in the major program districts. Both of these bonuses—which are substantial in both absolute terms and relative to pre-program school revenue and teacher salary estimates—are likely to create a pull from above for schools to make an effort to continue to raise their QAT performance. This mechanism is likely to produce learning gains by shifting as well as stretching the QAT score distribution to the right.

A fourth potential mechanism is the per-student subsidy benefit itself, which may provide higher revenues to the program school if pre-program tuition rates were lower than the subsidy benefit and/or school size has expanded. The regular, full receipt of the per-student subsidy benefit (rather than reliance on tuition payments from parents with low and/or unstable paying capacity) also raises the program school's confidence in the predictability of future cash flow. Both these factors may facilitate larger and longer-term investments in resources to improve teaching and learning at the school, and, hence, produce learning gains which are realized across the QAT score distribution.

A fifth potential mechanism is the tuition-free schooling condition for program benefit eligibility. Household expenditures previously allocated for tuition and fees are freed up for other household consumption and investment priorities. One potential outcome is the reduction in the opportunity cost of child time, allowing the student to attend school more regularly, stay for the full school hours, and spend after-school hours on homework assignments. Another potential outcome is households redirect (a portion of) the freed-up resources to investments and expenditures that may directly or indirectly improve student learning such as better nutrition, uniforms, books and stationery, transportation to school, and tutoring services.

The above potential mechanisms can generate real gains in student learning. However, the observed gains could also be the product of gaming behavior by schools, resulting in artificial gains in QAT scores. For example, program schools may respond to the QAT-related requirement for program continuation and additional benefits by systematically dropping low-performing students and/or screening in and admitting high potential or high-performing

[15] With ties broken by using mean QAT scores.

students, cheating on the QAT, manipulating which students are present for the QAT (which is announced in advance) and testing conditions, and devoting resources and directing staff and student effort to QAT-taking strategies as well as reorienting teaching and learning towards the subject and content matter covered in the QAT, at the possible expense of a more extensive and richer teaching and learning agenda (see, e.g., Jacob 2005 and Jacob 2007 for reviews).

These mechanisms likely work simultaneously to produce the observed learning gains under the program. Decomposing their individual contributions is likely to be difficult without structural modeling; and, even then, the data may not be sufficiently rich to permit the identification of the (full set) of structural parameters. However, it is still possible to partially unpack the black box of learning-gains production under the program via reduced-form estimation. This opportunity emerges from the combination of peculiar circumstance and program construction which allow the credible independent identification and estimation of the average causal effects of two mechanisms on defined subsets of program schools: (1) the threat of program exit for program schools that just failed the QAT a first time and (2) the prospect of receiving teacher bonuses in the next award round for program schools that just missed acquiring bonus eligibility in the preceding award round.

A. Threat effect on learning

As mentioned, design and circumstance combine to yield an opportunity and empirical strategy to investigate whether accountability pressure under the FAS program induces student learning gains. First, in terms of the enabling design, FAS program schools have to maintain a minimum QAT pass rate of 67% for continued program eligibility. Schools that fail to pass a given QAT are offered a second chance to pass the QAT the next time it is administered. In the intervening period, PEF may impose additional rules and regulations on first-time failers such as freezing their enrollment counts for benefit amount calculations. The school is permanently disqualified from the program if it fails the QAT two consecutive times. This appears to be a serious turn of event: Observations reported by PEF suggest that a disqualified school is significantly worse off relative to a low-cost private school that had remained continuously outside the program, with disqualified schools experiencing student and staff flight and closure. Thus, the high-stakes nature of the QAT introduces powerful incentives for first-time failers to boost learning in order to avoid the threat of program exit.

In terms of circumstance, in phase 3 of the program's expansion (the first major expansion of the program in terms of the number of schools enrolled), 514 low-cost private schools who applied to the program attained the minimum pass rate in the SLQAT, the final qualification step for program entry. Out of these schools, 482 schools (93%) signed their program participation agreements with PEF in July-August 2007 and began receiving the monthly subsidy benefit. In November 2007, three to four months after phase-3 program schools entered the program, QAT 4 (the first QAT for phase-3 program schools) was administered by PEF in all program schools. Out of 479 phase-3 program schools that took QAT 4, only 234 (49%) passed it. The high failure rate was unanticipated by PEF—such a rate had not been observed in preceding QAT rounds nor has it been repeated to date in subsequent QAT rounds. For example, in phase 4, the next major program expansion, of the 425 phase-4 program schools that took QAT 5 (which represented the first QAT for phase-4 program schools), 412 (97%) passed it.

The precise QAT pass rate cutoff for determining program eligibility combined with the unusually high failure rate among phase-3 program schools in QAT 4 provides a data design with adequate sample size to investigate whether the threat of program exit precipitated by failing QAT 4 has a causal effect on mean scores in QAT 5 (the ultimatum or threat QAT) as well as later QAT rounds (indicating effect persistence) for first-time failers near the cutoff among phase-3 program schools. Specifically, we fit a sharp RD design to the data to essentially compare mean scores in subsequent QAT rounds of phase-3 program schools that just failed QAT 4 (referred to as marginal failers), and, hence become subject to the threat, to those of corresponding schools that just passed QAT 4 (referred to as marginal passers).

To briefly delineate the identification strategy, let y_s denote the mean QAT score in the core subjects of Urdu, English, and mathematics in phase-3 program school s, and let the indicator variable d_s denote treatment (threat) assignment, where one denotes that the school failed QAT 4, and zero otherwise. In addition, let y_{0s} and y_{1s} denote the potential outcomes of school s in the untreated and treated states, respectively. Treatment status is assigned based on the decision rule

$$d_s(z_s) = 1\{z_s < c\}, \tag{6}$$

where z_s denotes school s's QAT 4 pass rate which is perfectly observed (z is more generally referred to as the assignment variable), c the known, distinct pass rate cutoff of 67% (more precisely, 66.67%), and 1 an indicator function. Thus, the conditional probability of treatment as a function of the assignment variable z, $\Pr\left[d_s = 1 \middle| z_s = c\right]$, dives discontinuously from one to zero at the cutoff, yielding a sharp RD design (Trochim 1984).

Let e denote an arbitrarily small number. Under the assumption that (i) the zero-limit of the conditional expectation of the counterfactual untreated outcome for treated schools is well defined, (ii) the conditional expectation of the outcome variable exhibits local smoothness at the cutoff in the absence of the treatment, (iii) and the density of the assignment variable z is positive in the neighborhood of the cutoff, the difference in mean outcomes between marginal passers and marginal failers identifies

$$\alpha_y \equiv E\left[\alpha_s \middle| z_s = c\right] = \lim_{e \downarrow 0} E\left[y_{1s} \middle| z_s = c - e\right] - \lim_{e \downarrow 0} E\left[y_{0s} \middle| z_s = c + e\right] \tag{7}$$

which represents the average treatment effect of the treated (ATT) or threat effect at the cutoff for phase-3 program schools (Hahn et al. 2001, Todd 2006). A consistent estimate of the threat effect at the cutoff is given by

$$\hat{\alpha}_y = \hat{\alpha}_y^- - \hat{\alpha}_y^+, \tag{8}$$

where $\hat{\alpha}_y^-$ and $\hat{\alpha}_y^+$ denote the conditional expectations of the outcome variable y at the cutoff from below and above, respectively.

Given that we are interested in flexibly estimating the treatment effect at a single point using observations in its neighborhood, an attractive method is local smoothing using nonparametric regression. We use local linear regression (a local polynomial of order one) to individually estimate the two conditional expectations at the cutoff in (5); the selection of this estimator is motivated by its faster convergence rates at boundaries relative to standard kernel regression (Fan and Gijbels 1996; Porter 2002). The practical implementation of local linear regression requires the specification of the kernel $K_h(\cdot)$, the weighting function, and the bandwidth h, the window width in which the kernel function is applied. We select the triangular kernel given that it is boundary optimal and thus well suited to regression-discontinuity designs (Cheng et al. 1997). We select the optimal bandwidth h by applying the plug-in method developed by Imbens and Kalyanaraman (henceforth, I-K) (2009) specifically for use in

regression-discontinuity settings, with *h* determined by minimizing the mean squared error (MSE) using observations only around the cutoff.[16] Following the guidance in Imbens and Wooldridge (2009), we also examine the sensitivity of our inference results to setting the bandwidth to half and twice the optimal bandwidth generated from the I-K method for each of our estimations.

The local estimation via RD aids in controlling for phenomena/events that may have biased our results if we had opted for a more global estimation method. We note three potential sources. First, it is conceivable that some program schools may place in the bottom of the score distribution in a given QAT round due to a transitory bad draw but can expect to make gains towards the mean in subsequent QAT rounds (Kane and Staiger 2002). Thus, the initial mean QAT score of a low-performing school may be a misleading measure of its true QAT performance. This implies that, in our case, the learning gains experienced by first-time failers may be attributable in part to both accountability pressure and regression to the mean. However, any potential mean reversion tendency is expected to be locally smooth at the pass rate cutoff (Chay et al. 2005). Second, the estimated RD effects are net of any natural student learning depreciation that may affect gains in mean scores (Andrabi et al. 2009), with the conditional mean level (and the rate) of learning depreciation expected to be locally smooth at the pass rate cutoff. Third, PEF responded to the high QAT 4 failure rate by organizing training in January 2008 on QAT content and format and targeting it at QAT 4 failers *as well as* marginal passers. Given this, while the training may have shifted the levels of the learning regression functions, the effect of the training on mean learning is expected to be locally smooth at the cutoff. Thus, the estimated RD effects at the pass rate cutoff net out any bias due to these sources and continue to identify the threat effect on marginal first-time failers.

We first discuss the findings from some basic model specification tests. First, we check if there is a positive mass of school observations in the neighborhood of the QAT 4 pass rate cutoff, given that this is a necessary condition for the identification strategy. This condition is verified by a visual inspection of the frequency distribution of phase-3 program schools by QAT 4 pass rate (see Figure 9). Second, although the following check does not necessarily pose a threat to identification, using Figure 9, we also visually inspect whether there is an unusual discontinuity in the density function of the QAT 4 pass rates at the cutoff, as it may indicate

[16] The I-K bandwidth selection algorithm is automated in `Stata` by Fuji et al. (2009).

strategic one-way manipulation of pass rates at the cutoff (McCrary 2008). The inspection suggests that the QAT 4 pass rate density function appears to be generally naturally characterized by a jagged structure; furthermore, this structure does not appear to be unique to the QAT 4 pass rate density for phase-3 program schools as other QAT rounds exhibit a similar pattern. Third, the RD analysis is performed on schools for which data could be linked across QAT rounds. As discussed in Section 3, the success rate in linking schools was not perfect, particularly between QAT 4 and QAT 5, and the loss of the unlinked schools may serve as a source of potential sample selection bias in the RD analysis. However, a comparison between the frequency distribution of all phase-3 program schools by QAT 4 pass rate with that of only linked phase-3 program schools (Figures 9 and 10), provides no telltale signs of dissimilarities, suggesting that the problem is negligible.

Fourth, we test for local smoothness in the conditional expectation of pre-program covariates at the QAT 4 pass rate cutoff using data from submitted program application forms. Although consistent evidence of discontinuities in the conditional means at the cutoff for these covariates does not necessarily undermine the identification strategy, it does cast doubt on its plausibility. Table 8 presents summary statistics on the pre-program number of students, number of teachers, maximum monthly teacher salaries, number of classrooms, gender, location, level, registration status, and mean SLQAT scores, as well as the RD estimates of the effects at the cutoff using local linear regressions. The findings strongly suggest that the null hypothesis of local regression smoothness in conditional expectations at the cutoff cannot be rejected at standard significance levels for the full set of pre-program covariates examined.

Fifth, we test if the conditional mean pre-program outcome (mean QAT 4 scores) exhibits local smoothness at the cutoff. This directly tests a necessary condition for identification under RD. Table 9 shows that the RD estimate of the threat effect on normalized mean QAT 4 scores at the cutoff are is −0.003 standard deviations, with a standard error of 0.775 standard deviations. The local linear regressions are depicted in Panel 1 in Figure 11. Thus, the finding suggests that the null hypothesis of local smoothness in mean QAT 4 scores at the pass rate cutoff cannot be rejected at standard significance levels.

Column 1 in Table 9 presents the RD estimates of the threat or "stick" effect on mean QAT scores for marginal failers among phase-3 program schools. Optimal bandwidth sizes for the local linear regressions, obtained using the I-K method, range between 20–37 percentage

points depending on the QAT round. To limit the influence of outliers in outcome values, school observations with QAT 4 pass rates within 15 percentage points of the cutoff with mean QAT score values smaller or larger than the 1st and 99th percentile values respectively were discarded from the analysis. Figure 11 depicts the local linear regressions of conditional mean QAT scores.

The estimated stick effect on mean scores at the cutoff in QAT 5, the threat QAT, is 0.664 standard deviations and statistically significant. We also find a RD stick effect on mean QAT 6 scores: the estimated effect is 0.561 standard deviations and significant. The inference findings for these QAT rounds are robust to setting the bandwidth to half and twice the optimal bandwidths for the estimations. The estimated RD effect on mean QAT 7 scores, two QATs removed from the threat QAT, drops in magnitude to 0.349 standard deviations and loses significance. The estimated RD effect on mean QAT 8 scores, three QATs removed from the threat QAT, is 0.089 standard deviations; this effect is also not statistically different from zero. The inference results are however not robust to setting the bandwidth to twice the optimal bandwidths for the estimations—when the bandwidths are doubled, the RD effects grow in size and gain significance. However, the observed degressive pattern in the RD effects over QAT rounds remains. Thus, the RD estimates collectively suggest the presence of a strong stick effect on learning for marginal failers. We also find weaker evidence of persistence in the stick effect but this effect appears to dissipate over time once the threat of program exit is no longer immediately present.

The detected RD stick effects on mean QAT scores may be due to QAT 4 failers manipulating the composition of student test takers to increase their chances of attaining the minimum pass rate in QAT 5. We see two opportunities for test-taking pool selection. First, the program school can take advantage of the minimum attendance requirement of 80% for testing to (temporarily) systematically rid itself of its poorest learners before the QAT is administered. Second, the program school can screen in and admit better (potential) learners in anticipation of their stronger performance in the QAT. QAT 4 failers may have more pursued both strategies more intensively than QAT 4 passers. While we cannot directly examine whether the composition of QAT takers have systematically changed over QAT rounds in a way that suggests positive test-taking pool selection, we can indirectly examine this question by studying whether we find any local discontinuities in the conditional expectation of the *number* of QAT takers at the pass rate cutoff.

Column 2 in Table 9 presents the RD estimates of the stick effect on the mean number of QAT takers. Optimal bandwidth sizes for the local linear regressions range between 12–31 percentage points depending on the QAT round. School observations with QAT 4 pass rates within 15 percentage points of the cutoff with test-taker values smaller or larger than the 1st and 99th percentile values respectively were discarded from the analysis. Figure 12 depicts the local linear regressions of conditional mean QAT takers.

The estimated RD effect on QAT 4 takers is –6.1 students; this effect is not statistically different from zero. Likewise, we do not find evidence of a RD effect on QAT 5 takers: the estimated RD effect is –4.2 students. The inference findings for QATs 4 and 5 are robust to setting the bandwidth to half and twice the optimal bandwidths for the estimations. In QAT 6-8, at the optimal bandwidths for the respective estimations, we also fail to reject the null hypothesis of local smoothness in the mean number of test takers at the cutoff. The inference findings are however not robust to setting the bandwidth to twice the optimal bandwidths for the estimations: the magnitude of the RD effects increases to roughly –15 students and the effects are significant at standard significance levels. Given that there is consistent evidence of local smoothness in mean test takers at the cutoff in QAT 5 (the threat QAT), we discount the sensitivity of the results in the post-threat QAT rounds and read the collective results as suggesting that test-taking pool selection is unlikely to be an important gaming strategy pursued by failers in order to raise mean QAT scores.

B. Carrot effect on learning

The FAS program also awards annual individual cash bonuses to a fixed number of teachers in qualifying program schools. With each annual round of bonuses, PEF has revised the qualifying criteria, largely motivated by the need to limit the number of bonus recipients. The first round of teacher bonuses was offered in January 2007, some 10 months into the academic year for program schools. The program schools that qualified for the bonus were those that achieved a minimum student pass rate of 90% on QAT 2, administered in November 2006. There were a total of 23 schools that qualified for teacher bonuses, which represented 43% of the program schools that took QAT 2. The second round of teacher bonuses was offered in January 2008. The qualifying schools were those that achieved *consecutive* minimum pass rates of 90% on QAT 3 and QAT 4, administered in March 2007 and November 2007, respectively. There were a total of

24 schools that qualified for teacher bonuses in the second round, which represented 12% of the program schools that took QAT 3. The third round of teacher bonuses was offered in January 2009. The qualifying schools were those that achieved the minimum pass rate of 90% in QAT 5, administered in February 2008, as well as obtained a minimum value of 100 in a simple average of the number of students tested in the QAT and the percent mean QAT score across all tested subjects. There were a total of 42 schools that qualified for teacher bonuses, which represented 4% of the program schools that took QAT 5.

Given the small number of schools that qualify for teacher bonuses, the sample for the analysis is expanded to include all program schools that took QAT 5. There were 1,083 schools that took QAT 5. If bonus qualification was solely based on obtaining a pass rate of 90%, 870 schools (80.3%) would have qualified. In contrast, if bonus qualification was solely based on obtaining a minimum score of 100 in the composite measure, 49 schools (4.5%) would have qualified. Applying the two criteria together yields 42 qualifying schools. Consequently, facilitating the application of a RD design to the data, it turns out that the general binding constraint for determining bonus qualification is the minimum score on the composite measure. Figure 13 depicts the independent effects of the two thresholds as well as their combined effect. Thus, the analysis of the carrot effect on learning is conducted on program schools which obtained the minimum pass rate of 90% in QAT 5 and are in the neighborhood of the minimum QAT 5 composite score. Discussions with PEF and a review of program documents indicate that the same thresholds are not used to determine eligibility for any other benefits (or penalties).

Analogous to the threat effect investigation, a sharp RD design is applied to examine the carrot effect for marginal bonus nonqualifiers among program schools who obtained the minimum QAT 5 pass rate of 90%. Given that the composite score is constructed from both learning and test participation measures, the carrot effect may manifest itself in mean QAT scores and/or the number of students that take the QAT (QAT takers). Thus, let y_s denote either the mean QAT score in the core subjects of Urdu, English, and mathematics or the number of QAT takers in program school s. Let the indicator variable d_s denote treatment assignment, where one denotes that the school failed to attain the minimum QAT 5 composite score, and zero otherwise. Finally, let y_{0s} and y_{1s} denote the potential outcomes of school s in the untreated and treated states, respectively. Treatment status is assigned based on the decision rule

$d_s(z_s) = 1\{z_s < c\}$, where z_s denotes school s's QAT 5 composite score which is perfectly observed, c the known, distinct pass rate cutoff of 100 points, and 1 an indicator function. The conditional probability of treatment as a function of the assignment variable z, $\Pr[d_s = 1 | z_s = c]$, dives discontinuously from one to zero at the cutoff.

Under the assumptions for sharp RD identification stated before, the carrot effect of the teacher bonus at the cutoff for program schools, α_y, is identified by the difference in post-bonus mean outcomes between marginal nonqualifiers and marginal qualifiers. A consistent estimate of the carrot effect at the cutoff is given by $\hat{\alpha}_y = \hat{\alpha}_y^- - \hat{\alpha}_y^+$, which we estimate via local linear regressions on either side of the cutoff with kernel K_h and bandwidth h fixed following the same procedures as in the threat effect investigation. In all estimations, sensitivity of the inference results is checked by setting the bandwidth to half and twice the optimal bandwidth. The carrot effect on learning remains identified in the potential presence of mean reversion from the top of the distribution as well as natural learning depreciation as the conditional expectations of these phenomena are assumed to be locally smooth at the composite score cutoff.

We begin by discussing some basic model specification tests. Figure 14 depicts the frequency distribution of program schools over QAT 5 composite scores. Each bin is two points wide and the vertical line represents the bonus qualification cutoff of 100 points. A visual inspection of the frequency distribution shows a nontrivial mass of observations in the neighborhood of the cutoff. The inspection also shows that there is no unusual discontinuity in the number of schools at the cutoff. Table 10 presents summary statistics on the pre-program number of students, number of teachers, maximum and minimum monthly teacher salaries, number of classrooms, gender, location, level, registration status and mean SLQAT score, as well as the sharp RD estimates of local discontinuities in the conditional expectations of these covariates at the cutoff using local linear regressions. Given that program application data are only available for phase-3 and phase-4 program schools, the total sample size for the local smoothness checks for pre-program covariate means decreases from 870 to 658 schools; importantly, only 30 schools qualify for bonuses in the sample. In contrast to the threat effect analysis, we do not find uniform evidence of local smoothness in the conditional means of pre-program covariates at the composite score cutoff: the RD effects for the means of number of teachers, share of schools that is rural, and the share of schools that is coeducational are

statistically significant. However, these findings are not robust to setting the bandwidths to half and twice the I-K derived optimal bandwidths for the estimations.

We test if the conditional expectations of mean scores and test participation in QAT 5 (the bonus assignment QAT) exhibit local smoothness at the composite score cutoff. This directly tests a necessary condition for identification under RD. Table 11 shows that the RD estimate of the carrot effect on mean QAT 5 scores at the cutoff is −0.60 standard deviations, with an estimated standard error of 0.60 standard deviations. The local linear regressions are depicted in Panel 1 in Figure 13. Similarly, the RD estimate of the carrot effect on QAT takers at the cutoff is –2.56 students, with an estimated standard error of 3.81 students. The local linear regressions are depicted in Panel 1 in Figure 15. The findings suggest that the null hypothesis of local smoothness in mean learning and test participation levels at the QAT 5 composite score cutoff cannot be rejected at standard significance levels.

Turning now to the results, Table 11 presents the RD estimates of the carrot effect for marginal bonus nonqualifiers among program schools that satisfy the QAT 5 minimum pass rate of 90%. Columns 1 and 2 present RD estimates for the carrot effect on mean QAT scores and QAT takers, respectively. RD effects controlling for pre-program covariates are not estimated given the limited degrees of freedom for the estimations in the cutoff neighborhood. Optimal bandwidth sizes for the local linear regressions, obtained using the I-K method, were roughly 5 points for mean QAT scores and 17 points for QAT takers. Figure 15 depicts the local linear regressions of the conditional mean QAT scores, separately by QAT round. Similarly, Figure 16 depicts the local linear regressions of the conditional mean QAT takers, separately by QAT round.

A priori, we expect no carrot effect in QAT 6, the QAT round just preceding the bonus award, and a carrot effect in QAT 7, the QAT round just following the bonus award, with possible persistence in the effect in QAT 8, the next QAT round. In terms of QAT performance, the estimated effect on mean scores at the cutoff in QAT 6 is −0.87 standard deviations. Although large in size, this estimate is not statistically different from zero. The estimated RD effect on mean scores in QAT 7, the immediate post-bonus QAT, is 0.21 standard deviations and not statistically different from zero. The estimated RD effect on mean scores in QAT 8, the post-bonus QAT once-removed, is 0.49 standard deviations and also not statistically different from zero. In terms of QAT participation, the estimated RD effect on QAT 6 takers is –2.6 students;

this effect is not statistically different from zero. Likewise, we find no evidence of a RD effect on QAT takers in QAT 7. In QAT 8, we find a significant, negative RD effect on the number of QAT takers—the finding is however not robust to setting the bandwidth to twice the optimal bandwidth for the estimation.[17]

The evidence collectively suggests that, despite their size, group-based teacher bonuses do not provide sufficiently strong incentives for marginal nonqualifiers to raise either QAT performance or participation in order to increase their chances for qualifying for bonuses in the next award round. It is plausible that given that PEF has revised the bonus eligibility criteria for every award round (without communicating the revisions well in advance), it may have generated uncertainty on the applicable criteria for the next award round, discouraging the effort of marginal disqualifiers to raise learning and test participation. Independent of this conjecture, it is plausible that gains in learning and test participation by marginal qualifiers may have been masked by gains in the corresponding measures among marginal nonqualifiers. Estimation issues may also be relevant: the inability to detect any effects at the composite score cutoff could be due to inadequate statistical power.

8. Conclusion

Low student learning is a consistent finding in much of the developing world and the search for solutions is an active research and public policy agenda. This paper uses a relatively unique dataset of five semiannual rounds of standardized test data to characterize and explain the short-term evolution of student learning in primary and secondary schools. The data are collected as part of the quality assurance system for a public-private partnership program which offers public subsidies conditional on minimum learning levels to low-cost private schools in Pakistan. Apart from a large positive distributional shift in learning between the first two test rounds, the learning distributions over test rounds show little progressive movement. School-level panel regressions essentially show that pre-program learning levels and the level of the school are associated with the level and evolution of conditional mean learning. Variance component decomposition shows that between 60–70% of the cross-sectional student-level test score variation is attributable to between-student variation, with no discernible trend in either the level or component shares of

[17] The effect remains statistically significant when the selected bandwidth is set at half the optimal bandwidth.

variation over test rounds. The share of test score variation attributable to between-grade variation is only somewhat smaller than the share attributable to between-school variation.

Schools are ejected from the program if they fail to achieve a minimum pass rate in the test in two consecutive attempts, making the test high stakes. Sharp regression discontinuity (RD) estimates show that the threat of program exit on marginal first-time failers induces large learning gains. The large change in learning between the first two test rounds is likely importantly attributable to this accountability pressure given that a large share of new program entrants failed in the first test round. Schools also qualify for substantial annual teacher bonuses if they de facto achieve a minimum score in a composite measure of student test participation and mean test score. Sharp RD estimates however do not show that the prospect of future teacher bonus rewards induces learning gains for marginal bonus nonqualifiers. Thus, the evidence collectively suggests that, apart from the pressure from below to maintain a minimum level of learning for program participation, program schools do not face any effective incentives to *continuously* raise learning.

References

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2008. A dime a day: The possibilities and limits of private schooling in Pakistan. *Comparative Education Review* 52(3):

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2009. Do value added estimates add value: Accounting for learning dynamics. Policy Research Working Paper No. 5066. Washington, DC: World Bank.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, Tara Vishwanath, Tristan Zajonc, and the LEAPS team. 2007. *Pakistan Learning and Education Achievements in Punjab Schools (LEAPS): Insights to inform the education policy debate*. Washington, DC: World Bank.

Araar, Abdelkrim, and Jean-Yves Duclos. 2009. "User Manual for Stata Package *DASP: Version 2.1*", PEP, World Bank, UNDP and Université Laval.

Baltagi, Badi. 2005. *Econometric analysis of panel data*. New York: Wiley.

Barrera-Osorio, Felipe, and Dhushyanth Raju. 2009. Evaluating a test-based subsidy program for low-cost private schools: Regression-discontinuity evidence from Pakistan. Manuscript.

Bloom, Benjamin S. 1956. *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain.* New York: David McKay Co Inc.

Chay, Kenneth Y., Patrick J. McEwan, and Miguel Urquiola. 2005. The central role of noise in evaluating interventions that use test scores to rank schools. *American Economic Review* 95(4):1237-1258.

Cheng, Ming-Yen, Jianqing Fan, and J. S. Marron. 1997. On automatic boundary corrections. *Annals of Statistics* 25(4):1691-1708.

Chiang, Hanley. 2009. How accountability pressure on failing schools affects student achievement. *Journal of Public Economics* 93(9-10): 1045-1057.

Das, Jishnu, Priyanka Pandey, and Tristan Zajonc. 2006. Learning levels and gaps in Pakistan. Policy Research Working Paper No. 4067. Washington DC: World Bank.

Davidson, Russell. 2009. Testing for restricted stochastic dominance: Some further results. *Review of Economic Analysis* 1:34-59.

Davidson, Russell, and Jean-Yves Duclos. 2006. Testing for restricted stochastic dominance. Discussion Paper No. 2047. Bonn: IZA.

Fan, Jianqing, and Irene Gijbels. 1996. *Local polynomial modeling and its applications*. New York: Chapman & Hall.35

Figlio, David N., and Cecilia Elena Rouse. 2006. Do accountability and voucher threats improve low-performing schools. *Journal of Public Economics* 90:239-255.

Fuji, Daisuke, Guido Imbens, and Karthik Kalyanaraman. 2009. Notes for Matlab and Stata Regression Discontinuity Software. Manuscript. http://www.economics.harvard.edu/faculty/imbens/files/rd_software_09aug4.pdf

Gauri, Varun, and Ayesha Vawda. 2003. Vouchers for basic education in developing countries. Policy Research Working Paper No. 3005. Washington, DC: World Bank.

Glewwe, Paul, and Michael Kremer. 2006. Schools, teachers, and education outcomes in developing countries. Chapter 16 in *Handbook of Economics of Education*, ed. Eric A. Hanushek and Finish Welch. Volume 2. Elsvier: 945-1017.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2003. Teacher incentives. NBER Working Paper No. 9671. Cambridge, MA: NBER.

Glewwe, Paul, Alaka Holla, and Michael Kremer. 2009. Teacher incentives in the developing world. In *Performance Incentives: Their Growing Impact on American K-12 Education*, edited by Matthew G. Springer. Washington, DC: Brookings Institution Press, pp. 295-325.

Government of Pakistan (Ministry of Education). 2009. *National Education Policy 2009*. Islamabad: Government of Pakistan.

Goyal, Sangeeta, and Corinne Siaens. 2009. Findings from the Bhutan Learning Quality Survey. South Asia Human Development Unit Report No. 21. Washington DC: World Bank.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69(1):201-209.

Hanushek, Eric A., and Ludger Wößmann. 2007. The role of education quality in economic growth. Policy Research Working Paper No. 4122. Washington, DC: World Bank.

Imbens, Guido, and Karthik Kalyanaraman. 2009. Optimal bandwidth choice for the regression discontinuity estimator. Working Paper No. 3995. Bonn: IZA.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47(1): 5-86.

Imbens, Guido W., and Thomas Lemieux. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142(2):615-635.

Jacob, Brian A. 2007. Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments. National Bureau of Economic Research Working Paper No. 12817.

Jacob, Brian A. 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.

Kane, Thomas J., and Douglas O. Staiger. 2002. The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16(4):91-114.

Lavy, Victor. 2009. Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review* 99(5):1979-2021.

Lavy, Victor. 2002. Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of Political Economy* 110(6):1286-1317.

Malik, Allah Bakhsh. 2007. *Freedom of Choice: Affordable Quality Education in Public-Private Partnership*. Lahore: Maqbool Academy.

McCrary, Justin. 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142(2):698-714.

Muralidharan, Karthik and Venkatesh Sundararaman. 2009. Teacher performance pay: Experimental evidence from India. NBER Working Paper No. 15323. Cambridge: NBER.

Porter, Jack. 2002. Asymptotic bias and optimal convergence rates for semiparametric kernel estimators in the regression discontinuity model. Harvard Institute of Research Working Paper No. 1989.

Rockoff, Jonah E., and Lesley J. Turner. 2008. Short run impacts of accountability on school quality. Manuscript.

Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio. 2007. Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. Education Research Section Working Paper No. 24. Princeton: Princeton University.

Schaffer, Mark E., and Steven Stillman. 2010. xtoverid: Stata module to calculate tests of overidentifying restrictions after xtreg, xtivreg, xtivreg2 and xthtaylor. http://ideas.repec.org/c/boc/bocode/s456779.html.

Social Policy and Development Centre. 2003. *Social Development in Pakistan: Annual Review 2002-03*. Karachi: Social Policy Development Centre.

Thistlethwaite, D., and D. Campbell. 1960. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology* 51:309-317.

Trochim, William. 1984. *Research design for program evaluation: The regression-discontinuity approach*. Beverly Hills CA: Sage Publications.

van der Klaauw, Wilbert. 2007. Regression-discontinuity analysis. Forthcoming in *The New Palgrave Dictionary of Economics*.37

West, Martin R., and Paul E. Peterson. 2006. The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments. *Economic Journal* 116:C46-C62.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

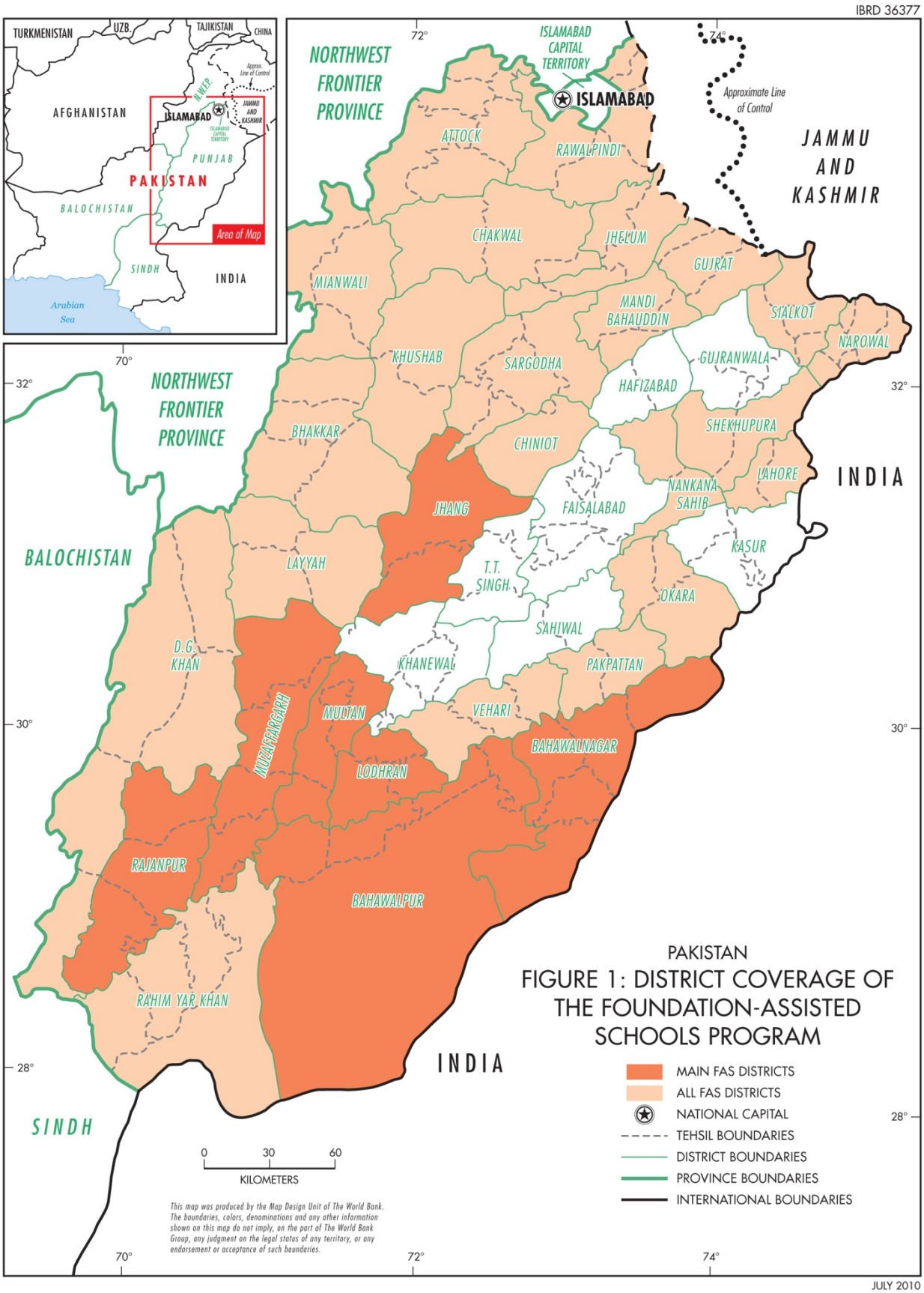

IBRD 36377
TURKMENISTAN
UZB.
TAJIKISTAN
CHINA
AFGHANISTAN
ISLAMABAD
JAMMU AND KASHMIR
PUNJAB
PAKISTAN
BALOCHISTAN
SINDH
INDIA
Arabian Sea
Area of Map
NORTHWEST FRONTIER PROVINCE
ISLAMABAD CAPITAL TERRITORY
ISLAMABAD
Approximate Line of Control
JAMMU AND KASHMIR
ATTOCK
RAWALPINDI
CHAKWAL
JHELUM
GUJRAT
MIANWALI
MANDI BAHAUDDIN
SIALKOT
NAROWAL
KHUSHAB
SARGODHA
GUJRANWALA
HAFIZABAD
SHEKHUPURA
BHAKKAR
CHINIOT
LAHORE
INDIA
NANKANA SAHIB
JHANG
FAISALABAD
KASUR
BALOCHISTAN
LAYYAH
T.T. SINGH
OKARA
SAHIWAL
D.G. KHAN
KHANEWAL
PAKPATTAN
MUZAFFARGARH
MULTAN
VEHARI
BAHAWALNAGAR
LODHRAN
RAJANPUR
BAHAWALPUR
RAHIM YAR KHAN
INDIA
SINDH
70°
72°
74°
32°
30°
28°
0 30 60
KILOMETERS
This map was produced by the Map Design Unit of The World Bank. The boundaries, colors, denominations and any other information shown on this map do not imply, on the part of The World Bank Group, any judgment on the legal status of any territory, or any endorsement or acceptance of such boundaries.
PAKISTAN
FIGURE 1: DISTRICT COVERAGE OF THE FOUNDATION-ASSISTED SCHOOLS PROGRAM
MAIN FAS DISTRICTS
ALL FAS DISTRICTS
NATIONAL CAPITAL
TEHSIL BOUNDARIES
DISTRICT BOUNDARIES
PROVINCE BOUNDARIES
INTERNATIONAL BOUNDARIES
JULY 2010

Table 1: Summary statistics for FAS program schools

Characteristic	Phase						
	1	2	3	4	5	6	All
Share primary	0.023	0.039	0.030	0.063	0.227	0.126	0.095
Share middle	0.295	0.406	0.678	0.760	0.671	0.575	0.642
Share secondary	0.682	0.555	0.291	0.177	0.102	0.299	0.263
Share girls only	0.068	0.047	0.009	0.010	0.007	0.023	0.016
Share boys only	0.068	0.023	0.015	0.012	0.007	0.026	0.017
Share coeducational	0.864	0.930	0.976	0.978	0.987	0.951	0.966
Share rural	0.636	0.547	0.524	0.575	0.628	0.661	0.591
Share urban	0.364	0.453	0.476	0.425	0.372	0.339	0.409
Share registered	1.000	1.000	1.000	1.000	0.997	[redacted]	0.758
Enrollment, boys	315.636	327.742	295.837	265.218	195.076	202.210	251.713
Enrollment, girls	320.091	316.031	236.109	206.308	150.987	156.360	203.247
Enrollment, total	635.727	643.773	531.946	471.527	346.063	358.570	454.961
Girls-boys enrollment ratio	1.495	1.029	0.910	0.974	0.947	1.253	1.034
Teachers	24.114	24.664	19.391	17.129	12.477	13.986	16.877
Classrooms	21.386	22.492	18.648	16.320	12.020	12.075	15.734
Student-teacher ratio	26.772	26.243	27.496	27.540	27.653	26.125	27.094
Student-classroom ratio	29.982	28.558	28.392	28.847	28.837	30.900	29.229
Schools *N*	44	128	460	412	304	428	1776
Districts *N*	6	13	8	8	8	22	29

Figure 2. QAT administration in FAS program schools

Table 2. Main combinations of grades test in QAT

Grade									Proportion of schools
2	3	4	5	6	7	8	9	10	
					1. QAT 4				
—	—	X	—	—	X	—	—	—	42.2
—	—	—	X	—	X	—	X	—	36.6
—	—	X	X	—	—	—	—	—	11.9
—	—	—	X	—	X	—	—	—	2.1
					2. QAT 5				
—	—	X	—	—	X	—	—	—	46.5
—	—	X	—	X	—	—	X	—	22.5
—	—	X	—	X	—	—	—	—	14.7
—	X	X	—	—	—	—	—	—	9.1
					3. QAT 6				
X	—	X	—	—	X	—	—	—	52.6
X	—	—	—	—	X	—	X	—	30.7
X	—	X	—	—	—	—	—	—	8.2
X	—	X	—	X	—	—	—	—	4.0
					4. QAT 7				
—	X	X	—	—	X	—	—	—	52.2
—	X	—	—	—	X	—	X	—	31.3
—	X	X	—	—	—	—	—	—	7.9
—	X	X	—	X	—	—	—	—	4.2
					5. QAT 8				
X	—	—	X	—	—	X	—	—	54.5
—	—	—	X	—	—	X	—	X	23.6
X	—	—	X	—	—	—	—	—	11.1
X	—	—	X	—	X	—	—	—	5.8

Notes: *N*: QAT 4: 675; QAT 5: 1083; QAT 6: 1082; QAT 7: 1079; QAT 8: 1324.

Table 3. FAS program school participation in the QAT, phase-3 and phase-4 program schools

Phase	Program entrants	QAT							
		1	2	3	4	5	6	7	8
3	482	—	—	—	479	479	479	479	464
4	425	—	—	—	—	422	422	422	413
Total	907	—	—	—	479	901	901	901	877

Table 4. Number of students tested in QAT, by grade, all program schools

Grade	QAT				
	4	5	6	7	8
2	—	—	44,895	165	45,638
3	—	1,757	244	39,682	31
4	10,259	25,703	20,088	19,810	1,021
5	9,876	—	—	—	37,095
6		10,537	1,028	1,114	15
7	11,774	6,832	21,700	21,575	1,090
8	—	—	4	8	22,919
9	8,703	7,849	10,225	12,901	111
10	—	—	266	102	9,525
Total	40,612	52,678	98,450	95,365	117,445

Table 5. Summary statistics for mean QAT scores for FAS program schools, by QAT, subject, and phase of program entry

QAT	Observations	All core subjects		English		Math		Urdu	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
1. Phase-3 program schools									
4	442	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
5	479	2.90	1.10	2.44	1.27	2.19	1.00	2.11	0.88
6	479	2.97	1.19	2.79	1.04	2.47	0.97	1.70	1.05
7	479	2.55	1.17	1.93	1.14	2.12	1.18	1.90	0.90
8	464	2.61	1.15	2.60	1.14	2.14	1.02	1.39	0.91
2. Phase-4 program schools									
5	422	2.51	1.17	2.19	1.30	1.73	1.14	1.89	0.91
6	422	2.82	1.26	2.74	1.09	2.34	1.09	1.54	1.08
7	422	2.61	1.12	1.78	1.12	2.26	1.10	2.05	0.87
8	413	2.87	1.08	2.94	1.02	2.21	1.00	1.57	0.88

Notes: Mean QAT scores are normalized by using the distribution for QAT 4 mean scores for phase-3 FAS program schools.

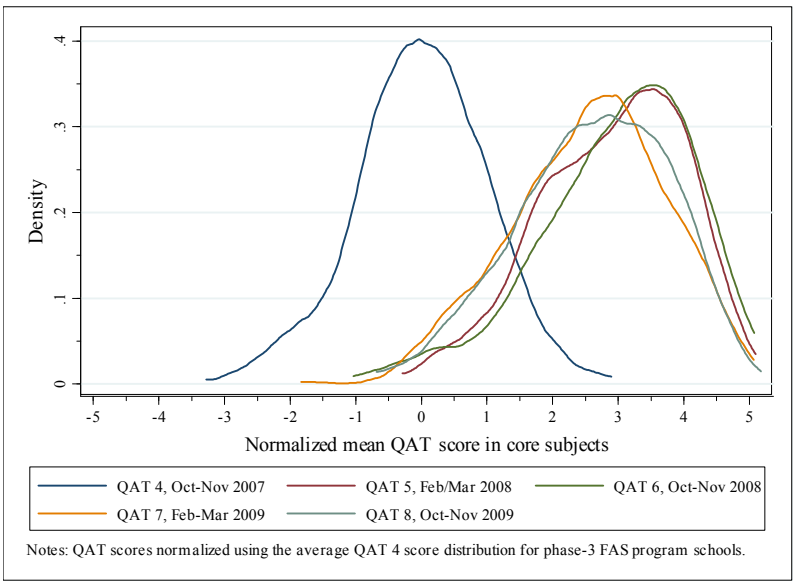


Figure 3. PDFs of mean QAT scores in core subjects, by QAT, Phase-3 FAS program schools

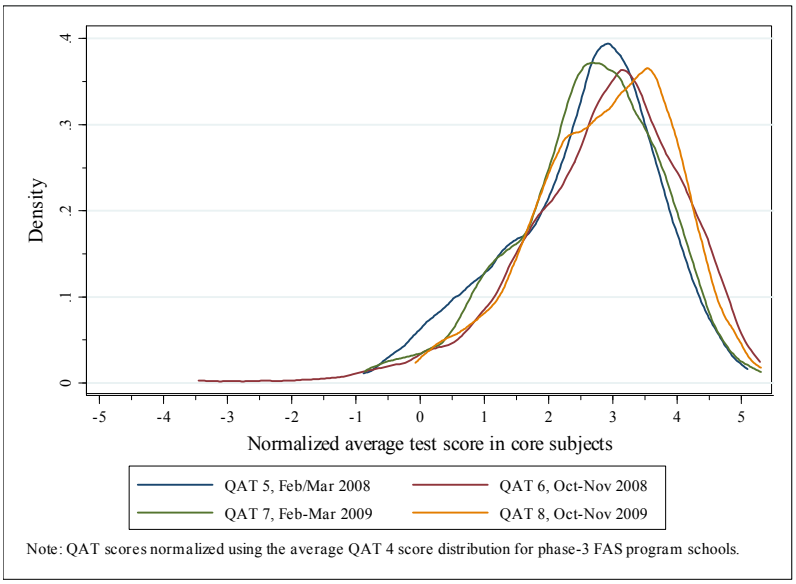


Figure 4. PDFs of mean QAT scores in core subjects, by QAT, Phase-4 FAS program schools

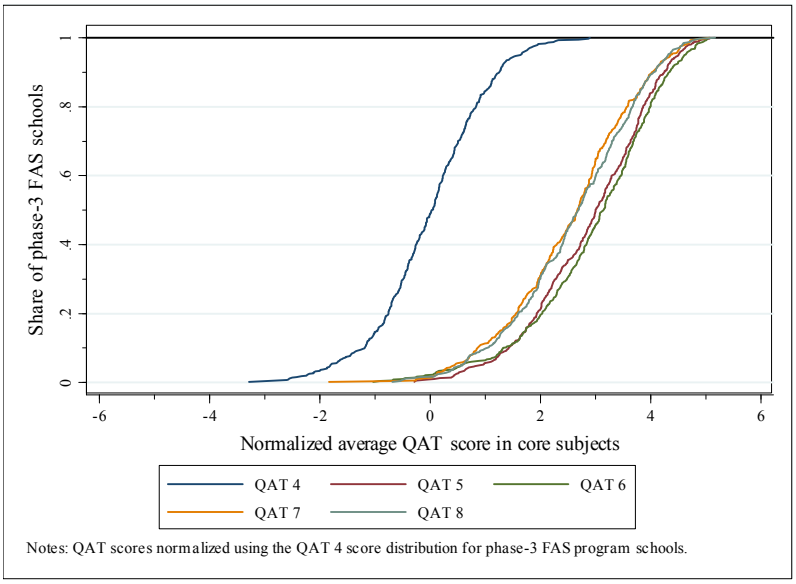


Figure 5. CDFs of mean QAT scores in core subjects, by QAT, phase-3 FAS program schools

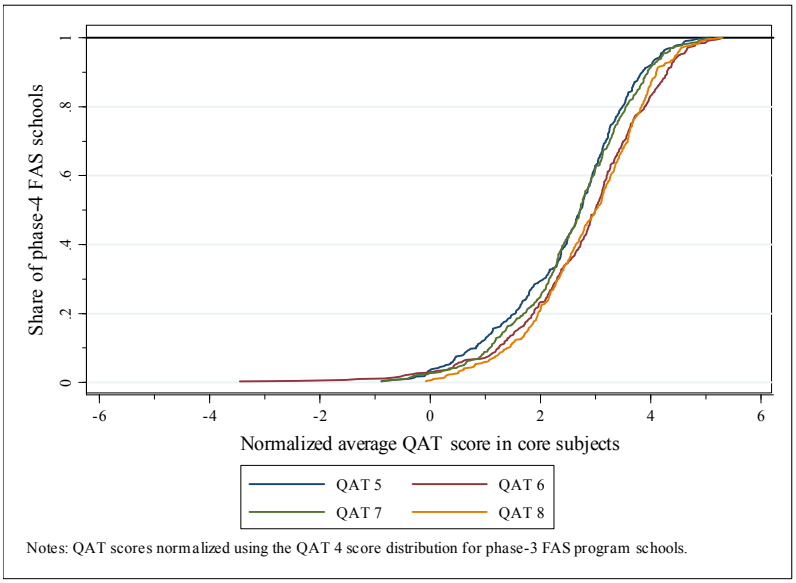


Figure 6. CDFs of mean QAT scores in core subjects, by QAT, phase-4 FAS program schools

Table 6. Decomposition of student QAT score variability

Decomposition dimension	(1)	(2)	(3)	(4)	(5)
	Proportion of total variance				
	QAT 4	QAT 5	QAT 6	QAT 7	QAT 8
1. Phase-3 FAS program schools					
2-level variance-components model					
Between schools	0.290	0.361	0.351	0.341	0.283
Within schools, between children	0.710	0.639	0.649	0.659	0.717
Total variation	320.55	282.23	359.24	343.94	294.78
4-level variance-components model					
Between districts	0.013	0.013	0.047	0.045	0.036
Within districts; between schools	0.196	0.277	0.255	0.250	0.200
Within districts and schools; between grades	0.151	0.152	0.164	0.160	0.167
Within districts, schools, and grades; between children	0.640	0.548	0.534	0.544	0.597
Total variation	318.38	284.46	366.08	355.72	302.58
N	23,092	21,414	43,910	43,186	46,532
2. Phase-4 FAS program schools					
2-level variance-components model					
Between schools	—	0.353	0.355	0.315	0.262
Within schools; between children	—	0.647	0.645	0.685	0.738
Total variation	—	339.85	410.52	339.13	269.43
3-level variance-components model					
Between districts	—	—	0.017	0.023	0.008
Within districts; between schools	—	0.270	0.269	0.223	0.189
Within districts and schools; between grades	—	0.156	0.174	0.184	0.186
Within districts, schools, and grades; between children	—	0.574	0.541	0.570	0.617
Total variation	—	338.11	411.01	341.67	274.69
N	—	15,958	32,493	30,747	35,137

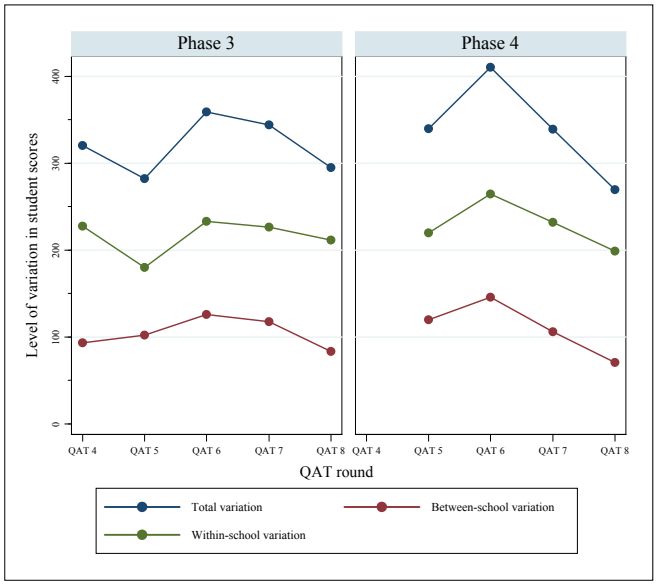


Figure 7. Evolution of variation levels in student QAT scores, by phase

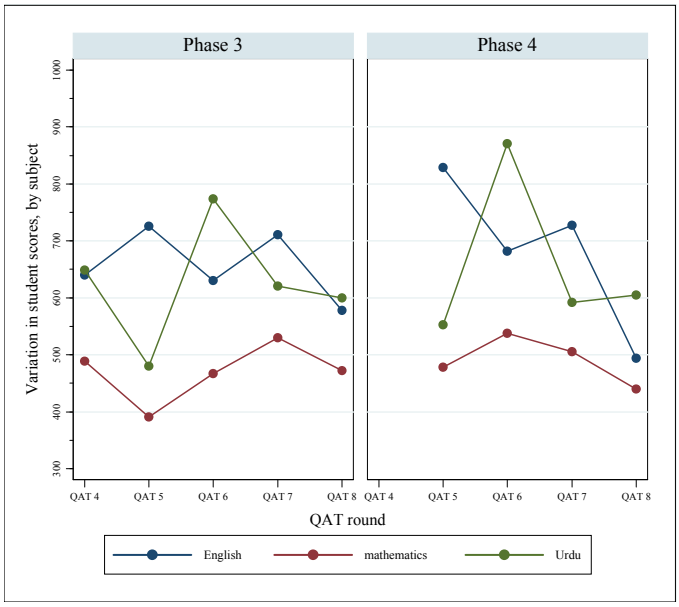


Figure 8. Evolution of variation levels in student QAT scores,
by core subject and program entry phase

Table 7. Learning growth curve regression estimation results, phase-3 FAS program schools

Dependent variable: Mean QAT score in core subjects			
Covariate	(1)	(2)	(3)
	OLS	School fixed-effects	Random-effects
QAT 5	4.151***	4.112***	4.173***
	(0.452)	(0.460)	(0.448)
QAT 6	4.953***	4.950***	5.011***
	(0.525)	(0.525)	(0.500)
QAT 7	4.128***	4.061***	4.129***
	(0.511)	(0.514)	(0.479)
QAT 8	5.282***	5.043***	5.111***
	(0.441)	(0.459)	(0.486)
Middle	0.073	—	0.045
	(0.245)		(0.195)
QAT 5 × Middle	–0.432	–0.365	–0.426*
	(0.294)	(0.299)	(0.245)
QAT 6 × Middle	–0.516	–0.463	–0.506*
	(0.317)	(0.310)	(0.275)
QAT 7 × Middle	–0.689**	–0.622**	–0.683***
	(0.275)	(0.278)	(0.262)
QAT 8 × Middle	–0.665**	–0.584**	–0.630**
	(0.261)	(0.282)	(0.267)
Secondary	0.061	—	0.025
	(0.255)		(0.208)
QAT 5 × Secondary	–0.850***	–0.780**	–0.836***
	(0.303)	(0.307)	(0.261)
QAT 6 × Secondary	–0.769**	–0.756**	–0.785***
	(0.331)	(0.323)	(0.293)
QAT 7 × Secondary	–1.213***	–1.149***	–1.203***
	(0.291)	(0.291)	(0.279)
QAT 8 × Secondary	–1.304***	–1.316***	–1.343***
	(0.284)	(0.302)	(0.284)
Coeducational	0.065	—	0.042
	(0.129)		(0.123)
QAT 5 × Coeducational	–0.050	–0.060	–0.050
	(0.153)	(0.152)	(0.153)
QAT 6 × Coeducational	–0.076	–0.094	–0.075
	(0.176)	(0.173)	(0.171)
QAT 7 × Coeducational	–0.011	–0.016	–0.005
	(0.156)	(0.153)	(0.163)
QAT 8 × Coeducational	0.213	0.216	0.226
	(0.175)	(0.178)	(0.167)
Registered	0.351*	—	0.357**
	(0.185)		(0.158)

Table 7. Learning growth curve regression estimation results, phase-3 FAS program schools

Dependent variable: Mean QAT score in core subjects			
Covariate	(1)	(2)	(3)
	OLS	School fixed-effects	Random-effects
QAT 5 × Registered	–0.030	–0.129	–0.046
	(0.228)	(0.224)	(0.194)
QAT 6 × Registered	–0.306	–0.417*	–0.319
	(0.241)	(0.237)	(0.213)
QAT 7 × Registered	–0.060	–0.157	–0.070
	(0.232)	(0.226)	(0.206)
QAT 8 × Registered	–0.345*	–0.284	–0.234
	(0.205)	(0.213)	(0.211)
School size	–0.000	—	–0.000
	(0.000)		(0.001)
QAT 5 × School size	0.001**	0.002***	0.002**
	(0.001)	(0.001)	(0.001)
QAT 6 × School size	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)
QAT 7 × School size	0.000	0.001	0.000
	(0.001)	(0.001)	(0.001)
QAT 8 × School size	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)
Teachers	0.016	—	0.012
	(0.020)		(0.019)
QAT 5 × teachers	–0.031	–0.036	–0.033
	(0.023)	(0.023)	(0.024)
QAT 6 × teachers	–0.007	–0.011	–0.010
	(0.026)	(0.026)	(0.026)
QAT 7 × teachers	–0.002	–0.006	–0.004
	(0.024)	(0.024)	(0.025)
QAT 8 × teachers	0.017	0.013	0.016
	(0.026)	(0.026)	(0.026)
Classrooms	0.005	—	0.008
	(0.018)		(0.017)
QAT 5 × classrooms	–0.025	–0.024	–0.024
	(0.020)	(0.020)	(0.021)
QAT 6 × classrooms	–0.025	–0.030	–0.028
	(0.024)	(0.024)	(0.024)
QAT 7 × classrooms	–0.035	–0.034	–0.034
	(0.023)	(0.024)	(0.023)
QAT 8 × classrooms	–0.076***	–0.072***	–0.073***
	(0.023)	(0.023)	(0.023)
Rural	–0.107	—	–0.102
	(0.096)		(0.094)
QAT 5 × Rural	–0.077	–0.032	–0.066
	(0.118)	(0.118)	(0.118)

Table 7. Learning growth curve regression estimation results, phase-3 FAS program schools

Dependent variable: Mean QAT score in core subjects			
Covariate	(1)	(2)	(3)
	OLS	School fixed-effects	Random-effects
QAT 6 × Rural	–0.087	–0.075	–0.106
	(0.130)	(0.128)	(0.131)
QAT 7 × Rural	–0.103	–0.064	–0.097
	(0.126)	(0.126)	(0.126)
QAT 8 × Rural	–0.117	–0.129	–0.150
	(0.127)	(0.129)	(0.128)
Mean SLQAT score	0.025***	--	0.025***
	(0.005)		(0.005)
QAT 5 × Mean SLQAT	–0.008	–0.006	–0.008
	(0.006)	(0.006)	(0.006)
QAT 6 × Mean SLQAT	–0.013**	–0.011*	–0.013**
	(0.006)	(0.006)	(0.006)
QAT 7 × Mean SLQAT	–0.008	–0.007	–0.008
	(0.006)	(0.006)	(0.006)
QAT 8 × Mean SLQAT	–0.020***	–0.019***	–0.021***
	(0.006)	(0.006)	(0.006)
Maximum teacher salary	0.000	--	0.000
	(0.000)		(0.000)
QAT 5 × Max teacher salary	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)
QAT 6 × Max teacher salary	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)
QAT 7 × Max teacher salary	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)
QAT 8 × Max teacher salary	–0.000	–0.000	–0.000
	(0.000)	--	(0.000)
Minimum teacher salary	–0.000		–0.000
	(0.000)	(0.000)	(0.000)
QAT 5 × Min teacher salary	0.000	–0.000	0.000
	(0.000)	(0.000)	(0.000)
QAT 6 × Min teacher salary	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)
QAT 7 × Min teacher salary	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
QAT 8 × Min teacher salary	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)
Constant	–2.158***	–0.059	–2.074***
	(0.384)	(0.044)	(0.369)
District dummies	Yes	Yes	Yes
N	2,123	2,123	2,123
R–squared statistic	0.565	0.713	
Schools	442	442	442

Table 7. Learning growth curve regression estimation results, phase-3 FAS program schools

Dependent variable: Mean QAT score in core subjects			
Covariate	(1)	(2)	(3)
	OLS	School fixed-effects	Random-effects

Notes: Panel robust standard errors reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.10$ (two-tailed significant tests).

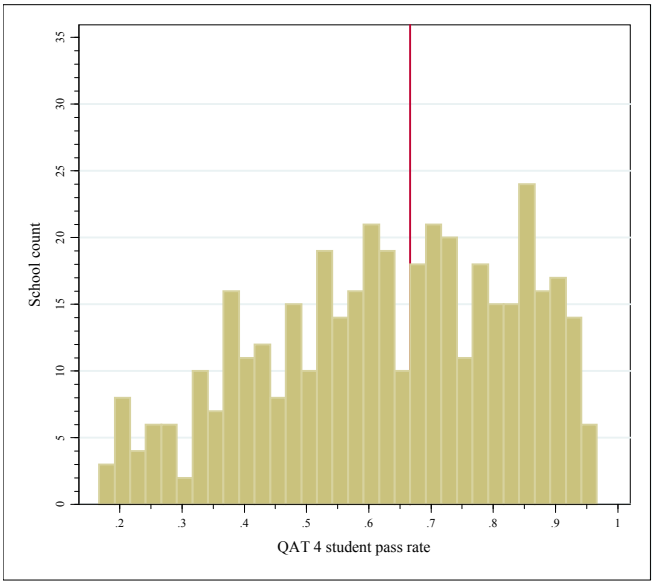


Figure 9. Distribution of schools by QAT 4 student pass rate,
Linked phase-3 program schools

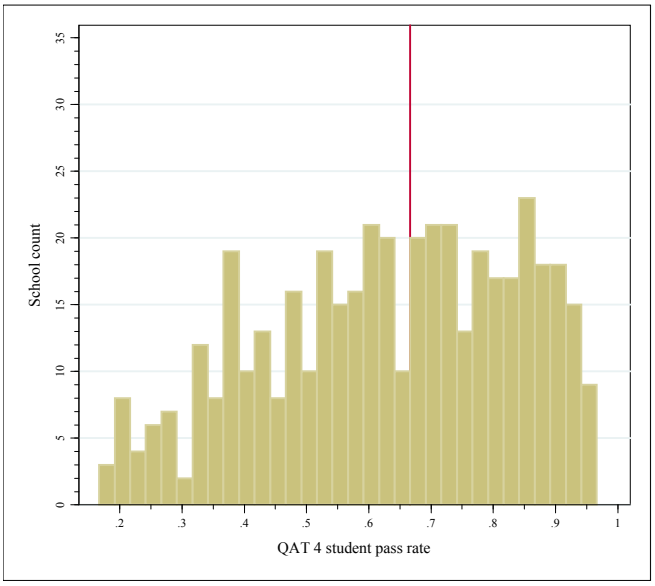


Figure 10. Distribution of schools by QAT 4 student pass rate,
All phase-3 program schools

Table 8. RD estimates of local smoothness in conditional means for pre-program covariates at the 67% QAT pass rate

Pre–program covariate	(1) Mean (s.d.)	(2) RD estimate (s.e.)
Number of students	266.883	37.199
	(160.847)	(35.451)
Number of teachers	10.686	1.019
	(5.121)	(1.041)
Maximum teacher salary (Rs.)	2985.501	307.958
	(1439.235)	(374.319)
Minimum teacher salary (Rs.)	1457.93	210.615
	(615.930)	(176.330)
Number of classrooms	9.981	0.709
	(4.904)	(1.276)
Share rural	0.525	0.060
	(0.500)	(0.113)
Share coeducational	0.813	(0.066)
	(0.391)	(0.101)
Share secondary	0.326	0.026
	(0.469)	(0.121)
Share registered	0.904	0.074
	(0.295)	(0.069)
Mean SLQAT score	52.516	0.170
	(10.121)	(2.305)

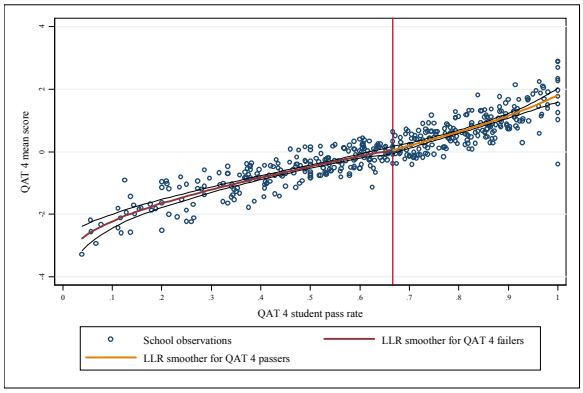
Notes: N=427 for all covariates. *** p<0.01, ** p<0.05, * p<0.10 (two-tailed significant tests). Standard deviations in parentheses in (1). Analytical standard errors in parentheses in (2). I-K method derived optimal bandwidths range from 18 to 30 percentage points, depending on the covariate.

Table 9. RD estimates of the stick effect on outcomes for marginal failers, phase-3 FAS program schools

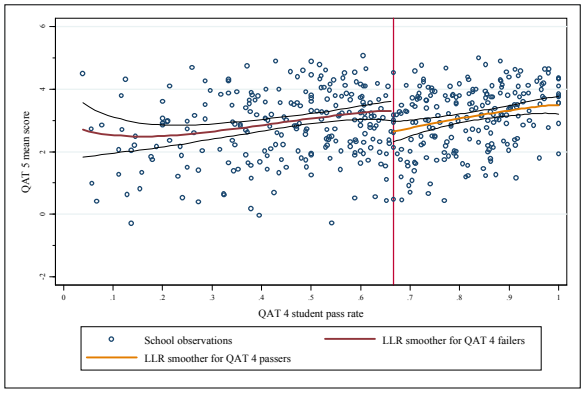
QAT	(1) Mean QAT score	(2) QAT takers
QAT 4 (pre-threat)	–0.003 (0.775)	−6.077 (6.317)
QAT 5 (threat/ultimatum round)	0.664 *** (0.232)	−4.213 (5.961)
QAT 6 (immediate post-threat, one degree separation)	0.561 *** (0.209)	−5.667 (11.055)
QAT 7 (post-threat, two degree separation)	0.349 (0.265)	−0.997 (10.148)
QAT 8 (post-threat, three degree separation)	0.089 (0.249)	−1.080 (14.888)

Notes: *** $p<0.01$, ** $p<0.05$, * $p<0.10$ (two-tailed significant tests). Analytical standard errors in parentheses. I-K method derived bandwidths range between 10 and 37 percentage points, depending on the specific outcome measure.

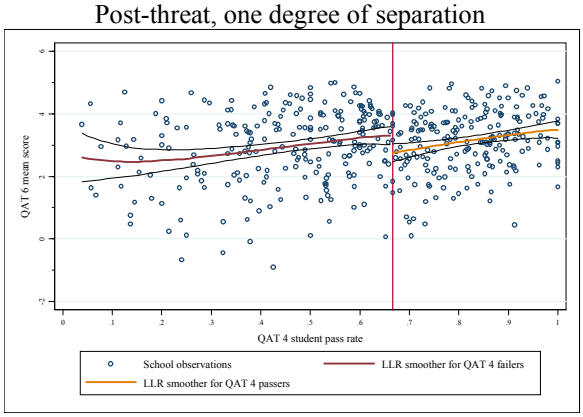
1. QAT 4
Pre-threat



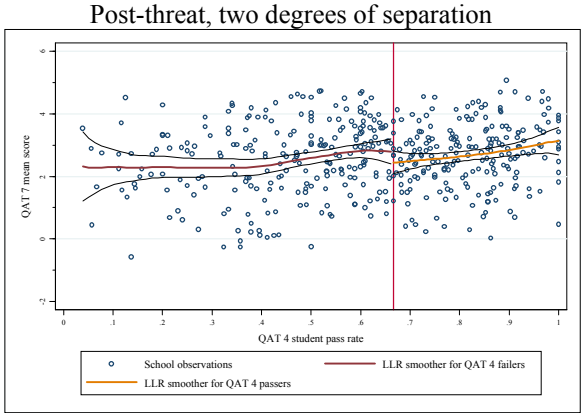
2. QAT 5
Threat



3. QAT 6
Post-threat, one degree of separation



4. QAT 7
Post-threat, two degrees of separation



5. QAT 8
Post-threat, three degrees of separation

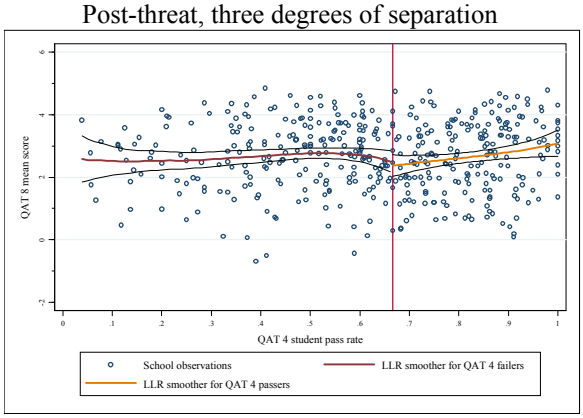
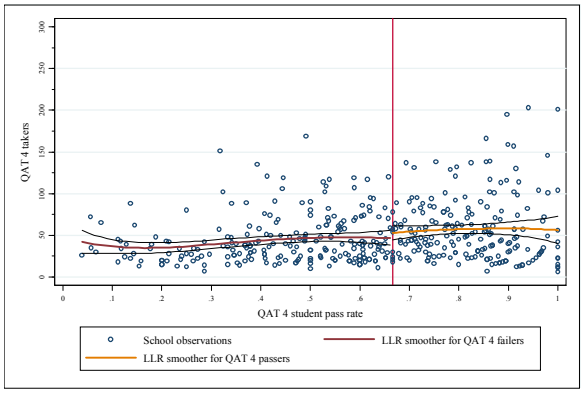
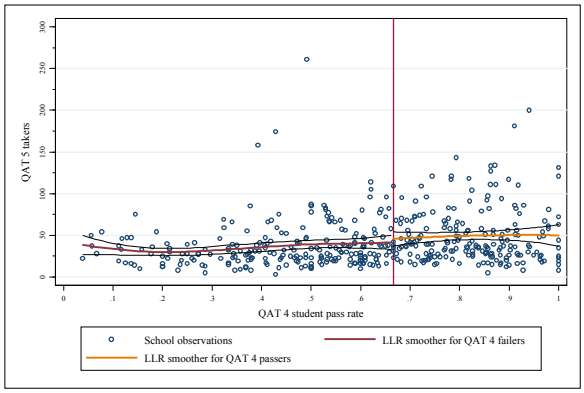


Figure 11. LLR plots of the stick effect on mean QAT scores
for marginal failers, phase-3 program schools

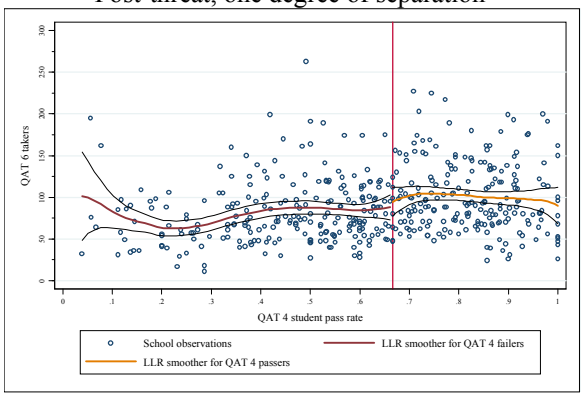
1. QAT 4
Pre-threat



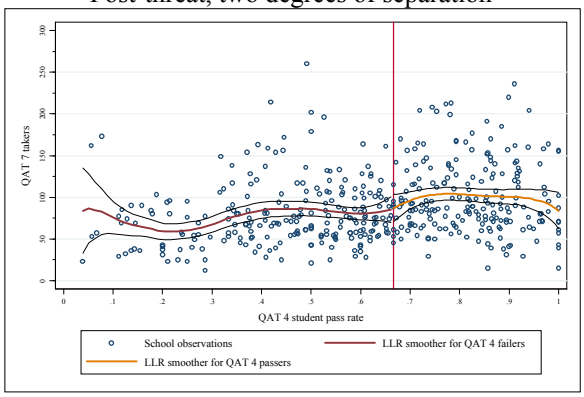
2. QAT 5
Threat



3. QAT 6
Post-threat, one degree of separation



4. QAT 7
Post-threat, two degrees of separation



5. QAT 8
Post-threat, three degrees of separation

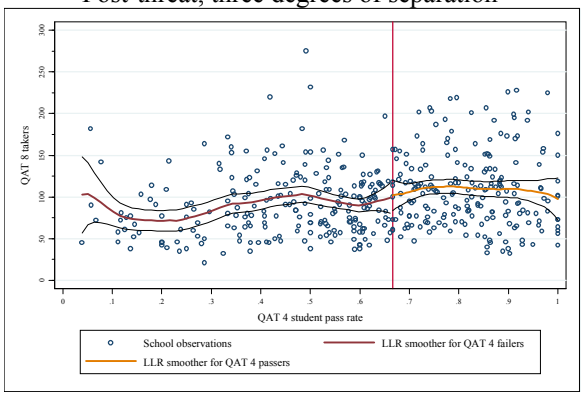


Figure 12. LLR plots of the stick effect on the number of QAT takers
for marginal failers, phase-3 program schools

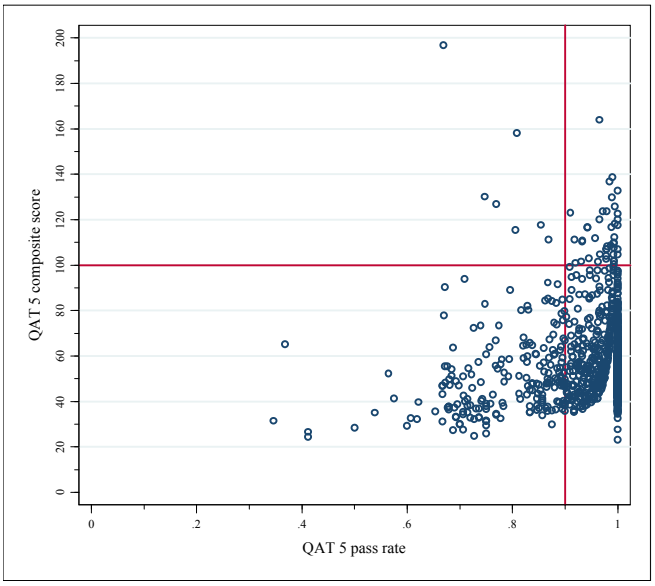


Figure 13. Scatter plot of QAT 5 composite scores against QAT 5 pass rates, FAS program schools

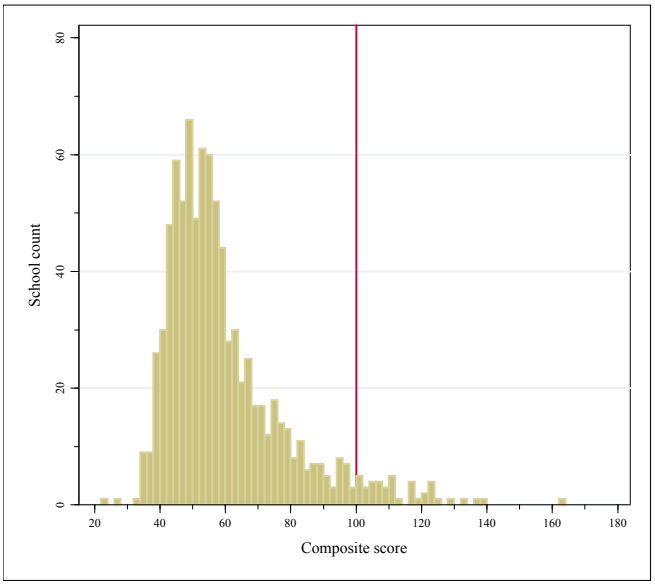


Figure 14. Distribution of schools by QAT 5 composite score, FAS program schools

Table 10. RD estimates of local smoothness in conditional means for pre-program covariates at the QAT 5 composite score cutoff

Pre-program covariate	(1) Mean (s.d.)	(2) RD estimate (s.e.)
Number of students	255.224	−4.979
	(145.956)	(119.030)
Number of teachers	10.305	−9.560*
	(4.708)	(5.416)
Maximum teacher salary (Rs.)	2881.969	246.696
	(1383.879)	(720.944)
Minimum teacher salary (Rs.)	1418.122	186.121
	(594.040)	(251.343)
Number of classrooms	9.614	−1.515
	(4.639)	(6.816)
Share rural	0.549	1.567**
	(0.498)	(0.720)
Share coeducational	0.828	−0.767*
	(0.378)	(0.463)
Share secondary	0.236	0.184
	(0.425)	(0.133)
Share registered	0.850	0.225
	(0.357)	(0.156)
Mean SLQAT score	52.139	–2.236
	(10.225)	(9.307)

Notes: Sample size varies between 658 and 688 schools, depending on the covariate. *** $p<0.01$, ** $p<0.05$, * $p<0.10$ (two-tailed significant tests). Standard deviations in parentheses in (1). Analytical standard errors in parentheses in (2). I-K method derived optimal bandwidths in the local linear regression estimations range from 3 to 43 points on the QAT 5 composite score, depending on the pre-program covariate under examination.

Table 11. RD estimates of the carrot effect for marginal bonus nonqualifiers, FAS program schools

QAT	(1) Mean QAT score	(2) QAT takers
QAT 5 (bonus determination round)	−0.600 (0.597)	−2.555 (3.809)
QAT 6 (immediate pre-bonus distribution)	−0.869 (0.641)	−10.326 (13.188)
QAT 7 (immediate post-bonus distribution)	0.211 (0.830)	−3.779 (12.802)
QAT 8 (post-bonus distribution, once-removed)	0.492 (0.811)	−26.667* (13.840)

Notes: *** $p<0.01$, ** $p<0.05$, * $p<0.10$ (two-tailed significant tests). Analytical standard errors in parentheses. The I-K method derived bandwidth for the local linear regression estimations is roughly 5 composite score points for mean QAT scores and roughly 17 points for the number of QAT takers.

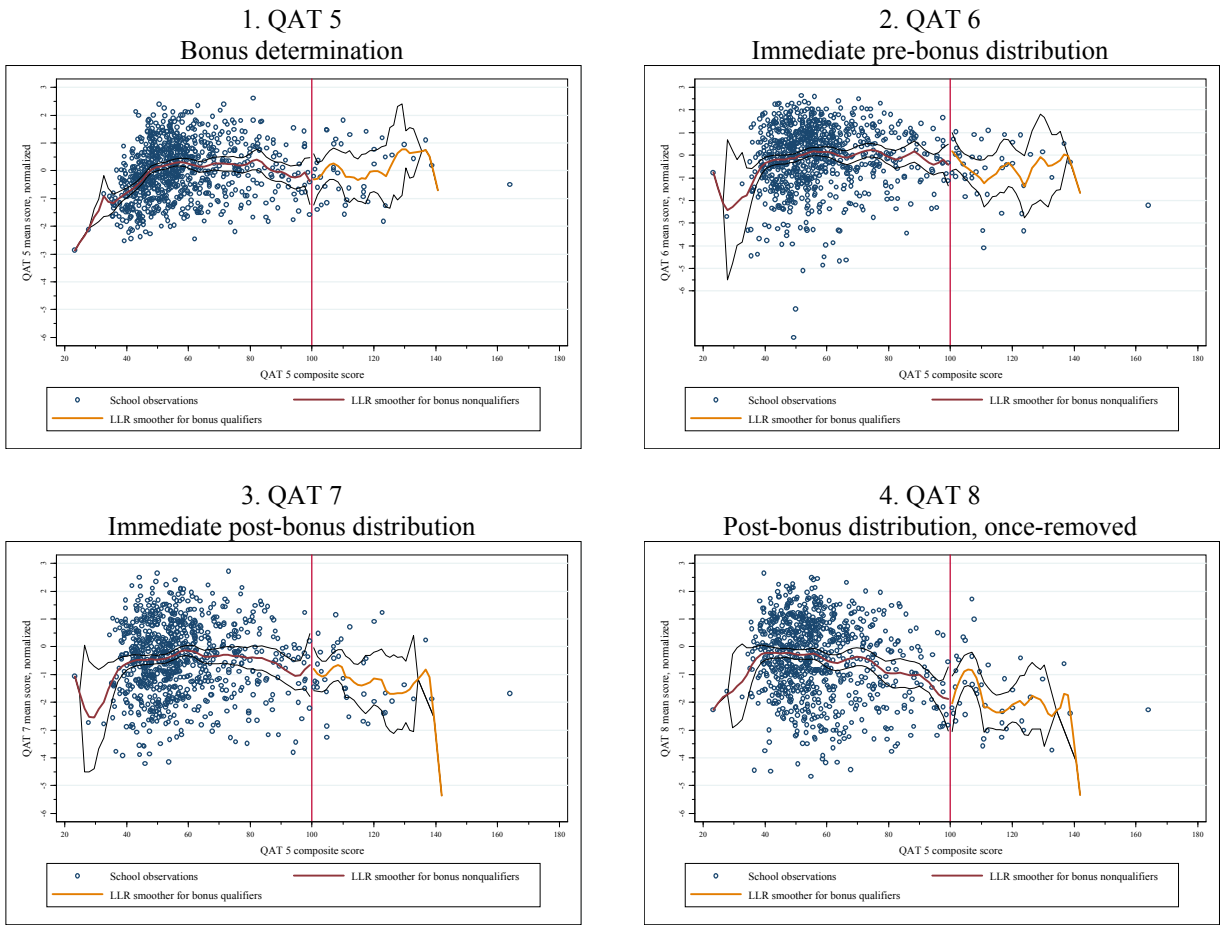


Figure 15. LLR plots of the carrot effect on mean QAT scores
for marginal bonus nonqualifiers, FAS program schools

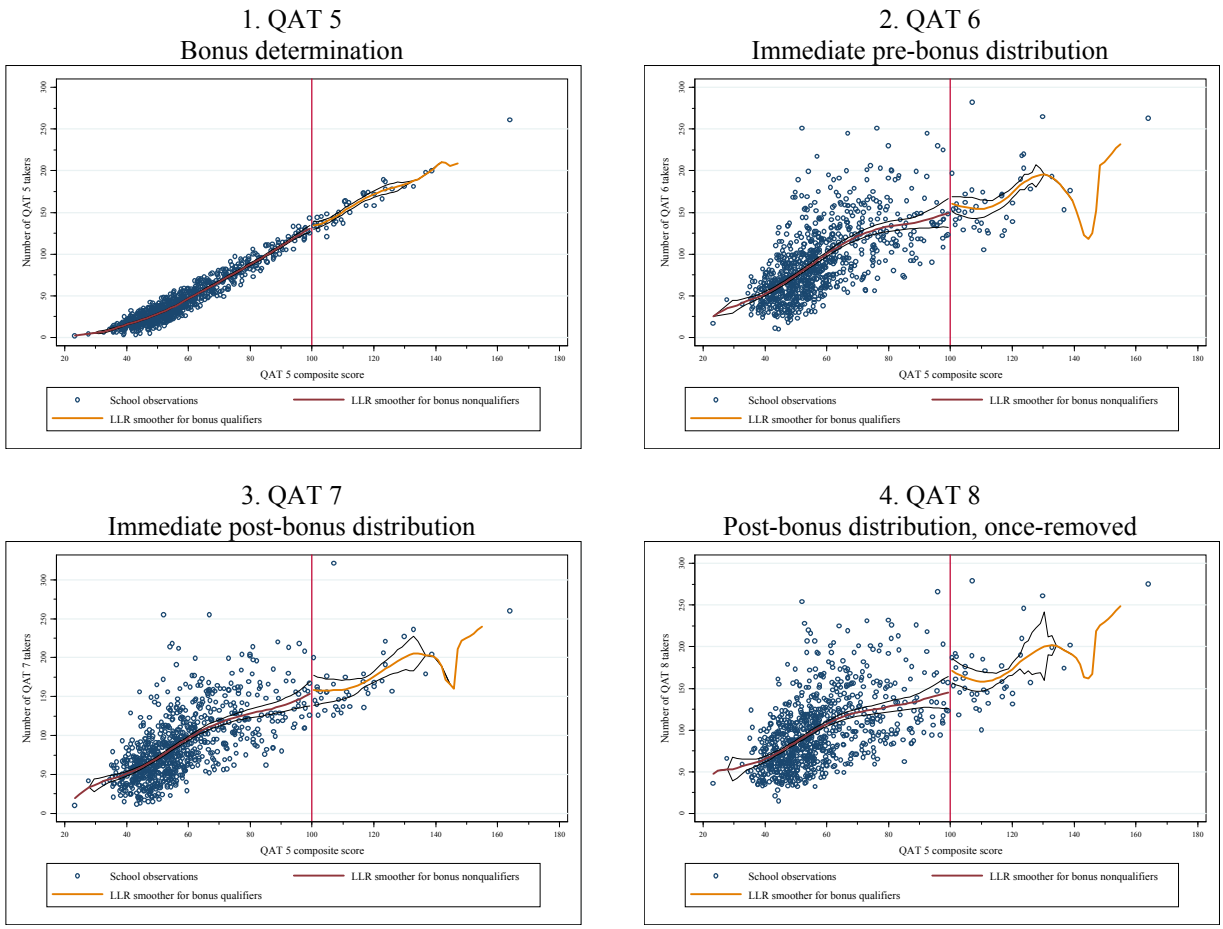


Figure 16. LLR plots of the carrot effect on QAT participation
for marginal bonus nonqualifiers, FAS program schools

Table A1. Learning growth curve regression estimation results over QAT 5–QAT8, phase-3 FAS program schools

Dependent variable: Mean QAT score in core subjects			
Covariate	(1)	(2)	(3)
	OLS	School fixed-effects	Random-effects
QAT 6	2.717***	2.735***	2.730***
	(0.392)	(0.387)	(0.557)
QAT 7	1.948***	1.905***	1.936***
	(0.376)	(0.374)	(0.555)
QAT 8	3.016***	2.808***	2.956***
	(0.350)	(0.349)	(0.560)
Middle	–0.098	—	–0.105
	(0.150)		(0.195)
QAT 6 × Middle	–0.314	–0.303	–0.312
	(0.221)	(0.213)	(0.308)
QAT 7 × Middle	–0.476***	–0.452***	–0.469
	(0.169)	(0.168)	(0.305)
QAT 8 × Middle	–0.449**	–0.409**	–0.437
	(0.174)	(0.174)	(0.308)
Secondary	–0.310*	—	–0.315
	(0.164)		(0.208)
QAT 6 × Secondary	–0.347	–0.375*	–0.352
	(0.235)	(0.226)	(0.328)
QAT 7 × Secondary	–0.760***	–0.740***	–0.756**
	(0.186)	(0.183)	(0.325)
QAT 8 × Secondary	–0.840***	–0.886***	–0.852***
	(0.198)	(0.197)	(0.330)
Coeducational	0.043	—	0.045
	(0.101)		(0.122)
QAT 6 × Coeducational	–0.054	–0.068	–0.056
	(0.132)	(0.129)	(0.191)
QAT 7 × Coeducational	0.006	0.005	0.006
	(0.118)	(0.117)	(0.190)
QAT 8 × Coeducational	0.214	0.218	0.215
	(0.142)	(0.141)	(0.193)
Registered	0.233*	—	0.240
	(0.125)		(0.153)
QAT 6 × Registered	–0.167	–0.202	–0.170
	(0.169)	(0.166)	(0.233)
QAT 7 × Registered	0.062	0.041	0.057
	(0.164)	(0.159)	(0.234)
QAT 8 × Registered	–0.209	–0.091	–0.178
	(0.154)	(0.152)	(0.240)
School size	0.000	—	0.000
	(0.000)		(0.001)
QAT 6 × School size	–0.000	0.000	–0.000

Table A1. Learning growth curve regression estimation results over QAT 5–QAT8, phase-3 FAS program schools

Dependent variable: Mean QAT score in core subjects			
Covariate	(1)	(2)	(3)
	OLS	School fixed-effects	Random-effects
	(0.001)	(0.001)	(0.001)
QAT 7 × School size	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.001)
QAT 8 × School size	–0.000	–0.000	–0.000
	(0.001)	(0.001)	(0.001)
Teachers	–0.001	—	–0.001
	(0.016)		(0.019)
QAT 6 × teachers	0.010	0.009	0.009
	(0.020)	(0.020)	(0.029)
QAT 7 × teachers	0.014	0.013	0.014
	(0.019)	(0.018)	(0.029)
QAT 8 × teachers	0.032	0.031	0.032
	(0.020)	(0.020)	(0.029)
Classrooms	–0.006	—	–0.005
	(0.014)		(0.017)
QAT 6 × classrooms	–0.013	–0.019	–0.014
	(0.017)	(0.017)	(0.026)
QAT 7 × classrooms	–0.023	–0.022	–0.022
	(0.018)	(0.018)	(0.026)
QAT 8 × classrooms	–0.060***	–0.058***	–0.059**
	(0.017)	(0.017)	(0.026)
Rural	–0.134*	—	–0.138
	(0.073)		(0.093)
QAT 6 × Rural	–0.047	–0.059	–0.050
	(0.099)	(0.096)	(0.146)
QAT 7 × Rural	–0.062	–0.048	–0.058
	(0.094)	(0.093)	(0.146)
QAT 8 × Rural	–0.073	–0.103	–0.081
	(0.099)	(0.098)	(0.147)
Mean SLQAT score	0.021***	—	0.021***
	(0.003)		(0.005)
QAT 6 × mean SLQAT	–0.011**	–0.010**	–0.011
	(0.005)	(0.005)	(0.007)
QAT 7 × mean SLQAT	–0.006	–0.006	–0.006
	(0.005)	(0.005)	(0.007)
QAT 8 × mean SLQAT	–0.017***	–0.018***	–0.018**
	(0.005)	(0.005)	(0.007)
Maximum teacher salary	0.000	—	0.000
	(0.000)		(0.000)
QAT 6 × Max teacher salary	–0.000*	–0.000	–0.000
	(0.000)	(0.000)	(0.000)
QAT 7 × Max teacher salary	–0.000	–0.000	–0.000

Table A1. Learning growth curve regression estimation results over QAT 5–QAT8, phase-3 FAS program schools

Dependent variable: Mean QAT score in core subjects			
Covariate	(1)	(2)	(3)
	OLS	School fixed-effects	Random-effects
	(0.000)	(0.000)	(0.000)
QAT 8 × Max teacher salary	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)
Minimum teacher salary	–0.000	—	–0.000
	(0.000)		(0.000)
QAT 6 × Min teacher salary	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)
QAT 7 × Min teacher salary	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
QAT 8 × Min teacher salary	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)
Constant	–2.722***	–1.287***	–2.709***
	(0.286)	(0.024)	(0.367)
District dummies	Yes	Yes	Yes
N	2,123	2,123	2,123
R-squared statistic	0.219	0.224	
Schools	464	464	464

Notes: Panel-robust standard errors reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.10$ (two–tailed significant tests).

Table A2. Learning growth curve regression estimation results over QAT 5–QAT8, phase-4 FAS program schools

Dependent variable: Mean QAT score in core subjects			
Covariate	(1)	(2)	(3)
	OLS	School fixed-effects	Random-effects
QAT 6	0.466	0.457	0.469
	(0.496)	(0.498)	(0.397)
QAT 7	0.315	0.291	0.302
	(0.379)	(0.376)	(0.397)
QAT 8	1.256***	1.216***	1.234***
	(0.420)	(0.419)	(0.401)
Middle	–0.285	—	–0.264*
	(0.176)		(0.148)
QAT 6 × Middle	0.021	–0.053	–0.013
	(0.203)	(0.201)	(0.163)
QAT 7 × Middle	–0.007	0.029	0.013
	(0.192)	(0.181)	(0.164)
QAT 8 × Middle	0.073	0.069	0.067
	(0.190)	(0.179)	(0.164)
Secondary	–0.983***	—	–0.963***
	(0.209)		(0.190)
QAT 6 × Secondary	0.333	0.303	0.327
	(0.241)	(0.238)	(0.211)
QAT 7 × Secondary	0.107	0.151	0.131
	(0.229)	(0.219)	(0.211)
QAT 8 × Secondary	–0.020	–0.070	–0.058
	(0.235)	(0.225)	(0.213)
Coeducational	–0.281**	—	–0.283**
	(0.122)		(0.131)
QAT 6 × Coeducational	0.368**	0.411**	0.394***
	(0.177)	(0.176)	(0.143)
QAT 7 × Coeducational	0.274*	0.269*	0.271*
	(0.141)	(0.140)	(0.143)
QAT 8 × Coeducational	0.207	0.207	0.205
	(0.146)	(0.145)	(0.144)
Registered	0.134	—	0.135
	(0.140)		(0.120)
QAT 6 × Registered	0.135	0.160	0.152
	(0.177)	(0.177)	(0.132)
QAT 7 × Registered	0.193	0.173	0.182
	(0.156)	(0.154)	(0.132)
QAT 8 × Registered	–0.007	0.020	0.013
	(0.153)	(0.153)	(0.133)
School size	0.000	—	0.000
	(0.001)		(0.001)
QAT 6 × School size	–0.001	–0.001	–0.001
	(0.001)	(0.001)	(0.001)

Table A2. Learning growth curve regression estimation results over QAT 5–QAT8, phase-4 FAS program schools

Dependent variable: Mean QAT score in core subjects			
Covariate	(1)	(2)	(3)
	OLS	School fixed-effects	Random-effects
QAT 7 × School size	–0.001	–0.001	–0.001
	(0.001)	(0.001)	(0.001)
QAT 8 × School size	–0.000	–0.000	–0.000
	(0.001)	(0.001)	(0.001)
Teachers	0.009	—	0.009
	(0.022)		(0.024)
QAT 6 × teachers	0.007	0.011	0.009
	(0.029)	(0.029)	(0.026)
QAT 7 × teachers	0.008	0.004	0.006
	(0.026)	(0.025)	(0.026)
QAT 8 × teachers	–0.012	–0.017	–0.015
	(0.026)	(0.026)	(0.026)
Classrooms	0.032*	—	0.031
	(0.019)		(0.020)
QAT 6 × classrooms	–0.004	0.002	–0.001
	(0.024)	(0.025)	(0.022)
QAT 7 × classrooms	–0.011	–0.008	–0.009
	(0.023)	(0.022)	(0.022)
QAT 8 × classrooms	–0.039**	–0.032*	–0.035
	(0.020)	(0.019)	(0.022)
Rural	–0.094	—	–0.106
	(0.100)		(0.098)
QAT 6 × Rural	0.012	0.071	0.042
	(0.130)	(0.131)	(0.108)
QAT 7 × Rural	0.056	0.081	0.070
	(0.103)	(0.101)	(0.108)
QAT 8 × Rural	0.016	0.052	0.038
	(0.106)	(0.105)	(0.108)
Mean SLQAT score	0.021***	—	0.021***
	(0.005)		(0.005)
QAT 6 × Mean SLQAT	–0.012*	–0.012*	–0.012**
	(0.006)	(0.006)	(0.005)
QAT 7 × Mean SLQAT	–0.008*	–0.008*	–0.008
	(0.005)	(0.005)	(0.005)
QAT 8 × Mean SLQAT	–0.011*	–0.010*	–0.011*
	(0.006)	(0.006)	(0.005)
Maximum teacher salary	–0.000	—	–0.000
	(0.000)		(0.000)
QAT 6 × Max teacher salary	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
QAT 7 × Max teacher salary	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)

Table A2. Learning growth curve regression estimation results over QAT 5–QAT8, phase-4 FAS program schools

Dependent variable: Mean QAT score in core subjects			
Covariate	(1)	(2)	(3)
	OLS	School fixed-effects	Random-effects
QAT 8 × Max teacher salary	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Minimum teacher salary	–0.000	—	–0.000
	(0.000)		(0.000)
QAT 6 × Min teacher salary	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)
QAT 7 × Min teacher salary	–0.000	–0.000	–0.000
	(0.000)	(0.000)	(0.000)
QAT 8 × Min teacher salary	–0.000	–0.000**	–0.000*
	(0.000)	(0.000)	(0.000)
Constant	–0.912**	0.013	–0.909**
	(0.426)	(0.033)	(0.459)
District dummies	Yes	Yes	Yes
N	1,596	1,596	1,596
R–squared statistic	0.170	0.085	
Schools	411	411	411

Notes: Panel robust standard errors reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.10$ (two-tailed significant tests).