NBER WORKING PAPER SERIES

SCHOOL TURNAROUNDS:
EVIDENCE FROM THE 2009 STIMULUS

Thomas Dee

Working Paper 17990
http://www.nber.org/papers/w17990

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2012

I would like to thank seminar participants at Stanford University and the University of Virginia for helpful comments including Sean Reardon, Susanna Loeb, Ed Haertel, Michael Kirst, Eric Hanushek, James Wyckoff, Daphna Bassok, Dan Player, Luke Miller, and Rick Hess. I would also like to thank Scott Latham for excellent research assistance. The usual caveats apply. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

School Turnarounds: Evidence from the 2009 Stimulus
Thomas Dee
NBER Working Paper No. 17990
April 2012
JEL No. H52,I2

ABSTRACT

The American Recovery and Reinvestment Act of 2009 (ARRA) targeted substantial School Improvement Grants (SIGs) to the nation's "persistently lowest achieving" public schools (i.e., up to $2 million per school annually over 3 years) but required schools accepting these awards to implement a federally prescribed school-reform model. Schools that met the "lowest-achieving" and "lack of progress" thresholds within their state had prioritized eligibility for these SIG-funded interventions. Using data from California, this study leverages these two discontinuous eligibility rules to identify the effects of SIG-funded whole-school reforms. The results based on these "fuzzy" regression-discontinuity designs indicate that there were significant improvements in the test-based performance of schools on the "lowest-achieving" margin but not among schools on the "lack of progress" margin. Complementary panel-based estimates suggest that these improvements were largely concentrated among schools adopting the federal "turnaround" model, which compels more dramatic staff turnover.

Thomas Dee
Frank Batten School of Leadership & Public Policy
and Department of Economics
University of Virginia
235 McCormick Road
P.O. Box 400893
Charlottesville, VA 22903
and NBER
dee@virginia.edu

1. Introduction

The challenges of improving organizational performance, particularly in the public sector, are a broad and recurring theme in the economics literature. These concerns are especially prominent with respect to public elementary and secondary schools, which are tasked with promoting human-capital development both generally but also with an emphasis on supporting the socioeconomic mobility of disadvantaged youths. In a June 22, 2009 speech, the U.S. Secretary of Education, Arne Duncan, called for a nationwide focus on "turning around" the nation's most chronically underperforming public schools (i.e., approximately 5,000 schools or 5 percent of the total), stating "we want transformation, not tinkering." The Secretary broadly outlined three different models for achieving school turnarounds in addition to the option of simply closing underperforming schools (Gewertz 2009). The U.S. Department of Education soon encouraged the implementation of these school-reform models with an unprecedented amount of funding appropriated by the American Recovery and Reinvestment Act of 2009 (ARRA). More specifically, the 2009 stimulus package added $3 billion to the $546 million already appropriated for School Improvement Grants (SIGs).

New federal guidance (U.S. Department of Education, 2010a; 2010b) subsequently outlined how states should identify their SIG-eligible schools and what would be required of schools accepting these awards. More specifically, using federal rules, states identified their "*persistently lowest-achieving*" (PLA) schools, which then had highly prioritized eligibility for SIGs (i.e., up to $2 million *per school* annually for each of three years). The PLA definition was largely restricted to schools receiving (or eligible for) Title I assistance whose baseline achievement placed them among the lowest five-percent of such schools in their state and who had made the least recent progress in raising student achievement. PLA schools that accepted a SIG were required to implement one of three federally prescribed, multi-faceted reform models (i.e., transformation, turnaround, or restart) or to close.[1]

This study examines the effects of these SIG-funded whole-school reforms using data from California, the state that made the most SIG awards. The research design used to identify

[1] Most schools, 74 percent, chose the least disruptive model, transformation, which requires replacing the principal and undertaking multiple instructional and personnel reforms (Hurlburt et al. 2011). Another twenty percent of SIG recipients who were required to implement a specific reform chose the "turnaround" model, which resembles the transformation model but also requires replacing at least 50 percent of the staff. Relatively few PLA schools (4 percent, n = 33) chose the restart option (reopening the school under the management of a charter or education management organization) and fewer still (2 percent, n = 16) used a SIG award to close. These models are described in more detail below.

these effects leverages the discontinuities in SIG eligibility generated by whether a school's baseline achievement placed them above or below the "lowest-achieving" threshold or the "lack of progress" threshold. The variation in the relevant baseline assignment variables around these two eligibility thresholds significantly influenced the likelihood of undertaking SIG-funded reforms but does not appear to be otherwise related to subsequent school performance. Therefore, these discontinuities provides compelling opportunities to identify the effects of this controversial federal program on important outcomes and to do so in specifications that support a strong causal warrant.

Understanding the effects of this stimulus-funded initiative is most obviously relevant because it is an expensive federal program designed to have broad effects on the practices and policies in schools serving the nation's most vulnerable students. However, the evidence from this study also constitutes a novel contribution to the extant literature examining earlier whole-school reforms (e.g., Borman et al. 2003). Furthermore, the SIG-funded school reforms studied here also have broader relevance because they have strong, contemporary parallels in other prominent federal initiatives that focus on similar school reforms and targeting them to the lowest-achieving schools (e.g., Race to the Top, State Fiscal Stabilization Funds, and "Priority Schools" in the NCLB waiver process). More generally, these SIG-funded reforms are a leading example of a long, historical trend in which the federal government has increasingly leveraged its comparatively small financial contribution to public K-12 education to bring about highly specified changes in school organizations and practices. Critics suggest that these bold and highly prescriptive federal requirements may constitute "counterproductive micromanagement" that stifles both true innovation and effective implementation of reforms (e.g., Hess and Darling-Hammond 2011).

The evidence from this initiative to target improvements among the least effective schools can also be situated in a broader, ongoing debate about the capacity of schools alone to act as agents of social mobility (e.g., "No Excuses" schooling versus the "Broader, Bolder" initiative). The policy initiative studied here has some relevance to this debate because it is an ambitious effort to improve school performance at some scale. However, the SIG-funded whole-school reforms could actually be viewed as a novel amalgam of elements from both the "No Excuses" and the "Broader, Bolder" perspectives in that they emphasize school leadership, culture, and effective instructional practices but also provide substantial resources to support such efforts and encourage schools to use various “social-emotional” and “community-oriented” wrap-around services to support students’ readiness to learn.

In section 2, I discuss the SIG program and school eligibility (both nationwide and in California) as well as the required school-reform models and the treatment contrast implied by the RD design. In section 3, I provide an overview of the relevant theoretical motivations and the background literature. In section 4, I discuss the data, the econometric specifications, and the various robustness checks for the validity of the RD design. Section 5 presents the results and section 6 concludes with emphasis on issues of cost-effectiveness, important external-validity caveats, and policy implications.

2. School Improvement Grants (SIG) and Whole-School Reforms

2.1 The SIG Program and School Eligibility

School Improvement Grants (SIGs) are authorized by Title I, section 1003(g), of the Elementary and Secondary Education Act (ESEA) of 1965. This program authorizes the U.S. Department of Education to provide SIGs to states for the purpose of awarding grants competitively to school districts that have eligible “schools in need of improvement” (SINI). This source of funding for school improvement was first authorized as part of the NCLB revisions to ESEA. However, it was unfunded until 2007 when it began receiving modest appropriations. Beginning in 2010, the U.S. Department of Education coupled the dramatic infusion of stimulus spending into this program with new eligibility criteria that narrowly targeted SIG funding to the nation’s most chronically underperforming schools. The revised procedures also mandated that schools accepting SIG awards undertake new and “radical steps” that were characterized as a “sharp contrast to the current free-flowing nature of Title I school improvement aid” (McNeil 2009).

During the first half of 2010, the U.S. Department of Education issued detailed guidance on how states should distribute the redesigned SIG grants and what actually would be required of schools receiving them (U.S. Department of Education, 2010a; 2010b, McMurrer, Dietz, and Rentner 2011). The first round of SIGs awarded under these criteria were in place during the 2010-11 school year. More specifically, these federal eligibility rules required state education agencies to identify their “persistently lowest-achieving” (PLA) schools and to give these schools the highest priority for SIG funding.

The pool of schools eligible for PLA status largely consisted of those receiving Title I aid and in improvement, corrective action, or restructuring under NCLB (i.e., "Tier 1" schools) and "secondary" schools eligible for but not receiving Title I aid (i.e., "Tier 2" schools).[2] State education agencies identified PLA schools from this pool using two key conditions. One is whether the school's baseline achievement in ELA and math placed it among the *lowest 5 percent* of schools in this pool. A second key condition was whether the school's ELA and math achievement met a "lack of progress" standard. As an aside, there were other mechanisms by which a school could either be labeled PLA or receive SIG funding. However, in general, these criteria had limited empirical relevance.[3] Similarly, the new federal regulations also defined a lower-priority "Tier 3" of schools that could receive SIG funding (but were not required to implement a school-improvement model). The prioritization of PLA schools (i.e., Tiers 1 and 2) for SIG awards is reflected in how the AY 2010-11 awards were actually allocated. Specifically, several large states such as California, New York, Pennsylvania made *all* of their SIG awards to Tier 1 and 2 schools (i.e., *none* to Tier 3 schools).[4]

2.2 SIG Eligibility in California

In sum, the federally mandated eligibility requirements as typically implemented by individual states suggest that both the "lowest-achieving" and "lack of progress" criteria may constitute empirically important and discontinuous assignment mechanisms that influenced the implementation of SIG-funded school improvement model. That is, they are candidate assignment variables in a regression-discontinuity design. This study focuses on schools in California, which had the largest number of SIG-eligible schools and made more SIG awards than any other state (i.e., 92 out of the 826 Tier 1 or 2 SIG awards nationwide; Hulbert et al. 2011).

[2] Nearly 30 states, including California, defined schools serving students in grades 6 or higher as "secondary" for purposes of Tier 2 designations (Hurlburt et al. 2011).

[3] For example, states could identify low-achieving schools as PLA under a flexible "newly eligible" standard or if they had persistently low graduation rates. However, most states (including California) labeled no schools as PLA under the "newly eligible" definition and only 4 percent of SIG-eligible schools nationwide become so under the graduation-rate criteria (Hulbert et al. 2011). In some states, alternative schools and schools with extremely low enrollments were also excluded from SIG eligibility.

[4] Similarly, over 90 percent of the SIG awards in Florida and over 70 percent of the SIG awards in Texas were to Tier 1 and Tier 2 schools. Nationwide, a small number of states that had relatively few schools meeting the Tier 1 and Tier 2 definitions (i.e., North Dakota, South Dakota, Kentucky, and Tennessee) accounted for virtually all the Tier 3 SIG awards (Hurlburt et al. 2011, Exhibits 11 and 12).

The California Department of Education identified its "persistently lowest-achieving" (PLA) schools out of more than 9,000 public schools using these federally mandated assignment rules. More specifically, from a pool of 3,652 schools eligible for PLA status (e.g., schools eligible for or receiving Title I aid, etc.), roughly five percent (i.e., 183 of 3,652 schools) were identified as PLA.[5] These 183 PLA schools were eligible for a SIG and roughly half received one. The California Department of Education based its "lack of progress" definition on its school-level, test-based Academic Performance Index (API), which is described in section 3.2 below. Specifically, for each of the 3,652 PLA-eligible schools, the state summed the annual API growth from five baseline years (i.e., AY 2004-05 through AY 2008-09). Schools for whom this summed growth measure was below 50 (or missing) were labeled as "lack of progress" schools. Roughly 40 percent of the schools in the PLA-eligible pool of schools met this definition (Table 1).

Federal guidance required that states use each school's combined reading and mathematics performance based on the "All Students" category to identify the lowest-achieving schools. Most states, including California, used 3 prior years of achievement data in forming this baseline measure. More specifically, the California Department of Education identified the lowest-achieving schools from the pool of PLA-eligible schools (n = 3,652) using each school's average math/ELA proficiency rate over the three prior years (i.e., 2007 to 2009).[6] In an effort to ensure that schools of different types were eligible for SIG awards, the California Department of Education initially planned to balance the five percent of schools within strata defined by Tier (1 or 2) and school level (elementary, middle, or high school). However, the State Board of Education (SBE) subsequently submitted a waiver to the U.S. Department of Education that redefined the Tier 2 pool. Specifically, Tier 1 schools that would not have been initially deemed SIG-eligible under the distribution of the eligibility slots across these strata were re-designated to the Tier 2, which was then resorted in order to identify and implement the cut score.

The practical upshot of this waiver is that the baseline AYP proficiency-rate thresholds that defined "lowest-achieving" schools varied by each of three school levels (but not by Tier): 29.97 percent or below for elementary schools, 22.44 percent or below for middle schools, and

[5] See section 3.2 for more details on the construction of the analytical sample. This pool excludes the n = 5 schools deemed PLA by a graduation-rate screen.

[6] This 3-year measure is based on the number of students achieving proficiency in these subjects relative to the number of students taking the relevant tests. This school proficiency measure is part of a school's "adequate yearly progress" (AYP) determinations under NCLB.

37.31 percent or below for high schools. Nearly 20 percent of the schools in California's PLA-eligible pool had baseline achievement at or below this threshold level (Table 1).

Among these schools, meeting both the "lowest-achieving" and the "lack of progress" criteria was a necessary condition for qualifying as a PLA school and undertaking SIG-funded school improvement. I present evidence below that the variation in the assignment variables (i.e., baseline proficiency rates and API growth) around their relevant thresholds generated sharp and substantial variation in the likelihood of receiving a 2010-11 SIG award. This study leverages these discontinuous assignment rules to identify the effects of implementing SIG-funded reforms. To be clear, each of these discontinuous assignment rules implies a "fuzzy" regression discontinuity (RD). That is, they meaningfully influence whether a school is eligible for a SIG but do not guarantee that an eligible school receives an award. School districts with eligible schools applied to the California Department of Education and SIG awards were approved on a competitive basis.[7]

Figure 1 uses the analytical sample described below to illustrate the "intent to treat" and treatment-compliance variation underlying the fuzzy RD designs. Specifically, Figure 1 shows scatter plots of the assignment variables: baseline proficiency rates, A_s, and baseline achievement growth measures, G_s, that have been standardized and centered on the SIG-relevant threshold values. Figure 1(a) shows this scatter plot only for SIG-ineligible schools, which are, by construction, totally unrepresented in the lower left quadrant. Figure 1(b) shows this scatter plot for SIG-eligible, which are exclusively isolated in this quadrant (i.e., meeting both the "lowest-achieving" and the "lack of progress" standards. The variation around both boundaries of the lower left quadrant implies the "intent to treat": credibly quasi-experimental variation in eligibility for SIG-funded reforms. Figure 1(c) illustrates the treatment compliance associated with this variation by showing that the SIG-awarded schools are also concentrated in this quadrant. Interestingly, Figure 1 suggests that the "compliers" with the intent-to-treat tend to be schools that are not too distal from the achievement thresholds, suggesting an external-validity caveat that is discussed in the concluding section.

The validity of an RD design that effectively leverages comparisons local to these boundaries turns on a variety of important assumptions that are discussed and examined in detail

[7] The potentially non-random selection of SIG recipients from the eligible schools is not an internal-validity threat to the RD design, which leverages SIG eligibility as an "intent to treat." However, selection into both applying for and receiving a SIG award may have policy-relevant external-validity implications, which are discussed below.

below. However, one important issue to consider in the context of California's eligibility procedures described here is that of manipulability. It is well understood that RD designs can lack internal validity when an assignment variable (or its thresholds) can be manipulated (Lee and Lemieux 2009). For example, if schools that were unique in some outcome-relevant way had a differential capacity to manipulate their status as "lowest-achieving" or "lack of progress" schools, any inferences based on variation around these thresholds would be suspect.

This sort of manipulation seems exceptionally unlikely in this context, at least at the school level, because eligibility determinations were conducted in 2010 using pre-determined school-level baseline data from 3 to 5 previous school years. Another possibility is that the state manipulated the threshold values to influence the eligibility of individual schools. While this seems highly unlikely, the possibility cannot be definitively ruled out. However, this purposive threshold choice would only be problematic if it collectively privileged schools that had similar values for the assignment variables but an unobserved propensity for improved outcomes. Robustness checks presented below (i.e., covariate balance around the threshold) suggest that this was not the case.

Another standard approach to assessing manipulation of an assignment variable is to examine the density of observations around the threshold. More specifically, McCrary (2008) introduced a test of whether the density of observations jumps discontinuously at a relevant threshold. I implemented this test separately for each of the two assignment variables (i.e., baseline proficiency rates and API growth, each standardized and centered over their threshold values). In both cases, the null hypothesis of no discontinuity at the threshold cannot be rejected (see Appendix Figures 1a and 1b). Furthermore, unrestrictive histograms indicate the absence of "heaping" for each of these assignment variables near their SIG-eligibility thresholds (Barreca, Lindo, and Waddell 2011).

2.3 The Treatment Contrast – Resources and School-Intervention Models

Schools whose prior achievement levels and growth were just low enough to meet the PLA definition were substantially more likely to receive a SIG award and to implement a federally prescribed school-improvement. This implies that the treatment contrast created by these discontinuities is between "business as usual" and a treatment that consists of SIG resources combined with one of three prescriptive school-reform models chosen by the school. In

2010-11, the SIG awards in California averaged nearly $1.5 million per school (Table 1) or roughly $1,500 per pupil. The three federally sanctioned school reforms to be supported by these resources consist of transformation, turnaround, and restart. Interestingly, the U.S. Department of Education has harmonized these models with the school reforms encouraged by several other prominent policy initiatives (i.e., "Race to the Top", "State Fiscal Stabilization Funds" and NCLB waivers).

What characterizes (and distinguishes) these whole-school reform models? The transformation model has multiple emphases including (1) teacher and principal effectiveness, (2) comprehensive instructional reform, (3) extended learning time and community engagement, (4) operational flexibility and support, and (5) the use of social-emotional and community-oriented services and supports (e.g., health and nutrition). Interestingly, under (1), the transformation model requires replacing the principal and introducing teacher evaluations that are based in part on student performance and used in personnel decisions (e.g., rewards, promotion, retention, and firing).[8] The transformation model also emphasizes data-driven and differentiated instructional strategies and extending the school day and year for students who need support in core academic subjects. Transformation also emphasizes embedded professional development for staff and technical support from the district, state, and outside providers.

The "turnaround" model consists of the school reforms that define the transformation model but, in addition to replacing the principal, it also requires replacing at least 50 percent of the school's prior staff. Under the restart model, the school reopens under the management of a charter-school operator, a charter-management organization, or an educational management organization. Unsurprisingly, the transformation model is often characterized as the least disruptive of these school reforms and it is by far the most popular intervention among SIG recipients (Klein 2011). Nationwide, nearly three quarters of SIG recipients who were required to implement a school reform during the 2010-11 academic year chose transformation. Another 20 percent of SIG recipients chose the turnaround model. Only 4 percent (i.e., 33 schools nationwide) chose the restart option and even fewer schools (2 percent, n = 16) accepted small SIG awards (typically, $50,000) to close. In California, the transformation model was somewhat

[8] Early anecdotal accounts suggest that the teacher evaluation component of the transformation model is being implemented slowly, in part because it can necessitate the renegotiation of collective bargaining agreements.

less popular (i.e., roughly 60 percent of SIG recipients) while roughly of third of SIG recipients adopted the turnaround model.[9]

3. Background Literature and Theoretical Considerations

The assumption behind the design of the SIG-funded school-reform models is that chronically underperforming public schools, which often serve communities with concentrated poverty, suffer from multiple, self-perpetuating problems (e.g., weak leadership and ineffective instructional practices reinforced by poor working conditions, high turnover and a lack of resources). The implicit motivation behind SIG-funded interventions is that effective reforms of such schools have to be extensive and multi-faceted rather than marginal or targeted. These dramatic changes include new leadership and staff, new instructional practices, and outcome-based staff evaluations coupled with resources and technical assistance. One dimension of the theoretical perspective implied by these reforms concerns imperfect information: principals and teachers in underperforming schools may have limited information on what constitutes effective practices as well as underpowered incentives to identify and implement them. Another implied theoretical assumption behind these reforms is that schools suffer from collective-action problems in aligning the efforts of principals and teachers to support a culture of school effectiveness.[10] Whole-school reforms like those supported by SIGs can then be viewed as an external effort to coordinate and sustain a larger and more efficient individual and collective provision of effective classroom and school-level practices.

However, the prior evidence on how underperforming schools can dramatically and quickly improve student outcomes is largely anecdotal. For example, a recent "practice guide" on school turnarounds, which was commissioned by the U.S. Department of Education, engaged this issue directly (Herman et al. 2008) and concluded that there was no known evidence that both focused explicitly on substantial improvements in chronically underperforming schools and met conventional standards for internal validity. However, drawing on case studies of successful school turnarounds in such schools, this report recommended reforms, which have some broad

[9] Only two California schools chose closure. These two schools are excluded from this analysis and this study's results are robust to excluding *all* schools from the districts in which they were located. And only 7 schools, all within the Los Angeles Unified school district chose restart. The results presented here are robust to excluding these schools as well as all schools from this large district.

[10] This is quite similar to the implicit theoretical motivation that motivates school-accountability reforms (e.g., Ladd 2007).

parallels in the school-improvement models supported by SIGs. In particular, the practice guide emphasized the importance of credibly signaling a commitment to meaningful improvement, strong and effective leadership, committed staff and a consistent focus on high-quality instructional practice.

Despite the lack of prior evidence on school turnarounds, the reforms supported by SIGs can be situated in a broader literature on whole-school reforms. One interesting example is the "school-wide program" (SWP) made available to eligible Title I schools beginning in 1978. Schools implementing SWP are allowed to use Title I funds on whole-school functions rather than targeting these resources to services for individually eligible students. Schools pursuing this option are required to conduct a needs-based assessment, to articulate a specific reform strategy and to conduct annual reviews and revisions of this strategy. The strategies adopted by SWP schools have been varied but often focused on class-size reductions, staff development activities, or the implementation of a whole-school reform model with the support of an outside vendor (Wong & Meyer, 1998; Sunderman, 2001, Wang, Wong, and Kim 1999). The available evaluation evidence on SWP activity is based on descriptive surveys and case studies and suggests that these programs generated non-existent or small gains in student achievement (Wong & Meyer, 1998).

A more recent and prominent example of federal efforts to promote whole-school reform began in 1998 with the introduction of the "Comprehensive School Reform Demonstration" (CRSD) program. This initiative provided 3-year grants to schools that could then be used to purchase the services of independent, school-reform developers using research-based designs. This grant program developed into a "leading strategy" of the U.S. Department of Education between 1998 and 2005 (Gross, Booker, and Goldhaber 2009) when nearly $2 billion were distributed to roughly 6,700 schools to support Comprehensive School Reform (CSR) efforts. Many additional schools also implemented CSR reforms using non-federal funding sources. However, federal support for this program ended with the 2007 fiscal year.

The U.S. Department of Education defined CSR models as consisting of 11 key components with a prominent emphasis on the use of "scientifically based" teaching and management methods and the school-wide integration of instruction, assessment, professional development, and school management (U.S. Department of Education 2010c). The evaluation evidence on the achievement effects of CSR is somewhat mixed. A federally sponsored evaluation concluded that CSR schools did not demonstrate larger achievement gains than comparison schools up to five years after receiving the award (U.S. Department of Education 2010c). Similarly, a recent study of CSR awards in Texas (Gross, Booker, and Goldhaber 2009) concludes that CSR awards led to only modest achievement gains among white students (0.04 effect size) and no detectable effects among minority students.

However, these studies acknowledge that they focus on the effects of receiving CSR funding rather than on the effects of which of the highly diverse CSR reform efforts were undertaken. Other evidence suggests that the quality of CSR implementation was sometimes uneven in ways that mattered for sustaining school improvement (Desimone 2002, Bifulco, Duncombe, and Yinger 2005, U.S. Department of Education 2010c). Interestingly, a meta-analytic review (Borman, Hewes, Overman, and, Brown 2003) that considered the efficacy of *specific* CSR models characterized three (e.g., Direct Instruction, Success for All, and the School Development Program) as having the "strongest evidence of effectiveness" in terms of the comparative quality and quantity of evidence suggesting meaningful impacts on student achievement. However, Borman et al. (2003) also suggest that CSR is more likely to have positive impacts when implemented over several years.

Overall, this background literature has several implications for evaluating school-turnaround efforts. First, while there is face validity to the theoretical motivations behind school turnarounds, there is effectively no prior evaluation evidence on the effectiveness of initiatives with these particular design features. Second, the broader literature on whole-school reforms provides relatively little encouragement that these initiatives can actually be effective at scale. Prior efforts to catalyze and sustain whole-school reforms broadly have not produced clear evidence of effectiveness, particularly in the short term. Third, the important role of implementation fidelity in prior whole-school reforms also raises questions about the likely effectiveness of SIG-funded school turnarounds. For example, schools and districts that pursued SIG awards largely because of their constrained resources during a recession may be unlikely to implement the required reforms well.

However, the fact that SIG awards mandated both new school leadership and unusually explicit, dramatic, and easily observable actions (e.g., extending learning time) could attenuate the weak implementation sometimes found in prior whole-school reform efforts. Furthermore, the required elements of SIG-funded reforms appear to track several defining features of other

CSR models with comparatively strong evidence of efficacy. These practices include the use of formative assessment and data-driven instruction (e.g., Success for All), school-wide planning and community engagement (e.g., the School Development Program), and differentiated instruction (e.g., Direct Instruction). In short, how SIG-funded school reforms may have influenced both school quality and its potential mediators is an open, empirical question and one that is engaged in the next two sections.

4. Data and Specifications

In this section, I outline the basic econometric specifications used in this study, the construction of the analytical sample and key variables, and several important robustness checks and extensions.

4.1 Multivariate Regression Discontinuity (MRD) Designs

The use of regression-discontinuity (RD) specifications (Cook 2008, van der Klauww 2008, Lee and Lemiuex 2009) has become increasingly popular in settings where values of a continuous variable relative to a specific threshold create a sharp and plausibly exogenous treatment contrast (e.g., the outcome of an election or eligibility for a program). One well-understood feature of correctly specified univariate RD designs is that they credibly identify average treatment effects that are local to the observations near the relevant threshold. However, the SIG-funded school-reform activities provide an example of a multivariate regression discontinuity (MRD) setting because *two* distinct assignment rules influence the likelihood of receiving a single treatment contrast (i.e., the implementation of SIG-supported school reform). Specifically, in this setting, a school's baseline achievement level, A_s, as well as its prior achievement growth at baseline, G_s, must both be below specific threshold values for a school to have an opportunity to receive a SIG.[11]

Individually, these two eligibility conditions imply "fuzzy" regression discontinuities in SIG eligibility and receipt (i.e., they influence but do not guarantee assignment to the treatment condition). To be clear, not every eligible school (i.e., those designated as "persistently lowest achieving") chose to apply for a SIG or received one conditional on applying. And those eligible

[11] There are other candidate discontinuities in how SIG eligibility is structured. These eligibility rules are discussed below but not used in this study because they influence relatively few observations and are underpowered.

schools that actually won SIG awards are quite likely to differ from other eligible schools in highly outcome-relevant ways (i.e., selection on unobservables at both the school and district level). However, the inferences presented in this study seek to avoid the bias implied by this potentially confounded selection by leveraging the credibly exogenous variation in whether a given school was even eligible to receive a SIG award (i.e., whether it just met the "lowest-achieving" criteria or just met the "lack of progress" criteria).

Despite the considerable methodological and practical attention recently directed to RD designs, the issues raised by multivariate RD settings such as this (i.e., where there is more than one assignment variable) have only recently begun to receive scrutiny (Reardon and Robinson 2012; Wong, Steiner, and Cook, forthcoming; Papay, Willett, and Murnane 2011, Imbens and Zajonc 2011). This literature outlines several possible strategies for undertaking estimation and inference in MRD settings. For example, Wong, Steiner, and Cook (forthcoming) introduce a "frontier" RD strategy that simultaneously estimates treatment effects along two frontiers and numerically integrates them to form a weighted average treatment. However, they note that this procedure obscures the treatment heterogeneity that may exist across different frontiers. Furthermore, this estimation strategy exhibits "metric" and "scaling" dependency (i.e., sensitivity to the units of measurement and the variance of the assignment variables, respectively). Both issues are likely to be relevant in this context given the conceptual and measurement differences in the two assignment variables (i.e., proficiency rates and API growth).

A "binding-score" or "centering" RD strategy collapses the assignment variables into a single assignment variable that measures an observation's overall proximity to the treatment threshold.[12] However, like the frontier RD approach, this strategy masks the treatment heterogeneity that may exist for different assignment variables. It should be noted this treatment heterogeneity could quite plausibly have practical relevance in this setting. Schools that are both lowest achieving and that made very little progress (i.e., distant from the "lack of progress" frontier) may realize particular gains from a dramatic intervention. In contrast, schools on the "lack of progress" frontier have made recent, non-trivial gains and, for them, interventions that change their leadership and instructional practices may be unproductive – or even actively disruptive - relative to their counterfactual (e.g., continued improvement).

[12] In this setting, the "binding-score" assignment variable would be $B_s = \max(A_s, G_s)$ where A_s and G_s are the "lowest-achieving" and "lack of progress" assignment variables, standardized and centered on their relevant thresholds.

This study privileges two strategies that transparently allow for treatment heterogeneity with respect to each assignment variable. First, a “response surface” (Reardon and Robinson 2012) or “IV” approach (Wong, Steiner, and Cook, forthcoming) involves focusing on each assignment variable in isolation in specifications that utilize the full data set (i.e., essentially treating each assignment variable as a conventional univariate “fuzzy” RD while ignoring the other assignment variable). Second, a related “univariate” (Wong, Steiner, and Cook, forthcoming) or “frontier” (Reardon and Robinson, 2012) approach similarly focuses on each assignment variable in isolation but uses only the subset of data for which the other assignment variable implies treatment eligibility (e.g., Jacob and Lefgren 2004). In this application, this implies treating the “lowest-achieving” assignment variable, A_s, in a conventional “fuzzy” RD framework but using only the data from schools meeting the “lack of progress” standard (i.e., $G_s \leq 0$) and vice versa for studying the “lack of progress” discontinuity. While each approach allows for treatment heterogeneity with respect to each assignment variable, which will provide more precision is an empirical question. The “response surface” may be comparatively well powered because it utilizes all available data. However, the “frontier” approach may gain precision by focusing on observations for which the assignment to treatment is sharper.

The general “first-stage” specification modeling SIG receipt by school s (i.e., SIG_s) takes the following general form:

(1) $$SIG_s = \alpha I(A_s \leq 0) + f(A_s) + \theta X_s + \varepsilon_s$$

where α identifies the discrete change in SIG receipt for schools meeting the “lowest-achieving” (i.e., $I(A_s \leq 0)$) conditional on its relationship with the assignment variable, $f(A_s)$. The variable, X_s, represents control variables characterizing a school, its teachers, and its students at baseline and υ_s is a mean-zero random error term. A similarly structured reduced-form specification is applied to the outcome measure, Y_s:

(2) $$Y_s = \gamma I(A_s \leq 0) + h(A_s) + \beta X_s + \varepsilon_s$$

In order to facilitate interpretation of this study’s RD results (i.e., identify the causal effect of undertaking SIG-funded school reforms), I also present 2SLS estimates of the following specification:

(3) $$Y_s = \pi SIG_s + k(A_s) + \phi X_s + \eta_s$$

where the discontinuity that influences SIG receipt (i.e., $I(A_s \leq 0)$) serves as an instrumental variable (IV). Under standard monotonicity (i.e., no “defiers”) and excludability assumptions, the

RD can function as valid IV (Hahn, Todd and van der Klaauw 2001). However, it is important to note that the resulting causal estimand should be interpreted as a local average treatment effect (LATE, Imbens and Angrist 1994). That is, it identifies the effects of using the SIG award for the schools whose uptake was affected by their change in SIG eligibility (i.e., "compliers"). The generalization of equations (1), (2), and (3) to the study of the "lack of progress" discontinuity (i.e., $I(G_s \leq 0)$) is straightforward.

One challenge of any RD design involves correctly specifying the form of the function relating the assignment variable to the relevant outcomes (e.g., $h(A_s)$). This study takes several approaches to examining the empirical relevance of these concerns, including visual inspections of the data and allowing the assignment variables to have effects that vary both above and below their threshold values as well as non-linearly. However, a particularly critical check is to examine the robustness of the results in subsets of the data defined by increasingly tight bandwidths around the discontinuity. We complement this ad-hoc, effectively non-parametric approach with alternative estimates based on the estimation and inference procedures recently developed by Imbens and Kalyanaraman (forthcoming) for "fuzzy" RD applications. Specifically, Imbens and Kalyanaraman (IK) outline a mean squared-error loss function and an accompanying algorithm for identifying asymptotically optimal bandwidths. They also incorporate this bandwidth choice within an estimation procedure that identifies the effect of treatment receipt (in this case, SIG-funded school reforms) on the outcome measure. This fuzzy-RD estimand is based on four separate, kernel-weighted local linear aggressions that identify the conditional expectations of both treatment receipt and outcomes on either side of the discontinuity. These IK estimates provide an important complement to the 2SLS estimates based on equation (3). Other critical robustness checks and extensions are outlined after introducing the key variables and sample construction.

4.2 The Academic Performance Index (API)

The Academic Performance Index (API) is a measure of school-level performance based on statewide student testing. As noted above, the state of California recently used this annually produced measure to construct a "lack of progress" assignment rule for whether a given school was eligible for an AY 2010-11 SIG. However, the state originally developed the API to serve as the "cornerstone of the state's accountability system" which was introduced over a decade earlier in the wake of the Public Schools Accountability Act (PSAA). Schools that achieve high API scores are eligible for distinctions such as a “California Distinguished School” designation while low-performing schools may participate in state intervention programs. The API is also one indicator of whether a school is making “adequate yearly progress” (AYP) and can avoid sanctions under the No Child Left Behind (NCLB) Act.

A school’s annual API can range from 200 to 1000 and is calculated by converting student performance on statewide tests covering the core academic subjects (i.e., advanced, proficient, basic, below basic, and far below basic) into values on the API scale. These calculations rely largely on student performance in the California Standards Tests (CSTs) in ELA, mathematics, social studies, and science. For example, for students in grades 2 through 11, CST scores in ELA are part of their contribution to their school’s API. The CST scores in history also contribute to a school’s API score for students in grade 8, grade 9 (U.S. History) and grades 9 through 11 (World History). The API also reflects student CST scores in science for grades 5, 8, and 10 as well as CSTs specific to particular high-school courses (e.g., chemistry, physics).

The aggregation of the relevant subject-specific and student-level test results into a single, school and year-specific API number reflects weighting that varies by the student’s grade level. More specifically, in schools with students in grades K through 8, performance on ELA and mathematics tests are heavily weighted in API calculations (e.g., 51 to 57 percent for ELA tests; 34 to 38 percent for mathematics tests). In contrast, the weighting applied to the performance of high-school students reflects more balance across core academic subjects (e.g., 23 percent for science and 14 percent for social studies). For high school students, performance on the California High School Exit Examination (CAHSEE) is also included in API calculations.

It should be noted that, in recent years, students with disabilities who meet state-defined eligibility criteria contribute to their school’s API through alternative subject-specific tests, the California Modified Assessments (CMAs), instead of the CST exams.[13] The CMAs are also linked to state content standards. Some commentators suggest that the use of the CMAs has led to an increase in measured API scores (e.g., Reese and Guiterrez 2011). In general, time-varying inflation in API scores is not a clear threat to the construct validity of this test-based performance index or an internal-validity threat of this study’s RD design. However, if first-year turnaround

[13] Similarly, the small number of students with significant cognitive disabilities contributes to their school API through their performance on the California Alternate Performance Assessment (CAPA), which correspond to select state standards.

schools in California were more (or less) to designate students as disabled for purposes of API testing, the RD estimates could be biased in a positive (or negative) direction. Whether this actually occurred is one of several important robustness checks for the RD design that are discussed below.

Overall, the API as designed has a certain normative appeal as a universal school outcome measure in that it puts a particular emphasis on more fundamental academic content areas in early grades (i.e., reading and mathematics). A second key feature of the API is that it has practical relevance for students, schools, and communities because of its use in state and federal accountability systems. And, third, the availability of a common, performance index for all schools (i.e., schools at all grade levels) makes it possible to increase the statistical power of this study's multivariate RD design.

4.3 Analytical sample and variables

The analytical sample used in this study effectively consists of the 2,892 California public schools that met the broad "Tier 1" and "Tier 2" criteria for receiving a 2010-11 SIG award. More specifically, California has over 9,000 public elementary and secondary schools. However, in this universe of schools, only 3,657 met the Tier 1 (n = 2,708) and Tier 2 (n = 949) criteria for a SIG award. Title I schools in program improvement were SIG-eligible as Tier 1 schools. Secondary schools that were eligible for Title I support but not receiving it were eligible as Tier 2 schools. Five of these 3,657 schools were determined to be SIG-eligible at this point by virtue of having graduation rates below sixty percent in each of the four academic years from 2004-05 to 2007-08.

California identified 183 other schools - roughly 5 percent of the remaining 3,652 schools - as the "persistently lowest achieving" (PLA) schools in the state. These 183 PLA schools were, therefore, eligible to receive a SIG award beginning in AY 2010-2011. However, the analytical sample was shaped by several further considerations. First, 422 schools were excluded because they had missing values for the two assignment variables.[14] An additional 106 schools were excluded because they were missing the key outcome variable (i.e., 2010-11 API scores) and 107

[14] Nearly all of these deletions involved schools that were missing baseline proficiency rates and were, therefore, SIG-ineligible. Most of these schools were "continuation" or "community day" schools serving children at high risk of failure or with serious behavioral issues. The schools with missing values for the API growth measure were deemed to have met the state's "lack of progress" screen.

schools were excluded because they had missing values for the baseline observables. One remaining school that used a small SIG award (i.e., $50,000) to close is also excluded. Also omitted are 98 remaining charter schools and 26 special-education schools.[15]

This construction leaves an analytical sample of 2,892 public schools of which nearly 3 percent (n = 82) received 2010-11 SIG awards. On average, each of these schools received a first-year grant of $1.48 million or, roughly, $1,500 per pupil. Most of these schools (n = 48) coupled these awards with the "transformation" model of school improvement. However, 27 schools chose the "turnaround" model and the remaining 7 schools chose the "restart" option.[16]

Nearly 19 percent of the schools in this sample met the "lowest-achieving" screen for SIG eligibility (Table 1). And nearly 40 percent met the "Lack of Progress" screen (i.e., baseline API growth over the previous 3 years less than 50 points). These two eligibility requirements are two key necessary conditions for being among the "persistently lowest achieving" (PLA) schools in the state. Interestingly, a small number of schools were excluded from PLA for two other reasons. Specifically, the state deemed ineligible the schools that had already reached or exceeded the state API target of 800 and the schools that had too few tested students to meet California's "n-size" requirement for API and AYP (adequate year progress) calculations. In theory, these additional criteria provide other assignment variables that can be used in a multivariate RD design. However, as a practical matter, these candidate variables are underpowered because they influenced relatively few schools. This study's key results are robust in specifications where these schools are excluded.

The control variables used in this study reflect the observed traits of students, teachers, and the schools themselves and were drawn from several state data files both for the baseline year (i.e., AY 2009-10) and the one available treatment year (i.e., AY 2010-11). Specifically, data on student traits in each of these years were drawn from the corresponding state API data files. These variables consist of the percent of students who were black, Hispanic, or Asian as well as the percent of students who were eligible for free or reduced-price lunches, the percent of students identified as English learners, and the percent of students identified as having a disability. The mean values of these student traits are listed in Table 1. Interestingly, the

[15] The empirical results are qualitatively unchanged in models that do not adopt these exclusions. However, this sample construction puts the focus on traditional, public schools that meet the PLA criteria and for whom data are available to conduct important internal-validity checks.

[16] As noted earlier, the results presented here are robust to excluding these 7 treatment schools as well as the single school district (Los Angeles Unified) in which they are located.

socioeconomic disadvantage of the students served by these schools is indicated by the fact that the percent eligible for free or reduced-price lunches averages 75%. Similarly, the mean 2010-11 API for schools in this sample has been standardized using the mean and variance for all public schools in the state. So, the schools in this sample have an average API that is 0.41 standard deviations (SD) below the state mean (Table 1).

I also constructed school-level measures of teacher traits for each of these academic years using the individual-level data available in California's "Staff Demographics" files. Specifically, I constructed school-level data on the number of teacher FTEs, teacher experience, and the percent of teachers who were female, black, Hispanic, and had graduate degrees, all weighted by FTE (Table 1). I also constructed a school-enrollment measure and combined this with the data on the number of teacher FTEs in each school to construct a pupil-teacher ratio. I also identified the urban classification (i.e., city, suburb, town, or rural) of each school using the state's "Public Schools Database" (Table 1).

In addition to these controls, I also report results based on school-level suspension and truancy rates. These measures could be viewed both as alternative, non-cognitive outcome measures and as relevant mediators of the reforms' main effects. The California Department of Education made these data available both for the 2010-11 post-treatment year and for the baseline 2009-10 school year.[17] The availability of data from the 2009-10 school year makes possible an additional falsification check like those outlined below.

4.4 Robustness Checks

The promise of an RD design to generate unbiased inferences that are "as good as random assignment" turns on several assumptions. Fortunately, there are several robustness checks and falsification exercises that can examine these assumptions directly (Lee and Lemieux 2009, Schochet et al. 2010). For example, the density tests discussed earlier (McCrary 2008, Barreca, Lindo, and Waddell 2011) provide evidence that the two baseline assignment variables were not manipulated. In section, I outline several other robustness checks that assess the validity of the RD design.

[17] I would like to thank Keric Ashley of the California Department of Education for her assistance in providing these data, which are available for 2,666 of the 2,892 sample schools.

First, I present graphical evidence that illustrate the study's key results under different assumptions and provide informal evidence on the functional form of the relationship between the assignment variables and key outcomes. Second, I also examine the robustness of the study's results to alternative specifications of how the assignment variables relate to the key outcomes (e.g., allowing heterogeneous and nonlinear relationships between the assignment variables and outcomes with respect to being above or below the thresholds for SIG eligibility).

Third, I also examine the robustness of this study's key results in specifications that focus exclusively on observations within increasingly tight local bandwidths around the threshold for for SIG eligibility. This "local linear regression" approach effectively provides a consistent and non-parametric way of estimating RD treatment effects. However, choosing a bandwidth around the cut point is not straightforward because it involves a tradeoff between unbiasedness and precision. I follow the standard practice of examining the robustness of our results to multiple bandwidths. However, I also implement a procedure recently developed by Imbens and Kalyanaraman (forthcoming) to identify the optimal bandwidth in an RD design (i.e., one which minimizes the mean squared error of the shift parameter of interest).

RD designs like those used here rely on the change in treatment status that occurs at a particular threshold. One heuristic way to assess whether the effects associated with particular thresholds are overstated with a given set of data and specification is to examine the effects associated with "placebo" discontinuities that had no practical relevance for the treatment status of schools. If these placebo discontinuities (e.g., $I(A_s \leq -0.1)$, $I(A_s \leq 0.1)$, etc.) appear to have had significant effects on API scores, it would cast doubt on whether the functional form had been correctly specified and on whether the RD strategy was generating reliable inferences. A fifth and important robustness check involves estimating auxiliary RD specifications where *baseline* school observables are the dependent variables. This evidence will suggest whether the outcome-relevant observed traits of schools are similar (i.e., except with regard to treatment status) on either side of the cut score.

As suggested earlier, another potential internal-validity threat to this RD design involves non-random student sorting. More specifically, it is possible that schools undertaking SIG-funded reforms experienced changes in the number of students entering or leaving the school. To the extent that a school turnaround encouraged students with higher or lower propensities for achievement to attend the school, the true effects of the reform on student performance would be respectively overstated or understated. However, it should be noted the role of student mobility is likely to be limited in this context (i.e., studying AY 2010-11 outcomes) simply because of the highly compressed schedule under which the stimulus-funded reforms were put in place. Specifically, California's SIG funding ($416 million, the largest of any state) was announced on June 24, 2010. The California Department of Education immediately made a Request for Applications available to school districts and began reviewing applications in July of 2010. The state's funded SIG awards were announced and in place in September of 2010. This timing implies that policy-endogenous sorting is unlikely because it would have to have been among students already enrolled in the school.

However, two other types of sorting could imply internal-validity threats. First, motivated parents may have been more likely to move their students out of SIG-eligible schools after the state list of "persistently lowest achieving" schools was announced in the spring of 2010. It should be noted that this sorting pattern probably implies that the RD strategy understates the effects of SIG eligibility and receipt (i.e., a PLA stigma could catalyze positive selection out of treatment-eligible schools). Second, as noted above, another potential confounder involves whether schools undergoing SIG-funded reforms were more likely to direct their students to the state's modified assessments for disabled students, possibly inflating their school's API. As a check on the empirical relevance of all of these concerns, I also present auxiliary RD estimates where *post-treatment* (i.e., AY 2010-11) school and student traits are the dependent variables (e.g., student observables, including disability status). Some of these 2010-11 school traits (e.g., pupil-teacher ratios, teacher traits) also provide evidence on the potential mediators of the SIG-funded reforms.

5. Results

5.1 Graphical Evidence

Figures 1(a), (b), and (c) provided clear evidence that schools meeting both assignment-rule conditions were more likely to undertake SIG-funded reforms. Figure 2 provides more explicit evidence on how the "lowest-achieving" discontinuity for SIG eligibility relates to the probability of actually undertaking SIG-funded reforms. Specifically, I organized the data into bins 0.1 standard deviations (SD) wide and defined for the 3-year baseline AYP proficiency rate, an assignment variable which is standardized and centered on its threshold value for a school's

SIG eligibility. This figure indicates that, for $A_s > 0$, the probability of receiving a SIG was 0. However, for schools whose baseline proficiency rate was just lower than this, the probability of undertaking SIG-funded reforms was approximately 20 percent.

Figure 2 provides clear evidence that the assignment variable implies a significant jump in the probability of receiving the treatment contrast. In Figure 3(a), I show the corresponding graph where the post-treatment outcome variable, 2010-11 API scores, is on the vertical axis. Unsurprisingly, this graph shows that 2010-11 API scores are related in a fairly linear way to the baseline achievement measure. However, this graph also suggests a distinct jump in the post-treatment API scores of the schools who just met the "lowest-achieving" standard for SIG eligibility. This discontinuous jump is consistent with a reduced-form treatment effect of roughly 0.1 SD.

Interestingly, the leftmost 4 bins in Figure 3(a) suggest that this treatment effect may have been more muted for the very lowest-achieving schools. This could reflect some plausible treatment heterogeneity or that fewer schools this far from the threshold received SIG awards (e.g., the leftmost two bins in Figure 2). However, observations like this may also exert an undue influence on the fitted line in Figure 3a and lead to an overstated discontinuity. In Figure 3(b), I exclude these distal observations and focus only on those in a ±0.5 bandwidth. This figure suggests a similarly large jump in post-treatment outcomes at the discontinuity. However, this figure may also be misleading because it reflects over-smoothing (i.e., the grouping of school-level observations into relatively few bins). In Figure 3(c), I present the same figure using the same ± 0.5 bandwidth but bin widths that have been halved to 0.05 SD. This figure also suggests a post-treatment jump in school performance at the "lowest-achieving" threshold for SIG eligibility. Figure 3(d), which reflects halving the bin width yet again to 0.025 SD, also suggests a school performance jump at the discontinuity associated with SIG eligibility.

In Figure 4, I present similar graphical evidence for the "lack of progress" assignment variable. Figure 4(a) demonstrates that no schools in the sample received a SIG award if they did not meet the "lack of progress" standard. And, for schools at or below the standard, the probability of undertaking SIG-funded reforms jumped to approximately 6 percent. In Figure 4(b), I present graphical evidence of how post-treatment API scores varied around the "lack of progress" threshold. This figure suggests that SIG eligibility on this threshold did not increase, and may have even reduced, subsequent school performance.

5.2 Main Results

In Table 2, I present regression-based versions of the graphical evidence from Figures 2 and 3 coupled with some important evidence on robustness. Specifically, in the top panel of Table 2, I show the estimated first-stage effects of the lowest-achieving discontinuity across a variety of specifications and both in the full sample and the smaller sample of schools meeting the "lack of progress" condition. The full-sample estimates consistently indicate that, at the lowest achieving threshold, the treatment probability jumped by 23 percentage points while, in the smaller sample of schools where $G_s \leq 0$, the treatment probability jumped by 53 to 55 percentage points. The baseline specification (column 1) controls for the assignment variable, a linear spline that allows it to have distinct effects above and below the threshold, and baseline 2010 and 2009 API scores. The subsequent specifications add to this model by conditioning both on the student, teacher, and school controls (Table 1) and on the square of the assignment variable, which is also interacted with a spline allowing it to vary above and below the threshold.

The lower panel of Table 2 presents the reduced-form estimates of the same 8 specifications (i.e., where 2010-11 API scores are the dependent variables). These results consistently indicate that schools whose baseline proficiency rate just met the lowest achieving threshold saw a statistically significant jump in post-treatment school performance. These treatment estimates range from roughly 0.07 to nearly 0.10 of a school-level standard deviation in the API measure.

Table 3 presents the results of parallel specifications that focus on the "lack of progress" discontinuity. These estimates consistently indicate that, in the full sample, schools just meeting this condition were 6 to 7 percentage points more likely to undertake SIG-funded reforms. In the much smaller sample of schools meeting the lowest-achieving condition (n = 542), this first-stage effect increases to roughly 40 percentage points. However, the lower panel of Table 3 indicates that this local variation in treatment eligibility and uptake did not have statistically precise effects on school performance in any specification or sample construction.

The comparative results in Tables 2 and 3 suggest a plausible pattern of treatment heterogeneity in which SIG-funded reforms improved the performance of schools on the lowest-achieving boundary but not among SIG-eligible schools that were already making some progress (i.e., those on the "lack of progress" boundary). To test whether this suggestion of treatment heterogeneity is statistically meaningful, I rely on the 2SLS estimates (equation (3)) implied by the first-stage and reduced-form estimates in Tables 2 and 3. More specifically, the estimated parameters from models focusing on the lowest-achieving margin (i.e., α^A and γ^A) and the "lack of progress" margin (i.e., α^G and γ^G) can be combined to test the hypothesis that the effects of SIG use generated by the "lowest-achieving" discontinuity (π^A) are the same as the effects of SIG use generated by the "lack of progress" discontinuity (π^G). The null hypothesis for this test reflects the "indirect least squares" structure of these estimates and takes the following form:

$$H_0 : \pi^A = \frac{\gamma^A}{\alpha^A} = \frac{\gamma^G}{\alpha^G} = \pi^G \tag{4}$$

The test for this nonlinear hypothesis is based on a Wald statistic and the covariance matrix for this test is based on the "seemingly unrelated" variance associated with the four simultaneously estimated equations. The p-value for this test statistic is 0.2537, indicating that the null hypothesis cannot be rejected. In other words, though treatment heterogeneity across the two boundaries is suggested, the sampling variation, particularly across the "lack of progress" discontinuity, is too limited to establish that these differences are statistically meaningful.

Table 4 summarizes the positive results based on the "lowest-achieving" discontinuity and provides initial evidence on their robustness by presenting 2SLS and IK estimates of the effects of SIG receipt on school performance. The 2SLS estimates indicate that the SIG-funded reforms catalyzed by the variation around this discontinuity improved school performance by 0.31 to 0.37 of a school-level standard deviation. These findings are robust across the use of alternative controls. Interestingly, different information criteria based on the residual sums of squares from these specifications privilege specifications that include the available controls and a linear spline for the assignment variable (Schochet et al. 2010). The non-parametric IK estimates in columns (5) and (6), which focus on the observations within an optimally determined bandwidth around the discontinuity (roughly 1 SD) and weight observations near the threshold more heavily, imply results similar to the 2SLS estimates.

5.3 Robustness Checks

The results in Tables 2, 3, and 4 indicate that schools who were just eligible for SIG-funded reforms on the "lowest-achieving" margin saw significant increases in subsequent school performance. However, there are not statistically precise effects associated with variation on the "lack of progress" margin. In light of these findings, the remaining analysis focuses on examining the robustness of the apparent effects associated with the "lowest-achieving" discontinuity. Table 5 presents important ad-hoc evidence on the validity of the performance gains on the lowest-achieving margin. More specifically, Table 5 presents reduced-form RD estimates associated with both the actual RD that meaningfully influenced treatment assignment and several "placebo" discontinuities, which had no such relevance. If the outcome measure varied significantly around the irrelevant placebo RDs, it would strongly suggest the presence of undiagnosed specification errors (e.g., incorrect functional form). However, the results in Table 5 consistently indicate that only the actual RD that influenced SIG receipt is associated with a statistically significant jump in post-treatment school performance. Notably, several of these placebo RD estimates have sufficient precision that, if their effects had been as large as the actual discontinuity, they would have been statistically significant.

Table 6 presents another set of important robustness checks. More specifically, Table 6 presents the estimated first-stage and reduced-form effects of the lowest-achieving discontinuity both for the full sample and across samples restricted to increasingly tight bandwidths around the threshold for SIG eligibility. This effectively non-parametric approach provides evidence on the extent to which the results in Tables 2 and 4 reflect the possibly distorting effects of functional-form assumptions or the spurious influence of observations that are distal from the threshold. The results in Table 6 indicate that both the first-stage and reduced-form estimates are quite robust as the sample shrinks to increasingly tight bandwidths around the threshold including a bandwidth quite close to the IK bandwidth and lower (i.e., within 1 and 0.75 standard deviations respectively). The reduced-form effect of the discontinuity does decrease somewhat in bandwidth-reduced samples. However, because the first-stage estimates also decrease somewhat, the implied 2SLS estimate associated with SIG-funded reforms is comparatively stable.

Table 7 presents another important set of robustness checks based on auxiliary RD regressions where baseline student, teacher, and school traits are the dependent variables. Each cell in this table presents an RD estimate from an individual regression. Collectively, these results provide important evidence on whether outcome-relevant covariates, in particular, baseline API scores, are balanced around the lowest-achieving discontinuity. The general absence of statistically significant estimates in Table 7 indicates that baseline traits were well balanced around the threshold that influenced assignment to treatment. Table 8 presents similar evidence but for the *post-treatment* (i.e., AY 2010-11) observables. This evidence also provides a critical robustness check because it indicates whether treatment eligibility led to potentially non-random student sorting. These estimates indicate that the schools on the treatment side of the lowest-achieving threshold did not see large or statistically significant changes in the share of students who were black, Hispanic, or Asian nor in the percent of students who were English learners, disabled, or eligible for free/reduced-price lunches. This pattern is consistent with the notion that the compressed, rapid rollout of SIG funding and reforms just prior to the 2010-11 school year did not leave time for empirically meaningful, policy-endogenous sorting. Similarly, the absence of an RD effect on the percent of students with disabilities also implies that schools undertaking SIG-funded reforms did not differentially sort students into modified assessments.

5.4 Treatment Mediators

The results in Tables 2 through 8 provide evidence that SIG-funded school reforms led to improvements in test-based school performance for schools along the lowest-achieving discontinuity. However, were there also effects on alternative measures that could be viewed as mediators of the treatment effects or as alternative outcome measures in their own right? And how might these effects have differed across schools that adopted the transformation, turnaround and restart models? Tables 9, 10, and 11 provide direct evidence on these questions. Specifically, Table 9 presents 2SLS and IK estimates that use the lowest-achieving discontinuity to identify how SIG-funded reforms influenced student suspension and truancy rates. The first two rows in Table 9 report these estimates for the 2009-10 (i.e., pre-treatment) suspension and truancy rates. The absence of statistically significant effects on these baseline measures is consistent with the causal warrant of the RD design. The final two rows of Table 9 provide weakly suggestive evidence that SIG-funded reforms reduced both student truancy and suspensions in 2010-11. These estimated effects are consistently negative in models that include the available controls and imply large effects relative to the sample means where significant. However, these results also exhibit sensitivity to the estimation method (i.e., 2SLS or IK).

Table 10 presents 2SLS and IK estimates of how SIG-funded reforms influenced 2010-11 mediators such as teacher traits and pupil-teacher ratios. I also constructed a complementary measure of whether there appears to have been a change in the principal of the school using pooled data from the state's School Accountability Report Cards (SARC).[18] These results provide consistent evidence that schools undertaking SIG-funded school reforms saw the average years of experience among their teachers drop by over two years, a decrease equivalent to roughly two-thirds of a standard deviation for the schools in this sample (see Table 2). This effect is consistent with the hypothesis that SIG-funded school reforms led to an influx of less experienced teachers who were new to the school. However, the other results in Table 10 suggest that these changes in the teaching staff did not influence teacher demographic traits or the share of teachers with a graduate degree. Table 10 also provides some evidence that SIG-funded reforms led to a substantial increase in the likelihood of a new principal. However, this effect is highly imprecise in the IK specification, which is limited to observations within 0.81 standard deviations of the assignment variable. Interestingly, the final row of Table 10 provides some evidence that schools undertaking SIG-funded reforms significantly reduced their pupil-teacher ratios by a substantial amount (i.e., nearly 5 students).

Another question of interest concerns whether the effectiveness of SIG-funded reforms varied among schools implementing the three federally prescribed models. Because the lowest-achieving discontinuity provides only one source of credibly exogenous variation in treatment assignment, this question cannot be engaged with the causal warrant of the RD design. However, the panel structure of the available data does make it possible to estimate this treatment heterogeneity in specifications that condition on time-invariant school traits. Table 11 presents such "difference in difference" estimates. Specifically, in these OLS specifications, the dependent variable is the first-difference in API scores (i.e., the difference between 2010-11 "growth" API scores and the corresponding 2009-10 "base" scores). This specification effectively compares the heterogeneous *changes* across the three types of treatment schools relative to the reference category: schools not adopting SIG-funded reforms. The full-sample results in the first row of Table 11 indicate that the achievement growth among SIG schools appears to have been largely concentrated in those that adopted the turnaround model. Interestingly, the null hypothesis that these three treatment effects are equivalent is easily rejected. Moreover, this result is quite robust across sample restrictions that effectively increase

[18] Specifically, this measure of principal turnover is based on data from SARC files from December 2009 and December 2011. Nearly 42 of schools in this sample experienced principal changes during this period.

the extent to which the control schools are similar to the treatment schools (i.e., restrictions that focus on the “lowest-achieving”, “lack of progress”, and SIG-eligible schools).

6. Discussion

The results presented in this study indicate that the stimulus-funded whole-school reforms targeting the nation’s chronically underperforming schools have led to statistically significant improvements in school performance, at least in California, the state where most of these turnaround schools were located. The results and the corresponding robustness checks affirm that these findings have the strong causal warrant typically associated with well-functioning regression-discontinuity designs. The direct relevance of these findings for this federal program and the support it received from the stimulus package is straightforward. However, these results also provide important, early evidence relevant to the broader debate about federal activism in education reform as embodied by several high-profile initiatives (e.g., Race to the Top and NCLB waivers) that encourage similar reforms.

As with any results based on an RD design, it is also important to underscore several caveats related to the generalizability of these results. Most obviously, these results rely on highly localized comparisons of schools that were just above and below the “lowest-achieving” eligibility thresholds so whether the effects of such reforms generalize to other types of schools (e.g., higher-achieving schools) is an open question. However, the fact that these inferences may have unambiguous salience only for the most underperforming schools should not be particularly relevant for policy makers interested in catalyzing improvements among such schools. An external-validity caveat that may have more relevance for the scalability of these results is that they identify the causal effects of these reforms for schools (and districts) that responded to their SIG eligibility by crafting a successful application (i.e., “compliers”). The effects of these funded reforms may differ if implemented in schools and districts that do not respond well or at all to such opportunities for improvement. For example, the district support that allowed struggling schools to successfully secure a SIG award may be closely related to the district traits necessary to support school turnarounds.

It is also notable that there were no statistically detectable effects on school performance around the eligibility threshold associated with “lack of progress.” This could reflect the irrelevance of such reforms for schools that were already improving to some extent. If so, such treatment heterogeneity would suggest that SIG eligibility could be more effectively targeted to schools meeting a more stringent “lack of progress” screen. However, the lack of a statistically significant effect around this particular eligibility threshold could also merely reflect the limited power of a discontinuity that had weaker first-stage effects.

A particularly critical question concerns how we interpret the size of the treatment effects implied by the variation around the “lowest-achieving” discontinuity. The results in Table 4 indicate that SIG-funded school reforms increased by roughly a third (e.g., 0.32 in column (6)) of the *school-level* standard deviation in this measure. This corresponds to roughly 34 points on the state’s API scale. One policy-relevant way to frame this effect is to note that average SIG-eligible school was, at baseline, roughly 150 points below the state’s performance target of 800. The reform-driven growth of 34 scale points, therefore, implies closing this gap by 23 percent. An alternative and important way to frame these effect sizes is to consider their comparative cost effectiveness. Such comparisons are important because, while the performance gains documented here seem quite large, these reforms are also fairly expensive (e.g., roughly $1,500 per pupil in California). A complication in making such cross-study comparisons is that the school-level performance results presented here were standardized with respect to the *school-level* standard deviation in test-based performance. The variance in test scores across schools is typically only 10 to 15 percent of the variance in student-level performance.[19] Therefore, using the school-level standard deviation overstates the *student-level* effect size by a factor of 2.6 to 3.2 (e.g., $(0.1)^{-0.5}$).

Under the conservative rescaling implied by assuming that the school-level variance is 10 percent of the student-level variance (e.g., dividing the estimates presented in Table 4 by 3.2), the resulting student-level effect size implied by undertaking SIG-funded reforms is roughly 0.10. One particularly attractive point of comparison for this treatment effect involves the short-term achievement gains associated with random assignment to smaller classes in the Project STAR study. Krueger (2003) argues that the short-term impact of random assignment to a small class is a 0.2 SD test-score gain at a cost that is approximately 47 percent of the overall spending per pupil. In California, spending per pupil is roughly $10,000 so the cost of a similarly large class-size reduction would be roughly $4,700 per pupil. This rough comparison suggests that the

[19] Unfortunately, because the API is an amalgam of multiple tests (and tests for which the statewide student-level data are not publicly available), replicating this standard result for this context is not straightforward.

SIG-funded school turnarounds are cost-effective relative to expensive class-size reductions, generating roughly half of the achievement gains associated with Project STAR but doing so at a third of the cost. Such cross-study comparisons are necessarily complicated by other factors such as the fact that the gains associated with school turnarounds involved not just kindergarten students but students in grades K through 12. A cost-effectiveness issue that may be particularly relevant in this context is that schools are expected to receive their SIG support for each of three years. Identifying whether turnaround schools will be able to sustain (or increase) the performance gains documented here as they continue to receive federal support will be an important issue to engage as the relevant data become available. Another important issue will be to understand how SIG-funded reforms were actually implemented. The prescriptive nature of the required reforms is fairly transparent. However, identifying the character and variance of their implementation will be an important feature of understanding this dramatic policy initiative and extending its lessons more broadly.

References

Barreca, A., Guldi, M., Lindo, J. M., Waddell, G. R, (2011). “Heaping-Induced Bias in Regression Discontinuity Designs,” NBER Working Paper No. 17408, September 2011.

Bifulco, R., Duncombe, W., & Yinger, J. (2005). Does whole-school reform boost student performance? The case of New York City. *Journal of Policy Analysis and Management, 24(1),* 47-72.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*, 125–230.

Cohen, D. K. & Moffitt, S. L. (2009). *The ordeal of equality: Did federal regulation fix the schools?* Cambridge, MA: Harvard University Press.

Cook, T.D., 2008. ‘‘Waiting for life to arrive’’: a history of the regression- discontinuity design in psychology, statistics and economics. Journal of Econometrics 142 (2), 636–654.

Desimone, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research, 72(3),* 433-479.

Gewertz, C. (2009). Duncan’s call for school turnarounds sparks debate. *Education Week, 28(37).*

Gross, B., Booker, K.T., & Goldhaber, D. (2009). Boosting student achievement: the effect of comprehensive school reform on student achievement,” *Educational Evaluation and Policy Analysis, 31(2), 111*-126.

Hahn, J., P. Todd, and W. van der Klaauw (2001). Identifcation and Estimation of Treatment Effects with a Regression-Discontinuity Design, Econometrica, 69, 201-209.

Herman, R., Dawson, P., Dee, T., Greene, J., Maynard, R., Redding, S., & Darwin, M. (2008). *Turning around chronically low-performing schools: a practice guide* (NCEE 20084020). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/ practiceguides.

Hess, F. & Darling-Hammond, L. How to Rescue Education Reform. The New York Times, December 5, 2011

Hurlburt, S., Le Floch, K. C., Therriault, S. B., & Cole, S. (2011). *Baseline analyses of SIG applications and SIG-eligible and SIG-awarded schools* (NCEE 20114019). Washington, DC: U.S. Department of Education.

Imbens, G. & Angrist, J. (1994). Identification and Estimation of Local Average Treatment Effects. Econometrica 62:467–475.

Imbens, G. & Kalyanaraman, K. forthcoming. Optimal bandwidth choice for the regression discontinuity estimator. Review of Economic Studies.

Imbens, G. & Zajonc, T. Regression Discontinuity Design with Multiple Forcing Variables. Working Paper, September 2011.

Jacob, B., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics, LXXXVI*, 226-244.

Klein, A. (2011). School improvement grant efforts face hurdles. *Education Week,* 30(29), 22, 26-27.

Krueger, A. B. (2003), Economic Considerations and Class Size. The Economic Journal, 113: F34–F63. doi: 10.1111/1468-0297.00098

Ladd, H. F. (2007). Holding schools accountable revisited. 2007 Spencer Foundation Lecture in Education Policy and Management, Association for Public Policy Analysis and

Management. Retrieved November 8, 2009 from https://www.appam.org/awards/pdf/2007Spencer-Ladd.pdf.

Lee, D.S. & Lemieux, T. (2009). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*, 281-355.

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

McNeil, M. "Tight Leash Likely on Turnaround Aid," Education Week 29(2), September 2, 2009, pages 1, 20-21.

McMurrer, J., Dietz, S., & Rentner D. S. (2011). Early state implementation of Title I school improvement grants under the Recovery Act. Washington, DC: Center on Education Policy.

Papay, J.P., Willett, J.B., and Murnane, R.J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. Journal of Econometrics 161, 203-207.

Reardon, S.F. and Robinson, J.P. (2012). Regression Discontinuity Designs with Multiple Rating Scoring Variables. Journal of Research on Educational Effectiveness 5(1), 83-104

Reese, Phillip & Guiterrez, Melanie. Growing use of simplified test inflates some California schools' scores. The Sacramento Bee, October 2, 2011, Page 1A.

Schochet, P. Z., Cook, T., Deke, J., Imbens, G., Lockwood, J.R., Porter, J., & Smith, J. (2010). Standards for Regression Discontinuity Designs. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf.

Sunderman, G. L. (2001). Accountability Mandates and the Implementation of Title I Schoolwide Programs: A Comparison of Three Urban Districts. *Educational Administration Quarterly, 37,* 503-532.

U.S. Department of Education. (2010a). *Guidance on school improvement grants under section 1003(g) of the Elementary and Secondary Education Act of 1965.* Washington, DC: Office of Elementary and Secondary Education. Retrieved from http://www2.ed.gov/programs/sif/legislation.html

U.S. Department of Education. (2010b). *Guidance on fiscal year 2010 school improvement grants under section 1003(g) of the Elementary and Secondary Education Act of 1965*. Washington, DC: Office of Elementary and Secondary Education. Retrieved from http://www2.ed.gov/programs/sif/sigguidance11012010.pdf

U.S. Department of Education (2010c). *Evaluation of the comprehensive school reform program implementation and outcomes: fifth-year report.* Washington DC: Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

van der Klaauw, W., 2008. Regression-discontinuity analysis: a survey of recent developments in economics. Labour 22 (2), 219–245.

Wang, M. C., Wong, K.K., & Kim, J. (1999). *A national study of Title I schoolwide programs: a synopsis of interim findings.* Philadelphia: The Mid-Atlantic Regional Educational Laboratory at Temple University Center for Research in Human Development and Education.

Wong, K. & Meyer, S. 1998. Title I schoolwide programs: a synthesis of findings from recent evaluation. *Educational Evaluation and Policy Analysis*, 20, 115-136.

Wong, V. C., Steiner, P. M., Cook, T. D. (forthcoming). Analyzing Regression-Discontinuity Designs with Multiple Assignment Variables: A Comparative Study of Four Estimation Methods. *Journal of Educational and Behavioral Statistics*.

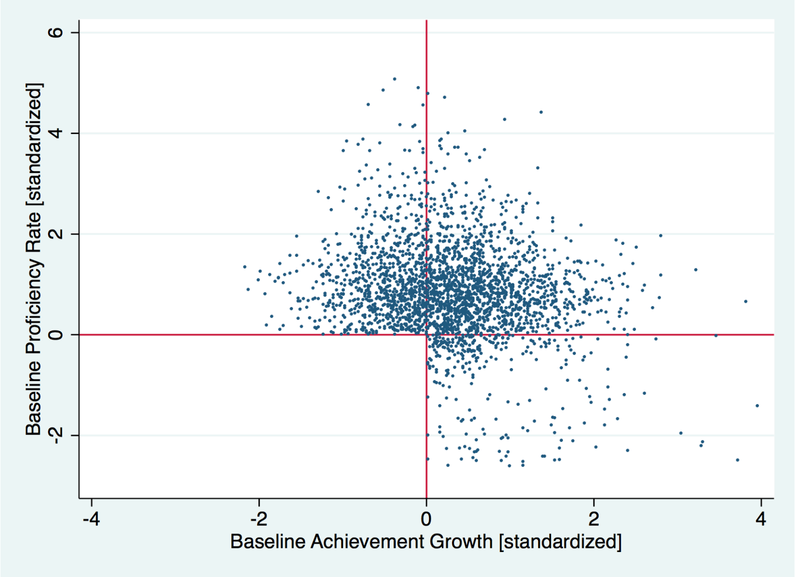


Figure 1(a) – Scatter Plot of Assignment Variables, SIG-Ineligible Schools

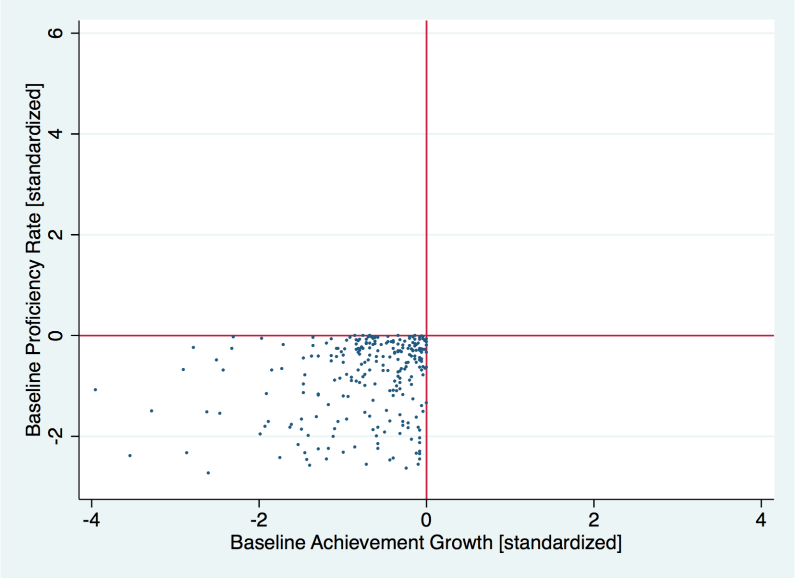


Figure 1(b) – Scatter Plot of Assignment Variables, SIG-Eligible Schools

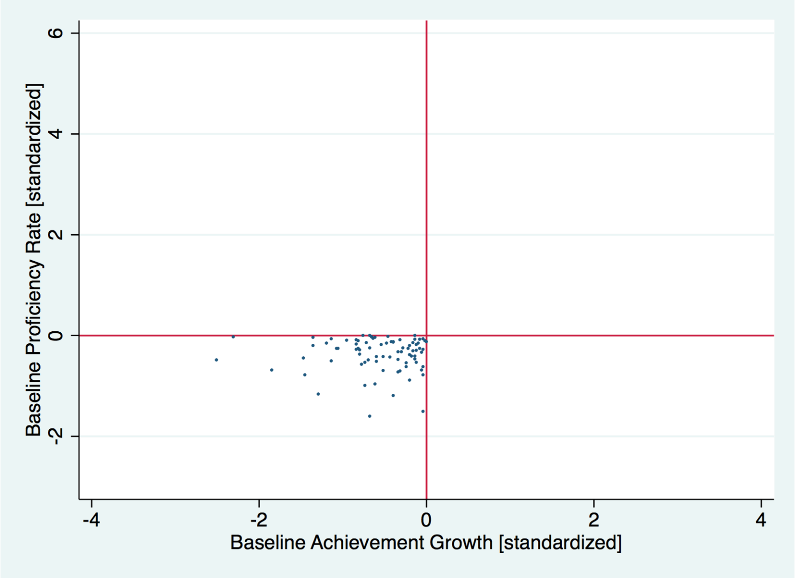


Figure 1(c) – Scatter Plot of Assignment Variables, SIG-Award Schools

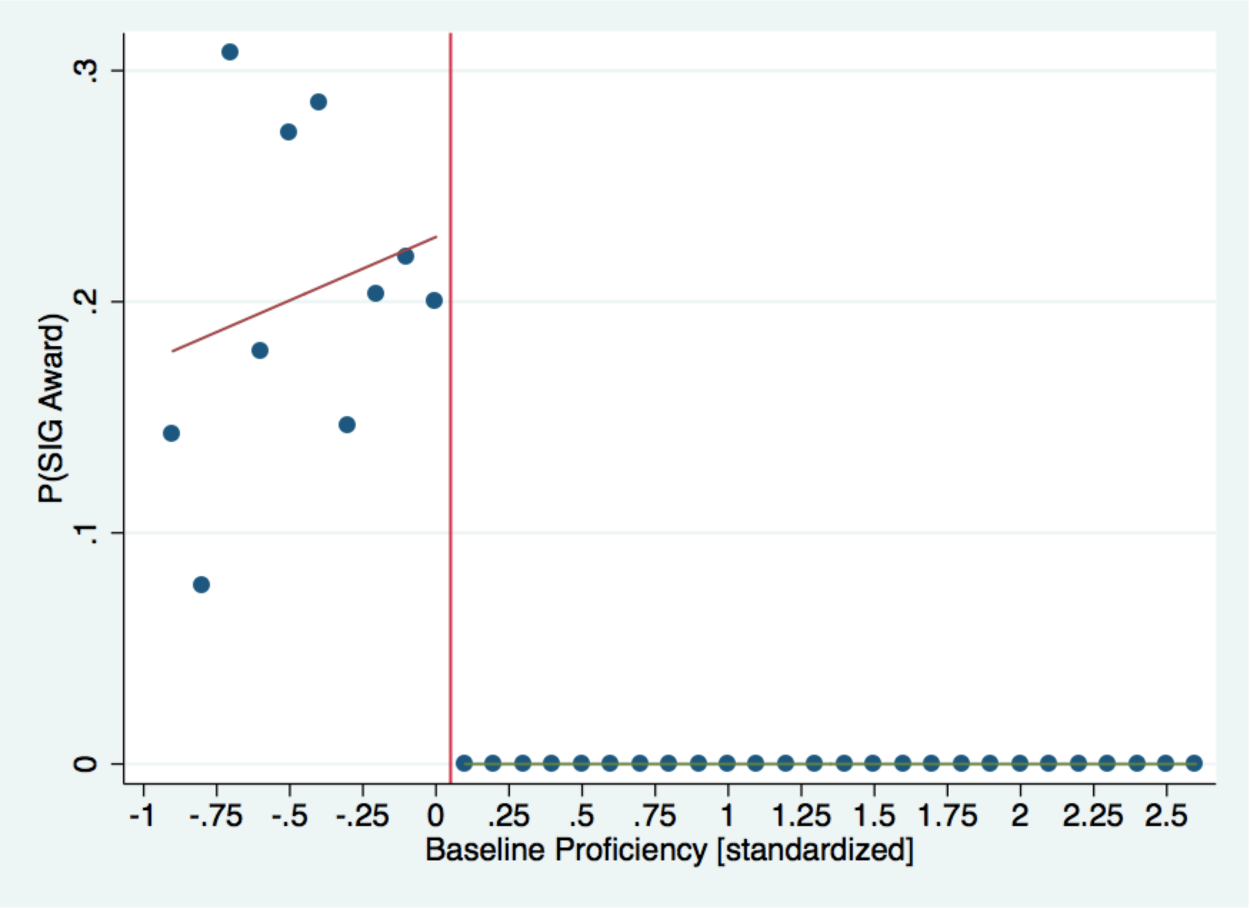


Figure 2 – SIG Awards and Baseline Proficiency Rates (bin width = 0.1)

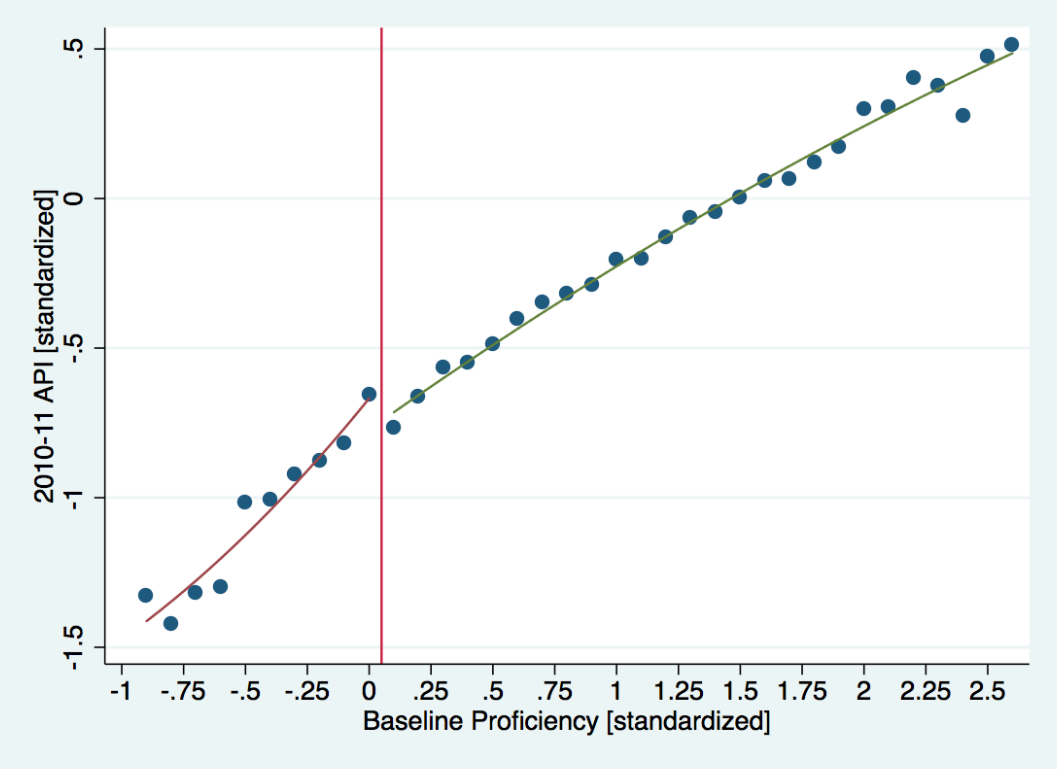


Figure 3(a) – 2010-11 API Scores and Baseline Proficiency Rates (bin width = 0.1)

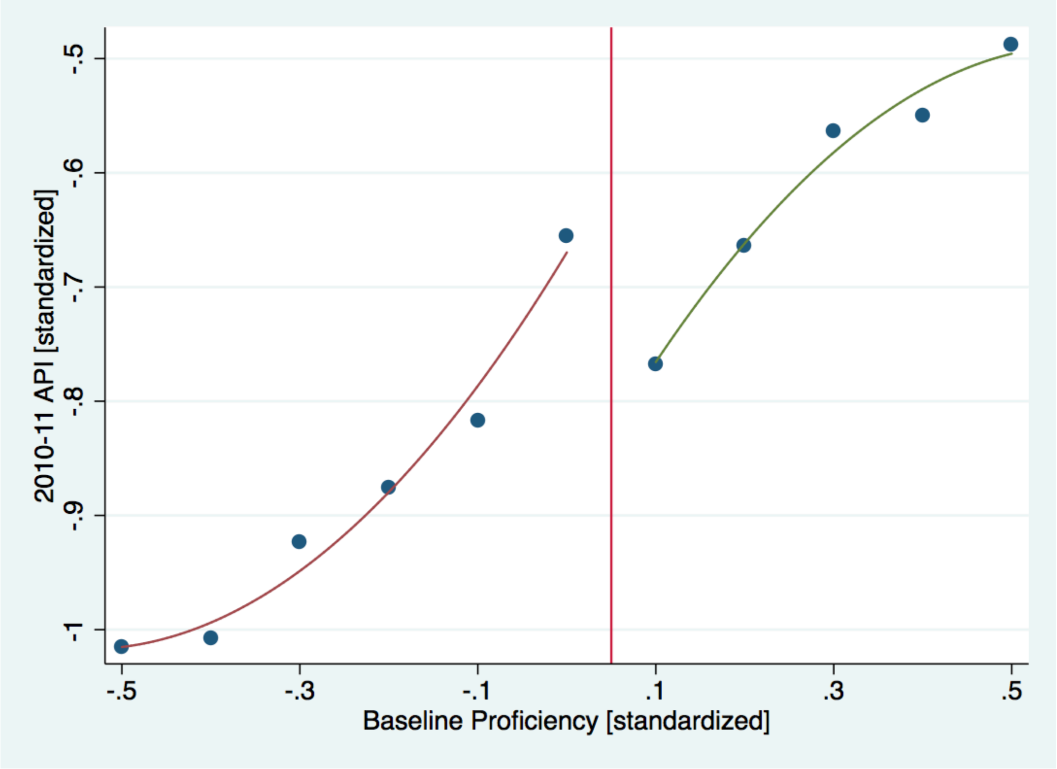


Figure 3(b) – 2010-11 API Scores and Baseline Proficiency Rates (bin width = 0.1, bandwidth = ±0.5)

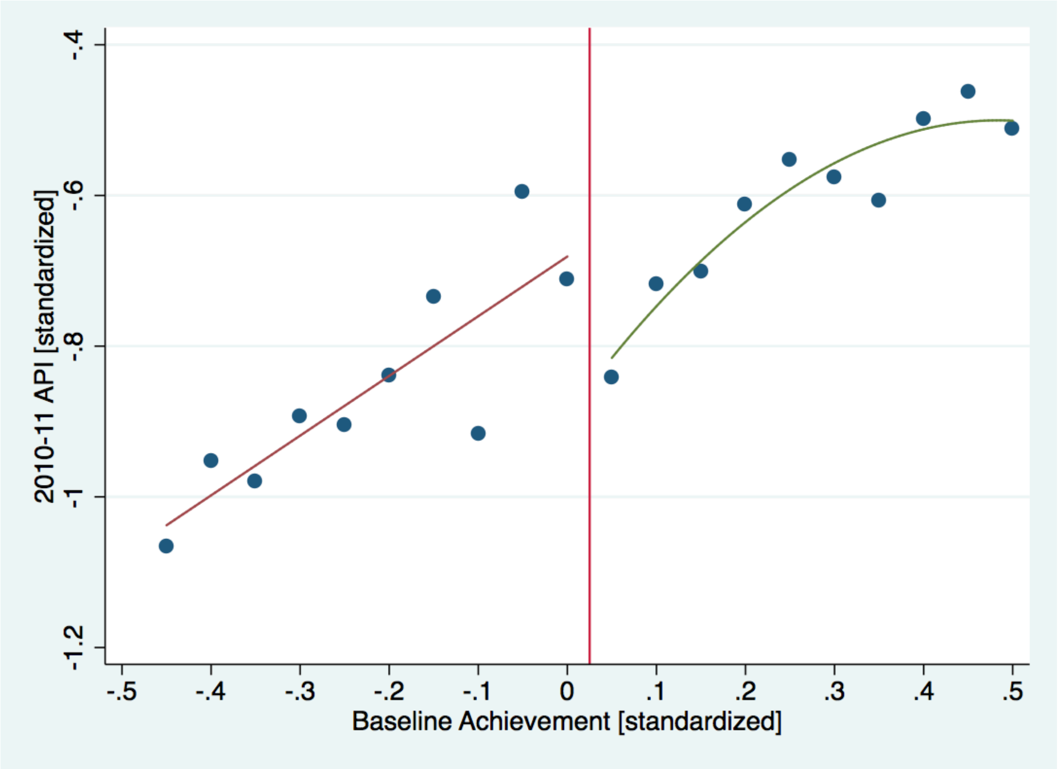


Figure 3(c) – 2010-11 API Scores and Baseline Proficiency Rates (bin width = 0.05, bandwidth = ±0.5)

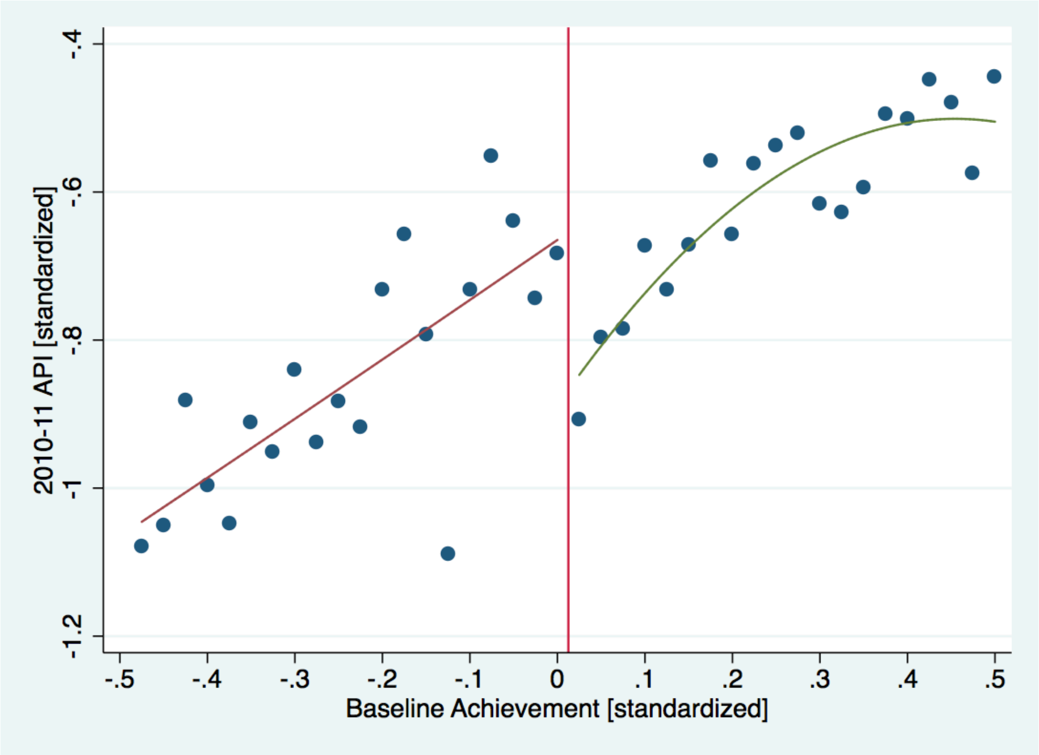


Figure 3(d) – 2010-11 API Scores and Baseline Proficiency Rates (bin width = 0.025, bandwidth = ±0.5)

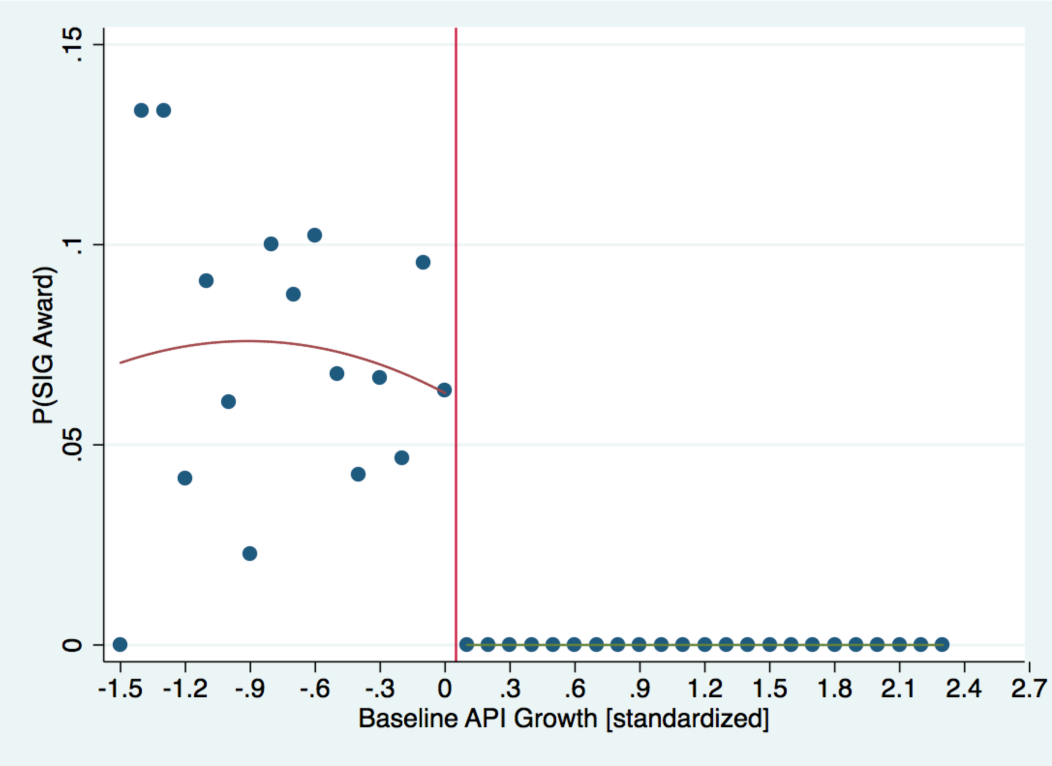


Figure 4 (a) – SIG Awards and Baseline API Growth (bin width = 0.1)

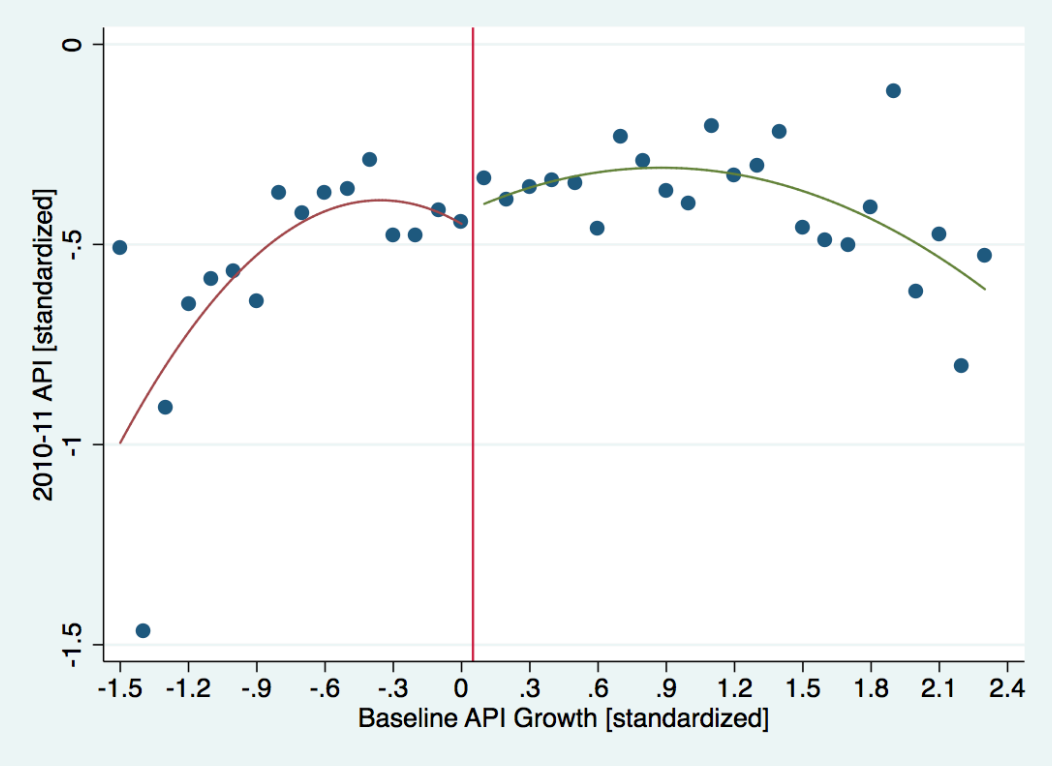


Figure 4 (b) – 2010-11 API Scores and Baseline API Growth (bin width = 0.1)

Table 1 - Variables and Descriptive Statistics

Variable	Mean	Standard Deviation
2010-2011 API (standardized)	-0.409	0.719
SIG Recipient	0.028	0.166
Lowest Achieving	0.187	0.390
Lack of Progress	0.399	0.490
Baseline Proficiency (standardized)	0.707	1.052
Baseline API Growth (standardized)	0.225	0.839
2009-2010 API (standardized)	-0.483	0.735
2008-2009 API (standardized)	-0.511	0.717
2009-10 Student Traits		
% Students Black	0.080	0.110
% Students Hispanic	0.658	0.243
% Students Asian	0.076	0.109
% Students Free/Reduced-Price	0.749	0.213
% Students English Learners	0.343	0.200
% Students with Disabilities	0.107	0.042
2009-10 Teacher Traits		
Average Teacher Experience	13.5	3.2
% Teachers with Graduate Degree	0.369	0.188
% Teachers Black	0.053	0.098
% Teachers Hispanic	0.230	0.185
% Teachers Asian	0.064	0.070
2009-10 School Traits		
Suburb	0.328	0.469
Town	0.101	0.301
Rural	0.092	0.289
Middle School	0.265	0.442
High School	0.210	0.407
Enrollment	843.8	648.0
Pupil-Teacher Ratio	21.3	3.9
2010-11 Student Traits		
% Students Black	0.079	0.109
% Students Hispanic	0.664	0.243
% Students Asian	0.076	0.109
% Students Free/Reduced-Price	0.756	0.216
% Students English Learners	0.331	0.203
% Students with Disabilities	0.109	0.042
2010-11 Teacher Traits		
Average Teacher Experience	13.712	3.445
% Teachers with Graduate Degree	0.375	0.196
% Teachers Black	0.052	0.099
% Teachers Hispanic	0.227	0.189
% Teachers Asian	0.060	0.075
2010-11 School Traits		
Enrollment	823.6	623.9
Pupil-Teacher Ratio	22.8	20.6

Notes: This simple consists of 2,892 California schools meeting the broad criteria for a 2010-11 SIG Award (see text for details). The 2010-11 SIG awards in this sample (n = 82) average $1.48 million (i.e., $1,506 per pupil).

Table 2 - First-Stage and Reduced-Form Lowest-Achieving RD Estimates

	Full Sample (n = 2,892)				$G_s \leq 0$ Sample (n = 1,155)			
Independent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable: SIG Recipient, 1st Stage Estimates							
Lowest Achieving	0.231*** (0.0238)	0.233*** (0.0240)	0.230*** (0.0313)	0.229*** (0.0315)	0.534*** (0.0441)	0.534*** (0.0441)	0.534*** (0.0441)	0.554*** (0.0582)
R^2	0.174	0.185	0.174	0.185	0.397	0.420	0.397	0.421
	Dependent variable: 2011 API Score (standardized), Reduced-Form Estimates							
Lowest Achieving	0.0760*** (0.0192)	0.0718*** (0.0197)	0.0855*** (0.0260)	0.0798*** (0.0258)	0.0963*** (0.0311)	0.0907*** (0.0320)	0.0830* (0.0430)	0.0809* (0.0431)
R^2	0.891	0.893	0.891	0.893	0.906	0.909	0.907	0.910
Baseline API Scores	yes	yes	yes	yes	yes	yes	yes	yes
Linear Spline	yes	yes	yes	yes	yes	yes	yes	yes
Quadratic Spline	no	no	yes	yes	no	no	yes	yes
Student Controls	no	yes	no	yes	no	yes	no	yes
Teacher Controls	no	yes	no	yes	no	yes	no	yes
School Controls	no	yes	no	yes	no	yes	no	yes

Notes: Heteroscedastic-consistent standard errors are reported in parentheses. See Table 1 for a description of the 2009-10 student, teacher, and school controls.

*** p<0.01, ** p<0.05, * p<0.1

Table 3 - First-Stage and Reduced-Form Lack-of-Progress RD Estimates

	Full Sample (n = 2,892)				$A_s \leq 0$ Sample (n = 542)			
Independent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable: SIG Recipient, 1st Stage Estimates							
Lack of Progress	0.0690***	0.0695***	0.0655***	0.0646***	0.406***	0.399***	0.400***	0.395***
	(0.0112)	(0.0111)	(0.0132)	(0.0131)	(0.0404)	(0.0411)	(0.0528)	(0.0546)
R^2	0.069	0.090	0.070	0.090	0.279	0.316	0.280	0.317
	Dependent variable: 2011 API Score (standardized), Reduced-Form Estimates							
Lack of Progress	1.28e-05	0.00277	-0.00208	-0.00333	-0.0293	-0.0226	0.0228	0.0308
	(0.0156)	(0.0154)	(0.0185)	(0.0184)	(0.0437)	(0.0450)	(0.0557)	(0.0575)
R^2	0.889	0.892	0.889	0.892	0.831	0.835	0.832	0.836
Baseline API Scores	yes	yes	yes	yes	yes	yes	yes	yes
Linear Spline	yes	yes	yes	yes	yes	yes	yes	yes
Quadratic Spline	no	no	yes	yes	no	no	yes	yes
Student Controls	no	yes	no	yes	no	yes	no	yes
Teacher Controls	no	yes	no	yes	no	yes	no	yes
School Controls	no	yes	no	yes	no	yes	no	yes

Notes: Heteroscedastic-consistent standard errors are reported in parentheses. See Table 1 for a description of the 2009-10 student, teacher, and school controls.
*** p<0.01, ** p<0.05, * p<0.1

Table 4 - 2SLS and IK Estimates of the Effect of a SIG Award on 2010-11 API Scores

	2SLS Estimates				IK Estimates	
Independent variable	(1)	(2)	(3)	(4)	(5)	(6)
SIG Award	0.329*** (0.0860)	0.308*** (0.0863)	0.372*** (0.119)	0.349*** (0.118)	0.282* (0.149)	0.319** (0.154)
Baseline API Scores	yes	yes	yes	yes	yes	yes
Linear Spline	yes	yes	yes	yes	yes	yes
Quadratic Spline	no	no	yes	yes	no	no
Student Controls	no	yes	no	yes	no	yes
Teacher Controls	no	yes	no	yes	no	yes
School Controls	no	yes	no	yes	no	yes

Notes: The instrumental variable is the "lowest achieving" discontinuity. Heteroscedastic-consistent standard errors are reported in parentheses for the 2SLS estimates. The IK estimates are based on the optimal-bandwidth procedure developed by Imbens and Kalyanaraman (2009) for fuzzy RD applications. See Table 1 for a description of the 2009-10 student, teacher, and school controls.

*** p<0.01, ** p<0.05, * p<0.1

Table 5 - Placebo and Actual Reduced-Form RD Estimates, 2011 API Scores

Independent variable	(1)	(2)
Placebo RD: $A_s \leq -0.5$	-0.00992	0.00112
	(0.0614)	(0.0610)
Placebo RD: $A_s \leq -0.4$	0.0193	0.0174
	(0.0601)	(0.0600)
Placebo RD: $A_s \leq -0.3$	0.0204	0.0164
	(0.0492)	(0.0502)
Placebo RD: $A_s \leq -0.2$	-5.42e-05	-0.00377
	(0.0446)	(0.0449)
Placebo RD: $A_s \leq -0.1$	-0.0392	-0.0436
	(0.0409)	(0.0404)
Actual RD: $A_s \leq 0$	**0.0865****	**0.0887****
	(0.0359)	**(0.0360)**
Placebo RD: $A_s \leq 0.1$	0.00640	0.00456
	(0.0295)	(0.0294)
Placebo RD: $A_s \leq 0.2$	-0.0354	-0.0297
	(0.0307)	(0.0309)
Placebo RD: $A_s \leq 0.3$	0.0236	0.0204
	(0.0374)	(0.0376)
Placebo RD: $A_s \leq 0.4$	0.0423	0.0389
	(0.0341)	(0.0340)
Placebo RD: $A_s \leq 0.5$	-0.0173	-0.0148
	(0.0210)	(0.0204)
Student Controls	no	yes
Teacher Controls	no	yes
School Controls	no	yes

Note: Each cell identifies the estimated effect of a "lowest achieving" discontinuity defined for different values of the assignment variable (i.e., the baseline proficiency rate). Heteroscedastic-consistent standard errors are reported in parentheses. Both models condition on 2009 and 2010 API scores and a linear spline of the assignment variable (n = 2,892)
*** p<0.01, ** p<0.05, * p<0.1

Table 6 - RD estimates, Full Sample and by Bandwidth

	Dependent Variable				
Sample	SIG Recipient	2011 API Score	SIG Recipient	2011 API Score	Sample Size
Full Sample	0.231***	0.0760***	0.233***	0.0718***	2,892
	(0.0238)	(0.0192)	(0.0240)	(0.0197)	
$\lvert A_s \rvert \leq 2$	0.236***	0.0838***	0.236***	0.0742***	2,587
	(0.0256)	(0.0214)	(0.0258)	(0.0213)	
$\lvert A_s \rvert \leq 1.75$	0.233***	0.0828***	0.232***	0.0718***	2,438
	(0.0270)	(0.0236)	(0.0273)	(0.0230)	
$\lvert A_s \rvert \leq 1.5$	0.235***	0.0735***	0.235***	0.0643***	2,252
	(0.0286)	(0.0232)	(0.0287)	(0.0227)	
$\lvert A_s \rvert \leq 1.25$	0.225***	0.0533**	0.224***	0.0490**	2,024
	(0.0310)	(0.0233)	(0.0309)	(0.0231)	
$\lvert A_s \rvert \leq 1.0$	0.215***	0.0447*	0.214***	0.0443*	1,692
	(0.0334)	(0.0253)	(0.0330)	(0.0256)	
$\lvert A_s \rvert \leq 0.75$	0.203***	0.0491*	0.204***	0.0520*	1,301
	(0.0373)	(0.0281)	(0.0369)	(0.0283)	
Student Controls	no	no	yes	yes	
Teacher Controls	no	no	yes	yes	
School Controls	no	no	yes	yes	

Notes: Each cell identifies the estimated effect of the "lowest achieving" discontinuity. Standard errors are adjusted for heteroscedasticity. All models condition on baseline API scores and a linear spline of the assignment variable, A_s. See Table 1 for a description of the 2009-10 student, teacher, and school controls.
*** p<0.01, ** p<0.05, * p<0.1

Table 7 - Auxiliary RD Estimates, Baseline 2009-10 Covariates

Dependent variable	(1)	(2)
2009-10 API Scores	0.0259	0.0265
	(0.0251)	(0.0248)
2008-09 API Scores	0.00772	0.00975
	(0.0190)	(0.0179)
2009-10 Student Traits		
% Black	0.00149	0.00155
	(0.0133)	(0.00714)
% Hispanic	-0.0223	-0.00593
	(0.0220)	(0.0124)
% Asian	0.00299	0.000588
	(0.00873)	(0.00714)
% Free/Reduced-Price	-0.0146	-0.0144
	(0.0157)	(0.0116)
% English Learners	-0.0126	-0.00314
	(0.0200)	(0.0114)
% with Disabilities	0.00485	0.00404
	(0.00398)	(0.00372)
2009-10 Teacher Traits		
Average Experience	-0.151	-0.100
	(0.318)	(0.287)
% with Graduate Degree	-0.0103	0.00266
	(0.0186)	(0.0167)
% Black	-0.00665	-0.00444
	(0.0143)	(0.00868)
% Hispanic	-0.0198	-0.00150
	(0.0178)	(0.0140)
% Asian	0.00417	0.00603
	(0.00664)	(0.00578)
2009-10 School Traits		
Suburb	0.0234	0.0201
	(0.0470)	(0.0466)
Town	0.0323	0.0296
	(0.0321)	(0.0305)
Rural	-0.00328	-0.0109
	(0.0279)	(0.0255)
Middle School	0.0589	0.0467
	(0.0377)	(0.0335)
High School	-0.0558	-0.0475*
	(0.0392)	(0.0260)
ln(Enrollment)	-0.0572	-0.00198
	(0.0678)	(0.0491)
Pupil-Teacher Ratio	0.154	0.336
	(0.390)	(0.357)

Note: Each cell identifies the estimated effect of the "lowest achieving" discontinuity on the baseline covariate. Heteroscedastic-consistent standard errors are reported in parentheses. All models condition on a linear spline of the assignment variable and baseline API scores, where defined. Model 2 also conditions on school, teacher, and student controls, where defined. These estimates are based on the observations (n = 1,671) within the Imbens-Kalyanaraman optimal bandwidth; see columns (5) and (6) in Table 5.
*** p<0.01, ** p<0.05, * p<0.1

Table 8 - Auxiliary RD Estimates, 2010-11 Covariates

Dependent variable	(1)	(2)
% Black	0.00471	0.00315
	(0.0131)	(0.00197)
% Hispanic	-0.0249	-0.00264
	(0.0217)	(0.00271)
% Asian	0.000819	-0.00240
	(0.00868)	(0.00155)
% Free/Reduced-Price	-0.0198	-0.0105
	(0.0170)	(0.0106)
% English Learners	-0.0147	-0.00113
	(0.0205)	(0.00626)
% with Disabilities	0.00552	0.00137
	(0.00421)	(0.00207)
ln(Enrollment)	-0.0352	0.0218
	(0.0679)	(0.0193)
Baseline API Scores	yes	yes
Linear Spline	yes	yes
Student Controls	no	yes
Teacher Controls	no	yes
School Controls	no	yes

Note: Each cell identifies the estimated effect of the "lowest achieving" discontinuity on the *post-treatment* covariate. Heteroscedastic-consistent standard errors are reported in parentheses. All models condition on a linear spline of the assignment variable and school, teacher, and student controls, where defined. These estimates are based on the observations (n = 1,671) within the Imbens-Kalyanaraman optimal bandwidth; see columns (5) and (6) in Table 5.
*** p<0.01, ** p<0.05, * p<0.1

Table 9 - 2SLS and IK Estimates of the Effect of a SIG Award on Suspension and Truancy Rates

	Dependent Variable				
Dependent variable	2SLS Estimate	IK Estimate	2SLS Estimate	IK Estimate	Dependent Mean
2009-10 Suspension Rate	-0.0652	0.122	-0.0809	-0.026	0.175
	(0.0831)	(0.159)	(0.0740)	(0.158)	
2009-10 Truancy Rate	0.119	-0.100	0.0698	-0.133	0.321
	(0.0936)	(0.199)	(0.0888)	(0.187)	
2010-11 Suspension Rate	-0.0891*	0.092	-0.120**	-0.059	0.151
	(0.0541)	(0.102)	(0.0484)	(0.092)	
2010-11 Truancy Rate	0.0504	-0.221	-0.0116	-0.230*	0.340
	(0.0811)	(0.140)	(0.0734)	(0.130)	
Student Controls	no	no	yes	yes	
Teacher Controls	no	no	yes	yes	
School Controls	no	no	yes	yes	

Notes: These dependent variables are available for n = 2,666 schools. The instrumental variable is the "lowest achieving" discontinuity. Heteroscedastic-consistent standard errors are reported in parentheses for the 2SLS estimates. The IK estimates are based on the optimal-bandwidth procedure developed by Imbens and Kalyanaraman (2009) for fuzzy RD applications. See Table 1 for a description of the 2009-10 student, teacher, and school controls.
*** p<0.01, ** p<0.05, * p<0.1

Table 10 - 2SLS and IK Estimates of the Effect of a SIG Award on Candidate 2010-11 Mediators

Dependent variable	2SLS Estimates	IK Estimates
New Principal	0.340**	0.0610
	-0.152	(0.325)
Average Teacher Experience	-2.075***	-2.152*
	(0.666)	(1.286)
% Teachers with Graduate Degree	0.00675	-0.0573
	(0.0299)	(0.061)
% Teachers Black	0.0108	0.0001
	(0.0126)	(0.026)
% Teachers Hispanic	0.0303	0.0058
	(0.0205)	(0.039)
% Teachers Asian	-0.00778	-0.0033
	(0.0138)	(0.028)
Pupil-Teacher Ratio	-4.733	-4.957**
	(3.237)	(2.448)
Baseline API Scores	yes	yes
Linear Spline	yes	yes
Student Controls	yes	yes
Teacher Controls	yes	yes
School Controls	yes	yes

Notes: The instrumental variable is the "lowest achieving" discontinuity. Heteroscedastic-consistent standard errors are reported in parentheses for the 2SLS estimates. The IK estimates are based on the optimal-bandwidth procedure developed by Imbens and Kalyanaraman (2009) for fuzzy RD applications. See Table 1 for a description of the 2009-10 student, teacher, and school controls.

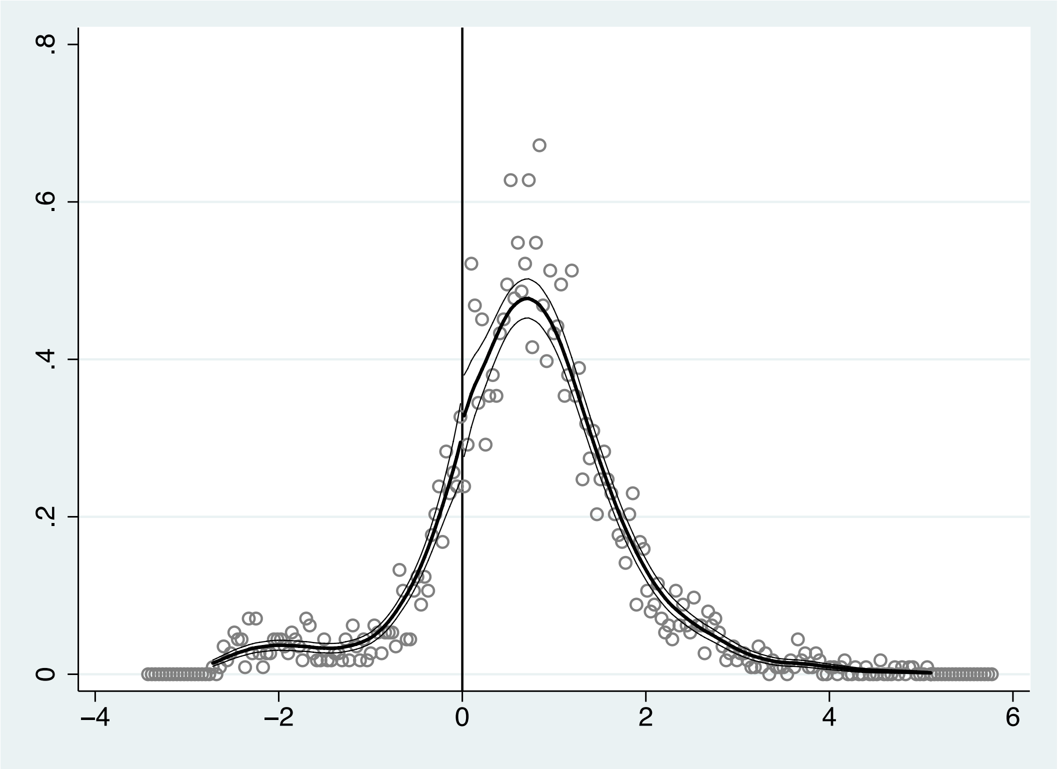
*** p<0.01, ** p<0.05, * p<0.1

Table 11 - Difference-in-Differences Estimates of the Effects of Specific SIG Reform Models

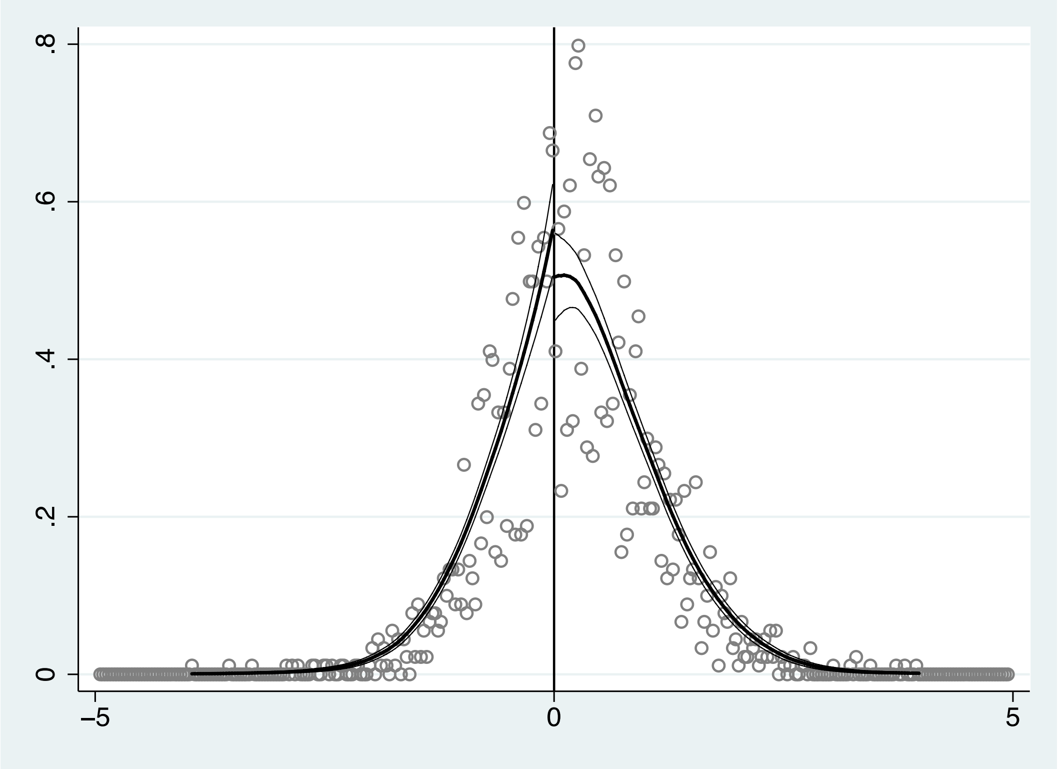
	SIG Model			
Sample	Transformation	Turnaround	Restart	Sample Size
Full Sample	0.044	0.268***	-0.109	2,892
	(0.037)	(0.063)	(0.091)	
Lowest-Achieving Schools ($\|A_s\| \leq 0$)	0.043	0.231***	-0.127	542
	(0.041)	(0.065)	(0.107)	
Lack of Progress Schools ($\|G_s\| \leq 0$)	0.030	0.255***	-0.129	1,155
	(0.040)	(0.064)	(0.095)	
SIG-Eligible Schools	0.014	0.228***	-0.127	168
	(0.051)	(0.073)	(0.123)	

Note: The dependent variable is API growth for AY 2010-11 (i.e., the difference between the 2011 "growth" API and the 2010 "base" API). Heteroscedastic-consistent standard errors are reported in parentheses. All models condition on 2009 API scores and the school, teacher, and student controls listed in Table 1.

*** p<0.01, ** p<0.05, * p<0.1



Appendix Figure 1(a) Density Test - Baseline Proficiency Rate



Appendix Figure 1(b) Density Test - Baseline API Growth

Table A1 - Estimated Effects of Student, Teacher, and School Traits on 2010-11 API Scores

Independent variables	(1)	(2)	(3)	(4)
2009-2010 API (standardized)	0.678***	0.678***	0.592***	0.592***
	(0.0297)	(0.0296)	(0.0528)	(0.0525)
2008-2009 API (standardized)	0.213***	0.215***	0.258***	0.261***
	(0.0427)	(0.0425)	(0.0838)	(0.0837)
Student Controls				
% Black	-0.134	-0.131	-0.184	-0.181
	(0.0964)	(0.0960)	(0.137)	(0.137)
%Hispanic	0.0387	0.0416	0.122	0.127
	(0.0522)	(0.0523)	(0.0781)	(0.0787)
% Asian	-0.0497	-0.0540	0.0592	0.0631
	(0.0573)	(0.0573)	(0.0889)	(0.0899)
% Free/Reduced	-0.0810**	-0.0782*	-0.0590	-0.0572
	(0.0400)	(0.0401)	(0.0661)	(0.0664)
% English learners	-0.0233	-0.0271	-0.0241	-0.0316
	(0.0431)	(0.0437)	(0.0741)	(0.0756)
% Disabled	0.0956	0.105	-0.0571	-0.0431
	(0.137)	(0.138)	(0.213)	(0.216)
Teacher Controls				
% with a Graduate Degree	0.0798***	0.0788***	0.0592	0.0575
	(0.0276)	(0.0277)	(0.0451)	(0.0453)
Average Teacher Experience	-0.00507***	-0.00503***	-0.00763***	-0.00744**
	(0.00193)	(0.00194)	(0.00296)	(0.00299)
% Black	0.0330	0.0307	0.105	0.0971
	(0.0866)	(0.0870)	(0.137)	(0.137)
% Hispanic	0.0218	0.0180	0.0221	0.0182
	(0.0376)	(0.0377)	(0.0693)	(0.0698)
% Asian	0.129	0.139*	-0.0947	-0.0905
	(0.0822)	(0.0820)	(0.133)	(0.134)
School Controls				
Suburb	0.00573	0.00584	-0.0240	-0.0233
	(0.00985)	(0.00985)	(0.0158)	(0.0159)
Town	-0.00249	-0.00176	-0.00329	-0.00148
	(0.0170)	(0.0171)	(0.0289)	(0.0291)
Rural	0.0206	0.0212	0.00483	0.00550
	(0.0189)	(0.0190)	(0.0270)	(0.0271)
Middle	-0.0352**	-0.0355**	-0.0404	-0.0415
	(0.0171)	(0.0170)	(0.0264)	(0.0264)
High	-0.0655***	-0.0642***	-0.0471	-0.0483
	(0.0227)	(0.0224)	(0.0361)	(0.0360)
Enrollment	0.0134	0.0135	-0.00502	-0.00504
	(0.0128)	(0.0128)	(0.0209)	(0.0210)
Pupil/Teacher ratio	0.000149	0.000150	0.00181	0.00185
	(0.00182)	(0.00182)	(0.00283)	(0.00281)

Notes: This table reports the estimated coefficients on the student, teacher, and school controls used in the reduced-form achievement equations (i.e.,models 2, 4, 6, and 8 from the bottom panel of Table 2)
*** p<0.01, ** p<0.05, * p<0.1