University of Pennsylvania
ScholarlyCommons

CPRE Policy Briefs

Consortium for Policy Research in Education
(CPRE)

9-2012

Learning From NCLB: School Responses to Accountability Pressure and Student Subgroup Performance

Elliot H Weinbaum
University of Pennsylvania

Michael J. Weiss

Jessica K. Beaver
University of Pennsylvania, jessica.k.beaver@gmail.com

Follow this and additional works at: http://repository.upenn.edu/cpre_policybriefs

Part of the Educational Administration and Supervision Commons, Educational Assessment, Evaluation, and Research Commons, Educational Leadership Commons, Educational Methods Commons, Education Policy Commons, and the Teacher Education and Professional Development Commons

Learning From NCLB: School Responses to Accountability Pressure and Student Subgroup Performance

Abstract

Much has been written in the last decade about the spotlight that the No Child Left Behind Act (NCLB) shines on schoolperformance. Proponents and opponents alike are quick to discuss the law's rigid definitions of school performance— exemplified by the classification of schools as making Adequate Yearly Progress (AYP) or not making AYP based largely on annual tests in reading and mathematics, disaggregating school performance by student subgroups, and requiring that all schools reach 100% proficiency. Yet for all its rigidity, the law has offered schools little guidance on how to make use of the performance data that the new systems provide or how to design improvement efforts. As policymakers discuss ways to change NCLB or design new federal education policies targeted at improving academic achievement, we present new research findings that can help to inform those discussions.

In this CPRE Policy Brief, we examine the extent to which the assumptions in the law manifest themselves in the actions that school leaders take. This brief asks and answers the question: How do school leaders—administrators and teachers— respond to the results of state assessment systems and the pressure of performance-based accountability? And do those responses seem to matter to achievement outcomes?

Disciplines

Educational Administration and Supervision | Educational Assessment, Evaluation, and Research | Educational Leadership | Educational Methods | Education Policy | Teacher Education and Professional Development

Comments

View on the CPRE website.

CPRE

Learning from NCLB: School Responses to Accountability Pressure and Student Subgroup Performance

By Elliot H. Weinbaum, Michael J. Weiss and Jessica K. Beaver

Graduate School of Education
University of Pennsylvania

Much has been written in the last decade about the spotlight that the No Child Left Behind Act (NCLB) shines on school performance. Proponents and opponents alike are quick to discuss the law's rigid definitions of school performance—exemplified by the classification of schools as making Adequate Yearly Progress (AYP) or not making AYP based largely on annual tests in reading and mathematics, disaggregating school performance by student subgroups, and requiring that all schools reach 100% proficiency. Yet for all its rigidity, the law has offered schools little guidance on how to make use of the performance data that the new systems provide or how to design improvement efforts. As policymakers discuss ways to change NCLB or design new federal education policies targeted at improving academic achievement, we present new research findings that can help to inform those discussions.

NCLB is based on the assumption that by using new data provided by testing, drawing public attention to student performance, and establishing sanctions for poor results, teachers and school leaders will be motivated and able to identify and adopt successful strategies for their students (Stecher, Epstein, Hamilton, Marsh, Robyn, McCombs et al., 2008; Hamilton, Berends, & Stecher, 2005; Haertel & Herman, 2005; Linn, 2005). In order for this assumption to be accurate, being identified as "in need of improvement" (the designation for schools that fail to meet AYP goals) would have to set off a chain reaction, wherein school or district personnel examine performance data, draw conclusions about where their challenges lie, search for programs and materials to address their challenges, and finally implement those new programs, balancing fidelity to the programs' designs with sensitivity to local context.

Prior to the mandatory testing and reporting required by NCLB, school improvement efforts were shown to lack coherence (Newman, Smith, Allensworth, & Bryk, 2001) and often included conflicting programs (Hatch, 2002). Part of the theory of performance-based accountability in general, and NCLB in particular, was based on the belief that regular and comprehensive evidence would help to focus these efforts. Research seems to indicate, however, that individuals vary widely in their assumptions about the value and purpose of evidence (Coburn & Talbert, 2006), that it is often difficult to identify students' challenges based on the resulting annual data (Black & Wiliam, 1998), and that searches for meaningful remedies to real or imagined problems are often extremely limited, somewhat chaotic, and frequently lead back to familiar practices as opposed to real change or innovation (Gross, Kirst, Holland, & Luschei, 2005).

In this CPRE Policy Brief, we examine the extent to which the assumptions in the law manifest themselves in the actions that school leaders take. This brief asks and answers the question: How do school leaders—administrators and teachers—respond to the results of state assessment systems and the pressure of performance-based accountability? And do those responses seem to matter to achievement outcomes?

Our findings are drawn from a three-year CPRE study of schools throughout the Commonwealth of Pennsylvania. Our data come from four main sources. First, we conducted phone interviews with 48 Pennsylvania principals in the fall of 2008 to determine the various strategies they employ to meet the performance targets required under NCLB. Second, using data from the telephone interviews, we developed and administered a survey to all elementary and secondary school principals in Pennsylvania in 2009, asking detailed questions about the types of strategies schools employ and the degree of effort they exert to support those strategies. Third, we conducted in-depth site visits at 11 schools—9 of which had been placed in "warning" status as a result of not making AYP following the 2007-2008 school year and two of which had made AYP. Each school was visited twice—once in the spring of 2009 and again in the spring of 2010—to get a sense of each school's improvement efforts and strategies over time. Finally, we looked at school-level achievement data to examine the link between selected improvement strategies and school performance. (See Appendix for more detail about the data collection).

Using previous research on school improvement (e.g., Gross & Goertz, 2005; Marzano, 2003; Newman et al., 2001; Learning First Alliance, 2003) and the initial principal interviews as guides, we identified nine categories of school improvement efforts on our principal survey. On the survey, principals were asked about each category as well as about 46 specific strategies related to the nine categories. (See Box 1 for a list of the nine categories.) For example, principals were asked how much effort they devoted to "introducing new instructional approaches" and also were asked specific questions about the use of strategies such as "common instructional practices in every classroom (e.g., differentiated instruction, responsive classrooms, student engagement strategies, formative assessment, etc.)" or the adoption of "literacy strategies used across the school (e.g., Reading Apprenticeship, guided reading, etc.)." While some of the categories are quite broad, the specific strategies listed on the survey all fall into one of these nine categories and are more narrowly defined. In addition, principals were asked whether they devoted "minor," "moderate," or "major" effort to each of the general categories and specific strategies during the current and previous school years.

Box 1

Nine Categories of Strategies for Improvement
1. New instructional approaches
2. New student and staff schedules
3. New or aligned curriculum
4. State test (PSSA) preparation
5. Remediation for struggling students
6. Data analysis to guide improvement
7. Outside expertise
8. Rewards and sanctions for performance
9. Efforts to address non-academic issues

Our objective is to examine three specific claims about NCLB to see if they hold true 10 years after the law's enactment. We first ask if schools that have been identified as "in need of improvement" implement more efforts or different kinds of efforts than those schools that meet performance targets. Some previous research suggests that while all schools are responsive to accountability pressure (Goertz & Gross, 2005), high-performing schools and low-performing schools select fundamentally different strategies (Rouse, Hannaway, Goldhaber, & Figlio, 2007; Mintrop, 2004; Hopkins, Harris, & Jackson, 1997). This research suggests that schools with large percentages of low-scoring students are more likely to focus on "teaching to the test," while schools that generally perform well but see problems with certain groups of students are more likely to consider and adopt more fundamental educational change (Haertel & Herman, 2005; Anagnostopoulos, Rutledge, Lynn, & Dreeben, 1999; Koretz, Barron, Mitchell, & Stecher, 1996; Darling-Hammond & Wise, 1985). To some extent, our research contradicts these claims from earlier research. We found that schools that make AYP and those that don't make AYP generally place major effort on the same types of reform strategies. Similar to some more recent research (Mintrop & Trujillo, 2007), we find that low- and high-performing schools were not clearly distinct in terms of the improvement strategies that they use. Furthermore, in our data, low-performing schools and high-performing schools all exert major efforts to support a large and varied number of reform strategies. The major difference that we found between high- and low-performing schools is that low-performing schools exert major efforts on *more* reform strategies than high-performing schools. Based on their self reports, low-performing schools are supporting more strategies with more effort than high-performing schools, but all schools are prioritizing efforts in very similar ways. Other recent research also finds common strategies adopted across diverse groups of schools, even while certain schools may emphasize particular strategies more (GAO, 2009).

Second, we examine whether disaggregation of data by the student subgroups required by NCLB (such as race, socioeconomic status, and special education) influences the selection of school improvement strategies. Advocates for including subgroup

reporting as part of NCLB generally argued one of two positions; either 1) that identifying challenges unique to particular student subgroups will help schools target their improvement strategies to subgroups of students who need the most support, or 2) that examining subgroup performance would help to identify schools that are not serving all students well and would otherwise have been deemed successful if only overall student performance was considered. Though the literature here is sparse, previous research on the use of subgroup data suggests that some schools do make use of subgroup data to target students (Booher-Jennings, 2005), sometimes in ways that do not benefit particular student populations (Valenzuela, 2004). While we do find that disaggregation of data by student subgroups identifies many schools as failing to make AYP that would not otherwise have been identified, our research does not find any claims of widespread use of data to target particular subgroups. We find that schools with different failing subgroups generally use improvement efforts in similar ways, regardless of the particular subgroup that is failing.

Third, we ask a question related to the fundamental goal of NCLB: Are the strategies that schools select for improvement related to gains in academic achievement? Although many educators and policymakers express negativity about schools that emphasize test preparation as a response to accountability pressure, we find that when examining nine categories of improvement strategies, test preparation is most positively associated with achievement gains. Though our research could not determine a causal relationship between the emphasis on test preparation and the improvement in test scores, it suggests that schools may be justified in choosing test preparation strategies because it helps them to meet the goals of the accountability system in which they work. Positive associations were also found between achievement and strategies targeted at providing remediation for struggling students and strategies focusing on using data, suggesting these strategies may also be successful ways for schools to respond to high-stakes accountability measures.

In sum, our findings deepen our understanding about schools' responses to accountability pressures. Looking at all schools in a single state, we find that schools at different achievement levels and with different types of subgroup failures all appear to

choose many varied strategies for improvement and prioritize those efforts in very similar ways. Additionally, we add to the evidence that test preparation is both extremely popular and seems to be the strategy most closely associated with testing gains. These findings have important implications for policymakers and practitioners alike. At the end of this Policy Brief, we consider the implications of these findings. First, we provide a bit more detail about our research and results.

Schools' Strategies for Improvement

Given that NCLB provides mechanisms for assessing student and school performance but offers no specific strategies for improvement, we were interested in looking deeper into the strategies that schools adopt. Specifically, we were interested in examining whether schools failing to make AYP adopt more or different types of improvement strategies than those that have made AYP; and whether schools with low performance in a particular student subgroup act differently, depending on the particular group that is failing.

Comparing Schools based on the AYP label. Using survey data in combination with state test performance data, we compared survey results for schools making AYP and those failing to make AYP. Table 1 shows that schools that fail to make AYP are more likely to report placing major effort on each one of the nine improvement categories (on average, 14 percentage points more likely). This may reflect a reality that low-performing schools truly are placing major effort on more school improvement strategies than high-performing schools, or it may reflect the effect of external pressure on school leaders. This pressure may result from noting areas of low performance in test results, from making those results public, and/or from specific state and district policies requiring certain actions of all schools or those that fail to make AYP (evidence to support this last point can be found in Padilla et al., 2006).

**Table 1. Current Reform Strategies by AYP Status -
Ranked by Percent of Schools Placing Major Effort on Strategy**

Reform Strategy	Full Sample		Made AYP		Failed to Make AYP	
	Rank	(%)	Rank	(%)	Rank	(%)
Use performance data to inform practice	1	72.0	1	69.5	1	78.2
Provide remediation for underperforming students	2	66.8	2	64.6	2	72.1
Introduce new instructional approaches	3	52.1	3	48.0	3	62.2
Improve the quality/alignment of curriculum	4	48.2	4	44.2	5	58.2
Allot time for PSSA preparation	5	45.8	5	39.9	4	60.7
Address non-academic issues	6	44.6	6	39.8	6	56.5
Change school and/or staff schedules	7	31.0	7	26.8	7	41.3
Bring in outside support and expertise	8	21.4	8	17.7	9	30.5
Create rewards and sanctions related to test performance	9	20.0	9	14.6	8	33.2
Sample Size	1900		1355		545	

Sources: Calculations from Principal Survey data.   Notes: Sample size varies by question, due to non-response.

According to the survey results, regardless of AYP classification, the relative popularity of categories of improvement efforts is extremely similar between the two groups of schools. The most popular strategies (using student performance data and providing remediation for struggling students) are popular among schools that make AYP as well as among schools that fail to make AYP. The least popular strategies (introducing new rewards and sanctions, bringing in outside support, and changing students' or teachers' schedules) are consistently less popular across the two types of schools. This suggests that some goals of NCLB are being met; data are being used and schools are trying to address the needs of students who are not achieving. Less popular strategies are those that are difficult to implement or would require new resources and policies.

It is also worth noting a couple of departures from the general similarity between the two groups of schools. First, more schools that failed to make AYP than schools that made AYP report exerting major effort on test preparation and addressing non-academic issues. While 61% of schools that failed to make AYP report placing major effort on test preparation, only 40% of schools that made AYP place major effort on this approach. However, it does not appear that the effort placed on test preparation and non-academic issues comes at the expense of efforts to improve curriculum or instruction. Even while they devote more effort to test preparation, low-performing schools still report engaging at similar levels to their higher performing peers in the kinds of deeper reforms that some have previously hypothesized are less likely to take place in low-performing schools. Second, more schools that failed to make AYP than schools that made AYP report creating rewards and sanctions related to test performance. In combination, these two differences may suggest that schools failing to make AYP are somewhat more attentive to issues directly related to test performance.

As is true with the general improvement categories, the popularity of specific improvement strategies is extremely similar among schools that failed to make AYP or made AYP. When looking at the 46 strategies that were listed on the survey (and are each related to one of the nine categories), principals prioritized efforts in similar ways regardless of the school's performance category. Certain reform strategies were generally popular. Benchmark testing, for example, was the most popular strategy reported as a major effort both by schools that made AYP (69%) and failed to make AYP (79%). Using benchmark testing results to identify students for remediation was a strategy for 60% of schools that made AYP and 74% of schools that did not make AYP. In combination, these demonstrate schools' commitment to using data for school improvement and to identify struggling students. Classroom observation of teachers was the third most popular strategy reported both by schools that made AYP (63%) and failed to make AYP (72%). Similarly, certain reform efforts were generally unpopular. Rewarding teachers for student performance, for example, was only done at 3% of schools making AYP and 10% of those failing to make AYP. Using external consultants was also near the bottom of the list for both school types.

The survey findings also highlight the fact that all school leaders report exerting major effort to support a large number of improvement efforts at the same time. On average, principals reported devoting moderate or major effort to 15 unique strategies (of the 46 options listed on the survey) to improve student performance. Figure 1 illustrates the distribution of schools that reported devoting major effort to various numbers of improvement strategies. Even recognizing that some of the strategies listed on the survey may overlap (for example, "introducing block scheduling" and "creating more time for reading instruction" were two separate strategies on the survey, but could be related in some cases), 15 is a large number of efforts to focus on in a given year. It is worth considering whether schools, particularly those schools that fail to make AYP, can adequately support so many efforts simultaneously.

Figure 1. Number of Reform Strategies on which Principals Reported Placing Major Effort
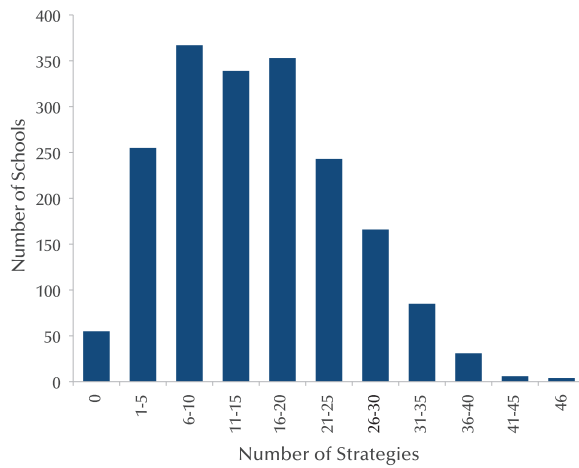
Our case-study visits to 11 schools—9 of which were designated as in "warning status" due to failing to make AYP for the first time in the year prior to our visit – helped to provide a picture of how schools respond to accountability pressure and may help to explain the high levels of effort reported in schools that failed to make AYP. We found that teachers and administrators were quite concerned about the "warning" label, reporting that the label often led them to redouble their efforts in certain ways. However, as the survey data suggest, staff did not describe initiatives that were materially different from their higher performing peers. Regardless of performance level, interviewees described trying everything they could think of that would help students and improve test scores. As one teacher at a low-performing elementary school stated, "When you are in warning, it's 'I better plug a little bit harder and go over things more, be more vigilant.'" This sentiment is strikingly similar to that expressed by the principal of a high-performing elementary school who said, "Our scores are good; they could be better…We do a lot of interventions in the classroom and we do a lot of [test] prep obviously."

In the 11 schools in our interview sample, it was not possible to detect a difference in the numbers or types of strategies that staff reported adopting. It is clear that all respondents were actively trying to improve student outcomes, with a special focus on outcomes on the state tests (known as the Pennsylvania System of School Assessment or PSSA). The case study interviews also helped to explain the overlapping nature of many of the

strategies that were most popular on the survey. Data use, for example, was often a starting point that led staff to pursue other strategies, such as modifying curriculum or targeting particular students. Similarly, efforts to provide remediation to struggling students often took the form of adopting new instructional programs or engaging in explicit PSSA preparation. In the interviews, the use of standardized data from PSSAs or benchmark assessments regularly preceded efforts in many of the other improvement categories identified on the survey. It is therefore not surprising that data use was most commonly reported, as it was often a precursor to other improvement efforts.

Looking at the Subgroups. As mentioned above, one of the novel aspects of NCLB is the disaggregation of performance data by student subgroup. Our findings indicate that the disaggregated data are not used in the straightforward manner that some may have thought; rarely did we find that those student subgroups that failed to reach performance targets were specifically targeted for additional support or assistance. Almost a quarter of Pennsylvania schools (24%) had one or fewer eligible student subgroups (in schools with a single student subgroup, a racial group was typically the subgroup), reducing the likelihood of any real targeting based on AYP categories. The reason for this is because of the common overlap between a single subgroup and the overall student population. For example, this most commonly occurs when a school has a subgroup because of the "White" student subgroup. When this occurs, Whites tend to account for the vast majority of students and thus "targeting" a part of the student body based on AYP information is not possible. Additionally, in about half of Pennsylvania schools, all eligible student subgroups made AYP, thus making it impossible to target subgroups based on AYP status. Only 22% of schools in the state had at least two eligible student subgroups where at least one subgroup made AYP and at least one did not. These would be the schools that would be most likely to use subgroup AYP status as an indicator of where they should focus their efforts.

Instead, we found, based on our survey data as well as our interview data, that regardless of the student subgroups that had been identified as failing, schools generally use and prioritize improvement efforts in

similar ways. Our research suggests that some of the similarity may be the result of reluctance on the part of respondents to acknowledge targeting particular demographic categories for attention. Other data suggest that the NCLB-defined categories are not highly relevant to school staffs. In our interviews, respondents were somewhat conflicted about whether or not the information about a subgroup's failure impacted the selection of a corresponding improvement strategy.

A conversation with one principal conveys these conflicted feelings. This principal was the leader of an elementary school where the overall student population was very close to the proficiency cut score but the school had been labeled as in "warning" due to the African American student subgroup. He commented,

> *"This year I am meeting with the 3rd-, 4th-, and 5th-grade teachers once a month to try to target students. That is what we have been trying to do, target students, work with those students. We missed AYP on one group last year and that was Black males. Without targeting, it sounds terrible, 'targeting Black males,' but that is what we are doing this year."*

Clearly, the principal is paying special attention to the group of students who, according to the policy and the school's analysis, had been identified as the reason for the school's failing label. However, just a few minutes later this same principal went on to say,

> *"I don't think of a student as 'oh, you are Black and economically disadvantaged so we have to do this for you.' It's, 'How are your grades? Are you struggling? OK we are going to get you help'…Maybe if I get into school improvement one (the next tier for accountability), I might think more that way, but to me every kid is a kid…I don't think of them in terms of categories."*

This principal was certainly aware that his school was labeled as being in "warning" and reported giving the failing subgroup some additional attention due to performance. However, he was not entirely comfortable with that approach and indicated that he also looked at all struggling students regardless of their inclusion in a designated "subgroup." This desire to group all struggling students together may limit the ability of school staff to select interventions that are specific to unique student subgroups. However, it may also focus school staff on the particular academic needs of students rather than their subgroup identification. Further research is needed to determine the cause and implications of staff reluctance to focus on subgroup affiliation.

Staff at all schools also understood that proficiency cutoffs were rising each year and that all schools are in danger of being labeled as failing in the future. In many cases, staff members spoke of identifying the "bubble kids"—those on the cusp of proficiency on the state test—and providing them with additional attention or assistance. [Other research has also identified the "bubble kid" phenomenon, e.g., Booher-Jennings (2005), Hamilton et al. (2007)]. The focus on bubble kids was an explicit strategy to avoid failing to make AYP in the future, but was not related to students' membership in a particular subgroup.

Finally, many schools reported supporting groups of students' identifiable academic needs, regardless of their test performance. Clearly, certain student subgroups, most notably the English Language Learners and Special Education subgroups, may require unique instructional efforts. However, our interview respondents discussed identifying and attempting to assist *all* students who struggled with vocabulary or reading or other academic skills, rather than focusing efforts based on subgroups. For all of these reasons, if identification of subgroups was intended to focus the attention or efforts of school staffs on externally defined subgroups of students, it did not seem fruitful.

Even while subgroup data were not used for targeting, 57% of schools that failed to make AYP failed due to the performance of one or more student subgroups, not due to the overall performance of the student body. This is the group of schools that can be cited by advocates of subgroup disaggregation who believed that it would help us to identify schools that would otherwise be viewed as performing adequately. Thus, disaggregation helps in labeling schools as failing to make AYP which, as shown above, is related to *more* action and focus, though not necessarily *different types* of action.

Policy Brief

As we saw with school performance overall, when we examine the efforts of schools that fail as the result of different student subgroups, we see very similar prioritization of efforts. Table 2 illustrates the prioritization of efforts across schools that failed to make AYP for different reasons. The columns of Table 2 show some of the most common causes of schools failing to make AYP in Pennsylvania and the percentages of those schools that are devoting major effort to each of the improvement categories.

Looking across these schools that failed for different reasons, and at those schools that passed, we again see relatively similar behaviors (r>.89 for all combinations). Even while low–performing schools are more likely to devote energy to test preparation, the data suggest that different groups of schools prioritize efforts in similar ways. This finding also suggests that providing schools with information about which subgroups have failed to make AYP does not affect behavior in a clear fashion.

Table 2. Current Reform Strategies by Pass / Reason for Fail – Ranked by Percent Placing Major Effort

| Reform Strategy	PASS	FAIL					
		IEP	Race & Econ	Econ	Race & IEP & Econ	Econ & IEP	Race
Use performance data to inform practice	1 (69.5%)	1 (84.4%)	1 (76.8%)	1 (69.3%)	1 (80.0%)	1 (76.9%)	1 (82.2%)
Provide remediation to underperforming students	2 (64.6%)	2 (75.8%)	2 (72.5%)	2 (66.3%)	3 (71.3%)	2 (73.1%)	2 (73.3%)
Introduce new instructional approaches	3 (48.0%)	3 (61.4%)	4 (66.7%)	5 (50.5%)	2 (73.8%)	3 (67.3%)	6 (53.3%)
Improve the quality/ alignment of curriculum	4 (44.2%)	4 (57.5%)	6 (56.6%)	4 (55.5%)	6 (63.8%)	5 (50.0%)	4 (62.2%)
Address non-academic issues	5 (39.9%)	6 (50.0%)	3 (71.7%)	6 (47.0%)	5 (65.0%)	6 (46.2%)	3 (64.4%)
Allot time for PSSA preparation	6 (39.8%)	5 (54.7%)	5 (60.2%)	3 (57.4%)	4 (68.8%)	4 (65.4%)	4 (62.2%)
Change school and/or staff schedule	7 (26.8%)	7 (36.2%)	9 (38.4%)	7 (37.6%)	7 (55.0%)	7 (42.3%)	7 (42.2%)
Bring in outside support and expertise	8 (17.7%)	8 (23.6%)	8 (41.4%)	8 (24.8%)	8 (40.0%)	9 (25.0%)	9 (26.7%)
Create rewards and sanctions related to test performance	9 (14.6%)	9 (21.4%)	7 (44.4%)	9 (23.8%)	8 (40.0%)	8 (39.2%)	8 (37.8%)

Strategy Selection and Achievement

Given that schools of varying performance levels seem to choose similar strategies for improvement, we examined the link between schools' strategy selections and their subsequent achievement levels. To do so, we used the survey results from Pennsylvania principals and student performance data from the PSSA exams, given in mathematics and reading in Grades 3-8 and again in Grade 11. The survey, given at a single point in time, asked principals to identify how much effort their schools were exerting on particular improvement initiatives in the current year (2009) and in the previous year (2008). In this way we attempted to capture changes in school behaviors across two years. To assess achievement, we used the total percentage of students (in schools that responded to the survey) performing at or above "proficient" on the mathematics and reading exams in 2008 and 2009.

To examine the relationship between changes in effort devoted to particular strategies and changes in student performance, we ran a difference-in-difference model, where the outcome was the *change* in proficiency levels from 2008 to 2009, and the independent variables were *changes* in the level of effort principals reported their schools exerting on each of the nine reform categories. In other words, we checked to see if the schools that said they increased levels of effort on particular reform strategies were also the schools where proficiency levels increased the most.

Though this method cannot support any causal conclusions about which strategies are effective, we found that an increased emphasis on test preparation stood out as being significantly associated with increased rates of proficiency in mathematics and reading. This was the only one of the nine categories where a *change* in the level of effort exerted was significantly related to a *change* in student performance in mathematics *and* reading.

Figure 2
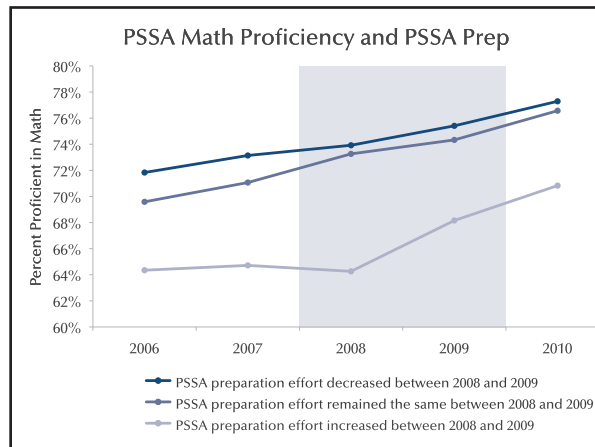
As can be seen in Figure 2, schools that increased their test preparation efforts between 2008 and 2009 (the bottom trend-line in Figure 2) also had an average associated increase of about 4 percentage points in their proficiency rates during that time. In contrast, schools that either kept their level of effort the same (the middle trend-line) or decreased their level of effort (the top trend-line) showed an average increase of only 1 and 1.5 percentage points respectively. It does not appear that the schools that increased their levels of effort on PSSA preparation between 2008 and 2009 were already on a steeper trajectory than those that did not—in fact, the opposite appears to be true. From 2006 to 2008 those schools that decided to increase their test preparation effort after 2008 showed little gain in proficiency, whereas schools that either maintained or reduced their levels of test preparation were showing gains in achievement. This contrast makes the change from 2008 to 2009 even more dramatic for those schools that increased their level of effort on PSSA preparation.

Furthermore, we found that a relationship may also exist between increased remediation efforts and mathematics proficiency, as well as increased use of student performance data and reading proficiency, although these latter findings were less robust than the relationship between test preparation and

performance. Though these three improvement efforts—test preparation, remediation, and data use—are not directly related to a content focus, our research suggests that they are related to increases in proficiency levels on the state tests and may warrant further study.

Policy Implications

Although NCLB has been characterized as shining a spotlight on student performance, that light is not as bright or as focused as one might think. The claims of accountability theory notwithstanding, our findings suggest that schools at different performance levels and with different types of student subgroups choose similar, and numerous, strategies for improvement. While it must be noted that our findings are based on self-report, this research contradicts some previous assertions that high-performing schools and low-performing schools respond very differently to accountability pressure and adds to the evidence that schools at all performance levels engage in a similar range of improvement efforts.

These findings also differ from the commonly accepted knowledge about the impact of performance-based accountability measures on schools, namely that schools will use their performance data to select a few choice strategies for improvement. Instead, we find that schools are selecting many varied strategies, which is somewhat like throwing many darts at a target and hoping that one of them hits the bulls-eye. If the intent was to have schools narrowly target specific efforts to specific groups of students, there appears to be a disconnect between the intent and reality of schools' responses to NCLB. As policymakers and practitioners consider new policies that may yield higher levels of student achievement, we offer several considerations for future policy designs.

Recognize that school performance is not directly related to strategy selection. Though there is evidence to dispute it, many believe that a school with high- achieving students is a fundamentally different place than a school with lower achieving students. While this may be true in some cases (for example, the quality of the teaching staff or the resources available to students), it does

not hold true in terms of the strategies that schools are prioritizing to help students achieve. Some strategies appear to be popular across the board (e.g., data use) and some are not (e.g., rewards and sanctions). And there is evidence that schools that fail to make AYP may devote more effort to explicit test preparation. However, even given these caveats, policymakers should take care to avoid characterizing schools with high proficiency rates and schools with low proficiency rates as schools that behave in unique ways—there's far more similarity than difference. All schools report using many strategies and prioritizing their efforts in very similar ways. This may be the result of commonly accepted practices thought to improve achievement or the result of state or district leadership on specific efforts, but the link between performance and types of effort does not appear to be particularly strong.

Given very limited use, consider the purpose of disaggregation of performance by student subgroup. Descriptive information about the performance of student subgroups may be interesting and informative for education stakeholders. In many cases subgroup performance was the sole trigger for causing a school to fail to make AYP. However, without clearer guidance on what to do with this information, it is unlikely that disaggregated results will be directly related to changes in school improvement efforts. Eighty-seven percent of all schools in this study reported doing no targeting based on NCLB-defined student subgroups. Educators in this study were much more likely to describe focusing on groups of students based on academic needs as opposed to membership in a pre-defined group.

"Teaching to the test" is a rational response to current policy. With a system that sets strict proficiency thresholds and uses standardized tests to measure performance, it is logical that schools at varying performance levels opt for data analysis, general test preparation, and individualized remediation as their main strategies for improvement. The research presented here shows these strategies are related to performance outcomes as measured on state tests. If policymakers would like schools to stress other improvement strategies at the school level, the accountability system must include measures that do not focus on the percentages of students meeting a proficiency

threshold and indicators that cannot be directly affected by test-focused strategies. Instead, policymakers could consider systems that assess instructional quality, student growth rather than status, and a host of non-cognitive measures that are increasingly found to be related to later success (see Farrington, Roderick, Allensworth, Nagaoka, Keyes, Johnson, & Beechum, 2012 for more research on these non-cognitive measures).

References

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5,* 7-74.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42(*2), 231-268.

Coburn, C.E., & Talbert, J.E. (2006). Conceptions of evidence use in school districts: Mapping the terrain. *American Journal of Education, 112*(4), 469–95.

Darling-Hammond, L., & Wise, A. E. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85*, 315-336.

Farrington, C.A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T.S., Johnson, D.W., & Beechum, N.O. (2012). *Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance:* A critical literature review. Chicago: University of Chicago Consortium on Chicago School Research.

Government Accountability Office. (2009). *Student achievement: Schools use multiple strategies to help students meet academic standards, especially schools with higher proportions of low-income and minority students* (GAO-10-18). Washington, DC: United States Government Accountability Office. Accessed online at: http://www.eric.ed.gov/PDFS/ED507150.pdf

Gross, B., & Goertz, M.E. (Eds.), Holding high hopes: *How high schools respond to state accountability policies.* Philadelphia: Consortium for Policy Research in Education.

Gross, B., Kirst, M., Holland, D., & Luschei, T. (2005). Got you under my spell? How accountability policy is changing and not changing decision making in high schools. In B. Gross & M. E. Goertz (Eds.), *Holding high hopes: How high schools respond to state accountability policies.* Philadelphia: Consortium for Policy Research in Education.

Haertel, E.H., & Herman, J.L. (2005). A historical perspective on validity arguments for accountability testing. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement.* Chicago: National Society for the Study of Education, 1-34.

Hamilton, L. S., Berends, M., & Stecher, B. M. (2005). *Teachers' responses to standards-based accountability.* Santa Monica, CA: RAND.

Hatch, T. (2002). When improvement programs collide. *Phi Delta Kappan, 83*(8), 626-634.

Hopkins, D., Harris, A. & Jackson, D. (1997). Understanding the school's capacity for development. *School Leadership and Management, 17*(3), 401–11.

Koretz, D., Barron, S., Mitchell, K., & Stecher, B. (1996). *The perceived effects of the Kentucky Instructional Results Information System.* Santa Monica, CA: RAND.

Learning First Alliance. (2003). *Beyond islands of excellence: What districts can do to improve instruction and achievement in all schools.* Washington, DC: Author.

Linn, R. (2005). Issues in the design of accountability systems. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement.* Chicago: National Society for the Study of Education.

Marzano, R. J. (2003). What works in schools: *Translating research into action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Mintrop, H. (2004). *How accountability works (and doesn't work).* New York: Teachers College Press.

Mintrop, H., & Trujillo, T. (2007). The practical relevance of accountability systems for school improvement: A descriptive analysis of California schools. *Educational Evaluation and Policy Analysis, 29*(4), 319-352.

Newman, F. M., Smith, B., Allensworth, E., & Bryk, A.S. (2001). Instructional program coherence: What it is and why it should guide school improvement policy. *Educational Evaluation and Policy Analysis, 23*(4),297-321.

Padilla, C., Skolnik, H., Lopez-Torkos, A., Woodworth, K., Lash, A., Shields, P.A., Laguarda, K., & David, J. L. (2006). *Title I accountability and school improvement from 2001 to 2004.* Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

Rouse C.E., Hannaway, J., Goldhaber, D., & Figlio, D. (2007). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *NBER Working Paper 13681.* Cambridge, MA: National Bureau of Economic Research.

Stecher, B. M., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., McCombs, J. S., Russell, J., & Naftel, S. (2008). Pain and Gain: Implementing No Child Left Behind in three states, 2004-2006. Santa Monica, CA: RAND Corporation.

Supovitz, J. A., & Klein, V. (2004). *Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement.* Philadelphia: Consortium for Policy Research in Education.

Valenzuela, A. (Ed.) (2004). *Leaving children behind: How "Texas-style" accountability fails Latino youth.* Albany, NY: SUNY Press.

About the Authors:

Elliot H. Weinbaum is a Senior Researcher at the Consortium for Policy Research in Education and is a Research Assistant Professor at the University of Pennsylvania. His work focuses on the development of education policy and its impact on teacher and administrator practice and school improvement. Dr. Weinbaum is co-editor of *The Implementation Gap: Understanding Reform in High Schools.* His research investigates state-led efforts to improve classroom instruction and central office efforts to scale-up reform practices. Elliot holds a B.A. from Yale University and a Ph.D. from the University of Pennsylvania.

Michael J. Weiss is a Research Associate at MDRC in the Young Adults and Postsecondary Education Policy Area. Weiss is currently the lead impact analyst on several MDRC random assignment evaluation projects, including the multisite Learning Communities Demonstration and the City University of New York's ASAP evaluation. He is the co-principal investigator on an Institute of Education Sciences–funded study of the long-term effects of two of MDRC's Opening Doors programs (learning communities and a student success course). Weiss formerly worked as a data analyst at the Educational Testing Services (ETS), where he analyzed student assessment data for the National Assessment of Educational Progress (NAEP) and the Early Childhood Longitudinal Study (ECLS-K). Weiss received his Ph.D. from the Policy, Measurement, and Evaluation Division at the University of Pennsylvania's Graduate School of Education.

Jessica K. Beaver is a doctoral candidate in the Education Policy program at the University of Pennsylvania, and she is an Institute of Education Sciences Pre-Doctoral Fellow. Her research interests include decision-making in educational organizations, as well as the impact of the No Child Left Behind Act on student academic achievement. Previously, Ms. Beaver worked for a Member of Congress on education policy and education appropriations issues, and before that for a government relations firm specializing in education advocacy. She holds a B.A. from Cornell University.

Appendix – Data Collection Information

Interviews. Interviews were conducted with 48 principals using the sampling matrix below. Using statewide data, all schools were categorized based on AYP status as well as whether they covered elementary or high school grades. Principals were randomly selected from each of the categories identified below and were contacted by telephone to participate in an interview of 30 to 45 minutes. Only one principal chose not to participate in the interview and was replaced with another randomly selected principal.

Table A1 – Interview Sampling Matrix

AYP Status	School Level		
	Elementary	High School	Total
Made AYP	4	4	8
Failed to make AYP - Whole School (Total)	4	4	8
Failed to make AYP - Special Education	4	4	8
Failed to make AYP - Economically Disadvantaged	4	4	8
Failed to make AYP - Racial subgroup	4	4	8
Failed to make AYP - Limited English Proficiency	4	4	8
TOTAL	24	24	48

Policy Brief

Survey. An invitation to complete an on-line survey was sent via email to all principals in Pennsylvania. Those who did not complete the survey on-line were sent paper copies. The table below provides information about the survey sample.

Table A2 - Survey Response Information

	Full Sample	Respondent	Non-Respondent
Proficient or Advanced in Math 2008 (%)	71.8	71.7	71.9
Proficient or Advanced in Read 2008 (%)	69.2	69.5	68.7
Made AYP (%)	78.3	78.8	77.5
School Size (number assessed)	300.0	306.0	287.9 **
Number of Subgroups Failing to make AYP	0.6	0.6	0.6
Grades Tested (School Type)			***
Grades 3-8	77.6	75.7	81.3
Grade 11	16.0	17.2	13.6
Mixed	6.4	7.1	5.1
Sample size	3,002	1,996	1,006

SOURCE: Pennsylvania Department of Education AYP Data.

NOTES: Calculations for this table used all available data for 3,002 schools. There were 18 schools that were not surveyed for reasons like the fact that some schools closed between 2008 and 2009. An additional 80 schools did not have any students in the tested grades (3,4,5,6,7,8, and 11).

A two-tailed t-test was applied to differences between the program group and control group for variables that are not mutually exclusive and mutually exhaustive (e.g., School Size). Levels for statistically significant differences between program and control groups are indicated as: * = 10% ; ** = 5%; and *** = 1%.

A chi-squared test was applied to differences between the groups of categorical variables that are mutually exclusive and mutually exhaustive (e.g., Grades Tested). Levels for statistically significant differences between program and control groups are indicated as: * = 10%; ** = 5%; and *** = 1%.

Missing values are not included in individual variable distributions.

Distributions may not add to 100% because of rounding.

Site visits. Site visits were conducted at 11 schools across the state of Pennsylvania. The 11 schools were randomly selected from 11 of the categories shown in Table A1. (There was no high school in Pennsylvania that failed to make AYP only as the result of the ELL subgroup, thus no school was visited in that category.) Schools were invited to participate and two declined from our original sample. These were replaced with two randomly selected schools. At each school, we spoke with the principal, assistant principal (if there was one), leadership staff involved in school improvement efforts, and a sample of teachers (invited by the school liaison) who taught in tested grades and subjects. Approximately, 10 to 15 individuals were interviewed at each school. Each school was visited twice, in the spring of 2009 and 2010. In total, we conducted 162 interviews with 118 individuals.

Policy Brief

About the Consortium for Policy Research in Education (CPRE)

Established in 1985, CPRE unites researchers from seven of the nation's leading research institutions in efforts to improve elementary and secondary education through practical research on policy, finance, school reform, and school governance. CPRE studies alternative approaches to education reform to determine how state and local policies can promote student learning. The Consortium's member institutions are the University of Pennsylvania, Teachers College-Columbia University, Harvard University, Stanford University, the University of Michigan, University of Wisconsin-Madison, and Northwestern University.

CPRE.ORG