THE EFFECT OF SCHOOL ACCOUNTABILITY SYSTEMS ON THE LEVEL AND DISTRIBUTION OF STUDENT ACHIEVEMENT

Eric A. Hanushek
Stanford University and
National Bureau of
Economic Research

Margaret E. Raymond
Stanford University and
CREDO

Abstract
The use of school accountability in the United States to improve student performance began in the separate states during the 1980s and was elevated through the federal No Child Left Behind Act of 2001. Evaluating the impact of accountability is difficult because it applies to entire states and can be confused with other changes in the states. We consider how the differential introduction of accountability across states affects growth in student performance on the National Assessment of Educational Progress (NAEP). Our preliminary analysis finds that: 1) accountability improves scores of all students; 2) there is no significant difference between simply reporting scores and attaching consequences; and, 3) while accountability tends to narrow the Hispanic–White gap, it tends to widen the Black–White gap in scores. The last finding suggests that a single policy instrument cannot be expected to satisfy multiple simultaneous goals. (JEL: I2, H7, J4)

1. Introduction

Adoption of statewide accountability systems for schools has been one of the most striking reforms in American education policy in the past twenty-five years. The change in focus away from inputs and processes and toward outcomes marks a dramatic shift in orientation. And yet we know little so far about how well these systems work. The lack of evidence on accountability is due in part to the way states have put these programs in place. States always have been the primary locus of education policy in the United States, adapting their programs to local circumstance and yielding a diverse array of programs across and within states. But, since each accountability program has invariably been implemented statewide at inception, it is impossible to ascertain the impact of any single state system. The variation in timing of implementation across states does nonetheless provide a way of estimating the impact of school accountability.

Starting in the mid-1980s many states in the United States voluntarily

E-mail addresses: Hanushek: hanushek@stanford.edu; Raymond: macke@stanford.edu

Journal of the European Economic Association April–May 2004 2(2–3):406–415

adopted accountability policies to measure the performance of their schools. These states began assessing the educational outcomes of their students and using these objective and external measures of performance as a way of gauging the effectiveness of their schools. While the federal government entered into accountability with the No Child Left Behind Act of 2001 that requires all states to have comprehensive accountability systems in place by 2006, many of the design details are left to states to decide.

Across states, however, the pattern of adoption of accountability over time makes it possible to study accountability as a policy writ large. This analysis is possible because of the extensive participation of states in the National Assessment of Educational Progress (NAEP). Called "The Nation's Report Card," NAEP is a federal program that has tested a random sample of students since roughly 1970. On a rotating scheduling, students in fourth, eighth, and twelfth grades have been tested approximately every four years in a range of subjects, including reading and math. Importantly, in 1990 NAEP began a program of state representative testing for states that volunteered to participate. Because students take the same test, NAEP provides an independent and consistent measure by which to compare academic achievement across states—something not possible using the states' own tests. The influence of accountability policies can be discerned by tracking changes in NAEP cohort performance over time as state accountability systems are introduced.

The increasing importance of testing and accountability systems in other nations around the world elevates the importance of this work. Countries mirror U.S. states in their ubiquitous adoption of accountability policies. Yet, because evaluation evidence must come from situations where there is variation in implementation and in characteristics of the accountability approach, national adoption of an accountability system generally precludes credible assessment of impacts. The decentralized adoption of accountability policies staggered over several years in the United States creates a rare opportunity to analyze the workings and effects of these policies.

2. Research Questions

Our prior work provided preliminary evidence that states that adopted some form of accountability produced on average higher achievement gains on the National Assessment of Education Progress (NAEP) tests than occurred in states that did not have such programs (Hanushek and Raymond 2003a, 2003b). These effects were estimated by identifying the years in which states' accountability systems became active and comparing the change in scores on previous and subsequent NAEP tests to similar changes in states that still did not have such systems. The effect held for both report card states—those whose accountability

program consists only of public disclosure of school-level scores—and consequence states, whose policies carry rewards and sanctions for schools depending on their performance.

Newly released NAEP test results expand the number of periods available to study the implementation and effects of state accountability systems. The updated analysis reported here reinforces earlier findings: Both report card states and consequence states show greater gains on average than nonaccountability states, even after better control for other influences.

A second question addressed in this paper concerns the consistency of effects from accountability programs on students in different race/ethnicity groups. The equity effects of education policy choices remain important considerations in the study of American educational policy. Equity of access to education has come haltingly to many communities. Gaps in achievement between Whites and other ethnic groups are long-standing and of great concern to policymakers. Many school districts continue to operate under court-supervised desegregation orders. Accordingly, beyond the effects of accountability policy decisions in the aggregate, this analysis pursues a complementary focus on the relative impacts for various subgroups of the student population. With more extensive data, we reopen an earlier analysis by Carnoy and Loeb (2002) showing that black and Hispanic students in states with accountability systems tended to improve even more than white students on the eighth-grade math NAEP after adjusting for other factors. Using a different identification strategy, we expand their analysis to include reading performance and later test data.

Finally, we test competing theories about the operative mechanisms of accountability systems. One possible explanation, driven by microeconomic theory of markets, suggests that accountability systems may work by virtue of their disclosure of information about schools. Creating common measures of performance and making them available to affected constituencies rectifies a significant market failure—in this case, asymmetric information about school performance.

An alternative explanation draws from research on motivation and strategies to shape behavior via external influences. Work in the field focuses on which strategies are most effective to achieve desired actions from individuals. Individuals are theorized to have unobservable states of motivation, subject to influence, that drive behavioral choices. From this perspective, accountability systems could be effective since they impose both positive and negative consequences to the results of teacher and administrator behavior. In the context of theories of motivation, teachers and administrators might be attracted to the rewards and/or be repelled by the prospect of sanctions.

The present analysis offers the chance to test these two competing theories. If the operative mechanism is largely motivational, then states with consequences would be expected to produce larger gains than states that have no

accountability systems or have report cards. (This assumes, of course, that disclosure of school performance without consequences is nonthreatening.) If however, the repair of imperfect information creates pressure to improve from a newly-informed community, states with report card accountability systems would be expected to do as well as consequence states but outperform states that have no accountability system.

3. Difficulties of Analyzing School Accountability

Analyzing the effects of accountability on student performance is difficult. First, because accountability systems are introduced across entire states, all local school districts in a state face a common incentive structure. Thus, the only possible variation comes from interstate differences in accountability, but states also differ in ways other than accountability. Second, samples are limited by the number of states that participate in national testing programs, making the available information even less. Third, while testing is designed to cover a random sample of the student population, various exclusion rules are applied for special education students and for limited English proficient students—and the application of these rules has changed over time.

Extensive analyses of educational production functions have been conducted, and they form the relevant background for this work. Those studies have concentrated on describing how various inputs to schools enter into the determination of student outcomes. As described elsewhere, however, these studies have not provided any consistent picture of how schools affect student performance (Hanushek 2003).

One reason frequently hypothesized for this lack of relationship relates directly to the prior discussion: without strong incentives, resources are not consistently and effectively transformed into outcomes. That fact provides the motivation for moving to greater accountability systems.

The difficulty is that little progress has been made in describing explicitly the different policies, regulations, and incentives that might be important in determining student performance. Educational policy is made at the state level and involves a wide range of factors including financial structure, collective bargaining rules and laws, explicit regulations on educational processes, and the like. The analytical complications are immediately apparent.

Consider a simple model of achievement such as:

$$O_{st} = f(X_{st}, R_{st}, \rho_s) \tag{1}$$

where O is the level of student outcomes in state s at time t, X is a vector of family and nonschool inputs, R is a vector of resources, and ρ captures the

policies of the state.[1] If one attempts, say, to understand the implications of different resources on student performance by regressing O on explicit measures of families and schools, the estimated effects will be biased to the extent that ρ is correlated with the included measures; that is, a standard model misspecification story.

This issue is nonetheless directly relevant to the analysis of accountability systems that we pursue here. While there are state data on student performance from NAEP, it is not possible to understand the impact of newly introduced accountability systems without considering the range of other possible impacts. A linearized version of this model is simply:

$$O_{st} = \beta_0 + \beta_X X_{st} + \beta_R R_{st} + (\rho_s + \varepsilon_{st}) \tag{2}$$

where the βs are unknown parameters of the educational process.[2] If, however, ρ is not observed and the βs are estimated with just information on X and R, correlations with ρ obviously lead to bias in the estimation. Now consider just adding A, a measure of whether or not accountability affects incentives and thus student performance.

$$O_{st} = \beta_0 + \beta_X X_{st} + \beta_R R_{st} + \gamma A_{st} + (\rho_s + \varepsilon_{st}) \tag{3}$$

The objective is to understand γ, but under almost all circumstances γ will also be biased through omission of relevant state policies.

Moreover, Hanushek, Rivkin, and Taylor (1996) demonstrate that the bias in any estimation will generally increase with the level of aggregation in situations like this. Specifically, when the omitted variable is relevant at the state level, estimation of the model across states will have the most bias. Note that this does not say anything about the direction of any bias, only that aggregation worsens the bias. In the case of measures of school resources, all evidence indicates that there is an upward bias from omitting state policies (Hanushek, Rivkin, and Taylor 1996; Hanushek 2003). It does not, however, give much indication of how any estimation of partial models of accountability would bias analyses of γ.

If, however, the state policies are constant over our observation period, a variety of estimation approaches are possible. In the simplest form, simply looking at outcome changes eliminates any state differences that are constant over the period t to t^*:

1. It does not matter for this discussion that we begin with aggregate outcomes for a state instead of building up from the individual student level (where the outcomes are presumably generated). The more general situation is discussed and developed in Hanushek, Rivkin, and Taylor (1996). Where the aggregation is important, we discuss the implications.

2. The linear form is not particularly crucial but simply makes the exposition easier. An alternative model where policies act as an efficiency parameter affecting the impact of resources is developed in Hanushek and Somers (2001). Within the limited data for this study, however, it is virtually impossible to distinguish between the alternative models. The results of estimating the alternative form, discussed below, are qualitatively very close to the included estimates.

$$\underset{t,t^*}{\Delta O_s} = \beta_X \Delta X_s + \beta_R \Delta R_s + \gamma \Delta A_s + \Delta \varepsilon_s \quad (4)$$

The key element is that effects of accountability systems are identified from changes in accountability across states over the sample period. A variant also pursued is to add a state fixed effect to the estimation. This provides much better control for other factors influencing performance growth but now estimates the effects of accountability entirely on the basis of the introduction of accountability systems within each state.

4. The Effects of State Accountability

The estimation of accountability effects uses two elements of the NAEP testing information. First, since the introduction of state level testing in 1990, NAEP has tracked performance over time for participating states. This testing provides directly useful data for two tests (mathematics and reading). The sampling/testing design of NAEP is particularly helpful because it has a basic four year testing cycle that involves testing fourth and eighth graders. Thus, fourth graders in 1992 are tested as eighth graders in 1996. While these are not the same students, a growth formulation holds constant common cohort experiences, and we can control for observed changes in the population. Moreover, because of consistent multiple testing in both math and reading, it is possible to create a panel with two time periods of achievement growth in each subject—thus permitting estimation that removes individual state fixed effects.

4.1. Overall State Impacts

From Table 1 that presents the combined performance of all students in tested states we find consistent evidence that introduction of state accountability had a positive impact on student math performance during the 1990s. In each case, state average NAEP scores in the eighth grade are related to the prior fourth grade performance along with measures of parental education, school expenditures, and the test exclusion rate over the relevant time period. The key finding is that having some kind of accountability system positively impacts on student performance and, except for reading by itself, is statistically significant. Moreover, the effects hold even when estimated in a fixed-effect framework. At the same time, while the point estimates suggest less influence of simply reporting the results as opposed to attaching consequences to them, the difference between the two types of systems is never statistically significant.

In terms of other factors, exclusions always have the expected effect on tests: More exclusions from a test increase the average score. Interestingly, however, the large differences in spending per pupil never influence scores,

TABLE 1. Effect of accountability on average state performance

	Without state fixed effects			With state fixed effects		
	Math	Reading	Combined	Math	Reading	Combined
Accountability	3.79	1.21	2.65	2.75	1.85	2.18
	(3.9)	(1.1)	(3.4)	(2.3)	(1.2)	(2.5)
Report card	−0.56	−1.44	−1.04	−1.38	−3.55	−1.58
	(−0.5)	(−1.3)	(−1.2)	(−0.8)	(−1.3)	(−1.1)
% High school or greater (pop > 25)	0.42	0.27	0.40	0.32	0.60	1.01
	(3.4)	(2.3)	(4.2)	(0.5)	(0.8)	(2.1)
Expenditure/pupil ($1,000)	−0.37	0.02	−0.28	−1.22	−0.54	−0.98
	(−1.1)	(0.0)	(−0.9)	(−0.5)	(−0.1)	(−0.5)
NAEP4 (math)	1.03		1.01	0.46		0.59
	(15.8)		(18.3)	(2.4)		(5.3)
NAEP4 (reading)		0.58	0.53		−0.005	0.16
		(10.7)	(−9.6)[a]		(−0.0)	(−8.8)[a]
Test exclusions	0.12	0.58	0.33	0.27	0.52	0.25
	(0.7)	(3.8)	(2.9)	(1.3)	(1.9)	(1.9)
Observations	70	68	138	70	68	138
States	39	38	42	39	38	42

Notes: Each regression includes separate intercepts for each observation period.
[a] t-test on difference in math and reading pretest score.

while higher parental education positively affects scores. Estimation with fixed effects control for constant state policies and for near constant population characteristics, but their introduction does not have a huge impact on estimation of accountability.

4.2. Racial and Ethnic Breakdowns

With separate information on race and ethnicity, we pool the separate observations. Prior conclusions are basically unchanged, although the estimation yields more precise estimates (Table 2). Accountability effects on reading performance are now significant even with state fixed effects. We begin to see, however, distinct differences in gains by blacks and Hispanics: Each shows roughly 10 points less growth on NAEP between fourth and eighth grade where the mean growth is 50 points.

The finding of lower black and Hispanic growth is particularly interesting in light of the narrowing of the achievement gap that occurred in the 1980s (Jencks and Phillips 1998). The lack of progress in the 1990s on aggregate tests (Hanushek 2001) shows up in the state details where there are controls for state policy, family backgrounds, and testing artifacts.

Importantly, we still find no differential by simple reporting versus consequences. Even with the greater estimation precision, report cards are not significantly different from consequential accountability. This finding indicates

TABLE 2. Effect of accountability on average performance by race/ethnicity

	Without state fixed effects			With state fixed effects		
	Math	Reading	Combined	Math	Reading	Combined
Accountability	3.15	1.83	2.41	5.03	2.62	3.54
	(2.8)	(1.95)	(3.1)	(3.5)	(2.2)	(4.0)
Report Card	−1.49	−1.05	−0.81	−3.23	−0.67	−1.65
	(−1.1)	(−1.0)	(−0.9)	(−1.4)	(−0.4)	(−1.2)
% High school or greater (pop > 25)	0.15	0.19	0.12	0.09	0.18	0.08
	(2.6)	(3.5)	(2.9)	(1.4)	(2.8)	(1.7)
Expenditure/pupil	0.39	0.28	0.39	−2.22	2.58	−0.13
	(1.0)	(0.8)	(1.2)	(−0.8)	(0.6)	(−0.1)
Black	−5.47	−11.7	−10.0	−7.38	−11.2	−10.7
	(−2.7)	(−8.7)	(−7.9)	(−3.1)	(−7.6)	(−7.8)
Hispanic	−5.77	−5.87	−8.22	−8.28	−5.6	−9.61
	(−3.2)	(−3.5)	(6.5)	(−4.0)	(−2.8)	(−6.9)
NAEP4 (math)	0.92	0.40	0.81	0.87		0.79
	(14.0)	(9.2)	(18.5)	(10.9)		(16.6)
NAEP4 (reading)			0.44		0.42	0.43
			(13.6)[a]		(8.6)	(13.2)[a]
Test exclusions	0.09	0.64	0.42	0.45	0.91	0.59
	(0.5)	(4.7)	(3.6)	(1.9)	(5.2)	(4.5)
Observations	188	160	348	188	160	348
States	39	38	42	39	38	42

Notes: Each regression includes separate intercepts for each observation period.
[a] *t*-test on difference in math and reading pretest score.

that the prime pathway for accountability to influence school performance is the provision of better information.

4.3. Details of Accountability

The impact of accountability may not have uniform effects on the separate groups (as constrained to do so in Table 2). Thus, we estimate the same basic models but permit the effects of accountability to differ by race and ethnicity. Table 3 presents the results for the models with state fixed effects. The first three columns are directly comparable to the previous, but they now indicate distinct differences by subgroup. Concentrating on the combined test results, we see that Hispanics are affected significantly more than Whites by having accountability. On the other hand, the estimates for blacks show accountability having a smaller marginal impact (although not significantly different from zero).

The last three columns provide further detail. When states introduce accountability systems, they may or may not disaggregate the test results by racial group. In the last columns we look at the differential impact of accountability for systems with subgroup disaggregation. While the point estimates are similar in magnitude, the combined estimates now reveal that Blacks perform the worst of all of the subgroups by a statistically significant amount.

TABLE 3. Details of accountability effects on average performance by race/ethnicity

	With state fixed effects					
	Math	Reading	Combined	Math	Reading	Combined
Accountability	4.25	1.99	2.94	4.31	2.19	3.09
	(2.7)	(1.7)	(3.0)	(2.9)	(2.0)	(3.5)
Accountability × black	−1.86	0.42	−1.55			
	(−1.2)	(0.3)	(−1.4)			
Accountability × Hispanic	2.76	0.89	2.63			
	(1.4)	(0.6)	(2.5)			
Disaggregated accountability × black				−2.47	0.30	−1.85
				(−1.7)	(0.3)	(−1.96)
Disaggregated accountability × Hispanic				2.88	0.33	2.10
				(1.94)	(0.3)	(2.1)
Black	−5.16	−11.5	−8.67	−5.18	−11.4	−8.76
	(−2.0)	(−5.9)	(−5.4)	(−2.2)	(−6.6)	(−5.9)
Hispanic	−8.80	−6.24	−10.2	−7.72	−5.9	−9.19
	(−4.1)	(−2.8)	(−7.0)	(−3.8)	(−2.7)	(−6.6)
Observations	188	160	348	188	160	348
States	39	38	42	39	38	42

Note: Each regression includes separate intercepts for each observation period, % high school or greater (pop > 25), NAEP4 (math and reading), indicators for specific time period, and test exclusions.

It is useful to put the detailed impacts into perspective. Accountability significantly increases the state achievement gain, particularly for Hispanics. However, because both Blacks and Hispanics show lower gains on each of the tests, accountability cannot close the gap in learning. Moreover, because whites gain more than Blacks after accountability is introduced, the racial achievement gap actually widens with the introduction of accountability.

5. Preliminary Conclusions

Accountability is important for students in the United States and in a variety of other countries that are pushing for better performance measurement. Regardless of any design flaws in the existing systems and of variations in design across states (Hanushek and Raymond 2003b), we find that they have a positive impact on achievement. We also find that the effect varies by subgroup, with Hispanics gaining most and Blacks gaining least. Finally, the impact appears to result primarily from the purely informational aspects of accountability and not from any explicit consequences.

The finding of differential effects raises a clear policy dilemma. A prime reason for the U.S. federal government to require each state to develop a test based accountability system involved raising the achievement of all students. These results suggest a beneficial effect on overall achievement but simultaneously that some gaps across subgroups could widen. We conclude from this

that additional policies are needed to deal with the multiple objectives. Again, as is frequently the case, a single policy cannot effectively work for two different objectives—raising overall student performance and providing more equal outcomes across groups.

References

Carnoy, Martin and Susanna Loeb (2002). "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." *Educational Evaluation and Policy Analysis* 24, 305–331.

Hanushek, Eric A. (2001). "Black–White Achievement Differences and Governmental Interventions." *American Economic Review* 91, 24–28.

Hanushek, Eric A. (2003). "The Failure of Input-Based Schooling Policies." *Economic Journal* 113, F64–F98.

Hanushek, Eric A. and Margaret E. Raymond (2003a). "Improving Educational Quality: How Best to Evaluate Our Schools?" In *Education in the 21st Century: Meeting the Challenges of a Changing World,* edited by Yolanda Kodrzycki. Federal Reserve Bank of Boston.

Hanushek, Eric A. and Margaret E. Raymond (2003b). "Lessons About the Design of State Accountability Systems." In *No Child Left Behind? The Politics and Practice of Accountability,* edited by Paul E. Peterson and Martin R. West. Brookings.

Hanushek, Eric A., Steven G. Rivkin, and Lori L. Taylor (1996). "Aggregation and the Estimated Effects of School Resources." *Review of Economics and Statistics* 78, 611–627.

Hanushek, Eric A. and Julie A. Somers (2001). "Schooling, Inequality, and the Impact of Government." In *The Causes and Consequences of Increasing Inequality,* edited by Finis Welch. University of Chicago Press.

Jencks, Christopher and Meredith Phillips, eds. (1998). *The Black–White Test Score Gap.* Brookings.