School Accountability & Students' Long Run Success:

Preliminary Evidence from the Panel Study of Income Dynamics

Sean Tanner
Goldman School of Public Policy
University of California, Berkeley

In this paper I present preliminary evidence on the long-run impacts of school accountability on students in the United States. Using data from the Panel Study of Income Dynamics, I find that attending a school that was under an accountability regime increases educational attainment and labor market performance for African American students, but does not for Caucasian students. However, the data places severe power restrictions on the estimation procedure and I am unable to distinguish the effects from artifacts of random sampling variance at any reasonable level of confidence. I discuss strategies to increase statistical power and identify data sets that will allow for a more powerful test of the impact of accountability on long-run student success.

Introduction

Teacher quality influences the pattern of human capital accumulation and plays a role in the inequality of life outcomes as well. The bulk of evidence agrees that a high quality teacher can dramatically improve cognitive achievement, as typically measured on standardized tests of math and verbal ability. For instance, an above average teacher can increase cognitive abilities by .22 standard deviations (Hanushek, Kain, Brien, & Rivkin, 2005, p. 14). More recently, panel data has allowed the estimation of teacher quality on distal outcomes such as college attainment and neighborhood quality. The results suggest that a high quality teacher improves income, college attainment, savings rates, and neighborhood quality and reduces the rate of teen pregnancy (Chetty, Friedman, & Rockoff, 2013). Estimates of the discounted present value of a high quality teacher hover around $400,000[1]. Such estimates are based only on the lifetime effects of cognitive skills gains and so likely underestimate the returns to a high quality teacher, as recent data reveals that non-cognitive skills are rewarded in the labor market independently of IQ and test scores (Almlund, Duckworth, Heckman, & Kautz, 2011). The inequitable distribution of high-quality teachers by race/ethnicity and socioeconomic status has been well documented (Adamson & Darling-Hammond, 2012; Reardon, 2011) and has been a formidable, perennial educational policy challenge. For example, poorly resourced school districts tend to serve students with greater needs, offer lower-quality working conditions, exhibit higher teacher turnover, and hire teachers with significantly lower qualifications. The frustrating aspect of these findings is that altering teacher quality has been difficult to achieve or redistribute with existing policy levers.

The historical track record of public policies' ability to significantly improve teacher quality and alter its distribution across poor and affluent schools can be characterized as dismal. Most policies regarding

[1] This is the discounted present value of the lifetime incomes of a class of 20 students that can be attributed to a teacher one standard deviation above average (84th percentile). It is also the discounted present value of replacing a teacher at the 5th percentile with one at the 50th percentile.

teacher quality have aimed to alter observable characteristics like certifications, tenure, and professional development. Believing these to be signals of teacher quality, policy makers attempted to purchase quality by increasing salaries for teachers that were more experienced and educated (Hanushek, 2003). Unfortunately, student achievement has not responded to such policies. Two primary critiques grew out of this situation: (1) inputs such as degree and tenure are weak signals of teacher quality and (2) policies aimed at inputs do not incentivize teachers to improve their effectiveness in the classroom. The critics of input-based policies argue that a system of accountability and performance pay is a far better tool for increasing teacher quality (Hanushek & Rivkin, 2010). Though short-term experiments have not shown a significant achievement return to performance pay (Springer et al., 2010), accountability policies more broadly have had a plausible, non-trivial effect on math and reading scores (Dee & Jacob, 2010, 2011), college performance (Donovan, Figlio, & Rush, 2006), and employment and wages by age 25 (Deming, Cohodes, Jennings, & Jencks, 2013). An important caveat to these findings is that the effects are heterogeneous in both magnitude and direction. In general, students whose demographic and scholastic characteristics make their performance consequential for their schools' accountability rating respond differently to accountability policies than do students whose characteristics make their scores trivial for their schools' accountability rating.

This paper expands on the prior literature by estimating the impact of accountability on educational attainment and labor-market performance for a nationally representative sample of children and young adults from the Panel Study of Income Dynamics. Using a fixed-effects, difference-in-difference estimation strategy, I find a similar pattern of heterogeneous effects across targeted and non-targeted groups of students (using race as a proxy for targeting). However, the data places severe power restrictions on the estimation procedure and I am unable to distinguish the effects from artifacts of random sampling variance at any reasonable level of confidence. I discuss strategies to increase

statistical power and identify data sets that will allow for a more powerful test of the impact of accountability on long-run student success.

Prior Literature

Output-based policies reward teachers and schools for their specific, contemporaneous effect on students. Prevailing economic theory suggests that explicitly rewarding student achievement will have at least three effects: (1) current teachers will re-allocate their time and schools will re-allocate their resources in order to optimize in the face of new incentives, (2) student achievement will increase, and (3) distal student outcomes relating to achievement will improve (Hanushek & Raymond, 2001, pp. 368–369). The estimators used in evaluating output-based educational quality can be broadly grouped into two categories: those that track changes in achievement data across successive cohorts of students and those that track achievement growth for individual students. The former usually goes by the label "cohort analysis" or "cohort-to-cohort gain" approach (Mccaffrey, Lockwood, Koretz, & Hamilton, 2003, p. 1) and the latter is almost universally labeled a "value-added" approach (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). While each approach is distinct in the level of aggregation, data requirements, and modeling assumptions maintained, the overarching goal of each approach is to assess an individual school's or teacher's impact on student achievement.

Cohort analysis gained favor in the 1990s, as states implemented sanctions and rewards based on schools' change in average student achievement scores over time. By the time the No Child Left Behind Act (NCLB) mandated such systems for all public schools nation-wide, half of states had already implemented similar systems (Dee & Jacob, 2011, p. 423). Several studies have attempted to assess the impact of these state accountability systems, including those mandated by NCLB specifically.

Dee & Jacob's (Dee & Jacob, 2011) work is the most comprehensive effort at pinning down the effects of NCLB on student achievement at the national level. Though NCLB required Title I schools in all states to

implement accountability plans, roughly half of states had consequential accountability prior to the law. These states constitute a reasonable counterfactual group for a comparative interrupted time series research design. Dee & Jacob use this method to uncover an estimate of the impact of accountability on student achievement using low-stakes testing data for 4th and 8th grade students. Table 1 presents some of the salient results from one of the primary tables in the analysis (Dee & Jacob, 2011, pp. 439–440), transformed into approximate effect sizes.[2]

Table 1: The Impact of NCLB on Student Achievement

	4th Grade		8th Grade
Sub-Group	Math	Reading	Math
White	0.157**	0.149**	0.048
Black	0.470**	-0.024	0.232
Hispanic	0.316**	0.007	0.216**
Male	0.246**	.062*	-.045
Female	0.240**	0.021	0.169
Free-Lunch	0.258**	0.069	0.415**
Non-Free Lunch	0.045	-0.133	0.026

Effect sizes calculated from Dee and Jacob (Dee & Jacob, 2011)

These results suggest a strong effect on 4th grade math, with little to no effect on 4th grade reading and 8th grade math overall. These results come from low-stakes testing data from the National Assessment of Educational Progress (NAEP) that should be fairly immune to the "teaching to the test" phenomenon (Jacob & Levitt, 2003), so there is plausible evidence that students' cognitive abilities did actually increase rather than merely their test scores. Moreover, in separate analysis Dee and Jacob observed

[2] Cohen's d is calculated as the total effect of NCLB as of 2007 on the mean of achievement scores divided by the standard deviation. As specific standard deviations were not available for the sub-groups, the overall standard deviations for 4th grade math, 4th grade reading, and 8th grade math are used as denominators (Dee & Jacob, 2011, p. 434).

that schools increased teacher salaries and hired more teachers with masters degrees, while teachers re-allocated their time toward reading and (to a lesser extent) math (Dee & Jacob, 2010, pp. 184–187).

The research on how accountability systems, and NCLB specifically, impact distal outcomes is limited. Deming, Cohodes, Jennings, and Jencks (2013) use comprehensive data from Texas students to identify the impact of being in a school-cohort group that faced accountability pressure. They find that historically low-performing schools facing probable sanctions tend to increase student achievement, with a greater increase for low-performing students. These increases in high stakes test scores are also associated with increased probability of graduating high school, attending college, earning a bachelor's degree, and higher wages by their late 20s. However, historically high-performing schools facing incentives to achieve a high-performance rating tend to decrease performance among low-achieving students (Deming et al., 2013, pp. 17–24).

Wong (2008) also uses differences in cohorts' exposure to accountability to examine differences in educational attainment and wages. Using data from the Census and American Community Survey, she finds that cohorts of students who went through schools when their states had adopted accountability did not differ from cohorts of students who went through schools prior to their states adopting accountability. Most of the impacts of accountability are indistinguishable from zero. Those that might be labeled "marginally" significant are often not in the "correct" direction and are highly likely to be artifacts of random sampling variance rather than true effects.

Donovan, Figlio, and Rush (2006) use data from a public university to identify the impact of the strictness of accountability on college-bound students. They find that students who attend schools facing weak accountability pressure perform more poorly than their peers who attend schools with more stringent accountability pressure.

The three studies of the long-run consequences of accountability (Deming et al., 2013; Donovan et al., 2006; Wong, 2008) provide reason to believe that teachers and schools respond to incentives to target certain students, but that the overall impact of NCLB-style accountability is quite marginal and possibly wholly restricted to those students whose scores mattered for accountability ratings at schools that were close to performance thresholds. Moreover, the data used in these studies placed important restrictions on the researchers. The two studies that find effects of accountability do so for only particular regions of the country and accountability schemes, Texas in the case of Deming et al. (2013) and an unidentified state in Donovan et al. (2006). In both of these cases, the accountability regime was implemented prior to the onset of NCLB and in states that elected to implement them. Results from these states might generalize poorly to the nation as a whole, particularly to states that were forced to implement accountability regimes by NCLB. Data availability is another limiting factor in these studies. Wong (2008) uses a large sample from the Census and American Community Survey, but the variables and pooled cross-section nature of the data restrict her analysis. Similarly, only college attending students are present in the data from Donovan et al. (2006) and thus the authors are unable to assess the impact of accountability on students who do not attend the state college system. Only the Texas study contains panel data on students from middle school to early adulthood.

Data

Data from the Panel Study of Income Dynamics (PSID) can be used to replicate and expand the findings from these three studies by including rich, longitudinal data on family background for students who were educated in U.S. public schools from the 1980s to the 2000s. The PSID began interviewing families in 1968 and collects information on family structure, income and wealth, geographic mobility, sociological characteristics. From 1968 to 1997, families were interviewed annually, after which interviews were conducted biennially. When members of one of the original families moves out of the household, they are given their own family id and added to the interview schedule as a new family.

Consequently, the PSID now contains data on three, and in some cases four, generations. In 1997, children ages 0 to 13 of families in the PSID were entered into the Child Development Supplement (CDS). The CDS contains much richer data on cognitive, behavioral, and psychological growth for children up to their teenage years. The CDS is collected every five years. Once the children in the CDS enter 18, they are entered into the Transition to Adulthood (TA). The TA contains rich data on young adults as they enter college and the workforce and is collected biennially. Table 2 provides a summary of when each subset of the PSID in this analysis was interviewed.

Table 2: Interview Waves in the PSID, CDS, and TA

	1968-1997	1999-2013
Main PSID Families	Annually	Biennially
CDS	1997	2002, 2007, 2013
TA	--	2005, 2007, 2009, 2011, 2013

The analysis in this paper uses data on individuals in PSID families who were ages 19 to 42 by 2011, the most recent year for which data is available[3].

I estimate the impact of school accountability policy on several educational and economic attainment outcome measures using the following regression model:

[3] The 2013 data will be released in early 2015.

$$Y_{isb} = \beta_1(13 - (A - K))T_{isb} * D_{isb}1(13 - (A - K) < 0) +$$

$$\beta_2(13 - (A - K))T_{isb} * D_{isb}1(13 - (A - K) \geq 0) + \sum_{j=3}^{k} \beta_j X_{ij} +$$

$$S_s + C_b + e_i \qquad (1)$$

where Y_{isb} is the outcome variable of interest for the i^{th} individual from childhood state *s* and birth cohort *b*; T_{isb} is a treatment variable defined as the number of school-age years spent under an accountability regime for the i^{th} individual from state *s* and birth cohort *b*. I create a spline at T=0, so that B_1 estimates the trend in the outcome leading up to the onset of accountability. The coefficient of interest is β_2, which is the effect of an additional year of exposure to accountability on outcome *Y* net of the pre-accountability trend (B_1). The model includes state fixed effects (represented by *S*), which account for persistent average differences in outcomes of children from different states and helps to account for potential policy endogeneity by states in their choice of whether and when to enact school accountability regimes. I also include birth cohort fixed effects (represented by *C*), which account for nationwide trends in outcomes that would have occurred in the absence of school accountability policy changes. I also include an array of observable childhood family factors (as represented by X_{ij}: a vector of *K*-1 control variables), including race, gender, parental education, parental income, and family structure; and e_i is a stochastic error term. This simple, fixed-effects difference-in-difference model is used as an identification strategy to attempt to investigate the question: do school accountability systems influence children's academic and economic success? And if so, for whom?

Results

Table 3 contains the results from the main PSID data. Focusing on the point estimates alone, a clear picture emerges: students who were more likely to be targets of their states' accountability systems had increases high school completion, college attainment, and wages. The average effect of exposure to

accountability on African American students is a 2.8% increase in the probability of obtaining a high school degree, a 5% increase in the probability of going to college, a 0.5% percent increase in the probability of obtaining a college degree, 9% higher wages in 2011, and a slight decrease in the probability of being unemployed during 2011. Conversely, students who were not likely to be targets of their states' accountability system had lower levels of educational attainment and labor market performance. This is consistent with findings in the prior literature and what theory predicts: schools target students and shift resources accordingly.

Table 3: Impact of the Average Exposure to Accountability

	Full Sample	African American Students
High School Completion	-5%	2.8%*
College Attendance	-.8%	5%*
College Degree	-0.5%	0.5%
Income	-4%	9%*
Unemployment	0.6%	-0.1%

***Difference from reference group significant at .1 level. The rest of the coefficients are not statistically significant.**

These findings can only be considered suggestive, as the standard errors are too large to reject the null hypotheses of no treatment effects by any reasonable confidence level. Though some of the coefficients on the treatment effects for African American students are marginally significant, the combined effects for those effects (rather than the difference in effect from the reference group) cannot be distinguished from artifacts of random sampling variance.

Re-Analysis with the NLSY79 Children and Young Adults

These results merit further investigation. Fortunately, data exists which can help bolster statistical power. First, in early 2015 the PSID 2013 data will be available and will include two new cohorts in the younger CDS and TA samples. This will allow the panel structure of the data to be used more powerfully

in separating the effects of NCLB and pre-NCLB accountability regimes. Second, the National Longitudinal Survey of Youth (NLSY79) began surveying children in 1979. In 1986, a new study of the children born to the women in the NLSY79 was initiated. The new survey is given to children and young adults biennially and covers cohorts similar to those in the PSID, CDS, and TA data. The data available is broadly similar to the CDS and TA and, when combined with the PSID data, will double the sample size from seven to more than fourteen thousand. More encouragingly, restricted use data will enable a much more accurate measure of the degree to which a student would have been consequential to their school's accountability score. As Deming et al. (2013) reveals, it is not just the student demographics that matter but the match between the students and the schools they attend. By leveraging school district data available with the restricted use PSID and NLSY data, a more focuses powerful test of the impact of accountability will be possible. A power analysis confirms that the effects found here will be able to be distinguished from random sampling variance at standard levels of power and significance.

Works Cited

Adamson, F., & Darling-Hammond, L. (2012). Funding Disparities and the Inequitable Distribution of Teachers: Evaluating Sources and Solutions. *Education Policy Analysis Archives*, *20*(37).

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. D. (2011). *Personality Psychology and Economics* (No. 16822). Cambridge. Retrieved from http://www.nber.org/papers/w16822

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2013). *Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood* (No. 19424). Retrieved from http://www.nber.org/papers/w19424

Dee, T. S., & Jacob, B. (2010). *The Impact of No Child Left Behind on Students, Teachers, and Schools*. *Brookings Papers on Economic Activity* (Vol. 2010, pp. 149–194). Washington, D.C. doi:10.1353/eca.2010.0014

Dee, T. S., & Jacob, B. (2011). The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management*, *30*(3), 418–446.

Deming, D. J., Cohodes, S., Jennings, J., & Jencks, C. (2013). *School Accountability, Postsecondary Attainment and Earnings* (No. 19444). Cambridge. Retrieved from http://www.nber.org/papers/w19444

Donovan, C., Figlio, D. N., & Rush, M. (2006). *CRAMMING: THE EFFECTS OF SCHOOL ACCOUNTABILITY ON COLLEGE-BOUND STUDENTS* (No. 12628). Cambridge.

Hanushek, E. A. (2003). The Failure of Input-based Schooling Policies*. *The Economic Journal*, *113*(485), F64–F98.

Hanushek, E. A., Kain, J. F., Brien, D. M. O., & Rivkin, S. G. (2005). *The Market for Teacher Quality* (No. 11154). Cambridge.

Hanushek, E. A., & Raymond, M. E. (2001). The Confusing World of Educational Accountability. *National Tax Journal*, *54*(2), 365–384.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review*, *100*(2).

Jacob, B. A., & Levitt, S. D. (2003). Rotten Apples : An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics*, *118*(3), 843–877.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. a, & Hamilton, L. (2004). Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101.

Mccaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability*. Santa Monica, CA.

Reardon, S. F. (2011). The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations. In G. J. Duncan & R. J. Murnane (Eds.), *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances* (pp. 91–116). New York: Russell Sage Foundation.

Springer, M. G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J. R., McCaffrey, Daniel, F., … Stecher, B. M. (2010). *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*. Nashville, TN. Retrieved from https://my.vanderbilt.edu/performanceincentives/files/2012/09/Full-Final-Report-Experimental-Evidence-from-the-Project-on-Incentives-in-Teaching-2012.pdf

Wong, K. N. (2008). *Looking Beyond Test Score Gains: State Accountability's Effect on Educational Attainment and Labor Market Outcomes*. Irvine, CA.