Abstract Title Page

Title: Accountability and Teacher Practice: Investigating the Impact of a New State Test and the Timing of State Test Adoption on Teacher Time Use

Author(s): Erin F. Cocke, Jack Buckley, and Marc A. Scott

Abstract Body

Background / Context:

There is much debate over the impact of high stakes testing as well as a growing body of research focused on both the intended and unintended consequences of these tests. One claim of both the popular media and education researchers is that high stakes tests have led to curricular narrowing – the idea that school time is increasingly allocated to tested subjects to the detriment of non-tested ones (Dillon, 2006; Center for Education Policy, 2006; West, 2007). In order to investigate the effects of testing on the allocation of instructional time, we analyze changing trends in reported teacher time use in situations where testing in new subjects has been recently added. This study uses the three most recent waves (1999-2000, 2003-2004 and 2007-2008) of the Schools and Staffing Survey (SASS) data to explore the how the addition of tests in science and social studies over time have impacted teacher time use within states.

The unprecedented levels of state accountability introduced by the No Child Left Behind Act of 2001 (NCLB) are largely manifest in the form of high stakes tests and their associated incentives and sanctions. Critical research focused on this federal act and the subsequent changes in state, district and school level accountability systems typically either address whether NCLB is "working" or the unintended consequences of these high stakes tests (Dee and Jacob, 2009; West, 2007; Booher-Jennings, 2005; Hanushek and Raymond, 2005). Most recently, Dee and Jacob find that NCLB produced significant increases in fourth grade math achievement scores but no significant increases in either eighth grade math or reading achievement scores (2009). Still, some strongly assert that, in general, high stakes testing leads to increased student performance (Hamilton & Strecher, 2002).

While the stated goal of NCLB, and standards based accountability in general, is to increase student learning many argue that educators focus their efforts on imparting only the skills and content necessary to perform well on tests. Changes in instructional time may include weeks of practice test taking and test preparation, and a reduction in the breadth or depth of topic coverage (Dillon, 2006; Hamilton, Berends & Stecher, 2005; Ladd & Zelli, 2002). However, if high-stakes testing does contribute to increased performance, the mechanism behind a change in performance remains unknown. Many studies imply that if student achievement is increasing this improvement must be motivated by a change in classroom or teacher practices. In fact, there is very little evidence to support this assumption.

Recent work by Reback, Rockoff & Schwarz significantly contributes to our understanding of the impact of high stakes testing on teachers and schools (2010). This rigorous work provides evidence that reading and math teachers in NCLB tested grades, teaching in schools that are close to making AYP, spend significantly more time on test preparation. West (2007) also attempts to unpack the impact of state testing on teacher time use using SASS data and finds that not only has time spent on reading and math increased over the past fifteen years but that time spent on social studies and science has decreased. He combines this descriptive evidence with inferential evidence asserting that teachers in states with a test in science or social studies spend more time teaching these subjects than in states without these tests (2007). However, as West admits, due to a lack of causal design his work does not provide a confident causal link between state testing and teacher time use.

Ultimately, the goal of NCLB is to increase student learning, particularly for lower performing students. Despite decades of unsuccessful education reform aimed at changing instructional practice there is evidence that high stakes testing and NCLB has altered teacher practice (Booher-Jennings, 2006; Reback et al 2010). This work aims to better capture this change in behavior through causally examining the extent to which the introduction of new high stakes tests in social studies and science are associated with a change in teacher practice, specifically in how

teachers allocate their time across subjects.

Purpose / Objective / Research Question / Focus of Study:

The purpose of this study is twofold. First, we investigate whether the addition of new tests in social studies and science impact how much time per week general elementary school teachers in self-contained classrooms report spending on social studies and science instruction. Second, we further investigate differences in this impact for early state test adopters (states that added new tests between 2000 and 2004) versus later state test adopters (states that added new tests between 2004 and 2008) in order to better understand the mechanism linking testing and changes in teacher practice.

Setting:

This study employs secondary data analysis of Schools and Staffing Survey (SASS) data. This data and our sample will be discussed in more detail below.

Population / Participants / Subjects:

The ideal teacher for our sample teaches in grades 1 through 5 and teaches all four core subjects to the same group of students every day. Given that we are analyzing teacher self reports of time use - as opposed to principal or educational administrator reports which might be able to speak to school or district level resource shifts – we assert that the general elementary school teacher offers the best opportunity to capture changes in teacher behavior under new pressures. In the face of new testing pressure, only this type of teacher is responsible for allocating time to each of the four core subject areas thereby offering a unique perspective on instructional practices.

We therefore, restrict our final sample to public elementary school teachers who report teaching students in grades 1 through 5. Additional exclusion criteria for the sample are based on several survey questions. Our final sample includes teachers who described their class organization as self contained, who report their teaching field as "Elementary, general," and who report teaching in only one of grades 1 through 5. Further, we utilized the time use variables to select teachers who report teaching hours in at least three of the four subject areas. Our final sample contains a total of 14,557 teachers across the three waves, with just over 5,000 teachers reporting in years 2000 and 2004 and slightly fewer than 5,000 in 2008.

Intervention / Program / Practice:

As mentioned above, our study explores the impact of new state mandated tests in social studies and science. These state mandated tests can be thought of similarly to an intervention or program, however not in a traditional sense. In our case, some states had already begun testing before the federal mandate, others waited until the mandate deadline, and others still began testing, only to cease the state test and start again later. While this example does not represent a traditional intervention, we are able to exploit this variation in the addition of testing in order to assess its impact on teacher time use.

Research Design:

Our study is a secondary data analysis using SASS data merged with additional state testing data. The data and analysis is described in more detail below.

Data Collection and Analysis:

This longitudinal analysis uses public teacher self reports of time use from three consecutive waves of restricted-use SASS data (1999-2000, 2003-2004 and 2007-2008). SASS data provides both statewide and nationally representative samples of teachers from elementary and secondary schools, which is important in analyzing the impact of a federal mandate. Teachers were asked how many hours (rounded to the nearest whole hour) they spent teaching English, reading or language arts - and of those hours how many were designated for reading instruction, how many hours they spent teaching math, social studies or history and finally, how many hours they spent teaching science. This study focuses on two of these time use variables, hours spent in science and hours spent in social studies as our dependent variables.

SASS data was then combined with data from a yearly Education Week report, "Quality Counts," which contains information on state testing policies. This data not only specifies whether each state has at least one state assessment in all four subject areas but also whether the test is present at the elementary, middle, or high school levels (Education Week Press, 2000, 2004, 2008). While the "Quality Counts" report provided the main source of testing data additional checks with online state report cards were used to validate this data source. Our future work will add in additional state testing data from a yearly report by the Council of Chief State School Officers (CCSSO) for years prior to the start of the "Quality Counts" reports (1994 to 1999).

In order to investigate the presence of a causal link between the addition of new state testing and teacher time use we first utilize a state and year fixed effects analysis. We can employ this method due to the variation within states in the presence of a test in social studies or science over the three SASS waves (1999-2008). The inclusion of state fixed effects allows us to control for unobserved state specific characteristics that do not change over time. The inclusion of year specific effects adjusts for a time-constant, unobserved, confounding variable possibly affecting all states, such as the adoption of a new social studies curriculum by a national social studies teacher organization.

However, not all states experience a change in testing status in social studies or science over the three SASS waves. Thirty-one states experienced at least one change in their science testing status and fifteen states experienced at least one change their social studies testing status. While the fixed effects analysis offers a more convincing identification strategy, our results are only generalizable to states that experienced a change in testing status.

In addition to state and year fixed effects our models also adjust for classroom level student demographics, school level demographics, school organizational characteristics and teacher characteristics. We also included grade level indicator variables based on the theory that teachers of different grades may respond differently to a new state test. Theoretically and statistically motivated interactions between several covariates and our testing variables of interest were also included. Finally, while state fixed effects account for within state correlation in teacher time use we also used a robust estimator of standard errors for the regression coefficients, clustered within states[1]. Our final fixed effects models therefore, adjust for student, school and teacher characteristics and state and year fixed effects, and cluster standard errors to adjust for within state correlation.

[1] In order to account for the sampling strategy used to create the SASS sample we include all of the sampling variables in every model rather than using survey weights. Both Little (1993) and Winship and Radvill (1994) demonstrate that including these sampling variables leads to the same estimates and associated standard errors as would the appropriate survey weights.

Our current work utilizes difference-in-difference (DiD) models in order to determine whether the impact of state science testing on teacher time use differs by time of test adoption. These models are based on the theory that early science test adopters may differ in observable and unobservable ways from later science test adopters. Descriptively, we know that about half of the states that experience a change in science testing status, experience this change in either 2006 or 2007, presumably due to the federal mandate. As states in our data have not yet experienced a social studies test mandate, and only 15 states in total experience a change in social studies testing status, we only pursue these models for the addition of a test in science.

By running two DID models, one for the 2000 and 2004 time period and one for the 2004 to 2008 period we aim to better identify whether simply the presence of a state test in science impacts instructional time or whether unobservable state level efforts or differences may be driving the impact on day to day teacher practice. We also plan to include a lag-time by treatment interaction variable in all of our DiD models. The lag-time represents the number of years a state has been testing in the science. Finally, we also run a larger difference-in-difference model which examines the difference between time use in 2000 and 2008.

Findings / Results:

Descriptively, we find that higher grades teachers (third through fifth grade) and male teachers spend more time, on average, teaching science than lower grades (first and second grades) and female teachers, respectively. Interestingly, charter school teachers spend about the same amount of time, on average, as traditional public school teachers in science and math but less time in English and more time in social studies than traditional public school teachers (please insert figures 1 through 3 here).

The results from our fixed effects analysis indicate a significant positive impact on the time spent in science due to a state test in science[2]. For teachers in states that added a science test the addition of this test corresponds to a seven minute per week increase in the amount of time spent on science. While seven additional minutes is arguably not substantial it does represent approximately a five percent increase in the amount of time spent on science per week with the addition of a state test. There is no significant impact demonstrated from the addition of a test in social studies on the amount of time spent in social studies. We believe that the lack of impact from a new test in social studies is due to both the pattern of social studies test adoption (several states with social studies tests discontinued them) and the relatively few states that added a social studies test. Therefore, our discussion of results will center on the findings related to the addition of a science test (please insert figure 4 here).

The model assessing the impact of the addition of a science test on time spent in science contains several variables and coefficients of interest in addition to testing. The small but significant coefficient associated with school size, labeled school enrollment, indicates that as school size increases the amount of time spent on science very slightly decreases, given a test in science. The results also show that after accounting for the addition of a test in science, teachers of different grades report spending different amounts of time per week on science and that fifth grade teachers report the most time spent on science (.707, $p < .001$). More specifically, the significant main effect for fifth grade teachers indicates that after accounting for a test in science, forty-two more minutes are spent on science in 5th grade per week. However the 5th grade by test in science interaction did

[2] Due to a positive skew in the distribution of both hours spent in science and hours spent in social studies both models were run using unlogged and logged versions of the dependent variables. The results from the model assessing the impact of a test in science on the logged hours spent in science are very similar to those shown in figure 4 and are significant at the $p \leq .10$ level.

not yield significant results indicating that 5th grade teachers are not impacted differently by the addition of a test in science.

The charter school and test in science interaction is added in order to assess whether charter school teachers behavior is impacted differently by the addition of a test in science. Our results indicate that charter school teachers are impacted differently; with the addition of a state test in science charter school teachers subsequently report spending an additional 24 minutes on science per week than they would have spent had the state in which they teach not added a test in science. Finally, our results show that given the addition of a state test in science female teachers report spending about fourteen minutes less in science per week than male teachers.

Our preliminary DiD models suggest a much larger effect in the 2000 to 2004 time period than the 2004 to 2008 time period. As we would expect, the 2000 to 2008 DID model shows a much dampened impact of testing on teacher time use. Again, this finding is consistent with the theory that early science test adopters may differ in important unobserved ways from later science test adopters. However, these models are still preliminary as we have not yet included the lag time by treatment interaction variable. We are currently adding new state testing data from the before mentioned CCSSO report and will be incorporating lag time into our next models.

Conclusions:

This exploration of the impact of a state test in science and social studies on teacher time use indicates that for states that added a test in science there is a small impact of this test on reported teacher time use in science. In addition, there is no significant impact of a new test in social studies on teacher time use in social studies. These results are in contrast with prior work finding a significant impact of a test in social studies and science on reported teacher time in these subjects (West, 2007). One obvious conclusion is that the content of what teachers are teaching matters and is driving change in student test scores rather than the actual time spent teaching each subject. However, this small impact could also be due to the lack of federal pressure currently associated with social studies and science tests, as these tests do not yet impact whether a school meets Average Yearly Progress. Teacher behavior may understandably be more responsive to high pressure accountability than to accountability without sanctions attached.

Our current work using DiD models will further disentangle whether the finding of a small increase in science instruction is being driven by a select group of 'early adopter' states. The preliminary results suggest that there is a difference in impact between states that adopted a test in the 2000 to 2004 time period versus states that adopted a test in the 2004 to 2008 time period. However, while both our fixed effects and DiD models attempt to identify the mechanism driving increases in student test scores following increased state testing we still cannot specify why teachers may be spending more time in science instruction. It is possible that teachers are reacting individually to state testing pressures, or that districts or schools are requiring more time spent in the newly tested subjects, or finally some combination of these two scenarios. Also, as SASS samples a new group of teachers in every survey administration it is not possible to include district or school level fixed effects as there are not enough teachers from the same districts and schools over time. While state level fixed effects are important and necessary, this analysis is not able to account for the additional nested structure of educational data.

Appendices

Appendix A. References.

Booher-Jennings, J. (2005). Below the Bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal*, 42, 231-268.

Booher-Jennings, J. (2006). Rationing Education in an Era of Accountability. *The Phi Delta Kappan*, 87(10), 756-761.

Carnoy, M. & Loeb, S. (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis, *Educational Evaluation and Policy Analysis,* 24(4), 302-331.

Center on Education Policy. (2006). From the Capital to the Classroom: Year 4 of the No Child Left Behind Act. (Washington, DC: Center on Education Policy).

Cullen, J. & Raebeck R. (2006). Tinkering Toward Accolades: School Gaming Under a Performance Accountability System, *Working Paper*, National Bureau for Economic Research.

Dee, T. & Jacob B. (2009). The Achievement Consequences of the No Child Left Behind Act. *Working Paper*, National Bureau for Economic Research.

Dillon, S. (2006). Schools Cut Back Subjects to Push Reading and Math. *New York Times*, March 26, A1.

Figlio, D. & Getzler L. (2002). Accountability, Ability and Disability: Gaming the System. *Working Paper*, National Bureau for Economic Research.

Fuller, B., Wright J., Gesicki K., & Kang E. (2007). Gauging Growth: How to Judge No Child Left Behind? *Educational Researcher* 36(5), 268-278.

Hamilton, L. & Strecher, B. (2002). Improving Test-Based Accountability. In L. Hamilton, B. Strecher & Klein, S (Eds.), *Making Sense of Test-Based Accountability in Education* (p.121). Santa Monica, CA: Rand.

Hamilton, L., Berends, M., & Stecher, B. (2005). Teachers' Responses to Standards-Based Accountability. *Working Paper*, Rand.

Hanushek, E. & Raymond, M. (2005). Does School Accountability Lead to Improved Student Performance? *Journal of Policy Analysis and Management*, 24(2), 297-327.

Jacob, B. (2005). Accountability, Incentives and Behavior: the Impact of High-Stakes Testing in Chicago Public Schools. *Journal of Public Economics*, 89, 761-796.

Jacob, B. & Levitt, S. (2003). Rotten Apples: an Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics,* CXVIII (3), 843-878.

Jennings, J. & Beveridge, A. (2009). How Does Test Exemption Affect Schools' and Students' Academic Performance? *Educational Evaluation and Policy Analysis*, 31(2), 153-175.

Ladd, H. & Zelli, A. (2002). School-Based Accountability in North Carolina: the Responses of School Principals. *Educational Administration Quarterly*, 38, 494-529.

Little, R.J.A. (1993). Post-Stratification: a Modeler's Perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

Neal, D. & Schanzenbach, D. (2007). Left Behind By Design: Proficiency Counts and Test-Based Accountability. *Working Paper*, National Bureau for Economic Research.

Reback, R., Rockoff, J. & Schwartz, H. (2010). Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under *NCLB. Working Paper.* Barnard College and Columbia Business School.

West, M. (2007). Testing, Learning, and Teaching: The Effects of Test-Based Accountability on Student Achievement and Instructional Time in Core Academic Subjects. In Finn, C. & Ravitch, D. (Eds.) *Beyond the Basics: Achieving a Liberal Education for All Children*. Washington, DC: Thomas B. Fordham Institute, pp. 45–62.

Winship, C & Radbill, L. (1994). Sampling Weights and Regression Analysis. *Sociological Methods and Research*, 23(2), 230-257.

Appendix B. Tables and Figures

Not included in page count.

Figure 1.

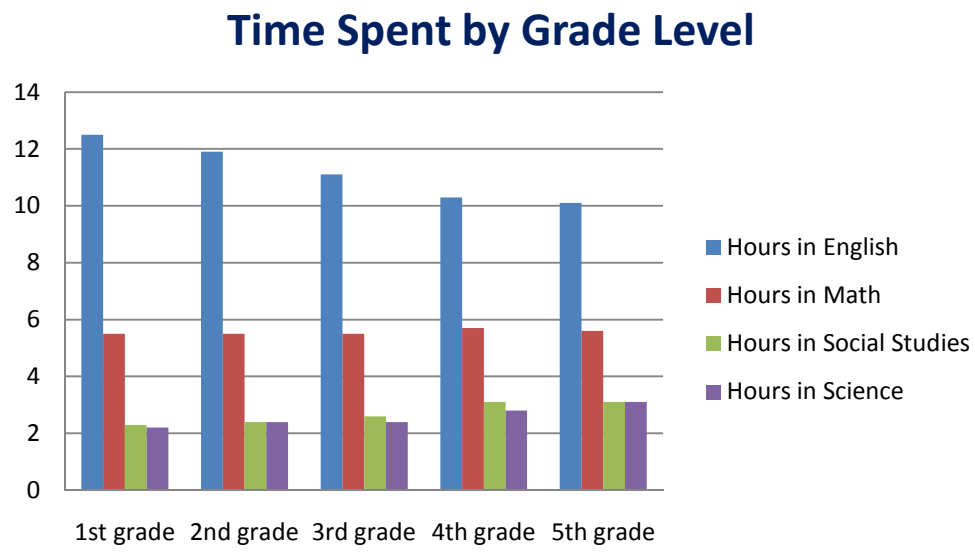

Figure 2.

Teacher Time Use by Gender

	Female Teachers (in hours)	Male Teachers (in hours)	Female – Male Difference (in minutes)
English	11.4	10.8	36
Math	5.5	5.7	-12
Social Studies	2.6	3.0	-24
Science	2.5	2.8	-18
N	13318	1239	-

Figure 3.

Teacher Time Use by Charter School Status

	Charter Schools (in hours)	Non-Charter Schools (in hours)	Charter – Non Difference (in minutes)
English	11.1	11.4	-18
Math	5.6	5.6	0
Social Studies	2.8	2.6	12
Science	2.5	2.5	0
N	653	13904	-

Figure 4.

The Impact of a Test in Science or Social Studies on Reported Teacher Time Spent on Science and Social Studies

	Science	Social Studies
Elementary Test in Science	0.1160* (0.0520)	—
Elementary Test in Social Studies	—	0.0320 (0.0710)
Suburban	-0.0317 (0.0424)	-0.0334 (0.0406)
Rural	0.0495 (0.0500)	-0.0021 (0.0505)
School Enrollment	-0.0005** (0.0002)	-0.0001 (0.0002)
Enrollment^2	0.0000** (0.0000)	0.0000 (0.0000)
2004	-0.4283*** (0.0447)	-0.5150*** (0.0511)
2008	-0.5600*** (0.0643)	-0.6572*** (0.0537)
2nd grade	0.1027* (0.0413)	0.1210** (0.0455)
3rd grade	0.3112*** (0.0557)	0.2828*** (0.0497)
4th grade	0.6805*** (0.0876)	0.8012*** (0.0671)
5th grade	0.7072*** (0.1094)	0.9448*** (0.0791)
5th grade x Test in Science	0.1731 (0.1148)	—
5th grade x Test in Social Studies	—	0.1625 (0.1654)
Percent free or reduced price lunch	-0.0010 (0.0008)	-0.0016* (0.0007)
Charter	-0.0155 (0.0617)	0.3661** (0.1160)
Charter x Test in Science	0.4001** (0.1416)	—
Charter x Test in Social Studies	—	0.1685 (0.1408)
Class Size	-0.0051 (0.0047)	0.0001 (0.0038)
Female	-0.2385*** (0.0643)	-0.3083*** (0.0595)
Teacher Age	-0.0000 (0.0010)	0.0033* (0.0012)
Constant	2.967*** (0.1295)	2.886*** (0.1169)
N	14557	14557