Teacher effectiveness on high- and low-stakes tests*

Sean P. Corcoran†
Jennifer L. Jennings
New York University

Andrew A. Beveridge
Queens College/CUNY

June 26, 2012

Abstract

A large literature finds substantial variation in teachers' effects on student achievement. Moreover, this research finds that little of this variation in effectiveness can be explained by traditional measures of quality, such as years of teaching experience. There remains, however, a gap in our understanding of how the choice of test measure—and teachers' own stake in the test's outcome—affects inferences about teacher quality. For example, test-based accountability policies may incentivize teachers to focus efforts on short-term, test-specific skills which may or may not generalize to other tests. In this paper, we use data from a large urban school district to estimate teacher effects on high- and low-stakes tests of the same content areas. We find that: (1) teacher effects are 15-31% larger on the high-stakes test, (2) teacher effects on the high-stakes test are moderate predictors of effectiveness on the low-stakes test, (3) returns to experience differ across tests in ways consistent with teachers' incentives to invest early in teaching skills and content specific to the high-stakes test, and (4) teacher effects on the high-stakes test decay at a faster rate than those on the low-stakes test.

*We would like to thank the Houston Independent School District for providing necessary data, and the IES pre-doctoral training program for providing research support. Jennings received additional support for this project from IES-AERA and Spencer Foundation dissertation fellowships. Rachel Cole provided expert research assistance. We thank seminar participants at NYU, the University of Notre Dame, Stanford University, and the Institute for Research on Poverty Summer Workshop for useful feedback. Dan McCaffrey, Doug Harris, Susanna Loeb, Sean Reardon, Matthew Wiswall, Julie Cullen, and two anonymous referees provided especially helpful comments on an earlier draft. All remaining errors are our own.

†Contact author. NYU Steinhardt School of Culture, Education, and Human Development, 665 Broadway Suite 805, New York, NY 10012. Phone: (212) 992-9468 E-mail: sean.corcoran@nyu.edu.

1 Introduction

A large literature finds substantial variation in teachers' effects on student achievement (e.g., Kane and Staiger, 2008; Nye, Konstantopoulos, and Hedges, 2004; Rivkin, Hanushek, and Kain, 2005). Moreover, this research finds that little of this variation can be explained by traditional measures of teacher quality, such as certification, degree attainment, and experience (e.g., Aaronson, Barrow, and Sander, 2007; Hanushek et al., 2005; Kane, Rockoff, and Staiger, 2008; Rockoff 2004). Taken together, these findings have fueled a recent movement to evaluate, promote, compensate, and dismiss teachers based at least in part on their estimated value-added to student achievement.[1] This movement was exemplified most recently by the federal Race to the Top competition, which rewarded states for implementing value-added systems of teacher evaluation based on standardized tests.

There are good reasons, however, to pay greater attention to the outcome measure on which inferences about teacher quality are based. First, teacher effects on a given standardized achievement test may only generalize to the domain of skills represented by that test (Koretz, 2008). Second, we have limited evidence on the extent to which teachers' short-run effects on achievement correspond to long-term impacts on achievement, attainment, and well-being (Chetty, Rockoff, and Friedman, 2011). Third, features of the tests themselves—such as their scaling, format, structure, and sensitivity to gains over the full range of the achievement distribution—may affect value-added estimates in predictable and unpredictable ways (e.g., Ballou, 2009; Briggs and Weeks, 2009; Koedel and Betts, 2009; Papay, 2011).

[1] In a 2009 speech, Bill Gates provided his own summary of this research: "A top quartile teacher will increase the performance of their class–based on test scores–by over 10 percent in a single year...That means that if the entire U.S., for two years, had top quartile teachers, the entire difference between us and Asia would go away. So, it's simple. All you need are those top quartile teachers...What are the characteristics of this top quartile?...You might think these must be very senior teachers. And the answer is no. Once somebody has taught for three years their teaching quality does not change thereafter."

Finally, for reasons described below, the attachment of high-stakes performance evaluation to a specific test could undermine that test's ability to draw valid inferences about teachers' effectiveness in increasing math and reading knowledge more broadly (Neal, 2011).

Research on school-level responses to high-stakes accountability offers some insight into the latter issue. It is now well-documented that gains on state achievement tests have significantly outpaced those on low-stakes national benchmark tests like the NAEP, with gains on some state tests nearly four times as large (Center on Education Policy, 2008; Fuller et al., 2007; Jacob, 2007; Koretz and Barron, 1998; Klein et al., 2000). These inconsistencies are at least in part attributable to strategic behaviors that inflate gains on high-stakes tests. For example, schools have been found to selectively exclude low-performing students from testing, both through suspension and re-classification into special education (Cullen and Reback, 2006; Figlio and Getzler, 2006; Figlio 2006; Jacob 2005; Jennings and Beveridge, 2009). Other evidence suggests schools re-allocate resources toward students on the margin of passing when given incentives to do so (Booher-Jennings, 2005; Reback, 2008; Neal and Schanzenbach, 2010). Coaching students to respond to the specific content or format of the state test—and in more extreme cases, cheating—are other explanations for inflated gains on high-stakes tests (Jacob 2005, 2007; Koretz and Barron, 1998).

In contrast, research on teacher quality has devoted less effort to understanding the impact high-stakes accountability has on inferences about teacher effectiveness. In this paper, we use data from the Houston Independent School District (HISD) to estimate teacher effects on high- and low-stakes tests of the same content areas. Since 1996, HISD has administered two standardized tests each spring: the TAAS/TAKS, required by the Texas state accountability system, and the nationally-normed Stanford Achievement Test (SAT).[2] The former

[2]The TAKS replaced the TAAS in 2002-03.

is a "high stakes" test in the sense that schools and teachers can be rewarded or punished according to students' progress on these tests, while the latter is a "low stakes" test intended as both an "audit" and a diagnostic assessment.[3] Because these tests are administered to the same students at roughly the same time of year, they provide an opportunity to understand whether and how value-added measures differ across tests, and the role accountability may play in generating these differences.

Answers to these questions are important for at least two reasons. First, as noted above, there is a gap in our understanding of how the choice of outcome measure—and teachers' own stake in the test outcome—affect inferences about teacher effectiveness. To the extent accountability policies alter teacher behavior, inferences about teacher quality will be influenced by these behavioral responses. Second, those adopting value-added systems to evaluate individual teachers do so under an assumption that these measures are broad and relatively consistent indicators of teaching effectiveness, and not highly sensitive to the choice of test instrument. Our paper is one of the few to examine these questions directly.

In line with prior research, we find large effects of 4th and 5th grade teachers on student achievement on both tests. However, we also find these measures differ in interesting and important ways across tests. For example, the magnitude of teacher effects—as estimated by their standard deviation—appears 15-31% larger on the high-stakes test than on the low-stakes test. Moreover, these measures are modestly correlated at the individual teacher level. Based on estimates using an identical sample of students and up to eight years of classroom data for each teacher, the correlation in teacher effects across the two tests is 0.50 in reading and 0.59 in math. Notably, we find the correlation in teacher effects is stronger *between subject areas* on the same test battery than across tests of the *same content*

[3] As we explain in Section 3, the "low-stakes" test is not entirely without consequences for students.

area. Consequently, these correlations yield inconsistent rankings of teachers. That is, many exceptional teachers on the state tests would be deemed less effective or ineffective on the low-stakes tests. The inconsistencies we observe across tests bears a strong resemblance to the pattern of year-to-year noise in teacher effects found in other work (e.g., McCaffrey et al., 2009).

We also find notable differences in the returns to teaching experience across the two tests. As others have found, on the high-stakes test we find the biggest increase in teacher effectiveness occurs within the first few years of a teacher's career, with minimal gains thereafter. In contrast, we find positive returns to experience on the low-stakes test over a longer time horizon, particularly in reading. An apparent depreciation in effectiveness among experienced teachers on the high-stakes math test is not observed on the low-stakes test. While these patterns are not sufficient evidence to demonstrate a role for accountability, they are consistent with teachers' incentives to invest early in teaching skills and/or content specific to the high-stakes test.

Finally, we use an empirical strategy proposed by Jacob, Lefgren, and Sims (2010) and Kane and Staiger (2008) to estimate the extent to which teacher effects measured by the two tests persist into students' later outcomes. If high-stakes tests incentivize teachers to focus efforts on short-term, depreciable skills, one would predict that effects on the high-stakes test dissipate more quickly than those on the low-stakes test. This is exactly what we observe in the data: persistence is higher on the low-stakes SAT than on the high-stakes TAAS/TAKS. We estimate that 60% of the prior year teacher's value-added on the SAT carries forward into the next grade, as compared with 40% for the TAAS/TAKS. Moreover, this is unlikely to be due to differences in test scaling or the continuity of tested content across grades. We find a teacher's value-added on the *low*-stakes assessment has more enduring effects on

high-stakes test outcomes than does value-added on the high-stakes test itself.

2 Existing evidence on teacher effect variability

In their seminal paper, Rivkin, Hanushek, and Kain (2005) demonstrated a substantial degree of variation between teachers in their effects on student achievement. Using data from Texas, they found that a one standard deviation (s.d.) increase in teacher value-added was associated with a 0.10 s.d. increase in reading test scores and a 0.11 s.d. increase in math scores. These estimates are consistent with those found in other contexts, including Chicago (Aaronson, Barrow, and Sander, 2007), Florida (McCaffrey et al., 2009), New Jersey (Rockoff, 2004), North Carolina (Rothstein, 2010), San Diego (Koedel and Betts, 2009) and elsewhere (Jacob, Lefgren, and Sims, 2010; Kane and Staiger, 2008; Papay, 2011). In a unique study of teacher effects under random assignment, Nye, Konstantopoulos, and Hedges (2004) found even larger effects of teacher quality on student achievement in the early grades (see also Chetty et al., 2011a).

An equally consistent finding in this literature is that observed characteristics of teachers, such as qualifications, training, and experience, do little to explain this variation. Using data from New York City, Kane, Rockoff, and Staiger (2008) estimated standard deviations in annual teacher effects of 0.21 and 0.20 in elementary math and reading. Measures of experience, college selectivity, and pathway into teaching (e.g., traditional vs. non-traditional routes) were only weakly related to these effects. They concluded that the vast majority of variation in teaching effectiveness is within groups, rather than between them. Others, including Leigh (2009), Rivkin, Hanushek, and Kain (2005), Rockoff (2004), and Buddin and Zamarro (2009), similarly found that teachers' effects on achievement plateau after only a few years of experience.

The impact of these findings on public education policy cannot be overstated. National, state, and local education leaders have called for an overhaul of systems for evaluating, rewarding, promoting, and dismissing teachers, with an explicit link to student achievement (most often, value-added to standardized tests used for school accountability). Academics and policymakers alike have invested considerable resources into the further development of these models to objectively evaluate and reward teachers based on performance (Goldhaber and Hansen, 2008; Gordon, Kane, and Staiger, 2006, Hanushek, 2009).

The rapid adoption of teacher value-added has been accompanied by a surge in research on these models' properties. This work has focused primarily on (1) whether or not value-added measures are unbiased, consistent estimators of the causal impact of teachers on achievement (e.g., Kane and Staiger, 2008; Rothstein, 2010), (2) the sensitivity of value-added models to model specification (e.g., Ballou, Sanders, and Wright, 2004; Harris and Sass, 2006; Lockwood et al., 2007), (3) the precision and intertemporal stability of estimated teacher effects (e.g., Goldhaber and Hansen, 2008; McCaffrey et al., 2009; Papay, 2011), and (4) the relationship between teacher value-added and other subjective measures of performance (e.g., Harris and Sass, 2007; Jacob and Lefgren, 2008).

Comparatively less attention has been given to the outcome measure itself. While some studies have examined the role test scaling plays in value-added, (e.g., Ballou, 2009; Briggs and Weeks, 2009; Koedel and Betts, 2009), fewer have validated teacher effects against other short- or long-run outcomes of interest. Notable exceptions to the latter include two studies by Chetty et al. that found strong evidence of teacher effects on long-run outcomes. Using data from Project STAR in Tennessee, Chetty et al. (2011a) found effects of kindergarten classroom teachers on students' earnings at age 27 and on college attendance, even as effects on achievement appeared to fade out after several years. Drawing on a much larger sample of

students in grades 3-8, Chetty et al. (2011b) similarly found effects of teachers on earnings, college attendance, college quality, savings, and the likelihood of a teenage birth. Importantly, in both papers, teacher effects were estimated on a low-stakes tests. Other studies have found fadeout of teacher effects on achievement, suggesting these effects dissipate in as few as two years (Jacob, Lefgren, and Sims, 2010; Kinsler, 2012; Rothstein, 2010). No study that we are aware of has considered how differences in incentives across tests impact value-added measures of effectiveness.

Only three papers we identified examined the consistency of teacher effects across achievement measures. Lockwood et al. (2007) estimated teacher effects separately for each math subscale of the SAT. They found that the choice of scale—that is, the decision to emphasize one set of tested skills over another—has a large impact on a teacher's perceived effectiveness. They concluded that teacher effects are sensitive to the weights assigned by a test to specific skills. In a second paper not directly interested in variation in teacher effects across tests, Sass (2008) found a correlation of 0.48 between teacher effects estimated on high- and low-stakes tests in Florida, very close to the one we find here.[4]

In the paper closest to ours, Papay (2011) estimated teacher effects using three different reading tests for a single urban district: a high-stakes state test, the SAT, and the SRI. He found weak to moderate correlations in teacher effects across tests administered to the same students, ranging from 0.15 to 0.58. He carefully explored several hypotheses for these differences, including differences in test timing, scaling, and content, although not the stakes associated with the test. Differences in the return to experience across tests were not explored, nor was the long-run persistence of effects. In the next section we offer a more

[4]Using the same data, McCaffrey et al. (2009) noted that there was little difference in year-to-year stability in teacher effects when using the high-stakes FCAT test or the low-stakes FCAT-NRT, a version of the Stanford Achievement Test.

detailed discussion of reasons why teacher effects might vary across tests.

3 Why teacher effects might vary across tests

Following the recent literature, we define a teacher effect as the extent to which her students' achievement differs on average from that predicted by students' past test performance and other student, family, classroom, and school influences on achievement. For example, a model for test outcome Y_{ijst} for student i of teacher j in school s in year t might be written:

$$Y_{ijst} = \beta X_{ijst} + \underbrace{\delta_j + \gamma_{jt}}_{\phi_{jt}} + u_{ijst} \tag{1}$$

where X_{ijst} is a vector of relevant fixed- and time-varying inputs into the achievement of student i in year t—including a measure of prior-year achievement Y_{ijst-1}, student and family inputs, school- and classroom-level factors, and so on.[5] ϕ_{jt} is an annual "teacher effect" for teacher j in year t, which is assumed to consist of a fixed (time-invariant) level of effectiveness δ_j that we refer to as the "stable" teacher effect, and year-to-year classroom variation in achievement γ_{jt}. The latter may represent the effects of idiosyncratic changes in classroom composition, common shocks, short-run variation in teacher performance, and the like. The goal of most teacher value-added models is to provide consistent estimates of the stable teacher effects δ_j, requiring multiple years of classroom data, though for some applications the annual effects ϕ_{jt} are of interest.

The relevant question for our purposes is why teacher effects estimated on assessment

[5]For a richer description of this model and its underlying assumptions, see especially Harris and Sass (2006), Kane, Staiger, and Rockoff (2008), McCaffrey et al. (2009), and Rothstein (2010). The covariate adjustment model outlined here is the predominant approach in recent work, although some authors expand the model to include student effects α_i where feasible.

A might differ from those estimated using assessment B. Among other things, our interest is in the overall magnitude of effects (summarized, as in other work, by their standard deviation) and their relative rankings of teachers (summarized by their covariance). Some of the variation in effects across tests will simply be due to noise. Student, classroom, and test-level error will attenuate correlation in teacher effects across tests, although some of this noise diminishes as the sample size of students and classrooms grows. Other sources of variation across tests are more systematic. For example, variation in test content or administration, or in student and teacher incentives can produce systematic differences in effects even with large samples.

In what follows, we elaborate on plausible reasons why teacher effects might vary across tests. We also highlight the extent to which these mechanisms may play a role in explaining variation we observe in this district. While we will not be able to cleanly apportion our observed differences among these hypotheses, several of these mechanisms will be unconvincing explanations in our context.

1. **Student-level noise**. One would not expect a perfect correlation in teacher effects even if the same test were administered twice to the same group of students. Because a student's observed score Y_{it} is a noisy estimate of his or her true achievement ($Y_{it} = Y_i^* + \epsilon_{it}$), random error will weaken the correlation of teacher effects across administrations of similar tests. At best, the correlation in teacher effects will be bounded by the reliability of the test itself. In our context, both the TAKS and SAT had a reported reliability of 0.85 to 0.91 during this period, depending on the subject, year, and grade level.[6]

[6]TAKS reliability estimates are reported in the TAKS Technical Digest, found here for 2006-07: http://www.tea.state.tx.us/student.assessment/techdigest/yr0607/

2. **Classroom-level noise**. Classroom-level factors that affect one test but not the other (e.g., a bout of illness, or the transfer of a disruptive student) will also weaken the correlation of teacher effects across tests. Test timing may influence the extent to which a classroom-level shock is likely to affect one or both tests: the further apart the tests, the more likely time-varying shocks will differentially affect the two tests. Fortunately, in this district the high- and low-stakes tests are administered relatively close together in the Spring (early March to mid-April).

3. **Differences in tested populations**. Some assessments purposefully exclude special populations, such as those with disabilities, limited English proficiency, or high rates of mobility. To the extent estimated teacher effects on tests A and B are based on different students, they will undoubtedly differ. In this district, the low-stakes test is administered almost universally, while a modest share of students are excluded from the high-stakes test. For our analysis, we restrict our estimates to the set of students taking both tests.

4. **Test content and scaling**. Tests vary in domain and scale, even within subjects and grade levels. For example, a state test is often designed to assess minimum proficiency in the state's content standards while a national achievement test may be written to test a broader domain. The criterion-referenced scale of the former can exhibit ceiling effects—or limits to growth at the top end of the distribution—which can in turn affect estimates of value-added, depending on the severity of the ceiling (Koedel and Betts, 2010). The national test's norm-referenced scale, on the other hand, is less likely to suffer from ceiling effects. These scaling differences apply in our setting, as the SAT is a norm-referenced test (without a ceiling) and the TAAS/TAKS is a criterion-

referenced test. [7] The TAAS in particular has a low ceiling. We assess the importance of differences in scaling for our results in two ways: by estimating models separately for the TAAS and TAKS years, and by re-estimating all models after imposing an artificial ceiling on the SAT scale, following Koedel and Betts, 2010.

5. **Student effort**. Students' own investment in a test may vary depending on their incentive to perform well. Many state accountability tests (including this one) are also high stakes for students, used for grade promotion or other rewards. Students may devote less effort to low-stakes tests that have little impact on them. In this district the low-stakes test is not entirely without consequences for students. It is used as one of many criteria for grade promotion, and is used to place students in specific programs, including gifted and special education. As an imperfect test for differential student effort, we calculated the correlation between students' scores in grade *g* and *g-1* on each test. Were students to put less effort into the low-stakes SAT in a manner that introduced greater noise (e.g. more guessing), one would predict a weaker correlation on the SAT than the TAAS/TAKS. In fact, the correlation is higher on the SAT (0.78 vs. 0.65 in math, and 0.77 vs. 0.60 in reading).

6. **Test timing**. Papay (2011) showed that test timing has large effects on teachers' value-added ranking. For example, teacher effects measured from June to June can look different than those measured from January to January or Fall to Fall. Timing may disproportionately impact teachers of disadvantaged students who suffer a "summer setback" relative to their more advantaged counterparts (e.g., Alexander, Entwisle,

[7] At least one analysis (Hoey, Campbell, and Perlman, 2001) mapped the standards on Texas' 4th grade TAAS math test to those covered on the SAT and found considerable overlap, with 83% of the Texas standards represented on the SAT. (The SAT was a bit more inclusive, with 74% of SAT standards represented on the TAAS)

and Olsen, 2001). As noted, the high- and low-stakes tests were administered at approximately the same time of year in this district.

7. **Teacher incentives**. In most accountability systems, educators are rewarded for increasing current test scores, not necessarily broader sets of skills. High stakes attached to a test may thus incentivize teachers to focus on test-specific instruction which may or may not generalize to other tests. For example, teachers can "teach to the test" or "teach to the format," altering their instruction to present content in the same manner as it appears on the test (e.g., Darling-Hammond and Wise, 1985; Shepard and Dougherty, 1991; Pedulla et al., 2003; Jennings and Bearak, 2010; Holcombe, Jennings, and Koretz, 2010).[8] Teachers aware of systematic omissions and repetitions can substantially inflate student scores by narrowly focusing their efforts towards these items. The TAAS/TAKS tests have long played a prominent role in the state's accountability system, and since 2000 have been a part of this district's performance pay system.

8. **Other**. Other factors that may contribute to variation in teacher effects across tests include differential time allotted to the test (especially for teachers of students affected by time constraints), test length, differences in test administration, accommodations, and so on. We investigated many of these issues for this district, and found few notable differences across tests.[9]

[8]To the extent students learn how to correctly answer questions when they are presented in a specific way, but struggle with the same skills when they are presented in another, such strategies will generate performance differentials across teachers and tests. A useful example reported in Shepard (1988) was a set of questions involving adding and subtracting decimals. When addition was presented in a vertical format on the state test, 86% of students answered these questions correctly, but in horizontal format, only 46% of students did; for subtraction the parallel figures were 78 and 30%.

[9]For example, the TAAS/TAKS and SAT tests in 4th and 5th grade are strictly multiple choice type exams, with similar numbers of test items. Although the SAT offers an open response test, HISD administers only the multiple choice section (personal communication with Sharon Bauknight, HISD Department of Student Assessment, March 3, 2011). The state test is untimed, as is the SAT-10. The SAT-9, however, was a timed test. Accommodations for students with special needs were identical on the two tests, as required

As noted, it will be difficult to precisely apportion observed differences in teacher effects among these competing explanations. We can be fairly confident that test reliability, student effort, exemptions, and timing do not individually play a central role. Test content and scaling are potentially more important issues, and we assess their influence to the extent possible. In addition to these, differences in teacher incentives across tests generate some useful predictions. First, to the extent high-stakes test results are used—formally or informally—to evaluate performance, teachers will have a strong incentive to invest early in their career in teaching skills and/or content specific to these tests. The pressure to perform well by these measures will be highest in the teacher's probationary period, and should weaken as the teacher accumulates experience and reputation.[10] If true, the experience-effectiveness gradient should vary according to the test's stakes at different points in the teacher's career.

Second, if high-stakes tests create incentives for teachers to devote more greater to short-run, depreciable skills, then one would predict that teacher effects on such tests will be less likely to carry forward into their students' outcomes in future years (Jacob, Lefgren, and Sims, 2010; Carrell and West, 2010; Kinsler, 2012). As described in the next section, we test this hypothesis, using an empirical strategy outlined by Jacob, Lefgren, and Sims (2010) and Kane and Staiger (2008).

by students' individualized education plans.

[10]While Texas is not a collective bargaining state and does not have tenure, the performance review in HISD is much more rigorous during the teacher's probationary period.

4 Data and empirical approach

4.1 Data

For this paper we compiled a longitudinal dataset of all students tested in Houston between 1998 and 2006, approximately 165,000 students per year.[11] HISD is the seventh largest school district in the country and the largest in the state of Texas. Fifty-nine percent of its students are Hispanic, 29% are black, 8% are white, and 3% are Asian. Close to 80 percent of students are considered by the state to be economically disadvantaged, 27% are classified as Limited English Proficient, and 11% receive special education.

An important feature of this dataset is its inclusion of multiple test scores for each student—both the Texas state assessments (the TAAS or TAKS) and the Stanford Achievement Test (SAT) battery.[12] The TAKS is administered to students in grades 3 to 11 in reading/ELA, mathematics, writing, science, and social studies, though reading and math are the only subjects tested annually in grades 3 to 8. The SAT is given to all students in grades 1 to 11 in reading, math, language, science, and social science. Eligible students are permitted to take Spanish-language versions of these tests. Using all reported scale scores, we standardized within subject, grade, year, and test version (English or Spanish).

Our interest in estimating teacher effects on multiple tests placed some restrictions on the data we could use. First, only grades and subjects covered by both the TAAS/TAKS and SAT were considered. This limited us to grades 3-8 reading and math. Second, the need for a lagged achievement measure eliminated grade 3 (the first tested on TAAS/TAKS) and 1998 (our first year of data). Third, an accurate match of students to classroom teachers

[11] Throughout the paper we refer to the *spring* of the school year. Details on the construction of this dataset can be found in the online appendix.

[12] The TAKS replaced the TAAS in 2003. For a discussion of the differences between TAAS and TAKS see Jennings and Beveridge (2009) and Koedel and Betts (2009).

was required. Teacher links were available for students in self-contained classes, and because many 6th graders were not in such classrooms we excluded grade 6.[13] Taken together, our analysis focused on reading and math achievement in grades 4 and 5 in 1999 to 2006, approximately 30,000 students per year. Third grade achievement was retained to serve as a lagged measure, and sixth grade data was retained for our analysis of persistence into future grade-level achievement. While we are limited to only two grade levels (4 and 5), these grades are by far the most used in teacher effect studies.

Table 1 provides summary statistics for the students included in our sample. Not all enrolled at the time of the test could be used to estimate teacher effects. While most (98%) had a nonmissing SAT score, only 87-91% of students had a TAAS/TAKS score, depending on the year and subject.[14] The remainder were either exempted, given an alternative assessment, or absent (Jennings and Beveridge, 2009). Imposing the requirement that students have a lag score for the same test further limited the sample to 76-79% of the base for the TAAS/TAKS, and 87% for the SAT. Finally, requiring a valid teacher match, a classroom with fewer than 25% special education students, and a teacher with at least seven students over time limits the sample to 65-73% of enrollment, as shown in panel A.

This loss of student observations is characteristic of all teacher value-added analyses, but particularly those in urban school districts with mobile populations. Panel B of Table 1 illustrates the effects of limiting the sample to students with sufficient data to contribute to teacher effect estimates. Because mobile students tend to be lower achieving, imposing these minimum data requirements increases average achievement in the sample to 0.13 - 0.16

[13] Teacher IDs were not consistently assigned over time. To create consistent IDs, we followed a procedure using full names, race/ethnicity, campus assignment, and experience to identify unique teachers and link them across years.

[14] 128,453 4th graders and 124,955 5th graders were enrolled at the time of one or both tests. For the small number of students repeating a grade, we used their last observed year in that grade.

s.d. above the grade level mean, depending on the test. Students with sufficient data on *both* tests averaged 0.25 s.d. above the mean on the SAT, due to the fact that the lowest achieving students tend to be excluded from the state test (but not the SAT).

Panel C of Table 1 provides average characteristics of the approximately 186,000 students generating our teacher effect estimates. Our sample roughly mirrors the population of 4th and 5th graders in Houston, although recent immigrants, special education, and black students are somewhat underrepresented. About 13% of the students in our sample took the Spanish version of the TAAS/TAKS and SAT tests, though they were much more likely to do so in 4th grade than 5th, a reflection of district policy (21% vs. 5%).

Table 2 provides average characteristics of the more than 3,700 teachers in our analytic sample. For the typical teacher-year, teacher experience averaged 10.2 - 11.2 years, although a large fraction (about 9-10%) were in their first year of teaching, and 25.5% were in their first three years. Twenty-two to 25% of teachers held a master's degree, and a minority (34 to 37%) were white. The average teacher was observed for just fewer than three years in our 8-year panel, and served 60-70 students with the minimum required data.

4.2 Empirical estimation of teacher effects

Following Gordon, Kane, and Staiger (2006), Kane, Staiger, and Rockoff (2008), Papay (2011), Jacob and Lefgren (2008) and others, we estimate individual teacher effects on achievement using a student-level value-added model that controls for prior achievement. Separately for each test (TAAS/TAKS and SAT) and subject (math and reading), we estimate the following model:

$$Y_{igjst} = \beta_g X_{it} + \alpha_g \bar{X}^c_{jst} + \xi_g \bar{X}^s_{jst} + \pi_{gt} + u_{ijst} \tag{2}$$

where Y_{igjst} represents a score for student i in grade g, classroom j and school s in year t. X_{it} is a vector of fixed and time-varying characteristics of student i; most importantly, X_{it} includes a cubic function of prior year achievement for student i in both reading and math on the same test battery (TAAS/TAKS or SAT), and an indicator of whether or not student i took the Spanish version of the test, interacted with grade. Other student-level covariates include indicators for gender, age, race, economic disadvantage, special education and LEP status, recent immigrants, migrants, and a school move in the prior year. $\bar{X}^c_{jst}$ and $\bar{X}^s_{jst}$ are vectors of average classroom and school characteristics for students in classroom j and school s, and π_{gt} is a vector of grade-by-year indicators. The coefficients β_g, α_g, and ξ_g are subscripted to indicate that they are allowed to vary by grade. In addition to classroom and school averages of the student covariates, we control for the number of students per class in classroom j, an indicator for grade-mixed classrooms, and a school average passing rate for the TAAS/TAKS in all subjects.

We assume the error term u_{ijst} in (2)—the extent to which student i's test score differs from that predicted given her prior achievement, individual, classroom, and school characteristics—can be decomposed into variation due to their teacher j's stable or long-run effectiveness (δ_j) and other unexplained variation (v_{ijst}): $u_{ijst} = \delta_j + v_{ijst}$.

For most applications our parameters of interest are the δ_j, or stable teacher effects. The mean student-level residuals $\bar{u}_j$ for each teacher j can be thought of as an estimator for δ_j, with variance $V_j = \sigma^2_v/n_j$, which approaches zero as the number of students observed under teacher j increases. Because estimates of δ_j are imprecise, the overall variation in estimated teacher effects overstates the true variation in stable teacher effects σ^2_δ. A standard approach is to "shrink" the individual estimates toward the average (normalized to zero) by multiplying by the estimator's reliability coefficient, below. The resulting teacher effect

estimator is known as the empirical Bayes, or shrinkage estimator (Raudenbush and Bryk, 2002; Jacob and Lefgren, 2008):

$$\hat{\delta}_j = \left(\frac{\sigma_\delta^2}{\sigma_\delta^2 + V_j} \right) \bar{u}_j \tag{3}$$

The larger the noise variance for teacher j (V_j) relative to signal variance (σ_δ^2), the more the estimated teacher effect is "shrunk" toward zero.

Rather than construct individual teacher effects from mean residuals (e.g. Kane, Staiger, and Rockoff, 2008; Kane and Staiger, 2008), we estimate the δ_j via maximum likelihood, assuming random teacher effects. The best linear unbiased predictions (BLUPs) of δ_j in this case will be the empirical Bayes teacher effect estimates. We alternatively estimate the teacher effects δ_j directly in (2) assuming fixed effects (and applying the reliability coefficient to account for sampling variation, as in (3)), though in most cases random and fixed effects estimates produce very similar results.[15]

In estimating the stable teacher effects δ_j in equations (2)-(3) we make use of all available student data for each teacher, which can include as many as 8 classroom years and 225 students, though in practice most teachers have fewer. We require a minimum of seven students with sufficient data to estimate a teacher effect. To examine the properties of time-varying (annual) teacher effects, we re-estimate (2)-(3) replacing the δ_j with a teacher-by-year (or classroom) effect ϕ_{jt}. These teacher-by-year effects are again estimated via maximum likelihood assuming random effects. Naturally, because fewer students contribute to the annual estimates, the variance of the ϕ_{jt} exceeds that of the δ_j.

In some specifications of (2) we replace the vector of school characteristics $\bar{X}_{jst}^s$ with

[15]The correlation between teacher effects estimated under fixed and random effects models is typically greater than 0.95.

school fixed effects θ_s. These school effects are intended to account for systematic differences in achievement across schools due to school leadership, unmeasured resources, parental inputs, and the like. Some (e.g., Gordon, Kane, and Staiger, 2006) argue persuasively that within-school estimates of teacher effectiveness—like those produced by a model with school effects—are inappropriate if teacher quality is unevenly distributed across schools. In other applications, the use of school fixed effects has become standard practice. For most of our results, this specification choice makes little difference. Where possible, we report results under both model specifications.

Finally, to estimate the returns to experience on the two tests, we include a vector of teacher experience indicators E_{jt} directly in equation (2) and omit individual teacher effects (e.g., Clotfelter, Ladd, and Vigdor, 2006):

$$Y_{igjst} = \beta_g X_{it} + \alpha_g \bar{X}^c_{jst} + \xi_g \bar{X}^s_{jst} + \theta_s + \pi_{gt} + \gamma E_{jt} + w_{ijst} \tag{4}$$

E_{jt} consists of indicators for 11 experience categories: 1-7 years (six categories, with year 1 omitted) and 8-10, 11-15, 16-20, and 21+ years. We also include an indicator of whether or not teacher j holds a master's degree or higher.

Including a full set of time-varying student, classroom, and school level controls in model (4), along with school fixed effects, serves to ameliorate the impact of non-random sorting of teachers to schools (and classrooms within schools). This approach does not account for sorting on any remaining unobserved factors, nor will it account for non-random exit from the profession (see Wiswall, 2011). However, as long as these processes do not vary differentially with the unobserved determinants of achievement on the two tests, the returns to teaching experience E_{jt} across tests should be comparable.

4.3 Estimating long-run persistence in teacher effects

In two recent papers, Jacob, Lefgren, and Sims (2010) and Kane and Staiger (2008) used an instrumental variables approach to identify the long-run impact of prior educational inputs (such as teachers) on achievement. Their approach uses various estimates of the coefficient on lagged achievement in a student-level model to construct an estimate of persistence—the fraction of past inputs that carry forward to a student's current achievement. In their formulation, observed achievement Y_t in year t reflects long-run ($y_{\ell,t}$) and short-run ($y_{s,t}$) components:

$$Y_t = \underbrace{\theta y_{\ell,t-1} + \mu_t^\ell + \eta_t^\ell}_{y_{\ell,t}} + \underbrace{\mu_t^s + \eta_t^s}_{y_{s,t}} \tag{5}$$

where μ_t^ℓ and η_t^ℓ are contemporaneous inputs that affect the stock of long-run skills and μ_t^s and η_t^s are inputs that have short-run effects but are perfectly depreciable. Importantly, teachers (and other inputs) have both long- and short-run effects. Long-run skills in the prior year ($y_{\ell,t-1}$) carry forward with some rate of decay $(1-\theta)$.

In constructing an estimator of persistence, Jacob et al. initially wish to estimate θ—the general rate of persistence of long-run skills. However, $y_{\ell,t-1}$ is unobserved to the researcher. Rather, one observes Y_{t-1}:

$$Y_{t-1} = y_{\ell,t-1} + \mu_{t-1}^s + \eta_{t-1}^s \tag{6}$$

Multiplying (6) by θ and subtracting from (5) yields:

$$Y_t = \theta Y_{t-1} + \underbrace{\left(\mu_t^\ell + \eta_t^\ell\right) + \left(\mu_t^s + \eta_t^s\right) - \left(\theta\mu_{t-1}^s + \theta\eta_{t-1}^s\right)}_{\epsilon_t} \tag{7}$$

OLS estimates of θ from (7) will be biased downward, because lagged achievement Y_{t-1} is correlated with ϵ_t through the effects of short-run inputs in the prior year. Jacob et al.

frame this as a form of measurement error: Y_{t-1} is a noisy measure of long-run skill—the cumulative effect of past inputs—and thus OLS estimates of θ will be attenuated the more variation in Y_{t-1} is comprised of short-run effects. They show that θ can be consistently estimated via instrumental variables, using twice-lagged achievement Y_{t-2} as an instrument for Y_{t-1}. This estimator, which we refer to as $\widehat{\theta_{IV}}$, purges Y_{t-1} of its short-run measurement error. As we do, Jacob et al. estimate a θ close to 1.

The θ parameter tells us the persistence of *all* long-run skills, however, not the persistence of specific inputs. Were one to instead instrument Y_{t-1} with a specific input from a prior period that has both long- and short-run components, say $M_{t-1} = \mu^{\ell}_{t-1} + \mu^{s}_{t-1}$, these authors show the second stage estimator $\widehat{\theta_M}$ converges to:

$$plim\left(\widehat{\theta_M}\right) = \theta\left(\frac{\sigma^2_{\mu_\ell}}{\sigma^2_{\mu_\ell} + \sigma^2_{\mu_s}}\right) \tag{8}$$

The term in parentheses is the fraction of the variation in input M that is due to long-run effects—that is, it is the long-run persistence of input M. Following these authors, we alternatively use a measure of value-added for the student's teacher in year $t-1$ and $t-2$ as our instrument M. Importantly, these value-added measures for student i are calculated using all years *other* than the one in which student i was in teacher j's class. We then estimate persistence as the ratio of $\widehat{\theta_M}$ to $\widehat{\theta_{IV}}$ (where the latter is approximately 1).

Jacob et al. elaborate on the conditions required for M_{t-1} to be a valid instrument. Chief among these is the requirement that assignment to teachers in $t-1$ be uncorrelated with current unobserved inputs (say, through dynamic tracking). This is difficult to test in practice, but Kane and Staiger (2008) find that persistence estimated under non-experimental conditions was very similar to that found under random assignment of students to teachers. Jacob et al. also derive (8) under more complicated assumptions about the correlation

between teachers' impact on long- and short-run knowledge, and show that if anything the estimator in (8) will *overestimate* persistence.

We use the method described above to estimate long-run persistence of teacher effects one and two grades into the future, separately for each test. If the effectiveness of the prior teacher as measured by the high-stakes test reflects comparatively greater emphasis on short-run skills (e.g. teaching to the test), we would expect less persistence of this input into future outcomes than the same measure from the low-stakes test.

One potentially important threat to this strategy is the influence of differential scaling and test content. For example, if the TAAS/TAKS 4th and 5th grade math tests were comprised of very different skills (e.g. geometry versus statistics), one might not expect strong persistence of 4th grade teacher effects on 5th grade outcomes. In contrast, if the low-stakes tests were more consistent from grade to grade then persistence should be higher. In this scenario, differences in teacher effect persistence across tests may not necessarily have anything to do with emphasis on short- versus long-run skills. We address this in two ways. First, we examine the persistence of teacher effects *as measured on the low-stakes test* onto *high-stakes* test outcomes, and vice versa. If the test's scaling and/or content across grades were driving differences in persistence, then one would not expect to see stronger persistence of low-stakes value-added onto high-stakes outcomes. Second, we estimate models that control for prior achievement as measured on the opposite test, to account for difference in test content across grades and test batteries.

5 Results

5.1 Variation in teacher effects on high- and low-stakes tests

We begin by describing the overall magnitude of teacher effects on each test, using their estimated standard deviation (s.d.) as a summary measure of their contribution to differences in student achievement (e.g., Rivkin, Hanushek, and Kain, 2005). Table 3 reports the s.d. of teacher effects under a number of different model specifications and student subsamples. Panel A reports the s.d. of stable teacher effects $\widehat{\delta_j}$, which are adjusted for sampling error and rely on all available years of data for each teacher j, while Panel B reports the s.d. for teacher-by-year effects $\widehat{\phi_{jt}}$. In all cases we rely only on estimates from students with scores reported for *both* the high- and low-stakes tests. As all scores have been normalized to mean zero and s.d. one, these effects are expressed in s.d. units of achievement.

Consistent with previous research, we find large effects of 4th and 5th grade teachers on achievement in both reading and math. These effects—which are on the high end of existing estimates—are present on both tests. Based on our baseline TAAS/TAKS model, we estimate that a one s.d. increase in teacher effectiveness is associated with a 0.205 s.d. increase in reading achievement and a 0.256 s.d. increase in math.

We observe, however, that the overall magnitude of teacher effects varies with the test. As seen in Table 3, there is uniformly greater variation in teacher effects on the high-stakes test than on the low-stakes test of the same subject. Across specifications in Panel A, teacher effects on the high-stakes reading test are 18 to 31% larger than those on the corresponding low-stakes test. Those on the high-stakes math test are 15 to 26% larger. In the baseline SAT model, we estimate that a one s.d. increase in teacher effectiveness is associated with a

0.169 s.d. increase in reading achievement, and a 0.218 s.d. increase in math.[16] The relative magnitude of the teacher-by-year effects (Panel B) is a slightly more modest.

As the state test changed when the TAKS was adopted in 2003, we were concerned that pooling TAAS and TAKS years might inflate the variation in teacher effects relative to what they would be if the test had remained stable over time. Rows (6) and (7) report the standard deviation in teacher effects separately for TAAS and TAKS years. These results are qualitatively similar, though the magnitude of teacher effects diverges most between tests in the TAAS years.[17]

If our estimates of teacher effects could be taken as causal effects on student achievement, the high- and low-stakes tests would offer somewhat different conclusions about the relative contribution of teachers to test scores. Using our baseline estimates of stable teacher effects, the difference in reading achievement between a student with a 25th percentile teacher and a 75th percentile teacher would be 0.248 s.d. on the high-stakes test. The corresponding difference on the low-stakes test is 0.198 s.d. While both are large effects, the SAT implies a 0.049 s.d. (or 20%) smaller impact of teacher quality on achievement.

5.2 Correlation in teacher effects across tests and subjects

Correlations between the two sets of teacher effect estimates indicate whether teachers deemed effective on the high-stakes test are similarly effective on the low-stakes test of the same subject. Any meaningful difference in teacher rankings that exists across the two tests are potentially important in light of proposals to reward the highest performing and

[16] When estimating teacher effects using the full population of SAT test-takers—not only those with both a TAAS/TAKS and a SAT score—the s.d. tends to be lower. For example, using our baseline model, the s.d. in teacher effects is 0.212 in math.

[17] As a sensitivity check, we also dropped observations from 2003 when the lag score for the TAKS would have been based on the TAAS. The results are very similar.

dismiss the lowest performing teachers based on their value-added (e.g., Gordon, Kane, and Staiger, 2006; Hanushek, 2009).

A scatter diagram plotting teachers' estimated value-added on the SAT reading test against their value-added on the TAAS/TAKS (Figure 1) indicates a moderate correlation between the two measures (r=0.499). Put another way, value-added on the high-stakes test is an imprecise predictor of value-added on the low-stakes test. The scatter diagram for math is comparable, though the correlation is higher (r=0.587).

Table 4 reports pairwise correlations for our estimated δ_j and ϕ_{jt} on the two sets of tests. For example, the first correlation coefficient in the upper lefthand corner of Panel A is the (Pearson) correlation between estimated teacher effects on the TAAS/TAKS reading test and those on the SAT reading test, using our baseline specification. In all cases the unit of observation is a teacher (Panel A) or teacher-year (Panel B). Sample sizes are reported in the third and fourth columns.

In reading, we find the correlation between stable teacher effects on the TAAS/TAKS and SAT range between 0.453 and 0.566, depending on the model and subsample.[18] In math, the correlation ranges between 0.560 and 0.616. Spearman rank correlations (not reported) are marginally higher. Unsurprisingly, these correlations—which make use of all available data and are less affected by year-to-year classroom noise—are stronger than the correlation in teacher-by-year effects, shown in Panel B. Here we observe correlations of 0.463 - 0.475 in reading and 0.528 - 0.542 in math. Interestingly, the correlation in teacher effects is stronger between subject areas on the same test than across tests of the same content areas, a finding consistent with differential teacher investments in tests. For example, on the TAAS/TAKS,

[18]In all cases, these correlations are weaker when estimating SAT teacher effects using the full sample of test takers. For example, the correlation in reading teacher effects based on the restricted sample (taking both tests) is 0.499, seen in Table 4. When using the full sample of SAT takers, this correlation drops to 0.485. The comparable numbers in math are 0.587 and 0.579.

the correlation between teacher effects in math and reading is 0.675. On the SAT, it is 0.625.

Both sets of correlations have implications for evaluating teacher effectiveness in practice. Teacher effects based on a single year's results may be used to award bonuses or identify teachers in need of improvement. Even when estimates from multiple years of data are preferred, a substantial share of teachers always have no more than a single year of results that can be used to estimate a teacher effect.[19] Given the correlations reported here, a teacher judged highly effective on the state test may be viewed differently when considering a low-stakes test of the same subject, even when those tests are administered to the same set of students at roughly the same time of year.

Figure 2 illustrates a policy implication of these results. For these graphs, we divide teachers into performance quintiles based on their stable teacher effect $\hat{\delta}_j$, separately for each test. We then show the percent of teachers in each quintile of the high-stakes teacher effect distribution (the horizontal axis) that ranked in the jth quintile of the low-stakes teacher effect distribution in the same subject (the bars). We find that 46% of teachers in the top quintile of effectiveness on the TAAS/TAKS reading test appear in the top quintile on the SAT reading test. More than 15% of these are in the bottom two quintiles on the SAT. The same asymmetry is observed for the bottom quintile of TAAS/TAKS teachers. Here 48% of bottom quintile reading teachers also appear in the bottom quintile of the SAT. One in eight (13%) ranked in the top two quintiles according to the SAT. A similar pattern is observed in math, though the quintile rankings are more consistent than in reading.

A comparison of quintile rankings for teacher-by-year effects ϕ_{jt} (not shown) produced qualitatively similar patterns, although as would be expected the rankings are less consistent. For example, 43% of those in the top quintile of effectiveness on the TAAS/TAKS reading

[19] In Houston, 14-22% of teachers had zero or one year of teaching experience when observed, depending on the year.

test were also in the top quintile on the SAT; 17% were in the bottom two quintiles.

Were one to set a threshold for "exceptionally low-" or "exceptionally high-performing" teachers based on value-added (such as the bottom 10% or top 5%), the percent of teachers who would meet this threshold on *both* tests of the same subject, or on all four tests, would be very small. For example, 40-42% of teachers who ranked in the bottom decile of one test also ranked in the bottom decile of the other test in the same subject (6 to 12% ranked above the median). 28% of those in the bottom 5% also ranked in the bottom 5% of the other test in the same subject. Almost no teachers (1.6%) ranked in the bottom decile of all four tests and only 17 of 3,677 teachers ranked in the bottom 5% of all four tests.

To summarize, were teachers to be rewarded for their classroom's performance on the state test—or alternatively, sanctioned for low performance—many of these teachers would have demonstrated different results on a low-stakes test of the same subject. Importantly, these differences need not be due to real differences in long-run skill acquisition, a point we return to in Section 5.4.

5.3 Returns to teacher experience on high- and low-stakes tests

As noted in Section 2, a consistent finding in the literature is that observed characteristics of teachers are only weakly correlated with teaching effectiveness. Rockoff (2004), for example, found that teachers' impact on math computation scores increased sharply in the first two years of teaching (0.1 s.d. units), but failed to rise much further. In reading comprehension, the marginal effect of accumulated experience remained positive for a longer period, but in vocabulary fell to zero after 3-4 years. In Table 5 and Figure 3 we report our estimates of the returns to teaching experience on high- and low-stakes tests in Houston.

In the first 6-7 years, we observe returns to experience that are similar to those found

in other work. On the high-stakes math test, for example, the biggest gains are immediate, occurring after the first year of teaching. We estimate the effect of a teacher with one year of experience is 0.089 s.d. greater than that of a teacher with no experience. This premium rises to 0.120 for teachers with two years, and 0.132 for teachers with four. After accumulating four years of experience, however, there appear to be no further gains.[20] The pattern is broadly similar on the SAT math test, with the biggest gains after year one (0.079 s.d.), and a plateau after four years (at 0.115 s.d.).

In reading, returns to experience are more gradual on both tests. On the TAAS/TAKS, the largest gains again are immediate (0.069 s.d. following year one). The premium continues to rise through year six, to a peak of 0.120. On the SAT, returns to experience in reading increase monotonically through year six, rising from 0.030 in the first year to 0.077.

For the more experienced teachers—those with more than seven years—the estimated returns diverge substantially, particularly in math. As seen in Figure 3, we observe a marked decline in effectiveness on the high-stakes math test. While the effect of a 7th year teacher is estimated to be 0.138 s.d. greater than that of a teacher with no experience, this premium falls nearly in half in later years, to 0.111 in years 8-10, and 0.073 in years 21 and higher.[21] Interestingly, this pattern is not observed on the low-stakes math test. Here the experience premium remains relatively constant at 0.115. If anything, teachers with 21 or more years of experience have the greatest differential over novices (at 0.131 s.d.).

A similar story holds in reading, if a less dramatic one. Here the experience premium

[20]The confidence intervals around the point estimates in Table 5 are wide enough that we cannot reject the hypothesis that these differentials are equal, though this is true in nearly all studies of returns to teacher experience.

[21]Note these are not within-teacher estimates of the returns to experience. Thus we cannot infer that teachers' effectiveness in math deteriorates over time. Rather, these differences may reflect selection or "vintage" effects. That is, teachers who are less effective in math may be those most likely to remain in teaching.

on the low-stakes SAT continues to rise even after year seven. We estimate the effect of a teacher in years 8-10 to be 0.071 s.d. above that of a first year teacher, a differential that rises to 0.081 in years 11-15, and a peak of 0.094 in years 16-20. In contrast, this pattern is not observed on the high-stakes test, where effectiveness appears to decline modestly after year seven. Consistent with almost all other research on the subject, we find no evidence of greater effectiveness among teachers with a MA degree, on any test. As seen in Table 5, the coefficient for teachers with a MA or higher is negative and statistically significant in 3 of the 4 tests, and is largest in math (-0.026 s.d.).

It is difficult to provide a causal explanation for the differences in returns to experience we observe across tests. However, the pattern is consistent with incentives facing teachers in Houston. While Texas does not offer tenure in the traditional sense, HISD's employment contract and performance evaluation varies with experience. New classroom teachers are hired under a probationary contract, renewable annually for two years.[22] After their third year, teachers may be offered a term contract, which automatically renews every 1-3 years. Probationary teachers are appraised using a thorough review protocol ("PDAS") that rates teachers along eight domains. Term teachers can opt for a modified review ("MPDAS"), in which teachers rated "proficient" or higher retain their scores from the prior year in 5 of the 8 domains, and waive a formal classroom observation. While none of the eight domains explicitly factor in classroom test scores, the review criteria strongly emphasize instruction aligned with the state standards, and recognize teachers' contribution to the school's TAKS test results.[23] Such an evaluation structure provides strong incentives for novice teachers to align their instruction with state standards, and invest early in skills and content specific to

[22] All details on HISD's employment policies can be found here: http://www.nctq.org/docs/32.pdf (last accessed March 16, 2011).

[23] In 2010, Houston adopted a policy in which value-added scores could be used to remove low performing teachers. See Ericka Mellon, "HISD moves ahead on dismissal policy," *Houston Chronicle*, January 14, 2010.

the high-stakes TAAS/TAKS test.

Another possible explanation is that estimated returns to experience on the high-stakes test are more sensitive to curriculum changes than those on the SAT. For example, a change in the state math curriculum could impact experienced teachers more than novices if experienced teachers were trained under the old standards. Our analysis—which spans both the TAAS and TAKS periods—may be particularly sensitive to this if experienced teachers were penalized by the test change. As a robustness check, we re-estimated our model from Table 5 separately by test period. Though the coefficient estimates were noisier, we found a qualitatively similar pattern as that shown in Figure 3 (in particular, the sharp drop in math effectiveness for experienced teachers appears on both the TAAS and TAKS).[24] The two notable differences were (1) a less immediate return in year one on the TAKS than the TAAS—perhaps because the TAKS was a new test with which existing teachers had no familiarity, and (2) a higher experience premium overall on the SAT in the post-2002 period.

Whatever the explanation, our results show that conclusions about the benefits of additional experience vary depending on the test. Were we to consider only the high-stakes test, for example, one might conclude that there are few benefits to retaining teachers beyond their 6th or 7th year (and in math, experienced teachers might even be deemed substantially less effective than novice teachers). The low-stakes test leads us to different inferences about the return to experience.

5.4 Persistence of teacher effects on high- and low-stakes tests

Finally, we use the empirical strategy outlined in Section 4.3 to compare the long-run persistence of teacher effects across tests. As shown in (8), persistence is constructed as the

[24] Results available upon request.

ratio of two estimates of the coefficient on lagged achievement: $\widehat{\theta_M}$, an IV estimator using a prior input as the instrument (here, a prior teacher's value-added), and $\widehat{\theta_{IV}}$, an estimator of the general rate of persistence in long-run skills, using twice-lagged achievement as the instrument. Our estimates of these parameters are reported in Tables 6 and 7, for persistence after one and two years, respectively.

In Table 6, each model regresses grade g achievement on lagged achievement, with a full set of student, classroom, and school covariates, and year dummies. Columns (1) and (5) report simple OLS estimates of the coefficient on lagged achievement (which are always less than one); (2) and (6) report $\widehat{\theta_{IV}}$ (which are much closer to one); and (3) and (7) report $\widehat{\theta_M}$. (We discuss columns (4) and (8) below). Persistence is calculated for the TAAS/TAKS as the ratio of columns (3)/(2) and as (7)/(6) for the SAT. Because twice-lagged achievement is required to estimate $\widehat{\theta_{IV}}$, our sample consists only of 5th graders with the necessary data.[25]

We find substantial differences in the persistence of teacher effects across tests. In both subjects, about 40% of the 4th grade teacher's value-added as measured by the high-stakes test carries over into the 5th grade. The rate of persistence of teacher effects as measured by the low-stakes test is closer to 60%. All of these estimates are statistically significant at conventional levels, as is a test for differences in these two estimates.

Table 7 reports the same estimates for persistence after two years. In this case our sample consists of 6th graders with the required data. These models are very similar to those used for Table 6, with a few exceptions. First, because we are interested in the carry-over of effects from two years prior, our coefficient of interest is that for twice-lagged achievement (i.e. 4th grade). Second, our instruments will be thrice-lagged achievement (3rd grade) for columns (2) and (6), and the 4th grade teacher's value-added in columns (3) and (7). Third, our

[25]Recall our first TAAS/TAKS test score is available in 3rd grade.

regression model must exclude classroom-level variables given that most 6th graders in this district are not in self-contained classrooms (we continue to include school-level controls).

Consistent with our findings in Table 6, we find substantial differences in the persistence of teacher effects after two years. In reading, about 32% of the 4th grade teacher's value-added as measured by the high-stakes test remains in the 6th grade, versus 53% of value-added based on the low-stakes test. In math, the comparable estimates are 28% and 46%.

As noted in Section 4.3, our persistence estimates may be sensitive to differences in test scaling or continuity of tested content across grade levels. For example, were skills tested on the high-stakes 5th grade test very different from those on the 4th grade test, one would be less likely to observe persistence of 4th grade teacher inputs onto 5th grade outcomes. In our context, if there is greater continuity on the sequence of low-stakes tests than high-stakes tests, persistence would necssarily be higher on the low-stakes tests. To address this possibility, columns (4) and (8) of Tables 6-7 report estimates of $\widehat{\theta_M}$ using the prior teacher's value-added as measured by the *opposite* test. That is, our high-stakes models in column (4) use teacher effectiveness on the low stakes test as the instrument, and the low-stakes models in column (8) use value-added on the high-stakes test as the instrument.

We find that a teacher's value-added on the low-stakes test has greater persistence on high-stakes achievement tests than does the teacher's value-added on the high-stakes test itself. For example, about 50% of the 4th grade teacher's value-added on the low-stakes SAT reading test carries forward to the 5th grade TAAS/TAKS, versus 40% using the TAAS/TAKS measure itself (as reported above). Persistence of the high-stakes measure on the low-stakes outcome does not similarly out-perform that for the low-stakes measure. However, at a rate of 51%, its persistence is somewhat larger than on the high-stakes test, suggesting some role for differences in test continuity across grades.

The same pattern roughly holds after two years: the low-stakes value-added measure in reading and math has greater persistence on high-stakes outcomes than do the high-stakes value-added measures themselves. Finally, we note that our concerns over differential continuity in tested skills is likely to be of greatest concern in math, where discrete changes in test content are more common. While we do find a modestly lower persistence of teacher effects on the high-stakes math test, our results indicate that teacher effects on high-stakes tests are less persistent in both subjects.

6 Conclusion

"Value added" measures of teacher effectiveness are the centerpiece of a national movement to evaluate, promote, compensate, and dismiss teachers in part based on their students' test results. Federal, state, and local policymakers have adopted these methods in an attempt to objectively quantify teaching effectiveness and promote and retain teachers with a demonstrated record of success. In this paper, we used data from Houston to examine whether, how, and to what extent value-added measures are sensitive to the test instrument used, particularly when those tests vary in the stakes attached to them.

As was found in other research, we observed large effects of teachers on student achievement, and these effects are present on both types of assessments. However, we also find that these measures differ in interesting and important ways across tests. For one, the magnitude of teacher effects appears larger on the high-stakes test than on the low-stakes test, with the former suggesting an impact of teacher quality that is 15-31% larger than that suggested by the low-stakes tests. Second, these measures are moderately correlated at the individual teacher level. That is, teachers deemed top performers on the high-stakes test are often average or low performers on the low-stakes test. Third, the returns to teaching experience vary

across tests in ways that are consistent with incentives to perform well by the high-stakes measure early in one's career. And fourth, teacher effects on the high-stakes test are less persistent into later outcomes than those on the low-stakes test.

Our results do not necessarily suggest that one test is superior to the other for constructing value-added measures. Nor do they suggest that an estimate that combines results from the two tests would be an unambiguous improvement over a single test battery.[26] Rather, they highlight the need for additional research on the impact that high-stakes accountability has on the validity of inferences about teacher quality.

References

[1] Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics* 25: 95-135.

[2] Alexander, Karl, Doris R. Entwisle, and Linda S. Olsen. 2001. "Schools, Achievement, and Inequality: A Seasonal Perspective," *Educational Evaluation and Policy Analysis* 23: 171-191.

[3] Ballou, Dale. 2009. "Test Scaling and Value-Added Measurement," *Education Finance and Policy*, 4: 351-383.

[4] Ballou, Dale, William Sanders, and Paul Wright. 2004. "Controlling for Student Background in Value-Added Assessment of Teachers," *Journal of Educational and Behavioral Statistics*, 29: 37-65.

[5] Booher-Jennings, Jennifer. 2005. "Below the Bubble: Educational Triage and the Texas Accountability System." *American Educational Research Journal* 42: 231-268.

[6] Briggs, Derek C, and Jonathan P Weeks. 2009. "The Sensitivity of Value-Added Modeling to the Creation of a Vertical Score Scale." *Education Finance and Policy* 4: 384-414.

[7] Buddin, Richard, and Gema Zamarro. 2009. "Teacher Qualifications and Student Achievement in Urban Elementary Schools," *Journal of Urban Economics* 66: 103-115.

[26] Interestingly, after our sample period Houston combined TAAS/TAKS and SAT measures for use in its ASPIRE teacher performance pay plan, effectively placing higher stakes on the SAT.

[8] Center on Education Policy. 2008. *Has Student Achievement Increased Since 2002? State Test Score Trends Through 2006-07.* Washington, DC: Center on Education Policy.

[9] Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011a. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project STAR." *Quarterly Journal of Economics* 126(4): 1593-1660.

[10] Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2011b. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." *National Bureau of Economic Research Working Paper* #17699.

[11] Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *Journal of Human Resources* 41: 778-820.

[12] Cullen, Julie B. and Randall Rebeck. 2006. "Tinkering Towards Accolades: School Gaming Under a Performance Accountability System." In Timothy J. Gronberg and Dennis W. Jansen (eds.) *Improving School Accountability (Advances in Applied Microeconomics, Volume 14)* Emerald Group Publishing Limited, pp. 1-34.

[13] Darling-Hammond, Linda, and Alfred.E. Wise. 1985. "Beyond Standardization: State Standards and School Improvement." *The Elementary School Journal* 85: 315-36.

[14] Figlio, David N., and Lawrence Getzler. 2006. "Accountability, Ability and Disability: Gaming the System." In Timothy J. Gronberg and Dennis W. Jansen (eds.) *Improving School Accountability (Advances in Applied Microeconomics, Volume 14)* Emerald Group Publishing Limited, pp. 35-49.

[15] Figlio, David N. 2006. "Testing, Crime and Punishment." *Journal of Public Economics* 90: 837-851.

[16] Fuller, Bruce, Joseph Wright, Kathryn Gesicki, and Erin Kang. 2007. "Gauging Growth: How to Judge No Child Left Behind?" *Educational Researcher* 36: 268-78.

[17] Goldhaber, Dan. 2008. "Teachers Matter, but Effective Teacher Policies are Elusive," in *Handbook of Research in Education Finance and Policy*, Helen Ladd and Edward B. Fiske (eds.), Routledge.

[18] Goldhaber, Dan, and Michael L. Hansen. 2008. "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance," Center on Reinventing Public Education Working Paper #2008-5.

[19] Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." Washington, D.C.: Brookings Institution.

[20] Hanushek, Eric A. 2009. "Teacher Deselection," in *Creating a New Teaching Profession*, Dan Goldhaber and Jane Hannaway (eds.) Washington, D.C.: The Urban Institute Press.

[21] Hanushek, Eric A., John F. Kain, Daniel M. O'Brien, and Steven G. Rivkin. 2005. "The Market for Teacher Quality," *National Bureau of Economic Research Working Paper* #11154.

[22] Hanushek, Eric A. and Steven G. Rivkin. 2006. "Teacher Quality," in *Handbook of the Economics of Education*, E.A. Hanushek and F. Welch (eds.), Elsevier.

[23] Harris, Douglas, and Tim R. Sass. 2006. "Value-Added Models and the Measurement of Teacher Quality," Working Paper, Florida State University.

[24] Harris, Douglas N, and Tim R Sass. 2009. "What Makes for a Good Teacher and Who Can Tell?" Working Paper 30, National Center for Analysis of Longitudinal Data in Education Research.

[25] Holcombe, Rebecca, Jennifer Jennings, and Daniel Koretz. 2010. "Predictable Patterns that Facilitate Score Inflation: A Comparison of New York and Massachusetts." Working Paper, Harvard University.

[26] Jacob, Brian A. 2005. "Accountability, Incentives, and Behavior: Evidence from School Reform in Chicago." *Journal of Public Economics* 89(5-6): 761-796.

[27] Jacob, Brian A. 2007. "Test-based Accountability and Student Achievement: An Investigation of Differential Performance on NAEP and State Assessments." *National Bureau of Economic Research Working Paper* #12817.

[28] Jacob, Brian and Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education," *Journal of Labor Economics* 26: 101-136.

[29] Jacob, Brian A., Lars Lefgren, and David Sims. 2010. "The Persistence of Teacher-Induced Learning," *Journal of Human Resources* 45: 915-943.

[30] Jennings, Jennifer L. and Andrew A. Beveridge. 2009. "How Does Test Exemption Affect Schools' and Students' Academic Performance?" *Educational Evaluation and Policy Analysis* 31: 153-175.

[31] Jennings, Jennifer L., and Jonathan M. Bearak. 2010. "Do Educators Teach to the Test?" Paper presented at the Annual Meetings of the American Sociological Association, Atlanta.

[32] Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City," *Economics of Education Review* 27: 615-631.

[33] Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," National Bureau of Economic Research Working Paper #14607.

[34] Kinsler, Joshua. 2012. "Beyond Levels and Growth: Estimating Teacher Value-Added and its Persistence." *Journal of Human Resources* 47: 722-753.

[35] Klein, Stephen, Laura Hamilton, Dan McCaffrey, and Brian Stecher. 2000. *What Do Test Scores in Texas Tell Us?* Santa Monica, CA: RAND.

[36] Koedel, Cory and Julian Betts. 2010. "Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation." *Education Finance and Policy* 5: 54-81.

[37] Koretz, Daniel M., and Sheila I. Barron. 1998. *The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND.

[38] Lockwood, J.R., Daniel F. McCaffrey, Laura S. Hamilton, Brian M. Stecher, Vi-Nhuan Le, and Jos Felipe Martinez. 2007. "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures," *Journal of Educational Measurement*, 44: 47-67.

[39] McAdams, Donald R. 2000. *Fighting to Save Our Urban Schools...and Winning!* New York: Teachers College Press.

[40] McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates," *Education Finance and Policy* 4: 572-606.

[41] Neal, Derek and Diane Whitmore Schanzenbach. 2010. "Left Behind By Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92: 263-283.

[42] Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. "How Large Are Teacher Effects?," *Educational Evaluation and Policy Analysis*, 26: 237-257.

[43] Papay, John P. 2011. "Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures" *American Education Research Journal* 48: 163-193.

[44] Pedulla, Joseph J., Lisa M. Abrams, George F. Madaus, Michael K. Russell, Miguel A. Ramos, and Jing Miao. 2003. *Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from a National Survey of Teachers*. Boston: National Board on Educational Testing and Public Policy.

[45] Reback, Randall. 2008. "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." *Journal of Public Economics* 92: 1394-1415.

[46] Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73: 417-458.

[47] Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review, Papers and Proceedings of the American Economics Association.* 94: 247-252.

[48] Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125: 175-214.

[49] Sass, Tim R. 2008. "Policy Brief: The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy," Washington, D.C.: CALDER.

[50] Shepard, Lorrie A. 1988. "The Harm of Measurement-Driven Instruction." Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

[51] Shepard, Lorrie A. and K.D. Dougherty. 1991. "Effects of High-Stakes Testing on Instruction." In *The Effects of High Stakes Testing,* ed. Robert L. Linn, Symposium presented at the annual meetings of the American Education Research Association and the National Council of Measurement in Education, Chicago, IL.

[52] Wiswall, Matthew. 2011. "The Dynamics of Teacher Quality." Working Paper, New York University Department of Economics.

A Online Data Appendix: HISD Longitudinal Data

We constructed our panel using source files provided by the Houston Independent School District (HISD), described below. The original data span 1997-98 through 2006-07, but for reasons we cite below we limited our panel to the 1998-99 through 2005-06 school years. To simplify notation, we hereafter refer only to the *spring* of the school year.

A.1 Student achievement measures

Between 1999 and 2006, HISD administered two standardized tests: the required Texas state assessments (TAAS/TAKS) and the Stanford Achievement Test (SAT). The TAAS (Texas Assessment of Academic Skills) was a minimum competency test given annually to students in grades 3-8 until 2003, when it was replaced by the TAKS.[27] TAAS includes math and reading tests in grades 3-8, a writing test in grade 4, and writing, science, and social studies tests in grade 8. The TAKS (Texas Assessment of Knowledge and Skills) is a standards-referenced exam given to students in grades 3-11. It includes math and reading/ELA tests in grades 3-11, writing tests in grades 4 and 7, a science test in grade 5, and science and social studies tests in grades 5, 8, 10, and 11. Spanish language versions of the TAAS and TAKS were available for grades 3-5 reading and math, grade 4 writing, and grade 5 science. An accommodated version is given to eligible students with special testing needs.

Under pressure from a local business task force that sought a nationally-normed benchmark test, HISD introduced the SAT-9 in 1996 (McAdams, 2000). The 10th edition (SAT-10) was adopted in 2004. The SAT is a battery of tests given to all eligible students in grades 1-11. The subject areas tested include reading, mathematics, language, science, and social science. All students receiving instruction in English, except those with serious disabilities, are required to take the SAT (Jennings and Beveridge, 2009). LEP students in grades 1-9 who receive reading and language arts instruction in Spanish are given the Aprenda, intended to be the Spanish language equivalent of the SAT. The TAAS, TAKS, and SAT-10 are untimed tests, while the SAT-9 had a recommended time limit. During this period, all three tests were administered in the Spring, from early March (SAT/Aprenda) to mid to late April (TAAS/TAKS).

The TAAS/TAKS is HISD's "high-stakes" test, for several reasons. First, passing rates on these tests have been an integral part of Texas' accountability system for years (Reback, 2008). Under this system–which served as the model for *No Child Left Behind*–schools and districts are labeled "exemplary," "recognized," "acceptable," or "low-performing" based on their pass rates in each subject area. In most of these years, monetary rewards were available for high-performing or improving schools, while low-performers were subject to sanctions, up to and including reconstitution and school closure. Second, HISD has operated a performance

[27] For a discussion of the differences between TAAS and TAKS see Jennings and Beveridge (2009) and Koedel and Betts (2009).

pay plan since 2000 that provides monetary rewards to schools and teachers for TAAS/TAKS results. Historically the district based these rewards on campus accountability ratings, but in recent years it has rewarded individual teachers based on their value-added to these tests. Third, Texas has required 3rd grade students to pass the TAKS reading test for grade promotion since 2003. From 2005, 5th grade students have been required to pass both the math and reading TAKS to be promoted.[28]

The SAT can be considered HISD's "low-stakes" test in that it is not tied to the state accountability system. However, the test plays several important roles in the district. For example, it is used as one criteria for grade promotion in grades 1-8. HISD students are expected to perform above a minimum standard on the SAT (e.g. at least one grade level below average) and the TAKS. In addition, the SAT is used to aid in the placement of students in specific programs, including gifted and special education. School-level results on the SAT are publicly reported in the local media, and in recent years value-added measures on the SAT were integrated into HISD's performance pay plan.

This first step in constructing our panel was compiling records for all students that were enrolled in grades 3-5 at the time of the TAAS/TAKS and/or SAT test, about 17,000 students per grade per year. 98 to 99 percent of these students had a test *record*, indicating they were enrolled at the time of the test. Not all, however, have a test score reported. Approximately 85-92 percent have a TAAS/TAKS score, depending on the subject, grade, and year. The 8-15 percent of students without scores were either exempted from the test, given an alternative assessment (such as the SDAA), or were absent on the day of the test. In contrast, nearly all students (98-99%) have a SAT score reported. This conforms to HISD's expectation that virtually all students be given this test.

As noted above, eligible students were permitted to take Spanish language versions of the TAAS/TAKS and the Aprenda. About 30 percent of all 3rd graders did so, dropping to 20% in 4th grade, and under 10% in 5th grade.

A.2 Student demographics and program participation

Using a 10-digit unique student identifier, we matched the above panel dataset to the student-level PEIMS (Public Education Information Management System). PEIMS data is recorded on or around October 30, and includes the following for each student: grade level (PK-12), 3-digit campus ID (school), date of birth, gender, race/ethnicity, LEP, ESL, and bilingual status, exceptionalities (special education and gifted), and immigrant and migrant status. Students from economically disadvantaged families are flagged, as are students deemed "at risk." In most grades and years, 95 percent or more of the students in our panel were matched to PEIMS data.

[28]Re-tests are offered in April and June for students who fail (and, since 2004, for students who were absent). For promotion grades and subjects, we use only the *first* observed score.

In the PEIMS, race/ethnicity includes five categories: American Indian or Alaskan Native, Asian or Pacific Islander, Black (not of Hispanic origin), Hispanic, and White (not of Hispanic origin). Indicator variables were created for each. The "LEP" variable identifies students classified as limited English proficient; "ESL" flags students participating in an intensive English instruction program; and "bilingual" designates participation in a full-time state-approved bilingual education program.

Texas uses a more inclusive measure of student poverty than most states. Students designated as "economically disadvantaged" (a) qualify for free or reduced price lunch, (b) are members of families that qualify for AFDC, or (c) fall into the "other economic disadvantaged" category. The latter includes families whose income qualifies them for free or reduced price lunch but who did not complete an application. Students deemed "at risk" in the PEIMS exhibit one or more of 13 criteria specified in Texas Education Code Section 29.081. These include: failure to advance from one grade to the next, failure to perform satisfactorily on the state assessments, expulsion, limited English proficiency, and homelessness.

A.3 Classroom identifiers and teacher characteristics

HISD provided a "crosswalk" designed to link students to classroom teachers via their 10-digit ID, campus, grade, and year. Nearly 100 percent of students were successfully linked to a classroom using this file. Teachers are identified with a unique 3-digit ID number within schools. Unfortunately, these IDs are not consistent over time. Teachers are assigned new numbers when moving between schools, and in some cases they change within the same school over time. The latter may reflect a change in assignment within the school; for example, a change from #410 to 510 may indicate a move from 4th to 5th grade.

The 3-digit teacher ID, campus, and year were sufficient to merge in teacher characteristics from the HISD personnel files. In most years these files included the following for each teacher: full name, gender, race/ethnicity, highest degree attained, certification (about 240 unique codes, with many teachers having multiple certifications), and total years of professional experience. Race/ethnicity includes five categories: Black, White, Hispanic, Asian, and Other.

Classroom IDs are sufficient for one-year teacher effect estimates, but we wanted the ability to match teachers over multiple years. Thus, to create a consistent teacher ID across all years, we used the following procedure. First, we stacked the teacher personnel files over all years (1998-2007) and assigned a tentative ID number to every distinct first and last name combination (27,467 unique combinations out of 138,447 total observations). This tentative ID is subject to one of two types of errors: (a) assigning a teacher more than one ID when she changes her name or her name is misspelled in some year, or (b) assigning two different teachers the same ID if they have the same name. Errors of the first type are not easily identified, but are likely to be rare. To address cases of multiple teachers with the same name, we did the following, in sequence:

1. Identified cases where a teacher name combination was observed multiple times in the same year with different values for experience (3,756 instances). Among these, assigned a new ID for teachers with distinct middle initials and ethnicity, which together with variation in experience we assumed was sufficient for identifying unique teachers (reduces the above to 567).

2. Identified cases where a teacher name combination was observed multiple times in the same year on different campuses (731 instances). Among these cases, assigned a new ID for teachers who had distinct middle initials and ethnicity, which together with variation in campus we assumed was sufficient for identifying unique teachers (reduces the above to 602). It is possible that some of the remaining teachers were assigned to more than one school.

3. Flagged cases where a teacher name combination was observed with different values for race/ethnicity (1,765 instances). Inspected these by hand (below).

4. Flagged cases where a teacher name combination was observed multiple times in the same year, in the same campus (30,704 cases). Inspected these by hand (below). It is likely that in many of these cases a teacher taught multiple sections, or a mixed grade classroom, with separate IDs for each grade.

Cases left unresolved by this procedure were inspected by hand and assigned new teacher IDs as needed. There were a small set of observations (1,286) involving duplicate teacher IDs in the same campus and year with *different names*. Why this is the case was not immediately obvious. Two hypotheses include: (a) teachers who leave during the year are replaced by a new teacher who is assigned the same classroom ID. A large number of these cases involved teachers with 30+ years of experience, who may have retired during the year, and (b) team-taught classes. For now, we have dropped duplicate teacher IDs in the same campus and year, keeping the most experienced teacher of the duplicates.

Altogether, about 90 percent of students were successfully matched to their classroom teacher's characteristics, in all years but 2007. In 2007, HISD moved to a 5-digit unique teacher identifier designed to be consistent across years. The 2006 personnel file included a crosswalk to the new (2007) teacher ID, for those teachers present in both years. Unfortunately, the source file for 2007 is missing a large number of teacher IDs, and only about 50 to 57 percent of students could be matched to teachers in that year. Thus we take 2006 to be the last year of our panel.

A.4 Final steps

Taken together, our panel includes student achievement, demographics, classroom identifiers, and teacher characteristics for all HISD students in 3rd, 4th, and 5th grade during the 1999 to

2006 school years. In rare cases where PEIMS data was missing, we attempted to impute this data using PEIMS information from other years for the same student. Missing race/ethnicity could also be imputed using a separate race variable in the Stanford test files.

TAAS/TAKS and Stanford scale scores were standardized within subject, grade, year, and test version (English or Spanish). Lagged standardized scores were created for 4th and 5th grade students.

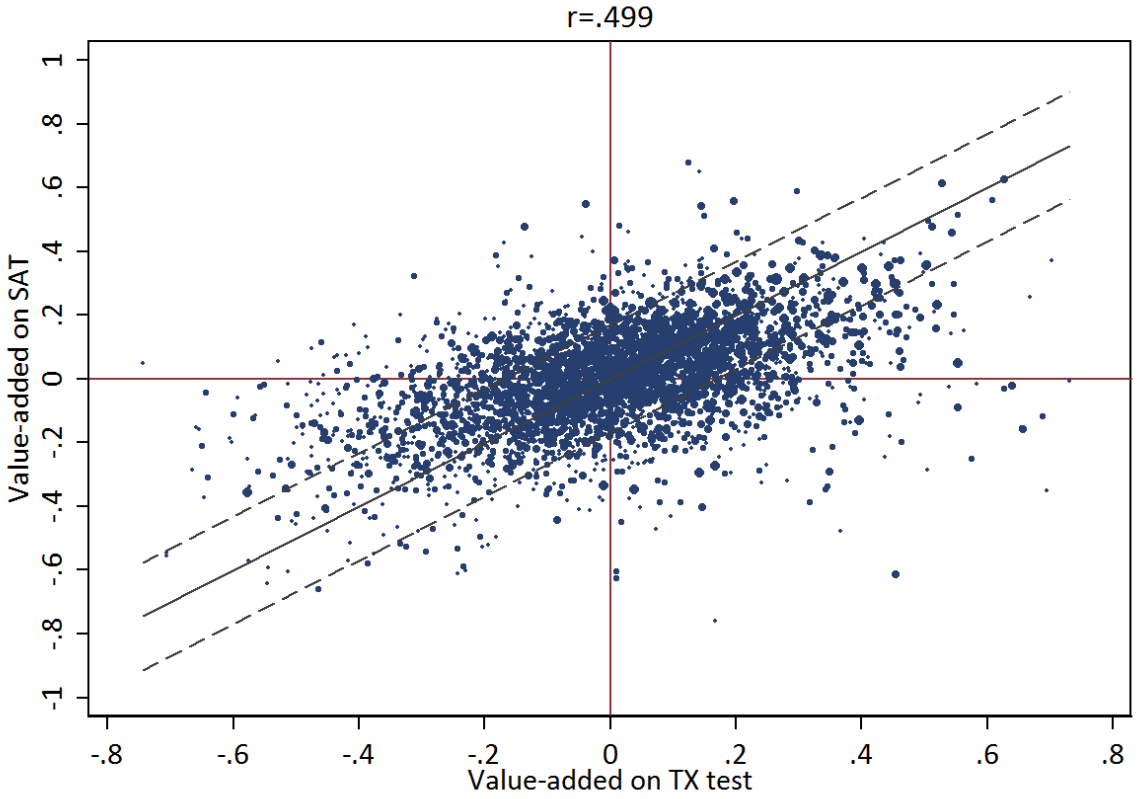

Figure 1: Correlation in stable teacher effects: reading

Notes: estimates from baseline model for stable teacher effects (using all years of available data) shown in equations (2) and (3), and using only students with both TAAS/TAKS and SAT reading scores. Scatterplot omits teacher effects greater than one in absolute value (n=16). Center diagonal is the 45-degree line, while the dotted diagonals are 1 s.d. above and below the 45-degree line (based on the distribution of SAT teacher effects).

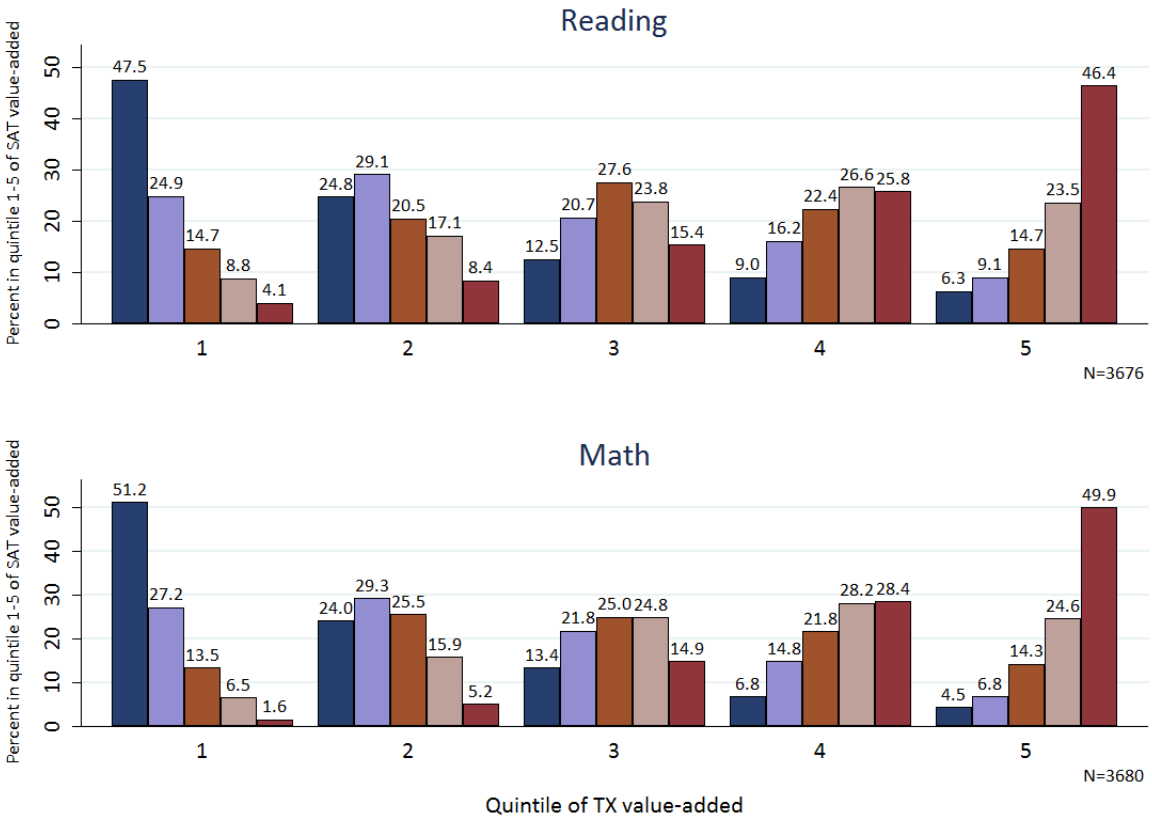

Figure 2: Quintile rankings of stable teacher effects: TAAS/TAKS and SAT

Notes: from baseline model for stable teacher effects (using all years of available data) shown in equations (2) and (3), and using only students with both TAAS/TAKS and SAT scores in the given subject.

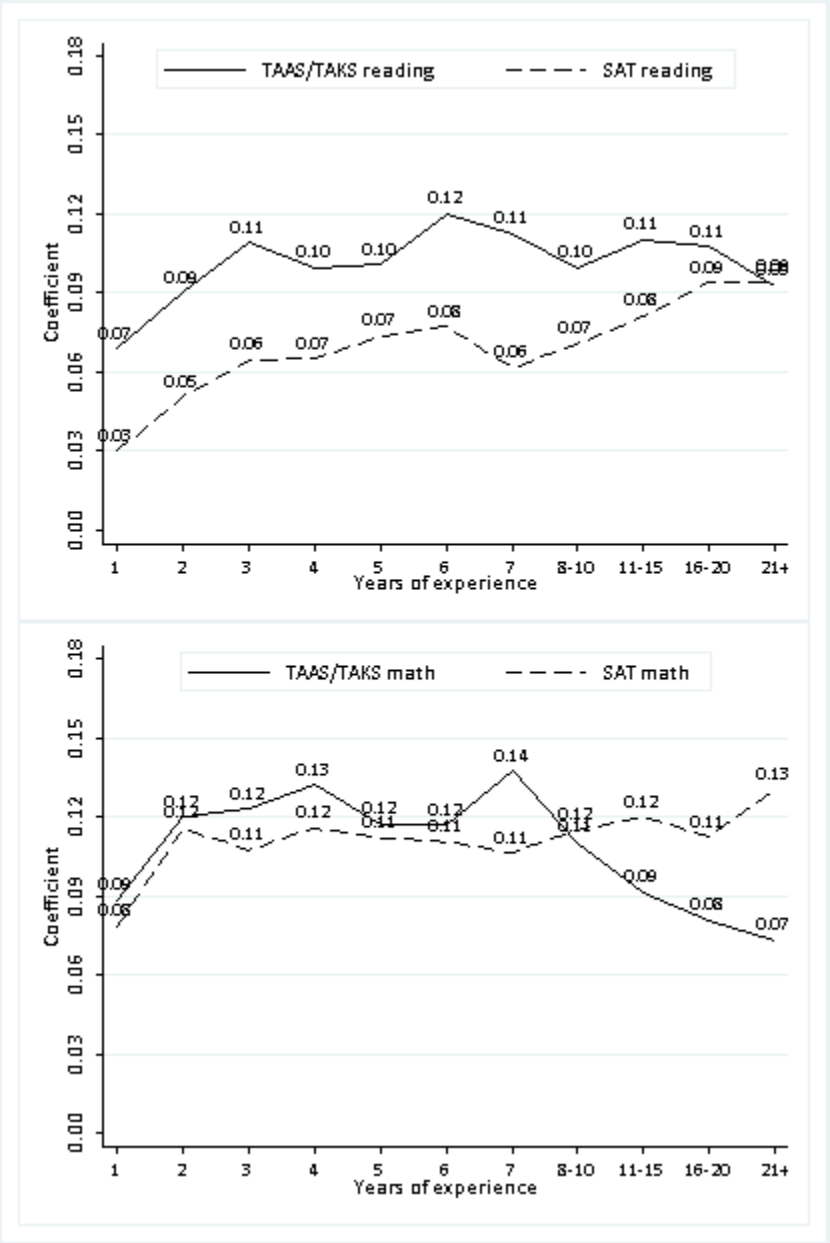

Figure 3: Returns to teacher experience: TAAS/TAKS and SAT

Notes: coefficient estimates from Table 5.

Table 1: HISD students contributing to teacher effect estimates, 1999-2006

A. Percent of all students:	Reading	Math
TAAS/TAKS	66.3	67.3
SAT	72.7	72.7
Both tests in subject	65.5	66.5

B. Mean z-scores:	Reading	Math
TAAS/TAKS	0.131	0.154
SAT	0.152	0.164
SAT: conditional on having both tests in subject	0.262	0.254
SAT: conditional on having this or any test	0.151	0.162
Percent took Spanish version of TAAS/TAKS	13.2	13.1
Percent took Spanish version of SAT (Aprenda)	13.1	13.1

C. Descriptive statistics:	Mean	SD
Age	10.746	0.766
LEP	0.308	0.461
Special education	0.088	0.283
Immigrant	0.043	0.202
Migrant	0.007	0.083
Economically disadvantaged	0.805	0.396
Black	0.282	0.450
Hispanic	0.580	0.494
Asian	0.031	0.172
White	0.106	0.308
Female	0.509	0.500
Changed school	0.133	0.340
Class size	21.627	4.251
Percent of students special ed in classroom	0.086	0.069
Mixed 4th and 5th grade class	0.024	0.153
Grade 5	0.492	0.500

Notes: percentages in (A) based on 253,408 students enrolled in 4th or 5th grade in the Spring of 1999-2006. Students contributing to teacher effect estimates have nonmissing current and lagged test scores in a given test and subject, a teacher ID, a teacher with at least 7 students tested in the given test and subject (over all years), and are in a classroom with no more than 25% special education students. N=186,508 in (C).

Table 2: Summary statistics: HISD teachers 1999-2006

	Grade 4	Grade 5	All
Years of experience	10.150	11.240	10.652
None	0.103	0.087	0.096
1 year	0.092	0.078	0.085
2 years	0.078	0.069	0.074
3 years	0.065	0.068	0.066
4 years	0.055	0.058	0.057
5 years	0.044	0.048	0.045
6 years	0.042	0.040	0.041
7 years	0.038	0.034	0.035
8-10 years	0.097	0.090	0.093
11-15 years	0.127	0.121	0.125
16-20 years	0.088	0.093	0.090
21 or more years	0.172	0.214	0.191
Masters degree or higher	0.225	0.245	0.232
Black	0.345	0.369	0.362
Hispanic	0.292	0.236	0.257
Asian	0.024	0.025	0.024
White	0.337	0.365	0.354
Students over all years	57.904	69.041	59.057
Number of years observed	2.992	3.194	2.892
N teachers	2,416	1,901	3,726
N teacher years	5,749	4,559	10,061

Notes: includes teachers for whom a teacher effect on any subject/test can be calculated. These teachers have at least 7 students tested in a given subject and test over all years, that are not in a classroom with more than 25% special education students. Mean experience based on teacher-by-year observations, while other means based on unique teachers.

Table 3: Standard deviations in teacher effects: TAAS/TAKS and SAT

	TAAS/TAKS Reading	SAT Reading	TAAS/TAKS Math	SAT Math
A. Stable teacher effects				
(1) Baseline specification	0.205	0.169	0.256	0.218
(2) w/o school covariates	0.229	0.173	0.284	0.227
(3) w/school effects	0.214	0.163	0.268	0.213
(4) Grade 4 only	0.212	0.178	0.267	0.222
(5) Grade 5 only	0.211	0.169	0.257	0.223
(6) TAAS years only	0.231	0.184	0.277	0.227
(7) TAKS years only	0.197	0.162	0.267	0.228
(8) Fixed effects (unshrunken)	0.284	0.241	0.317	0.274
(9) Fixed effects (shrunken)	0.226	0.199	0.269	0.232
B. Teacher-by-year effects				
(1) Baseline specification	0.233	0.195	0.291	0.254
(2) w/o school covariates	0.258	0.200	0.320	0.264
(3) w/school effects	0.241	0.189	0.300	0.246

Notes: With the exception of lines (8)-(9), the above table reports standard deviations of the best linear unbiased predictors (BLUPs) of the teacher effects. The baseline model (1) is that shown in equations (2) and (3). Specification (2) is the same model with $\bar{X}^s_{jst}$ omitted. Specification (3) replaces $\bar{X}^s_{jst}$ with school fixed effects θ_s, while specifications (4)-(7) estimate the baseline model on various subsamples: students in grade 4 only, grade 5 only, tested during the TAAS years only (1998-2002), and tested during the TAKS years only (2003-2006). Specifications (8)-(9) estimate teacher effects as fixed, rather than random, effects.

Table 4: Pairwise correlation of teacher effects: TAAS/TAKS and SAT

	Reading	Math	N(r)	N(m)
A. Stable teacher effects				
(1) Baseline specification	0.499	0.587	3,676	3,680
(2) w/o school covariates	0.521	0.616	3,677	3,681
(3) w/school effects	0.508	0.603	3,677	3,681
(4) Grade 4 only	0.520	0.581	2,320	2,323
(5) Grade 5 only	0.473	0.580	1,821	1,825
(6) TAAS years only	0.453	0.560	2,333	2,336
(7) TAKS years only	0.566	0.603	2,428	2,430
(8) Teacher FE (not RE)	0.514	0.585	3,676	3,680
SAT cross-subject: baseline	0.625	-	3,720	-
TX cross-subject: baseline	0.675	-	3,679	-
B. Teacher-by-year effects				
(1) Baseline specification	0.463	0.528	9,799	9,815
(2) w/o school covariates	0.475	0.545	9,806	9,822
(3) w/school effects	0.473	0.542	9,806	9,822
SAT cross-subject: baseline	0.587	-	9,945	-
TX cross-subject: baseline	0.622	-	9,801	-

Notes: With the exception of line (8), the above table reports correlations of the best linear unbiased predictors (BLUPs) of the teacher effects. The baseline model (1) is that shown in equations (2) and (3). Specification (2) is the same model with $\bar{X}^s_{jst}$ omitted. Specification (3) replaces $\bar{X}^s_{jst}$ with school fixed effects θ_s, while specifications (4)-(7) estimate the baseline model on various subsamples: students in grade 4 only, grade 5 only, tested during the TAAS years only (1998-2002), and tested during the TAKS years only (2003-2006). Specification (8) estimates teacher effects as fixed, rather than random, effects.

Table 5: Returns to teaching experience: TAAS/TAKS and SAT

	TAAS/TAKS Reading	SAT Reading	TAAS/TAKS Math	SAT Math
Years of experience:				
1	0.0688	0.0298	0.0883	0.0786
	(0.0151)	(0.0136)	(0.0171)	(0.0156)
2	0.0900	0.0505	0.1201	0.1156
	(0.0161)	(0.0132)	(0.0181)	(0.0159)
3	0.1091	0.0644	0.1230	0.1075
	(0.0167)	(0.0138)	(0.0192)	(0.0166)
4	0.0993	0.0651	0.1320	0.1156
	(0.0177)	(0.0151)	(0.0192)	(0.0174)
5	0.1008	0.0736	0.1172	0.1119
	(0.0182)	(0.0148)	(0.0208)	(0.0186)
6	0.1200	0.0772	0.1168	0.1106
	(0.0198)	(0.0158)	(0.0218)	(0.0187)
7	0.1125	0.0615	0.1375	0.1063
	(0.0199)	(0.0161)	(0.0228)	(0.0189)
8-10	0.0995	0.0705	0.1105	0.1150
	(0.0155)	(0.0130)	(0.0172)	(0.0146)
11-15	0.1101	0.0811	0.0916	0.1200
	(0.0142)	(0.0120)	(0.0160)	(0.0137)
16-20	0.1076	0.0938	0.0809	0.1124
	(0.0155)	(0.0129)	(0.0176)	(0.0148)
21 or more	0.0926	0.0940	0.0732	0.1306
	(0.0138)	(0.0115)	(0.0154)	(0.0134)
MA or higher	-0.0063	-0.0118	-0.0247	-0.0261
	(0.0076)	(0.0063)	(0.0087)	(0.0076)
Student covariates	YES	YES	YES	YES
Class characteristics	YES	YES	YES	YES
School characteristics	YES	YES	YES	YES
School fixed effects	YES	YES	YES	YES
Observations	166,775	164,520	167,563	165,305
R-squared	0.457	0.626	0.454	0.587

Notes: standard errors in parentheses clustered by classroom (teacher x year).

Table 6: Persistence of teacher effects: after one year

	TAAS/TAKS				SAT			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Reading								
Coefficient on	0.498	0.901	0.374	0.447	0.583	0.881	0.535	0.446
lagged achievement	(0.006)	(0.012)	(0.037)	(0.065)	(0.005)	(0.007)	(0.047)	(0.055)
1st stage F-statistic	-	788.0	352.3	321.3	-	1766.7	629.9	621.3
2nd stage R-squared	0.456	0.338	0.445	0.455	0.623	0.562	0.621	0.610
B. Math								
Coefficient on	0.523	0.876	0.347	0.412	0.644	0.928	0.568	0.472
lagged achievement	(0.006)	(0.011)	(0.030)	(0.041)	(0.005)	(0.008)	(0.032)	(0.037)
1st stage F-statistic	-	958.9	417.0	371.2	-	1600.6	496.9	467.2
2nd stage R-squared	0.429	0.331	0.405	0.420	0.574	0.515	0.570	0.552
Instrument	-	2nd lag	VA	VA:SAT	-	2nd lag	VA	VA:TX

Notes: all coefficients are from a model for current student achievement, with student, class, and school covariates, and year effects. Standard errors (in parentheses) clustered at the classroom level. N=57,686 in Panel (A) and N=58,625 in Panel (B). F-statistic $p<0.001$ in all first stage regressions.

Table 7: Persistence of teacher effects: after two years

	TAAS/TAKS				SAT			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Reading								
Coefficient on twice lagged achievement	0.456	0.877	0.289	0.424	0.534	0.823	0.433	0.431
	(0.004)	(0.011)	(0.027)	(0.041)	(0.004)	(0.007)	(0.027)	(0.035)
1st stage F-statistic	-	803.0	364.4	332.0	-	1882.3	693.0	673.5
2nd stage R-squared	0.444	0.317	0.424	0.444	0.605	0.546	0.598	0.597
B. Math								
Coefficient on twice lagged achievement	0.486	0.853	0.237	0.385	0.620	0.913	0.418	0.385
	(0.004)	(0.009)	(0.019)	(0.027)	(0.004)	(0.007)	(0.020)	(0.026)
1st stage F-statistic		938.0	388.9	333.8		1742.5	546.7	510.6
2nd stage R-squared	0.446	0.341	0.398	0.438	0.587	0.526	0.558	0.548
Instrument	-	3rd lag	VA	VA:SAT	-	3rd lag	VA	VA:TX

Notes: all coefficients are from a model for current student achievement, with student, class, and school covariates, and year effects. Standard errors (in parentheses) clustered at the classroom level. N=40,900 in Panel (A) and N=41,635 in Panel (B). F-statistic $p<0.001$ in all first stage regressions.