Abstract Title Page

Not included in page count.

Title: **Teacher Effectiveness on High- and Low-Stakes Tests**

Author(s): **Sean P. Corcoran, Jennifer L. Jennings, and Andrew A. Beveridge**

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

A large and growing literature demonstrates that teacher effects on academic achievement are substantial in size (Clotfelter, Ladd, and Vigdor, 2006; Rivkin, Hanushek, and Kain, 2005; Rowan, Correnti, and Miller, 2002). Moreover, this research finds that little of the variation in teacher effectiveness can be explained by observable characteristics such as certification, education, and experience (Aaronson, Barrow, and Sander, 2007; Hanushek, Kain, O'Brien, and Rivkin, 2005; Kane, Rockoff, and Staiger, 2008; Rockoff 2004). Motivated by these findings, policymakers have sought to require that teachers' evaluation, pay, and tenure be tied directly or indirectly to measures of their “value-added” to achievement on standardized tests.

Mounting evidence on school responses to test-based accountability, however, suggests that school behaviors can diminish the validity of test score gains associated with such systems. Following earlier analyses of trends on “high-stakes” tests (Klein et al., 2000; Koretz et al., 1991; Koretz and Barron, 1998; Linn, 2000), recent studies have concluded that the gains on these tests significantly outpace those on national benchmark tests such as the NAEP, with the gains in some cases almost four times as large (Center on Education Policy, 2008; Fuller et al., 2007; Koretz and Barron, 1998; Klein et al., 2000; Jacob, 2007). A potential explanation for these differences can be found in studies of strategic responses to test-based accountability. These studies address a wide range of activities that inflate perceptions of achievement gains, including the re-classification of students as requiring special education (Figlio and Getzler, 2002; Jacob, 2005), strategic exemption of students from testing (Cullen and Reback, 2006; Jacob, 2005; Jennings and Beveridge, 2009), re-allocation of resources toward students on the margin of passing (Booher-Jennings, 2005; Reback, 2008; Neal and Schanzenbach, 2007), suspension of low-scoring students near the test date (Figlio, 2005), and “teaching to the test” (Jacob 2005, 2007).

Perhaps surprisingly, these two bodies of research have remained largely separate. To date, research on teacher productivity has not investigated the potential impact of accountability systems on the validity of “high-stakes” test score gains as a primary measure of teacher effectiveness. Yet is plausible that teachers who appear effective on these tests may not be deemed similarly effective on a second, low-stakes test of the same subject, particularly when that test covers a broader and less predictable range of skills.

In this paper, we use data from the Houston Independent School District to estimate teacher effects on two different academic tests of the same subject areas, administered in the same school year to the same students at approximately the same time of year. The first is the statewide “high-stakes” test administered as part of the Texas accountability system, while the second is a nationally-normed “low-stakes” test intended as both an audit test and as a grade promotion tool.

Building on past work, we estimate the size of teacher effects on each of these tests, and examine how these effects relate to each other, and differentially persist over time.

Only three studies that we are aware of have considered how teacher effects vary across multiple achievement measures. In a paper similar to ours, Papay (forthcoming) estimated teacher effects in a large urban district on three different reading tests: the state accountability test, the Stanford Achievement Test, and SRI. He found weak to modest correlations in teacher effects across tests, ranging from 0.15 to 0.58, even when the tests were administered to the same sets of students. Papay carefully explored several hypotheses for these differences across tests and concluded that test timing played an important role (we discuss such mechanisms in the next section). Similarly, Lockwood et al. (2007) estimated teacher effects separately for the subscales of the Stanford math test. They found that differences in the choice of subscale measure have large effects on a teacher's effectiveness measure. Finally, in a paper not directly interested in teacher effects on different tests, Sass (2008) found a correlation of 0.48 between teacher effects estimated on the high- and low-stakes tests in Florida. Importantly, none of the papers cited here considered how the incentive effects of test-based accountability impact value-added measures of effectiveness.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

We use data from the Houston Independent School District to estimate teacher effects on two different academic tests of the same subject areas, administered in the same school year to the same students at approximately the same time of year. The first is the statewide “high-stakes” test administered as part of the Texas accountability system, while the second is a nationally-normed “low-stakes” test, intended as both an audit test and as a grade promotion tool. For reasons explained below, we focus on achievement in reading and math in the 4th and 5th grade.

Given these two effectiveness measures, we address the following questions: (1) Do these estimates of teacher effectiveness suggest a similar level of variation in quality across teachers? (2) How strongly are these two measures correlated? Is it the case that teachers who appear effective on a “high-stakes” state test are similarly effective on a “low-stakes” test of the same subject? (3) Is one measure of teacher effectiveness more stable from year to year than the other? (4) Are there differences in decay rates in teacher effects on high- and low-stakes tests? (5) To what extent does the high- and low-stakes nature of the test contribute to these differences?

Setting:

Description of the research location.

For this paper we drew from a longitudinal dataset of all students tested in the Houston Independent School District (HISD) between 1998 and 2006, approximately 165,000 per year. 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, while 29% are African-American, 8% are Caucasian, and 3% are Asian-American. Close to 80 percent of students are considered by the state to be economically disadvantaged, 27% are classified as Limited English Proficient, and 11% are classified as receiving special education.

HISD is an ideal setting for this study in that they administer multiple assessments each year to most students: the mandatory Texas state assessments (TAAS or TAKS) and the Stanford Achievement Test. 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 every year between grades 3 and 8. The Stanford Achievement Test is administered to all students grades 1 to 11 in reading, math, language, science, and social science.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features or characteristics.

Our interest in value-added measures on multiple tests restricts the data we are able to use. First, only grades and subjects where both the TAAS/TAKS and Stanford were given were considered (grades 3-8, reading and math). Second, the need for a lagged achievement measure on both tests 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 limited us to grade 4 and 5 students in self-contained classes. Fourth, students were excluded who were missing test scores—in most cases due to purposeful exclusion. Taken together, our analysis focuses on 4th and 5th grade students tested in math and reading on both tests, approximately 27,000 per year between 1999 and 2006. There are 2,100 to 2,600 unique teachers per grade represented in the analysis, depending on the grade and subject.

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.

Like all Texas school districts, HISD has administered annual state assessments (the TAAS or TAKS) since the 1980s. The TAAS was a minimum competency test given from 1991 until 2003, when it was replaced by the TAKS, a criterion-referenced test (Jennings and Beveridge, 2009; Koedel and Betts, 2009). In 1996, HISD added the Stanford Achievement Test under pressure from a business task force that sought a nationally-normed benchmark (McAdams, 2000). Since that time, all students have been required to take the Stanford.

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). Schools and districts are rewarded or punished based on these test results. Second, HISD has operated a performance pay plan since 2000 that provides monetary rewards to schools and teachers for TAAS/TAKS results. Third, Texas uses the TAKS to award grade promotion in grades 3 and 5.

The Stanford 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. In addition, the Stanford is used to aid in the placement of students in specific programs, including gifted and special education. School-level results are publicly reported in the local media, and in recent years value-added measures on the Stanford were integrated into HISD's performance pay plan.

Despite their disparate uses, reading and math skills covered on the TAKS and Stanford are broadly similar.

Research Design:
Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

This is a secondary analysis of longitudinal student-level achievement data from the Houston Independent School District, in which students are matched to their classroom teachers. Following earlier work on teacher value-added modeling, we estimate teacher effects on achievement using a covariate adjustment random effects model with prior year test scores included as a key explanatory variable. Additional details are provided in the next section.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

For this paper we constructed a longitudinal dataset of all students tested in the Houston Independent School District (HISD) between 1998 and 2006. Our analysis focuses on 4th and 5th grade students tested in math and/or reading on both the Texas state assessment (TAAS or TAKS) and the Stanford Achievement Test, approximately 27,000 students each year. All students are matched to demographic and program participation data, including age, gender, race/ethnicity, recent immigrant and migrant status, economic disadvantage, Limited English Proficiency, and special education status.

Following Gordon, Kane, and Staiger (2006), Kane, Staiger, and Rockoff (2008), Papay (forthcoming), Jacob and Lefgren (2008) and others, we estimate individual teacher effects on achievement using the following student-level model, separately for each test (TAAS/TAKS and Stanford) and subject (math and reading):

$$(1) \qquad Y_{ijt} = \beta_g X_{ijt} + \xi_g C_{jt} + \gamma_g S_{jt} + A_g W_{jt} + \pi_{gt} + u_{ijt}$$

where Y_{ijt} represents the score for student i in classroom j in year t. X_{ijt} is a vector of fixed and time-varying characteristics of student i, most importantly a cubic function of prior year achievement in both subjects on the same test series (TAAS/TAKS or Stanford). C_{jt} and S_{jt} are vectors of average classroom and school characteristics, and W_{jt} represents teacher characteristics including experience and highest degree. π_{gt} is a fixed grade-by-year effect. In some cases we include school fixed effects, and all coefficients are allowed to vary by grade.

We assume the error term u_{ijt} in (1)—the extent to which student i's test score differs from that predicted given her past score, individual, classroom, school, and teacher characteristics—can be decomposed into variation due to persistent teacher effectiveness δ_j and an unexplained component v_{it}:

$$(2) \qquad u_{ijt} = \delta_j + v_{it}$$

The parameters of interest are the δ_j, or the persistent "teacher effects." Although the δ_j can be estimated as fixed effects, we take the approach followed by Kane, Staiger, and Rockoff (2008) and others and treat these parameters as random effects, which we adjust for sampling variation using empirical Bayes shrinkage (Raudenbusch and Bryk, 2002; Jacob and Lefgren, 2008).

Our estimates of the δ_j make use of all available data for each teacher, which can include as many as 8 years of classroom data and 225 valid students. In examining properties of teacher effects such as inter-temporal stability and correlation with time-varying factors such as experience, we estimate a version of (1) using teacher-by-year (or classroom) effects for each subject and test. The resulting δ_j estimates are used to address the research questions outlined above.

Findings / Results:

Description of the main findings with specific details.

Our results indicate that teacher effects on the *high*-stakes test vary substantially more than those in the same subject on the low-stakes test (insert Figure 1 here). For the TAAS/TAKS, we find a standard deviation in teacher quality of 0.23 in reading and 0.28 in math. In contrast, the standard deviation is nearly half this size on the Stanford: 0.13 and 0.15, respectively. Teacher effects on different tests of the same subject in the same year are only modestly correlated, at 0.61 in math and 0.52 in reading. Figure 2 expresses this correlation another way, showing the proportion of teachers in each quintile of effectiveness on one test that ranked in quintiles 1-5 on the second test (insert Figure 2 here). As an illustration, we find only 48 percent of teachers in the top quintile of the TAKS math test were also in the top quintile of the Stanford test. A non-trivial share (13%) ranked among the *lowest* two quintiles of the Stanford. We find very little difference across the two tests in inter-temporal stability.

Perhaps more importantly, we find that teacher effects on the high-stakes TAKS test decay at a much faster rate than those on the low-stakes Stanford test. Using the method proposed by Jacob, Lefgren, and Sims (forthcoming), we estimated the persistence of teacher-induced gains on achievement in later grades and found that 34% of a teacher's effect on grade 4 mathematics carried through to grade 5, as measured by the Stanford test, while only 16% of her effect on achievement persisted as measured on the high-stakes TAKS test. The corresponding numbers in reading were 31% and 20%.

Finally, we find important differences in the impact of teacher observables on student performance across the two tests. The returns to teacher experience are compressed on the high-stakes test, such that the majority of the returns occur in the first 2 to 3 years. In contrast, we find positive returns to experience on the low-stakes reading test throughout the first 15 years of teachers' careers.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

If our estimates of teacher effects could be considered causal effects on student achievement, the high-stakes state assessments and low-stakes Stanford Achievement Test would offer very different evidence about the overall variation in teacher quality and the relative contribution of teachers to test outcomes. Moreover, differences in these sets of estimates have implications for value-added based systems in practice. Rewards and sanctions linked to student performance on one test may yield quite different results when applied to a different test of very similar content. More research is needed on the extent to which "high-stakes" testing alters teacher behavior (relative to a low-stakes test), such that value-added based estimates of effectiveness are compromised.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*, 95-135.

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

Center on Education Policy. (2008). Has student achievement increased since 2002? State test score trends through 2006-07.

Clotfelter, C.T., Ladd, H.F. & Vigdor, J.L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*, 778-820.

Cullen, J.B. & Reback, R. (2006). Tinkering towards accolades: School gaming under a performance accountability system. *National Bureau of Economic Research Working Paper* #12286.

Figlio, D.N. & Getzler, L. (2002). Accountability, ability and disability: Gaming the system. *National Bureau of Economic Research Working Paper* #9307.

Figlio, D.N. (2005). Testing, crime, and punishment. *National Bureau of Economic Research Working Paper* #11194.

Fuller, B., Wright, J., Gesicki, K. & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher*, *36*, 268-78.

Gordon, R., Kane, T.J. & Staiger, D.O. (2006). Identifying effective teachers using performance on the job. Washington, D.C.: Brookings Institution.

Hanushek, E.A., Kain, J.F., O'Brien, D.M. & Rivkin, S.G. (2005). The market for teacher quality, *National Bureau of Economic Research Working Paper* #11154.

Jacob, B.A. (2005). Accountability, incentives, and behavior: Evidence from school reform in Chicago. *Journal of Public Economics*, *89*(5-6), 761-796.

Jacob, B.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.

Jacob, B.A. & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26.

Jacob, B.A., Lefgren, L. & Sims, D. (forthcoming). The persistence of teacher-induced learning gains. *Journal of Human Resources*.

Jennings, J.L. & Beveridge, A.A. (2009). How does test exemption affect schools' and students' academic performance? *Educational Evaluation and Policy Analysis*, *31,* 153-175.

Kane, T.J., Rockoff, J.E., & Staiger, D.O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*, 615-631.

Klein, S., Hamilton, L., McCaffrey, D., & Stecher, B. (2000). What do test scores in Texas tell us? Santa Monica, CA: RAND.

Koedel, C. & Betts, J. (2009). Value-added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy*, *4*.

Koretz, D., Linn, R.L., Dunbar, S.B., & Shepard, L.A. (1991). The effects of high-stakes testing: Preliminary evidence about generalization across tests, in R. L. Linn (chair), *The effects of high stakes testing*, symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, IL.

Koretz, D.M. & Barron, S.I. (1998). The validity of gains on the Kentucky Instructional Results Information System (KIRIS). Santa Monica, CA: RAND.

Linn, R.L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Lockwood, J.R., McCaffrey, D.F., Hamilton, L.S., Stecher, B.M., Le, V.N., & Martinez, J.F. (2007) The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44,* 47-67.

McAdams, D.R. (2000). *Fighting to save our urban schools...and winning!* New York: Teachers College Press.

McCaffrey, D.F., Sass, T.R., Lockwood, J.R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*, 572-606.

Neal, D. & Schanzenbach, D.W. (2007). Left behind by design: Proficiency counts and test-based accountability. University of Chicago Working Paper.

Papay, J.P. (forthcoming). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Education Research Journal.*

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, *92*, 1394-1415.

Rivkin, S.G., Hanushek, E.A., & Kain, J.F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*, 417-458.

Rockoff, J.E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*, 247-252.

Rowan, B., Correnti, R., & Miller, R.J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104*, 1525-1567.

Sass, T.R. (2008). Policy brief: The stability of value-added measures of teacher quality and implications for teacher compensation policy, Washington, D.C.: CALDER.

Appendix B. Tables and Figures

Not included in page count.

Figure 1: Distribution of teacher effects: 4th and 5th grade mathematics and reading

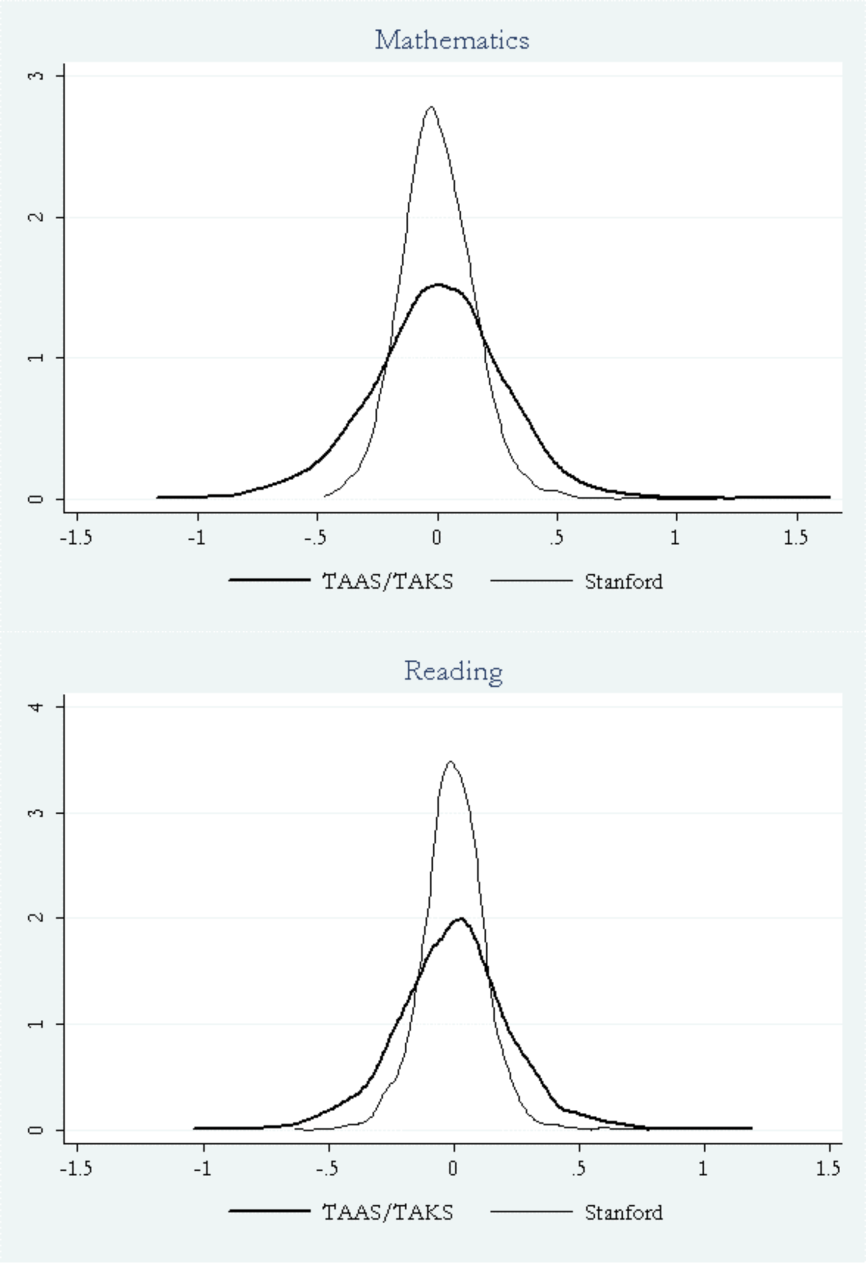
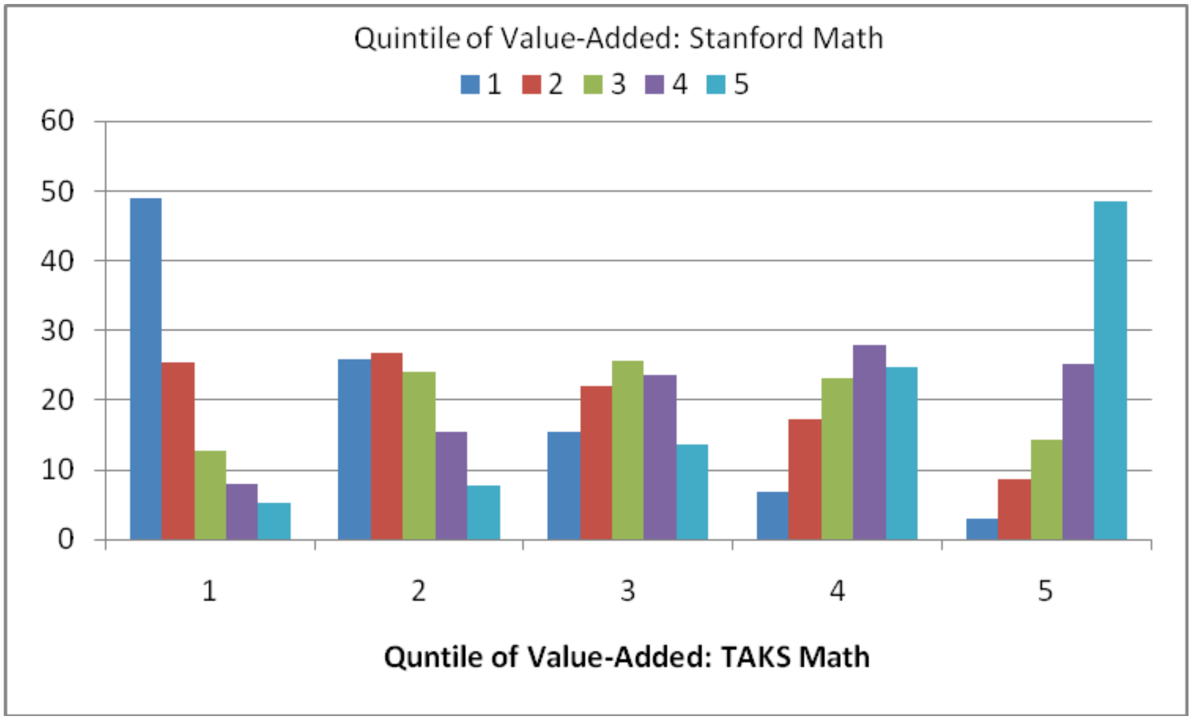


Figure 2: Quintiles of value-added on Stanford mathematics test, by quintile on TAKS math test


Quintile of Value-Added: Stanford Math
Quntile of Value-Added: TAKS Math