MEMORANDUM January 11, 2019

TO: Board Members

FROM: Dr. Grenita Lathan
Interim Superintendent of Schools

SUBJECT: **TEACHER INCENTIVE FUND STEM GRANT: PROGRAM EVALUATIONS**

CONTACT: Carla Stevens, 713-556-6700

The fourth cohort of the Teacher Incentive Fund federal grant competition ("TIF4") included special consideration for projects that would identify, develop, and utilize master teachers as leaders of STEM education (science, technology, engineering, and mathematics). In September 2012, HISD was awarded a TIF4 grant for $15.9 million over five years. The TIF4 project schools were among the HISD schools serving grades K–8 with the highest student economic disadvantage and the most risk factors for chronic absenteeism.

Attached are the three program evaluation reports associated with the TIF4 grant. A human capital approach to strengthening STEM education addressed the TIF4 project schools' need for high-quality supports for student learning, and the systemic challenges to teacher retention, development, and recruitment in hard-to-staff subjects. The first report in this series provided a descriptive overview of the grant-funded activities and interventions unique to the TIF4 project schools, setting the context for a meaningful discussion of programmatic impact.

The second report in the series addressed student outcomes for State of Texas Assessments of Academic Readiness (STAAR) Mathematics (grades three through eight) and STAAR Science (grades five and eight), during the grant period of 2012–2013 to 2016–2017. The TIF4 programming produced substantive, statistically significant results for science and for secondary mathematics. Key findings include:

- **STAAR Science, Grades 5 and 8.** Over the grant period, the cumulative impact of the TIF4 program on Grade 5 Science was an increase in student achievement of about a fifth of a standard deviation (0.20 SD). The impact on Grade 8 Science was about a quarter of a standard deviation (0.24 SD). Both estimates are statistically significant, although the evidence in eighth-grade science is less compelling. With a fifth of a standard deviation of improvement, a student initially at the 50th percentile would improve to the 58th percentile.
- **STAAR Math, Grade 6.** The point estimates suggest a cumulative impact over the grant period of about a fifth of a standard deviation (0.21 SD). These estimates were not considered statistically significant at conventional levels.
- **STAAR Math, Grades 7 and 8.** Over the grant period, the cumulative impact of the TIF4 program on Grade 7 Math was about half of a standard deviation of student achievement (0.49 SD). The impact on Grade 8 was about four-tenths of a standard deviation (0.39 SD). Both estimates were statistically significant at conventional levels. A half-standard-deviation increase would improve the achievement of a student at the 25th percentile to the 43rd percentile, or a student at the 50th percentile would then grow to the 69th percentile.
- **STAAR Math, Grades 3 to 5.** In grades three through five, the TIF4 program did not appear to have a large effect on mathematics achievement in any year of the grant period.

The third and final report overviews the performance-based compensation strategies implemented through the TIF4 grant, as well as situates that work in the context of HISD's challenges for teacher retention and mobility. Key findings include:

- The TIF4 schools paid out about ten $5,000 retention bonuses for each $10,000 recruitment bonus (178 Retention vs. 18 Recruitment). This suggests that effective math and science teachers at hard-to-staff HISD schools find retention bonuses to be meaningfully more compelling than larger recruitment bonuses.
- In Years Three, Four, and Five, the TIF4 schools retained 75% of their Effective and Highly Effective math and science teachers.
- During the grant period, HISD directed $3,330,781 of federal, state, and local resources into the ASPIRE Award at the TIF4 project schools. Over a thousand (1,012) ASPIRE Awards were paid to educators at the TIF4 campuses during this time. Every TIF4 school had at least one educator who received an ASPIRE Award during the grant.
- By the start of the third year after their initial hire, 46% of new teachers had left the HISD school where they started. This attrition rate is higher for new math (60.8%) and new science (61.2%) teachers.
- During this period, the top ten percent of HISD schools (90th percentile and upward) annually retained over 80% of all their high TADS teachers, regardless of subject area or years of experience.

Taken together, these findings strongly suggest that the high turnover among HISD's math and science teachers can be mitigated through investment in retention bonuses for effective and highly effective teachers already working at specific campuses.

Should you have any further questions, please contact Carla Stevens in Research and Accountability at 713-556-6700.

_________________________________GL

Attachments (3)

cc: Noelia Longoria, Interim Chief Academic Officer
Julia Dimmitt, Chief Human Resources Officer
Silvia Trinh, Chief of Staff
Annie Wolfe, Secondary Curriculum and Development Officer
Justin Fuentes, Assistant Superintendent for School Choice
Angela Brooks, Manager for Grants Development

HOUSTON INDEPENDENT SCHOOL DISTRICT

RESEARCH

Educational Program Report

TEACHER INCENTIVE FUND STEM GRANT IN HOUSTON ISD: A MATCHED-COMPARISON ANALYSIS OF MATH AND SCIENCE STAAR SCORES

HISD | Research and Accountability

ANALYZING DATA, MEASURING PERFORMANCE.

2018 BOARD OF EDUCATION

This work was supported by U.S. Department of Education Award No. S374B120011.

The opinions expressed are those of the authors and do not represent the views of the HISD Board of Education, or the U.S. Department of Education.

Suggested Citation:
Price, L.E., Christian, M., & Stevens, C.J. (2018). *Teacher Incentive Fund STEM Grant in Houston ISD: A Matched-Comparison Analysis of Math and Science STAAR Scores.* Houston ISD: Department of Research and Accountability.

Houston Independent School District
Hattie Mae White Educational Support Center
4400 West 18th StreetHouston, Texas 77092-8501

www.HoustonISD.org

Teacher Incentive Fund STEM Grant in Houston ISD: A Matched-Comparison Analysis of Math and Science STAAR Scores

Executive Summary

Program Description

The fourth cohort of the Teacher Incentive Fund grant competition (TIF4) included special consideration for projects that would identify, develop, and utilize master teachers as leaders of STEM education. A human capital approach to strengthening STEM education addressed the TIF4 project schools' need for high-quality supports for student learning, and the systemic challenges to teacher retention, development, and recruitment in hard-to-staff subjects. The previous report in this series provided a descriptive overview of activities and interventions unique to the TIF4 project schools, setting the context for a meaningful discussion of programmatic impact. This analysis addresses student outcomes for STAAR Mathematics (grades three through eight) and STAAR Science (grades five and eight), during the grant period of 2012–2013 to 2016–2017.

Highlights

Through a matched-comparison group design, a regression analysis was implemented to detect causal relationships between students' STAAR achievement and the school's participation in the TIF4 programming. Specifically, the annual dependent variable for each school was the mean scale score of all students in each grade level who took the STAAR exam in either English or Spanish. In grades three through five, the TIF4 program did not appear to have a large effect on mathematics achievement in any year of the grant period. However, this analysis demonstrates that the TIF4 grant did produce substantive, statistically significant results for science and for secondary mathematics.

- **STAAR Science, Grades 5 and 8.** Over the grant period, the cumulative impact of the TIF4 program on Grade 5 Science was an increase in student achievement of about a fifth of a standard deviation (0.20 SD). The impact on Grade 8 Science was about a quarter of a standard deviation (0.24 SD). Both estimates are statistically significant, although the evidence in eighth-grade science is less compelling.
- **STAAR Math, Grade 6.** The point estimates suggest a cumulative impact over the grant period of about a fifth of a standard deviation (0.21 SD). These estimates were not considered statistically significant at conventional levels.
- **STAAR Math, Grades 7 and 8.** Over the grant period, the cumulative impact of the TIF4 program on Grade 7 Math was about half of a standard deviation of student achievement (0.49 SD). The impact on Grade 8 was about four-tenths of a standard deviation (0.39 SD). Both estimates were statistically significant at conventional levels.

The TIF4 programming produced substantive, meaningful improvements in student achievement. With a fifth of a standard deviation of improvement, a student initially at the 50th percentile would improve to the 58th percentile. A quarter standard deviation improvement moves a student from the 50th percentile to the 60th percentile. A half-standard-deviation increase would improve the achievement of a student at the 25th percentile to the 43rd percentile, or a student at the 50th percentile would then grow to the 69th percentile.

Notably, these outcomes are meaningfully stronger than the findings of recent high-quality research on the effects of teacher coaching on student outcomes. This suggests that the complex programmatic aspects of the TIF4 program produced substantive results, where simpler programs may have fallen short. Future reporting in this series will investigate human capital outcomes for science and math teachers at the TIF4 project schools.

Table of Contents

Figures

Introduction

Since established by an Appropriations Act in 2006, the Teacher Incentive Fund (TIF) competitive grant program in the U.S. Department of Education (USDE) has supported human capital strategies "to ensure that students attending high-poverty schools have better access to effective teachers and principals, especially in hard-to-staff subject areas" such as science and math. Responding to the national agenda to improve STEM education, in 2012, the fourth cohort of the Teacher Incentive Fund federal grant competition (TIF4) included special consideration for projects designed to improve STEM education by identifying, developing, and utilizing master teachers as leaders of broader improvements (OESE, 2012a).

In September 2012, HISD was awarded a TIF4 grant for $15.9 million over five years (HISD Communications, 2012). The human capital strategies supported through TIF4 in Houston continue the successes and strategies of HISD's previous TIF grants (Price & Stevens, 2017), and are similar to strategies undertaken by the other 35 TIF4 grant recipients nationwide (OII, 2015). For more information about the Teacher Incentive Fund grant, see **Appendix A**.

HISD was one of just six TIF4 grantees to support a "comprehensive approach to improving STEM instruction" as part of their overall human capital strategy (OESE, 2012b). STEM grantees advanced the Absolute Priorities required of all TIF grantees — regarding human capital management systems, and educator evaluation — as well as a third Priority that incorporated "STEM master teachers" into their strategy for STEM improvement at the TIF4 project schools. In the verbiage of the TIF program officers, "STEM master teachers" are those educators "who serve as recognized leaders in STEM education improvement efforts regardless of their specific duties" (Zawaiza & Robinson, 2014). In HISD, the TIF4 grant supported twelve full-time positions for "STEM master teachers" — a STEM Curriculum Manager, ten STEM Teacher Development Specialists (TDS), and a STEM TDS Team Lead.

A human capital approach to strengthening STEM education addressed the project schools' need for high-quality supports for student learning, and the systemic challenges to teacher retention, development, and recruitment in hard-to-staff subjects. For a comprehensive overview of the supports for STEM teaching and learning at the TIF4 project schools, see the first report on TIF4 on HISD's website (Price, Provencher, & Stevens, 2018).

Theory of Action

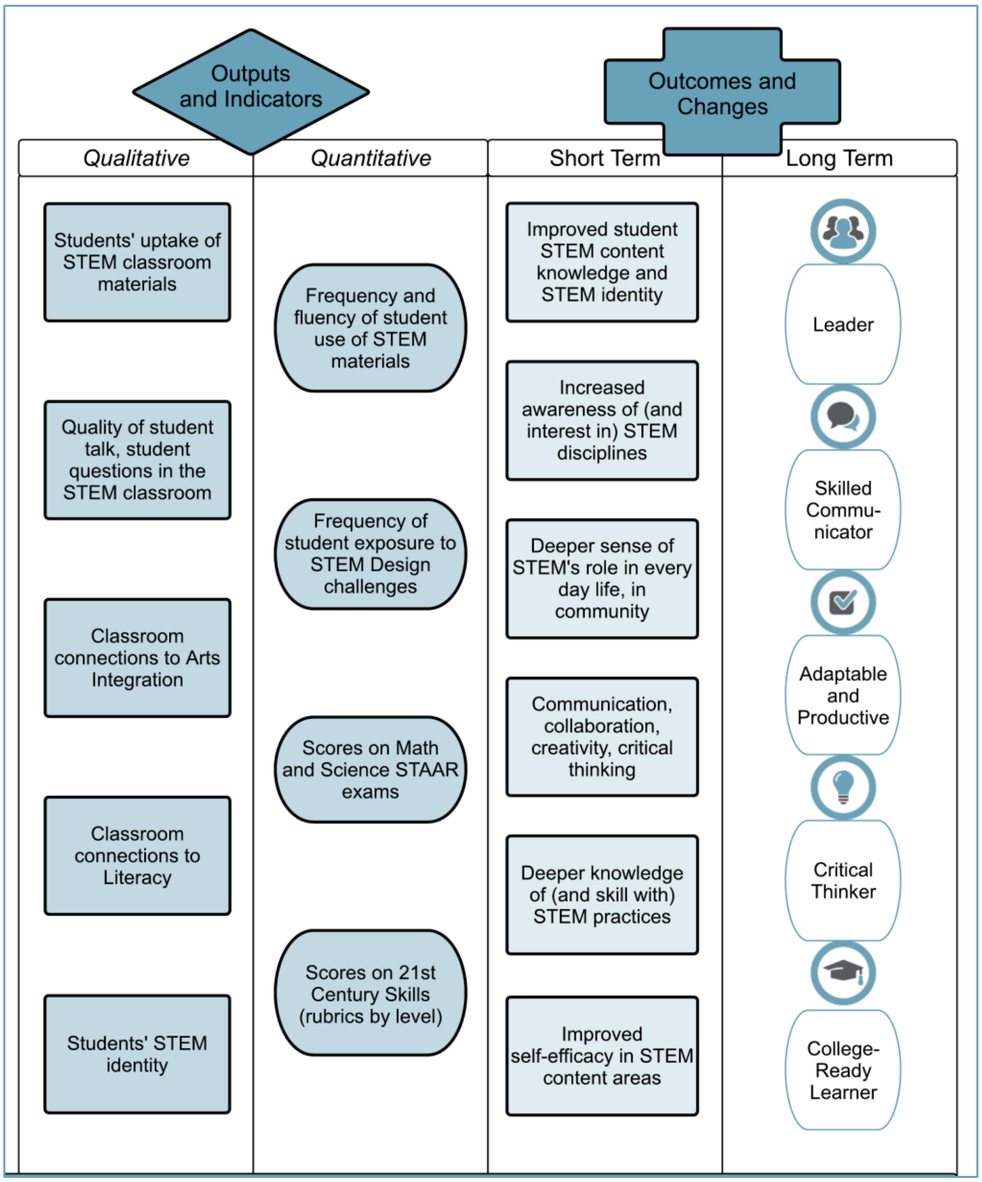
Under the assumptions guiding the TIF grant program, student outcomes are a function of human capital management (HCM) inputs — educator recruitment, retention, selection, assessment, professional development and supports, and performance-based compensation (Miller et al., 2015) — as mediated by teaching and learning behaviors. Through the TIF4 grant, HISD supported some HCM activities that addressed teaching and learning across all content areas, and some HCM activities that addressed teaching and learning only within the STEM content areas. Within this theory of action, the TIF activities focused explicitly on STEM teaching would affect students' science and mathematics learning at the project schools. Consequently, it is important to examine those outcomes, and to evaluate whether it is appropriate to make causal statements about the relationship between the TIF4 activities and the student outcomes at the grant schools.

Even under perfectly controlled experimental conditions, there are many intermediate steps between the efforts to shape teachers' professional activities and their students' learning outcomes; all of them need to succeed in order to see an effect in student outcomes. In other words, it is a complex theory of action with many mediating variables. In their August 2013 webinar to grantees, the TIF4 STEM Technical Assistance providers identified broad steps in this causal pathway, from: (1) Inservice Teacher Professional

Development, to (2) Teacher Knowledge, Skills, Beliefs, and Intentions, to (3) Classroom Practice, to, finally, (4) Student Outcomes (Weiss, 2013). Each are critical to the STEM instructional strategies employed at the TIF4 project schools.

Student exam scores are not the only outcomes of these interventions. As shown in **Figure 1**, students' math and science scores are just one of the indicators and outputs of the TIF4 strategies for STEM instruction in HISD: (1) Students' uptake of STEM classroom materials; (2) Quality of student talk, and student questions in the STEM classroom; Classroom connections to both (3) arts integration and (4) literacy; (5) Students' STEM identity; (6) Frequency and fluency of student use of STEM materials; (7) Frequency of student exposure to STEM Design Challenges; (8) student scores on Math and Science STAAR exams; and (9) student scores on 21st Century Skills rubrics, by Grade Level.

Figure 1. Student-Level Outcomes, Indicators, and Changes from TIF4 STEM Strategies



Despite the complexity of these mediating variables, sufficient high-quality research has been conducted so that it is possible to make some educated estimates about the impact of the "master teachers" approach in HISD supported through TIF4. A recently published meta-analysis of 37 high-quality studies on teacher coaching explored the complicated relationship between student outcomes and professional supports for teachers (Kraft, Blazar, & Hogan, 2018). The authors' theory of action — reproduced in **Figure 2**, from a pre-publication version — outlines dynamics between programmatic inputs (coaching, curricular materials, and training sessions/workshops), interim outcomes (teacher knowledge and teaching behavior), and the long-term student outcomes.

In their careful meta-analysis, the authors wrote candidly about the "strong supporting evidence" for a causal relationship between instructional practice and students' academic outcomes. However, they also cautioned readers to recognize the implications of this connection — that even modest changes in student achievement are the result of "relatively large improvements in instructional quality" (p. 22). This meta-analysis underlines the complexity of the work at hand: the grant-funded activities to improve STEM instruction at the TIF4 project schools must have surpassed a certain threshold of impact on teachers' instructional practice in order for a causal analysis to detect corresponding change in student outcomes.

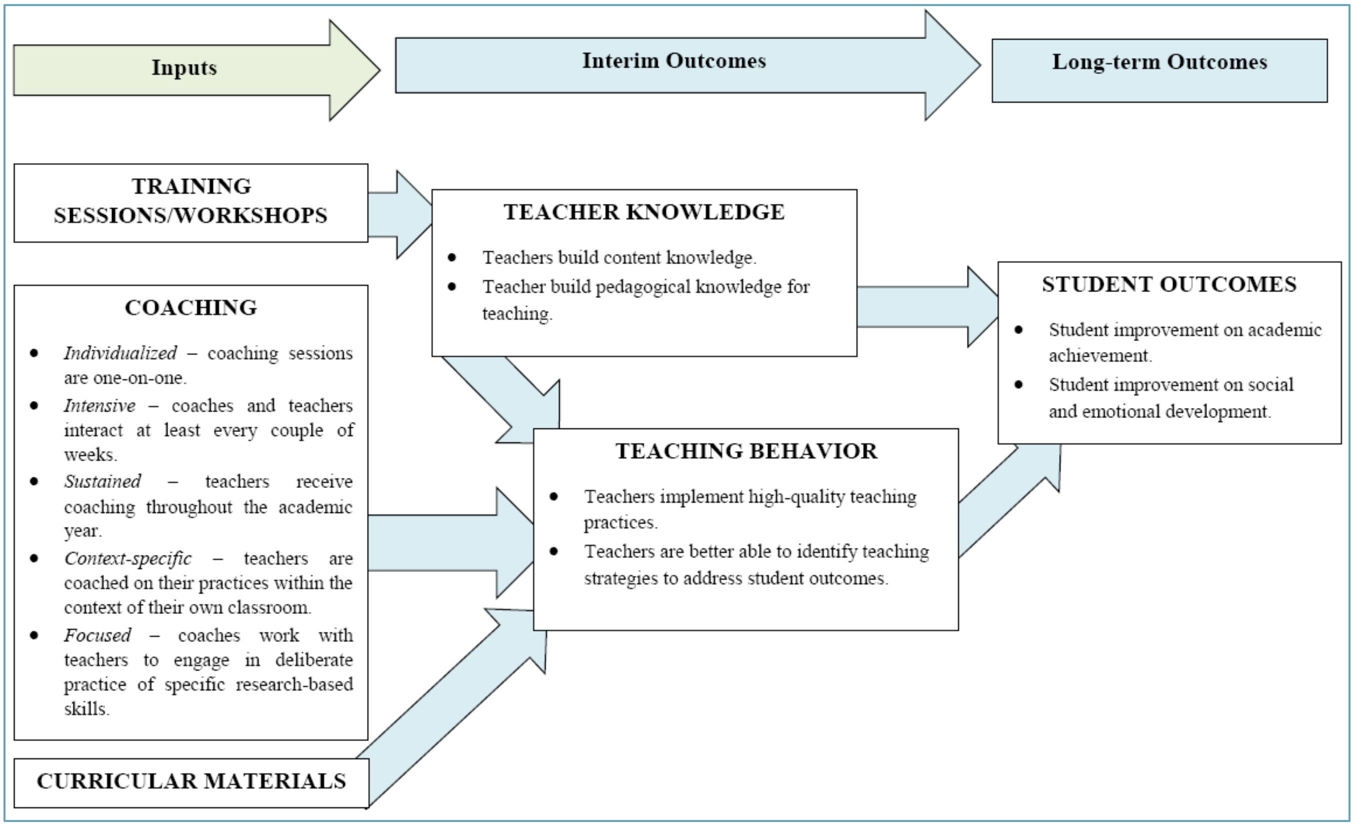
Purpose

Under the definitions[i] used in federal law (ESSA, 2015), the TIF4 STEM master teachers strategy can already be described as "evidence based" to improve instructional practices. However, this report represents the first investigation into the relationship between HISD's master teachers strategy, and students' math and science scores. If well-designed and well-implemented, this quasi-experimental[ii] study analysis could provide "Moderate Evidence" for the impact of the TIF-supported strategy on student learning outcomes, thereby making available additional funding opportunities for the District and also better informing leadership conversations about goals and priorities in an environment of limited financial resources.

The purpose of this report is to provide HISD leadership and USDE program staff with a detailed examination of the math and science student outcomes for schools participating in the TIF4 STEM grant (Award #S374B120011) from 2012–2013 through 2016–2017. The report addresses the grade-level scale scores used in the state-wide criterion-referenced STAAR (State of Texas Assessments of Academic Readiness) exams required under section 1111(b)(3) of the federal Elementary and Secondary Education Act, as well as the proficiency levels used in state accountability metrics (TEC § 39.023 and § 39.053). Wherever possible, this report was done in alignment with the standards and procedures of the What Works Clearinghouse™ (WWC). Established under the Education Sciences Reform Act of 2002, the WWC is an initiative of the U.S. Department of Education's Institute of Education Sciences (IES, 2017a).

Internal reports during the grant period suggested the project schools were experiencing meaningful gains in their students' math and science metrics — do these trends hold up to more rigorous analytic methods that could detect a causal relationship between student outcomes and the grant activities? Informal assessments during the grant period showed evidence of changes in teachers' own employment decisions, as well as positive changes in the instructional practice of specific STEM teachers — so if student outcomes could be attributed to the school's participation in the TIF grant, then it is reasonable to assume that the human capital strategies deployed through TIF were sufficient to impact student math and science metrics. Additional reporting in this series will evaluate those specific retention, compensation, development, and recruitment strategies at the TIF4 project schools.

Figure 2. Theory of Action for Teaching Coaching (Kraft, Blazar, & Hogan, 2016, p. 43)

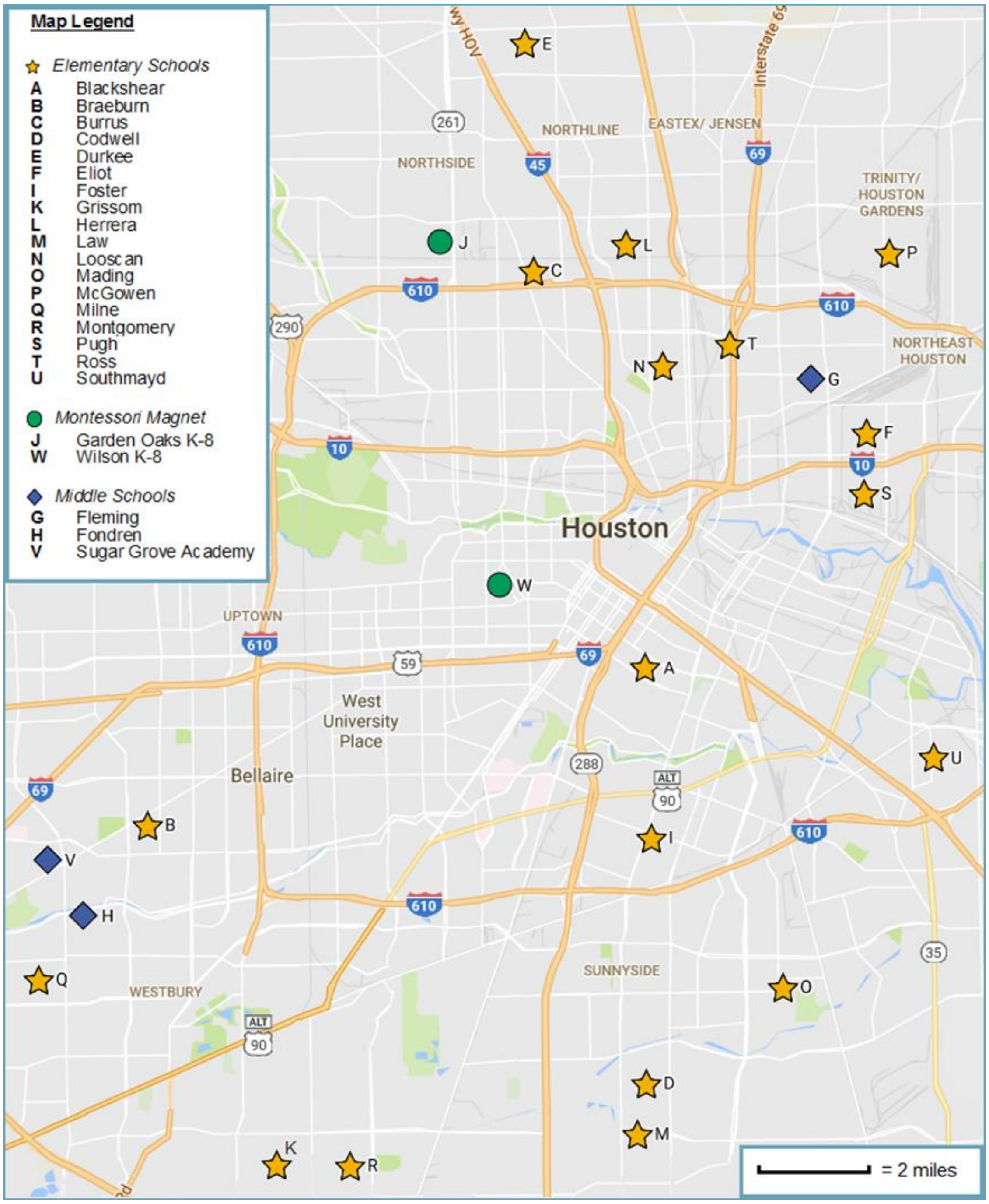


Methods

Research Design

In July 2012, HISD leadership identified specific schools to receive STEM programming through the TIF4 grant (HISD, 2012). Each year, these schools served approximately 7,500 students from pre-kindergarten through eighth grade — located in almost every quadrant of Houston (see **Figure 3**). Like most of the schools in HISD, the TIF4 project schools were considered "high-need" under the definitions in the U.S. Department of Education's Request for Application (OESE, 2012a). Additionally, the TIF4 project schools each had a persistent track record of underperforming on the science STAAR exams required under the Elementary and Secondary Education Act (NCLB, 2002). Their inclusion in the TIF4 grant was intended to address student learning and achievement in both math and science. The TIF4 project schools were identified based on their need for supports, rather than randomly. Consequently, HISD project staff were precluded from conducting a randomized controlled trial, which is considered to be the most rigorous research design for making causal inferences (Murnane & Willett, 2011).

Figure 3. Geographic Location of the TIF4 Project Schools



To appropriately account for the selection of the TIF4 schools in the assessment of impact, HISD project staff chose a matched-comparison group (MCG) research design. Considered to be a "rigorous design" for education research, a MCG design is comprised of a treatment group and a comparison group. When these two groups are highly similar at the beginning of the intervention, differences between the groups after the intervention are likely due to the intervention itself rather than some other pre-existing difference (Hanita, Ansel, & Shakman, 2017). Here, the MCG design allowed project staff to estimate the size of the TIF4 intervention on the math and science outcomes of those schools' students. To evaluate the impact of the

STEM interventions on math and science scores at the TIF4 project campuses, then, project staff set out to identify comparable schools that could be an appropriate comparison group.

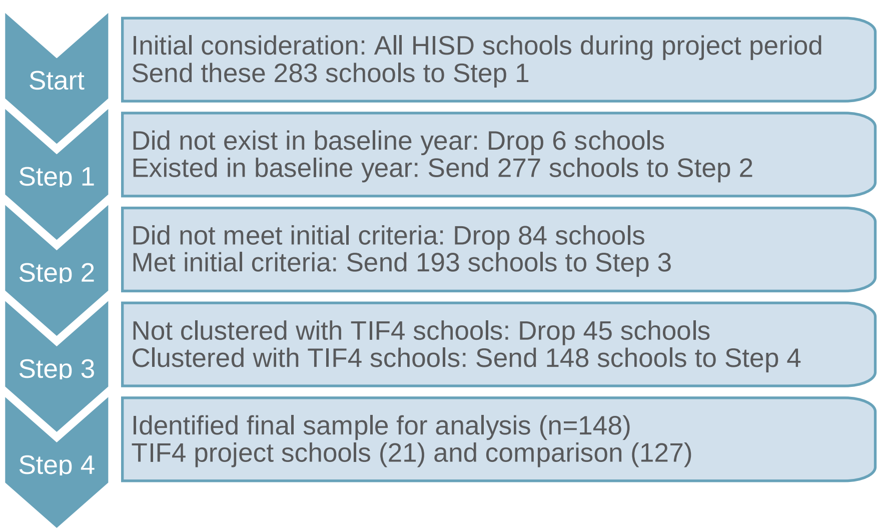
Identifying a Comparison Group

In short, a matched-comparison group research design requires a matched-comparison group. For this, project staff relied on the work published in the 2014 internal memorandum to HISD Chief School Officers, "Identification of Homogenous School Clusters" (Chang, 2014; see **Appendix B**). For details on the advantages and tradeoffs of this approach, see the "Limitations" section of **Appendix E**.

A "comparable school" was defined as a school serving the same grade levels, with similar enrollment size, and similar relationships between the following indicators in 2012–2013: Students identified as economically disadvantaged (%), students identified as at risk (%), annual student mobility rate (%), students who are zoned for the school (%), students identified as English Language Learners (%), students identified as African American, Hispanic, and White (%). **Figure 4** illustrates the steps to identify the matched-comparison group from among the 283 schools in HISD during the grant period (2012–2017).

- Step 1: From the 283 schools initially considered, drop the six schools that did not exist in the baseline year for student data (PEIMS, 2013).
- Step 2: From the 277 remaining, drop 84 schools that did not meet initial criteria for inclusion:
    - Did not serve grades K–8 (n=23)
    - Did not have comparable schools in HISD (n=60). *Note*: Garden Oaks Montessori and Wilson Montessori (K–8) were both dropped from the analytic sample due to this step, even though they participated in the TIF4 programming.
    - Did not have three years of student data (n=1; Dodson Elementary was closed after 2013 and its zoned students incorporated into the nearby Blackshear Elementary)
- Step 3: From the 193 remaining, drop 45 schools that were not comparable to the TIF4 schools.
- Step 4: The remaining 148 schools comprise the analytic sample for this analysis: 21 TIF4 project schools (also "treatment"), and 127 comparison schools:
    - 132 elementary schools (18 TIF4 schools and 114 comparison)
    - 16 middle schools (3 TIF4 schools and 13 comparison)

Figure 4. Steps in Identifying Matched-Comparison Group



For the names, clusters, and sample grouping of these 148 schools, see **Appendix C**. Any HISD school not named in Appendix C was not included in the sample as a treatment school or comparison school.

Assessing the Baseline Equivalence of the Analytic Sample

Identifying sample schools through the steps described above ensured that the Treatment and Comparison schools would be similar along the characteristics used in clustering. Project staff then examined the standardized mean difference between the groups in 2013, to gauge whether the groups were similar enough to be considered equivalent at baseline; under the WWC Procedures, a difference of $g \leq 0.05$ meets the criterion of the baseline equivalence[iii]. The standardized mean difference between the groups (Hedges' g) for these variables did not satisfy the baseline equivalence requirement ($g \leq 0.05$) for these variables, and so these variables were included as covariates (i.e., "controlled for") in the analysis best suited for detecting causal impact.

Table 1. School Characteristics at Baseline — Mean, Standard Deviation, and Effect Size

Variable in 2013	TIF4		Comparison		|g|
	Mean	SD	**Mean**	SD	
Percent African-American	45.1	(34.3)	27.0	(28.3)	0.62
Percent designated as Limited English Proficient or English Language Learner	5.4	(10.8)	3.4	(5.9)	0.29
Percent with Disabilities	7.88	(2.9)	7.2	(3.3)	0.19
Percent Economically Disadvantaged	94.9	(2.9)	91.7	(7.7)	0.45
Percent Immigrant	1.9	(3.3)	3.3	(3.7)	0.39
STAAR Reading, Grade 3	1373.3	(35.9)	1391.7	(40.1)	0.47
STAAR Reading, Grade 4	1442.4	(28.2)	1463.8	(40.1)	0.56
STAAR Reading, Grade 5	1492.8	(19.0)	1512.5	(34.6)	0.60
STAAR Reading, Grade 6	1489.6	(9.9)	1526.6	(55.8)	0.72
STAAR Reading, Grade 7	1554.3	(19.6)	1598.6	(51.5)	0.92
STAAR Reading, Grade 8	1601.9	(6.0)	1639.1	(48.4)	0.83
Note: Hedges' g corrected for uneven group sizes was calculated with Tannenbaum (2013).					

Project staff conducted additional testing of the sample balance, drawing on the internal report "A Better Picture of Poverty" (Reeves, McCarley, Mosier, & Carney, 2015). In this report, HISD staff used 2014 data and identified two dozen school and neighborhood risk factors that affect academic performance and correlate with chronic absenteeism. This additional analysis, along with variable definitions and sources, can be found in **Appendix D**. For the limitations in assessing baseline equivalence, see **Appendix E**.

Dependent Variable

This analysis addresses student outcomes for STAAR Mathematics (grades three through eight) and STAAR Science (grades five and eight), during the five-year grant period of 2012–2013 to 2016–2017. The 2012–2013 outcomes serve as pre-intervention baseline: even though the grant was awarded in October 2012, in-school supports for STEM did not begin until the 2013–2014 school year. Specifically, the annual dependent variable for each school is the mean scale score of all students in each grade level who took the STAAR exam in either English or Spanish[iv]. Analysis shown in **Table 2** illustrated that the TIF4 schools at baseline demonstrated a particular need for science and math intervention: the standardized mean difference between the groups (Hedges' g) for these variables does not satisfy the baseline equivalence requirement ($g \leq 0.05$) for the dependent variable. Note that the TIF project staff chose scale scores because the performance levels on the STAAR assessments changed during the grant period; by using scale scores, the modeling was not affected by changes in performance levels. See **Appendix E** for an overview of the STAAR performance levels, and the considerations given to various limitations within STAAR indicators.

Table 2. Difference between TIF4 and Comparison Schools in Baseline Year (2013)

2013 STAAR Exam	TIF4		Comparison		|g|
	Mean	SD	**Mean**	SD	
Math, Grade 3	1398.0	46.7	1438.3	50.8	0.80
Math, Grade 4	1456.8	45.7	1514.8	53.6	1.10
Math, Grade 5	1514.7	37.8	1554.7	50.5	0.82
Math, Grade 6	1533.3	31.9	1566.4	61.9	0.57
Math, Grade 7	1516.1	8.7	1559.4	27.5	1.70
Math, Grade 8	1620.0	9.9	1643.9	47.6	0.50
Science, Grade 5	3506.4	104.6	3671.3	160.9	1.06
Science, Grade 8	3547.0	159.9	3718.9	278.5	0.65
Note: Hedges' g corrected for uneven group sizes was calculated with Tannenbaum (2013).					

Unit of Analysis

This analysis focuses on school-wide metrics, not on the metrics of individual students and not on the aggregate metrics of students linked to a specific teacher.

- First, this is consistent with the program's theory of action: that the availability of job-embedded professional supports for STEM will improve science and math outcomes across all grade levels.
- Second, student mobility through regular grade promotion would confound a by-student analysis of four years of "treatment." This is simply due to typical grade promotion practice: a third grader at a TIF4 project school in 2013 would have moved up to another school for sixth grade by 2016, and not necessarily one of the three middle schools participating in the grant.
- Third, while all the TIF4 project schools experienced specific STEM activities, there was meaningful variation between schools in the exact order and manner in which those activities unfolded. Although components were targeted at specific teachers, the intervention was not identical for any two teachers.

In other words, the STEM master teachers required flexibility to meet each school's unique and evolving needs. Rather than prioritizing uniformity of implementation (as would befit a teacher-level or student-level analysis), they prioritized *differentiating each school's STEM supports based on the school's specific needs.* For more on the choice of dependent variable and unit of analysis, see **Appendix E**.

Three Phases of Analysis

The first phase of analysis simply compares the TIF4 project schools to themselves — specifically, the trends in their students' performance levels over the grant period. On their own, these performance levels would be insufficiently rigorous measures for making causal inferences. However, these trends can offer suggestive evidence for the impact of the TIF4 project. Additionally, they reflect the indicators that HISD reported to USDE program officers in annual performance reports. The second phase of analysis addresses the gaps in mean scale scores between TIF4 and comparison schools. If the TIF4 intervention was having an effect on students' math and science scores, then one point of evidence could be whether the TIF4 schools shrank the annual achievement gaps by outpacing the comparison schools during the grant period.

Both the first and the second analyses are insufficiently rigorous to make causal inferences about the effect of the TIF interventions, but they are important for other reasons: they underpin state accountability metrics, school leader appraisal scores, district-wide goals, and the TIF4 progress measures reported to the USDE each year. The third step of analysis employs a statistically sophisticated model to examine the causal effect of a school's participation in TIF4 on their school's science and mathematics scores, in each year and for each grade and subject. For details on the model, see **Appendix E**.

Results

Result 1: TIF schools saw meaningful change in their students' math and science proficiency levels. As detailed above, the first analysis addresses the trends in students' performance levels over the grant period. The cut scores for these performance levels are determined annually by the Texas Education Agency (TEA), and reflect the student's mastery of the content for their current grade level (Student Assessment Division, 2017). See **Appendix E** for an overview of the performance levels.

Table 3: Annualized Rate of Change, Count of TIF4 Students at Each Proficiency Level (2013–2017)

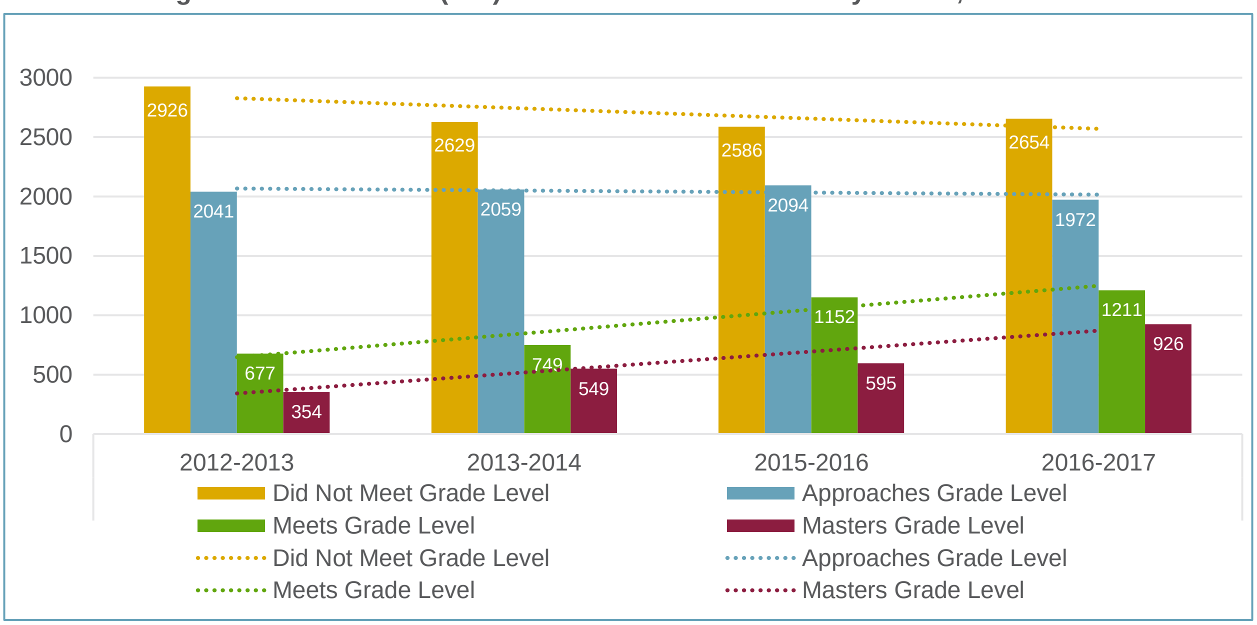
Subject and Exam	Did Not Master Grade Level	Approaches Grade Level	Meets Grade Level	Masters Grade Level
STAAR Math (Grades 3–8)	-85.9	-17.2	+200.0	+176.2
STAAR Science (5 & 8)	-35.5	+1.0	+48.4	+27.7
Algebra I EOC	-2.0	-6.3	+1.1	+8.4

STAAR Math, Grades 3–8

Figure 5 shows the number of students at the TIF4 project schools who scored at each proficiency level on the STAAR Math exam in English (grades 3–8) and in Spanish (grades 3–5) during the grant period. The linear trend for each level is represented with a dotted line in the same color; the first row of **Table 3** shows these linear rates of change as an annual figure. Over the grant period (2013–2017):

- The number of students at TIF4 schools at the Did Not Meet Grade Level standard on the STAAR Math exam decreased by 9.3% (272 students), at an average linear rate of -85.9 students per year.
- The number of students at TIF4 schools at the Approaches Grade Level standard on the STAAR Math exam decreased by 3.4% (69 students), at an average linear rate of -17.2 students per year.
- The number of students at TIF4 schools at the Meets Grade Level standard on the STAAR Math exam increased by 78.9% (534 students), at an average linear rate of 200 students per year.
- The number of students at TIF4 schools at the Masters Grade Level standard on the STAAR Math exam increased by 161.6% (572 students), at an average linear rate of 176.2 students per year.

Figure 5. STAAR Math (3–8) at TIF4 Schools: Proficiency Levels, 2013–2017



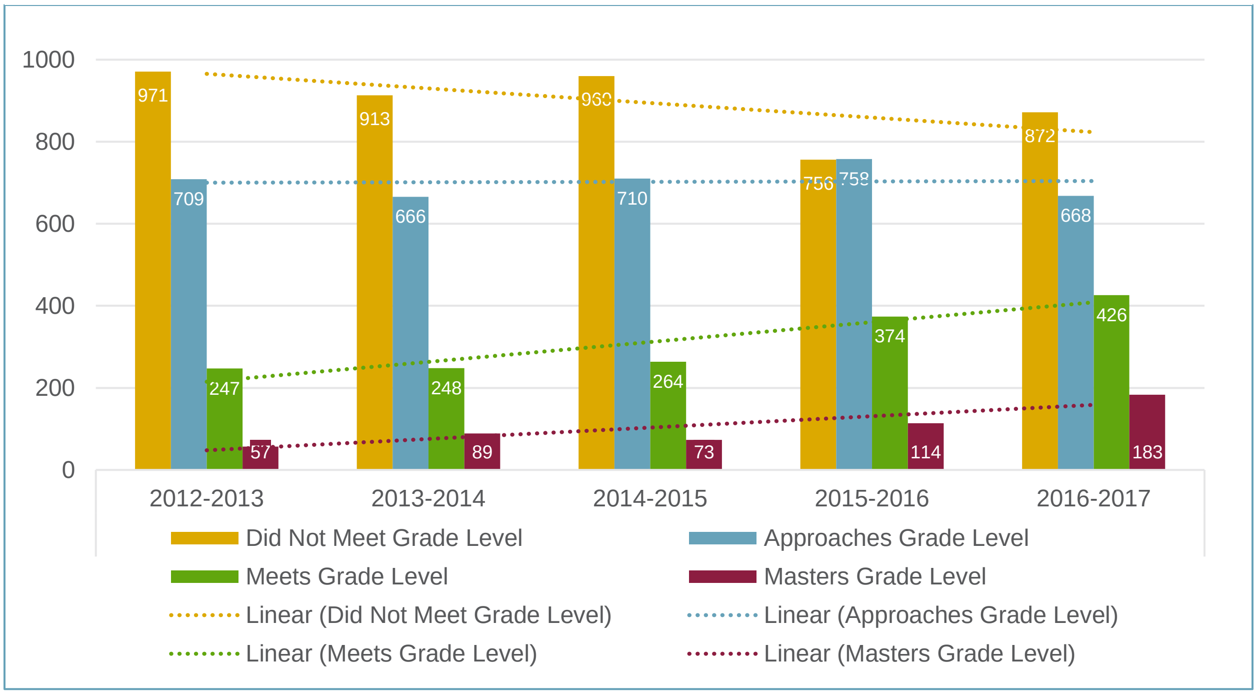
Note: The number of students at each proficiency level — as presented here — is mutually exclusive.

STAAR Science, Grades 5 and 8

Figure 6 shows the number of students at the TIF4 project schools who scored at each proficiency level on the STAAR Science exam in English and Spanish (grades 5 and 8) during the grant period. The linear trend for each level is represented with a dotted line in the same color; the second row of **Table 3** shows these linear rates of change as an annual figure. Over the grant period (2013–2017):

- The number of students at TIF4 schools at the Did Not Meet Grade Level standard on the STAAR Science exam decreased by 10.2% (99 students), at an average linear rate of -35.5 students per year.
- The number of students at TIF4 schools at the Approaches Grade Level standard on the STAAR Science exam decreased by 5.6% (41 students), but at an average linear rate of 1 student per year.
- The number of students at TIF4 schools at the Meets Grade Level standard on the STAAR Science exam increased by 72.5% (179 students), at an average linear rate of 48.4 students per year.
- The number of students at TIF4 schools at the Masters Grade Level standard on the STAAR Science exam increased by 68.9% (126 students), at an average linear rate of 27.7 students per year.

Figure 6. STAAR Science (5 and 8) at TIF4 Schools: Proficiency Levels, 2013–2017



Note: The number of students at each proficiency level — as presented here — is mutually exclusive.

STAAR Algebra I, Grade 8

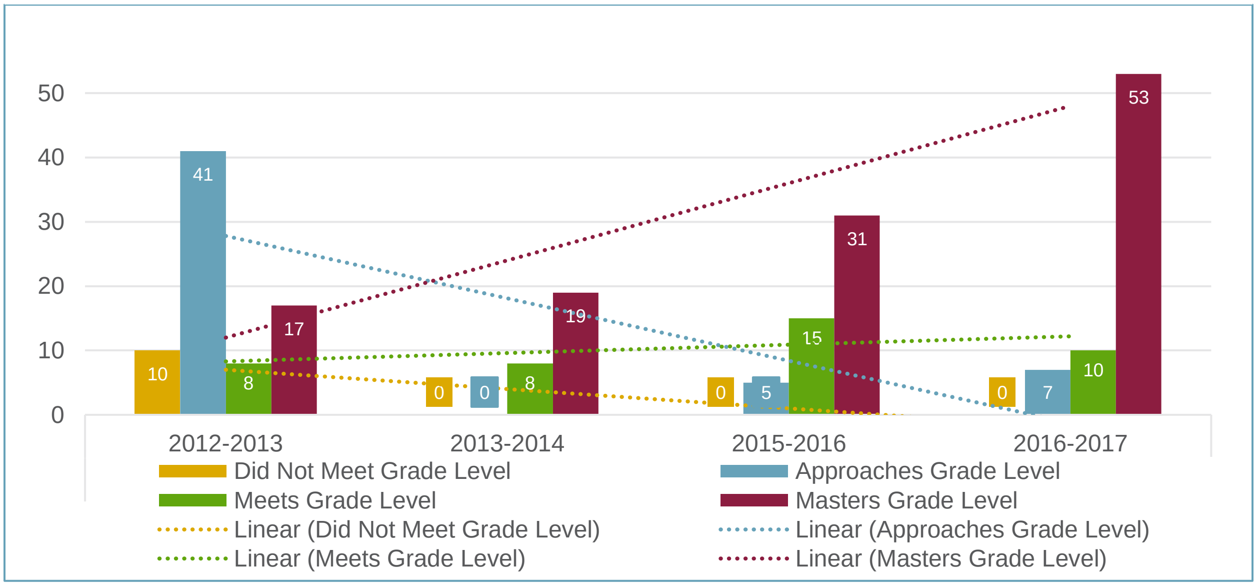
Figure 7 shows the number of students at the TIF4 schools taking the exam for the first time who scored at each proficiency level on the STAAR Algebra I End of Course (EOC) exam during the grant period. The linear trend for each level is represented with a dotted line in the same color; the third row of **Table 3** shows these linear rates of change as an annual figure. Although the EOC exams are not assigned to students by grade level, the students taking this Algebra I exam at these schools were all in the 8th grade. This EOC is only offered in English, whereas STAAR Math is offered in both English and Spanish for grades 3–5.

- Over the grant period, the number of students at TIF4 schools at the Did Not Meet Grade Level standard on the Algebra I exam decreased by 100% (10 students), at an average linear rate of -2 students per year. This annual rate is deceptive, however: **Figure 7** illustrates zero students at this level after 2013.

- Over the grant period, the number of students at TIF4 schools at the Approaches Grade Level standard on the Algebra I exam decreased by 64.2% (34 students), at an average linear rate of -6.3 students per year.
- Over the grant period, the number of students at TIF4 schools at the Meets Grade Level standard on the Algebra I exam increased by 20.0% (2 students), at an average linear rate of 1.1 students per year.
- Over the grant period, the number of students at TIF4 schools at the Masters Grade Level standard on the Algebra I exam increased by 67.9% (36 students), at an average linear rate of 8.4 students per year.

The changing number of students each year reflects changes in which schools offered Algebra I to their eighth graders: In 2013, all three middle schools offered Algebra I. In 2014, only one TIF4 school offered Algebra I; in 2016, two schools offered Algebra I, and by 2017, all three were again offering Algebra I. This also affected the number of eighth graders taking the STAAR Math exam, as addressed in the third analysis.

Figure 7. Algebra I EOC at TIF4 Schools: Proficiency Levels, 2013–2017



Note: The number of students at each proficiency level — as presented here — is mutually exclusive.

Result 2: Comparing scale scores over time, the TIF4 schools closed the gaps on every metric.

While certainly encouraging, the first results could be a function of factors other than TIF4 participation (e.g., changes in cut scores, or which students sit for which exams). If the TIF4 intervention was having an effect on students' math and science scores, then a point of evidence could be whether the TIF4 schools shrank the gaps in achievement by outpacing the comparison schools in their growth.

Elementary — Math, Grades 3 to 5

Figure 8 illustrates the average scale score for STAAR Math during the grant period (2013–2017) in grade 3 (blue), grade 4 (yellow), and grade 5 (green) for both comparison (circle) and TIF4 (triangle) schools. For both the TIF4 and Comparison schools, all three grade levels saw an increase in their average scale score during the grant period. This increase in average scale scores in both groups and across all grade levels is a good sign for student learning. However, as illustrated in **Figure 9**, it also means that the gaps between TIF4 and comparison schools showed only modest decreases: a decrease of 0.4% or -5.5 points for grade 3, a decrease of 1.4% or -19.7 points for grade 4, and a decrease of 0.4% or -5.7 points for grade 5. Note:

Appendix F Table 1 shows each grade level's average scale score, the standard deviation (in parentheses), and the number of students who took the exam each year.

Figure 8. Scale Score Trends for STAAR Math, Grades 3–5 (2013–2017)

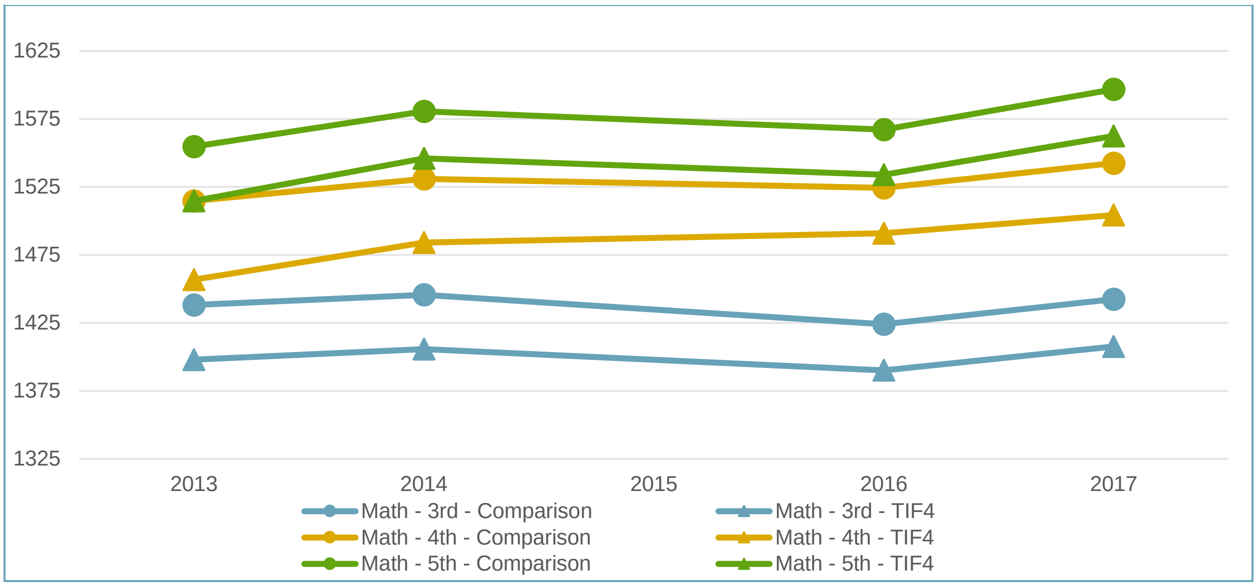
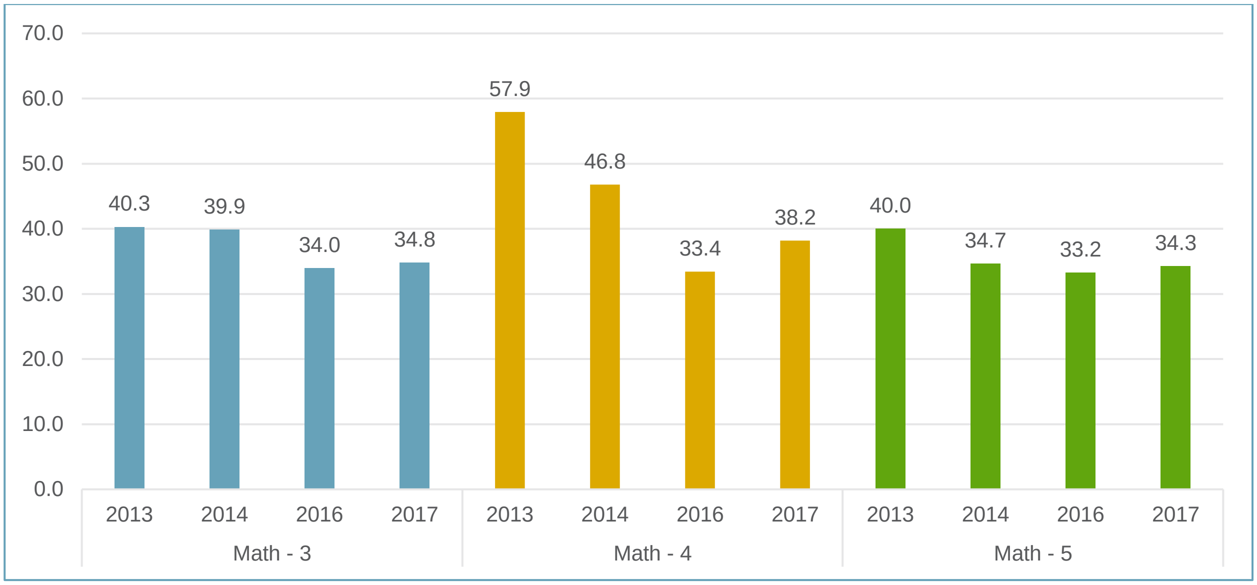


Figure 9. Annual Gap in Scale Score Points between TIF4 and Comparison Schools, STAAR Math 3–5



Science — Grades 5 and 8

Figure 10 illustrates the average scale scores for STAAR Science in grade 5 (blue) and grade 8 (yellow), for comparison (circle) and TIF4 (triangle) schools. Only Grade 8-Comparison did not experience real gains across the grant period. However, the linear trend in science proficiency levels (e.g., Figure 6) obscures the detail in that growth: on average, every grade experienced declines in scale scores between 2014 and 2015, and gains between 2015 and 2017. See **Appendix F Table 2** for each grade level's average scale score, the standard deviation (in parentheses), and the number of students who took the exam each year. Versus the comparison schools, the TIF4 schools decreased the scale score gap by about three percent over the grant period: a decrease of 2.7% or -93.7 points for grade five, and a decrease of 3.4% or -121.1 points for grade 8. The trend shown in **Figure 11** is generally downward over the grant period.

Figure 10. Scale Score Trends for STAAR Science, Grades 5 and 8

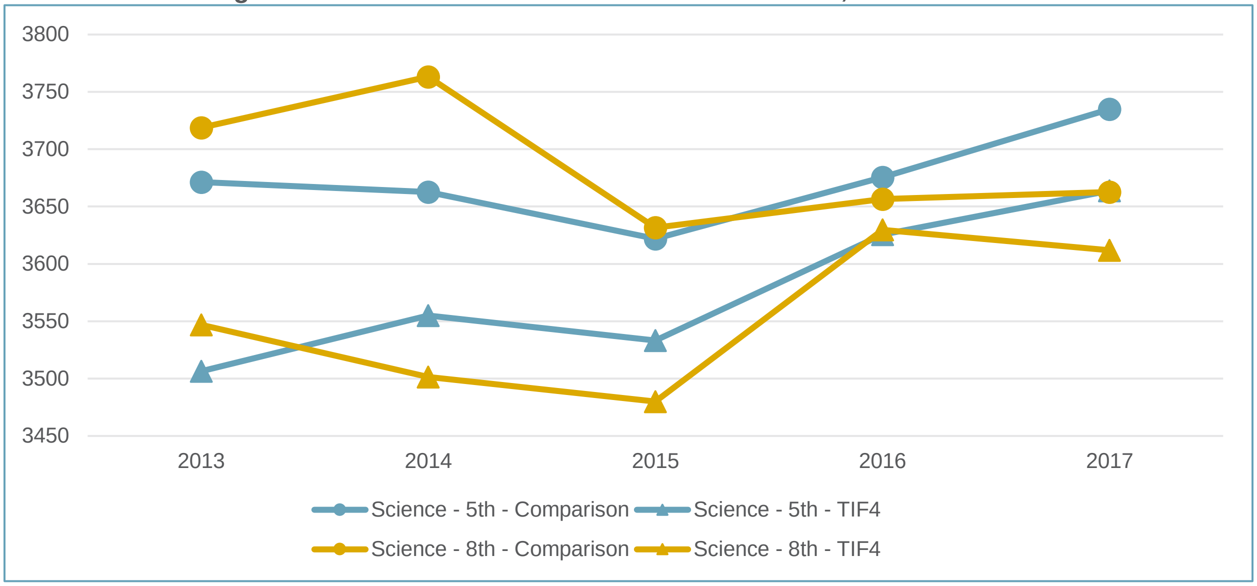
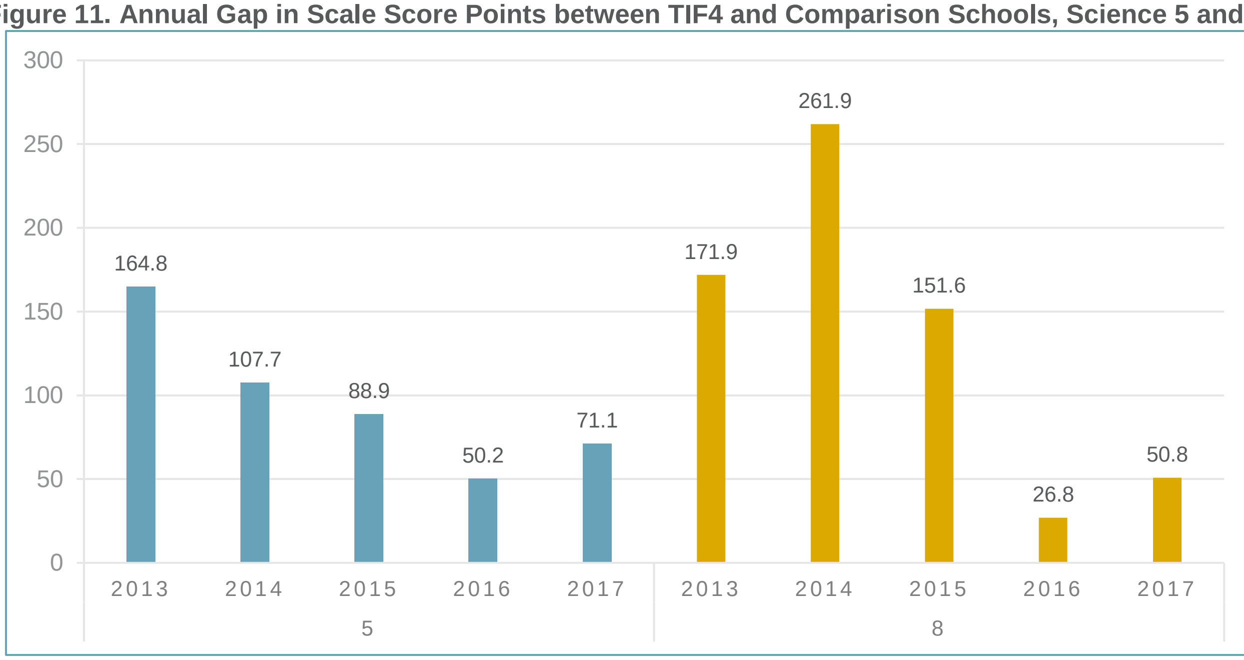


Figure 11. Annual Gap in Scale Score Points between TIF4 and Comparison Schools, Science 5 and 8



Middle — Math, Grades 6 to 8

Figure 12 illustrates the average scale score for STAAR Math during the grant period in grade 6 (blue), grade 7 (yellow), and grade 8 (green) for both comparison (circle) and TIF4 (triangle) schools. At the TIF4 schools, all three grade levels saw an increase in their average scale score during the grant period; at the Comparison schools, both 6th grade and 8th grade saw declines. Note: **Appendix F Table 3** shows each grade level's average scale score, the standard deviation (in parentheses), and the number of students who took the exam each year. The students at TIF4 schools overtook their counterparts, with a gap decrease of 2.8% or -42.9 points for grade six, a decrease of 5.2% or -79.2 points for grade seven, and decrease of 3.9% or -63.4 points for grade eight. In **Figure 13**, the years in which TIF4 students overtook their Comparison counterparts are shown as negative.

Figure 12. Scale Score Trends for STAAR Math, Grades 6–8

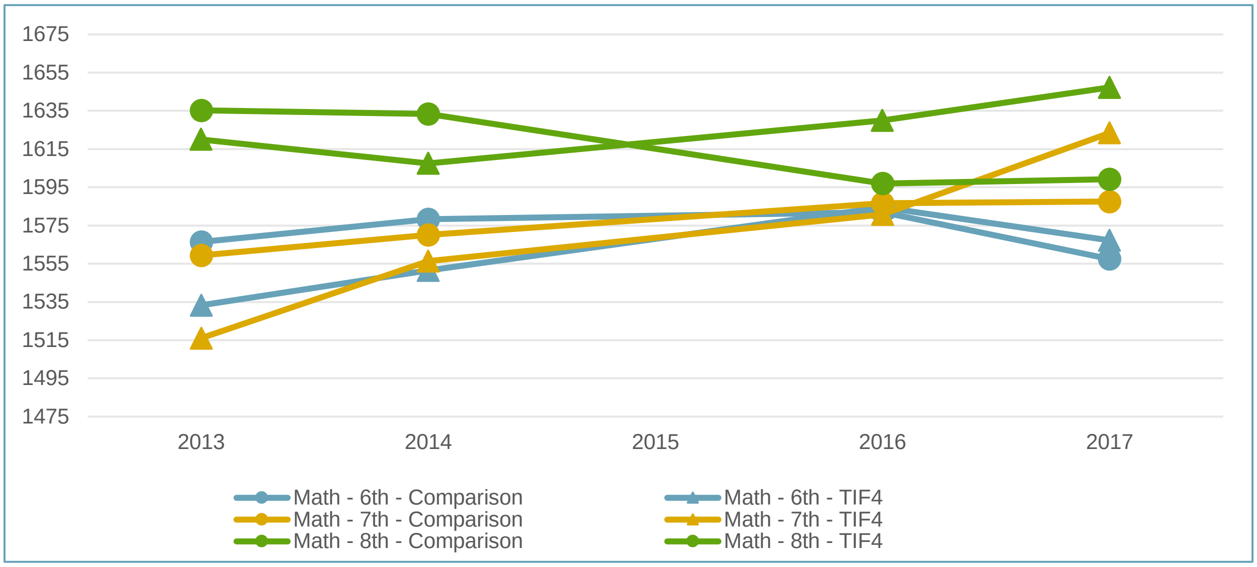
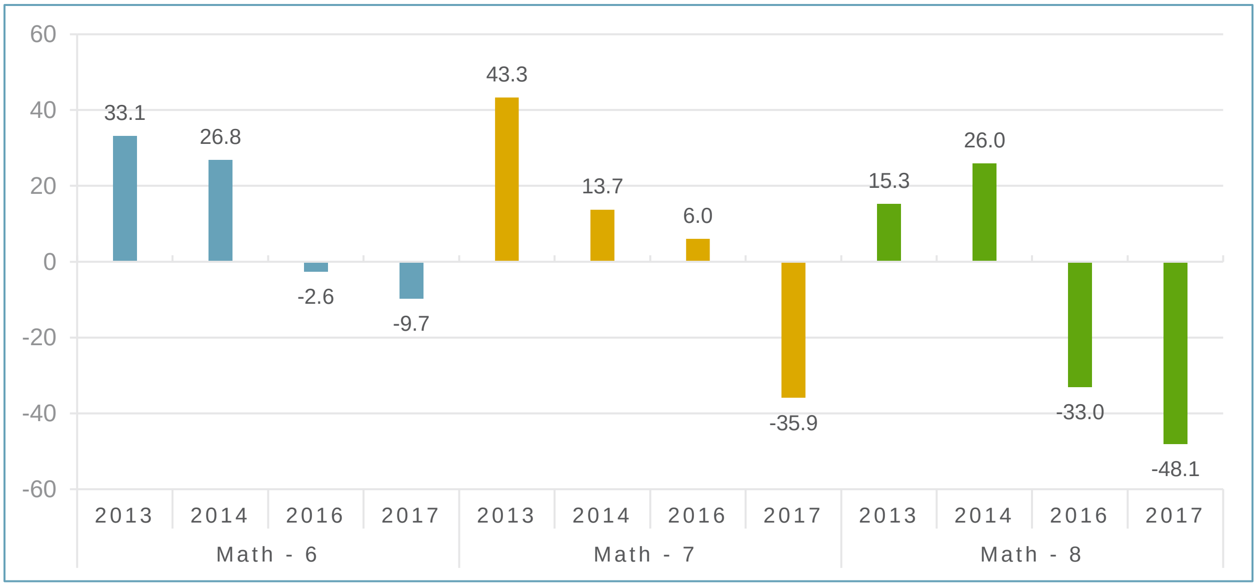


Figure 13. Annual Gap in Scale Score Points between TIF4 and Comparison Schools, Math 6–8



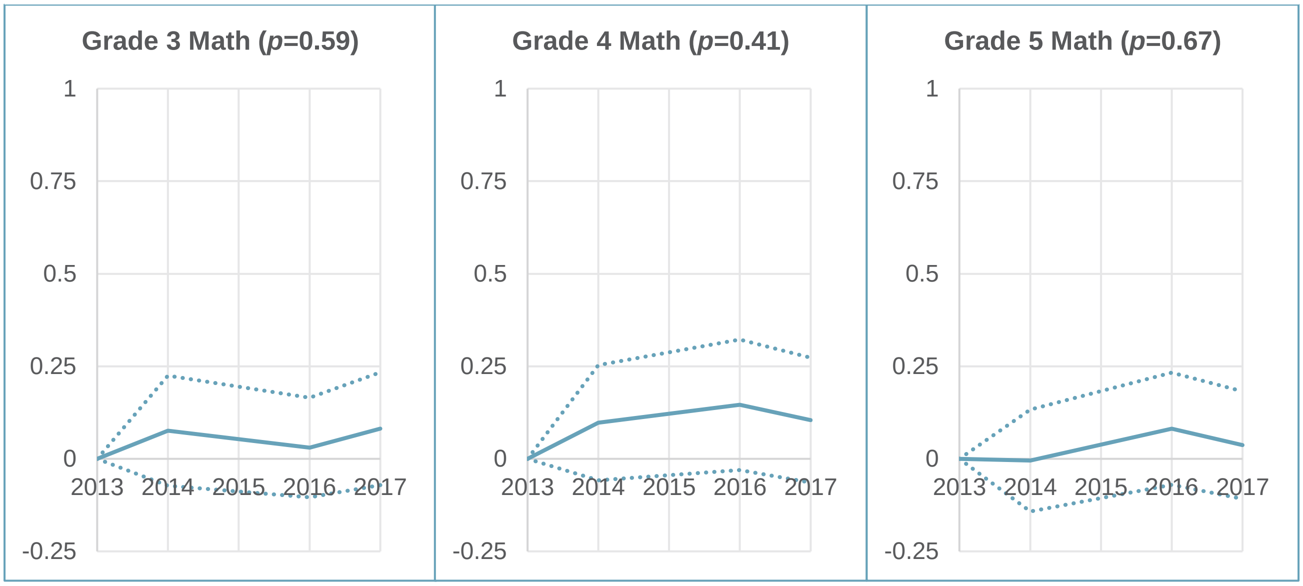
Result 3: Under analysis suited to isolate causal effects, some results are substantive.

The model used to evaluate the impact of the TIF4 program can be expressed as follows:

$$y_{jt} = \beta_{0j} + \beta_{1t} + \beta_{2t} TIF_j + \beta_{3t} X_{jt} + \varepsilon_{jt}$$

where y_{jt} is the average STAAR score in science or mathematics at school *j* in year *t*; β_{0j} is a fixed effect for school *j*; β_{1t} is a fixed effect for year *t*; TIF_j is an indicator variable that equals 1 if school *j* is a participant in the TIF4 program and 0 if school *j* is a comparison school; and X_{jt} is a vector of characteristics of school *j* in year *t*. For more details about this model, see **Appendix E**. Average STAAR scores are normalized by subject, grade, and year using the mean and standard deviation of STAAR scores across students in Texas. The plots present estimates of the year-specific β_{2t} coefficients on TIF_j for 2014, 2015, 2016, and 2017. The effect for the pre-TIF4 baseline year, 2013, is set to zero. These plot the estimated cumulative impact of having been in TIF4 since the start of the program up to that particular year. The impact is on student achievement in that particular year, measured in student-level standard deviations (details in Appendix E). Also included are error bands representing ±2.0 standard errors (an approximately 95% confidence interval). In the table that accompanies the plot is the *p*-value from an *F*-test of the hypothesis that all of the β_{2t} coefficients are equal to zero (**Tables 4, 5, and 6**). This *p*-value is the statistical significance of the results — the probability that the pattern observed would have been produced in the absence of any effect.[v]

Figure 14. Impact of TIF on School's Average STAAR Score, Math 3–5 (in Standard Deviations)



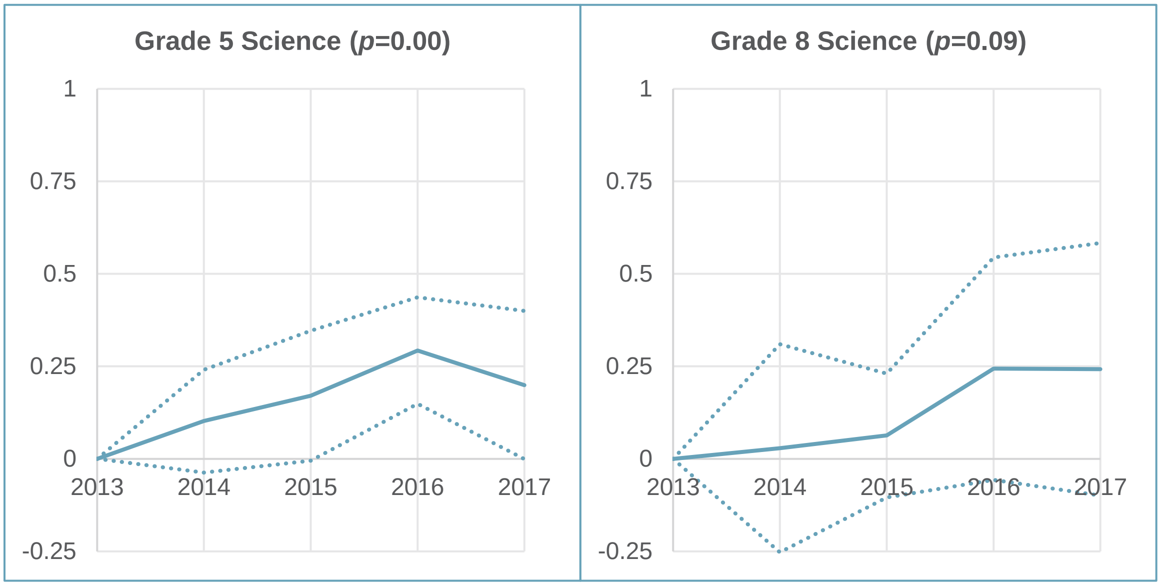
Elementary — Math, Grades 3 to 5

Figure 14 presents the estimated impact of TIF4 over the four years of its implementation on mathematics achievement in grades three through five. In grades three through five, the TIF4 program does not appear to have a large effect on mathematics achievement in any year (see coefficients in **Table 4**), and the estimated impacts are not statistically significant.

Table 4. Impact of TIF4 on Elementary Mathematics: No Large Effects

Grade	Subject	Year	Coefficient (SD)	SE	*p*-value
3	Math	2014	0.08	(0.07)	
3	Math	2016	0.03	(0.07)	
3	Math	2017	0.08	(0.08)	0.596
4	Math	2014	0.10	(0.08)	
4	Math	2016	0.15	(0.09)	
4	Math	2017	0.10	(0.08)	0.412
5	Math	2014	0.00	(0.07)	
5	Math	2016	0.08	(0.08)	
5	Math	2017	0.04	(0.07)	0.675

Figure 15. Impact of TIF on School's Average STAAR Score, Science 5 and 8 (in Standard Deviations)



Science — Grades 5 and 8

Figure 15 presents the impact of the TIF4 program on science achievement in fifth and eighth grades. We can see that, in both grades, the impact of participation in TIF4 accumulates positively over the first three years of implementation (2014 – 2016), and then levels out in the fourth year (2017). The total, cumulative impact of TIF over the course of the four years is an increase in student achievement of about a fifth of a standard deviation in grade five and about a quarter of a standard deviation in grade eight (see **Table 5**).

This is a substantive improvement. For example, with a fifth of a standard deviation of improvement, a student initially at the 25th percentile of achievement would improve to the 32nd percentile; one at the 50th percentile would improve to the 58th percentile; and one at the 75th percentile would improve to the 81st percentile. A quarter standard deviation improvement moves a student from the 25th percentile to the 34th percentile, from the 50th percentile to the 60th percentile, and from the 75th percentile to the 82nd.

Table 5. Impact of TIF4 on STAAR Science: Substantive Improvement

Grade	Subject	Year	Coefficient (SD)	SE	*p*-value
5	Science	2014	0.10	(0.07)	
5	Science	2015	0.17	(0.09)	
5	Science	2016	0.29	(0.07)	
5	Science	2017	0.20	(0.10)	0.003
8	Science	2014	0.03	(0.13)	
8	Science	2015	0.06	(0.08)	
8	Science	2016	0.24	(0.14)	
8	Science	2017	0.24	(0.16)	0.091

In fifth-grade science, the improvement in science STAAR scores among students in TIF4 schools is statistically significant. The evidence in eighth-grade science is less compelling, even given the substantive point estimate of the impact of the TIF4 program. This is because the sample of schools is sufficiently small that even a substantive measured impact is not necessarily statistically significant. See **Appendix E** for additional technical details about the model specifics for fifth grade science and eighth grade science.

Middle — Math, Grades 6 to 8

In contrast to the findings for grades 3 to 5, a more substantive effect of TIF4 is measured in mathematics in grades six, seven, and eight (see **Table 6** and **Figure 16**). See Appendix E for additional technical details. As shown in **Figure 16**, the point estimates suggest a substantive impact in sixth-grade mathematics — a cumulative impact over the four years of about a fifth of a standard deviation. These estimates are not sufficiently precise to be statistically significant at conventional levels (*p*=0.42).

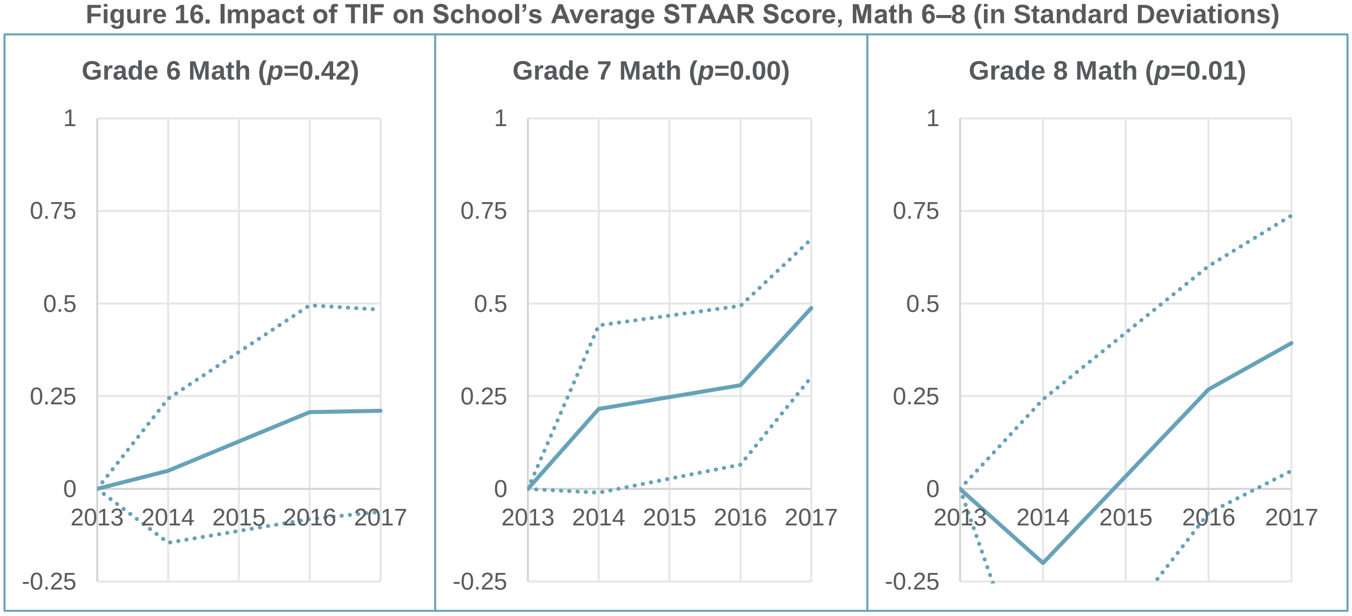
In seventh grade mathematics, the TIF4 program has an immediate effect of about one-fifth of a standard deviation of student achievement, which increases slightly to about a quarter of a standard deviation in the third year of TIF4. In the fourth year, the cumulative impact of the TIF4 program ticks upward to about half of a standard deviation of student achievement. A half-standard-deviation increase would improve the achievement of a student at the 25th percentile to the 43rd percentile; that of a student at the 50th percentile to the 69th percentile; and that of a student at the 75th percentile to the 88th percentile.

In eighth grade mathematics, we see achievement dip among TIF4 schools relative to non-TIF4 schools in the first year, only to recover in the third year to a level of about one-quarter of a standard deviation higher among TIF4 schools than among non-TIF4 schools, and to further improve to about four-tenths of a standard deviation higher in the fourth year. This suggests that, while we do not measure any positive immediate effect in the first year, we measure a substantive, significant cumulative effect by the end.

Table 6. Impact of TIF on STAAR Math 6–8: Substantive Improvement

Grade	Subject	Year	Coefficient (SD)	SE	*p*-value
6	Math	2014	0.05	(0.10)	
6	Math	2016	0.21	(0.14)	
6	Math	2017	0.21	(0.14)	0.424
7	Math	2014	0.22	(0.11)	
7	Math	2016	0.28	(0.11)	
7	Math	2017	0.49	(0.09)	0.001
8	Math	2014	-0.20	(0.22)	
8	Math	2016	0.27	(0.17)	
8	Math	2017	0.39	(0.17)	0.011

Figure 16. Impact of TIF on School's Average STAAR Score, Math 6–8 (in Standard Deviations)



Conclusion

Supporting the federal priority to improve STEM education, the fourth cohort of the Teacher Incentive Fund grant competition (TIF4) included special consideration for projects that would identify, develop, and utilize master teachers as leaders of STEM education. As a comprehensive intervention, the TIF4 approach to STEM education in HISD supported program activities that reached students, teachers, and school-wide systems — in short, the key programmatic aspects necessary to impact student outcomes as outlined in Figure 2 (Kraft, Blazar, & Hogan, 2016).

Kraft, Blazar, and Hogan (2018) found that coaching generally resulted in only weak improvements to student achievement (0.11 SD), because generally the changes to instructional practice were not sufficient to affect student outcomes. The evidence presented in this report strongly suggests that a school's participation in the TIF4 grant did impact teachers' instructional practice strongly enough for a causal inference analysis to detect subsequent changes in student outcomes.

Indeed, these findings comprise compelling evidence that the coaching-centered TIF4 STEM intervention caused substantive improvement in four areas of student achievement: fifth grade science (0.20 SD, $p<0.00$), eighth grade science (0.24 SD, $p<0.09$), seventh grade mathematics (0.49 SD, $p<0.00$), and eighth grade mathematics (0.39 SD, $p<0.01$). The evidence for TIF4 impact on sixth grade mathematics was also strong (0.21 SD) but not statistically significant at any traditional level of certainty ($p<0.42$). Notably, the TIF4 results for elementary mathematics were more in line with those found in Kraft, Blazar, and Hogan (2018): In grades three through five, the TIF4 program did not appear to have a large effect on mathematics achievement cumulatively or in any single year, and the estimated impacts are not statistically significant. This analysis did not include a specific investigation into possible reasons for the difference between elementary and middle school math TIF4 outcomes.

On the whole, this report suggests that the complex programmatic aspects of the TIF4 program produced substantive and reproducible results for student achievement through human capital strategies. Additional reporting in this series will investigate human capital outcomes for science and math teachers at the TIF4 project schools — including whether the implementation of TIF4 human capital strategies were meaningfully different between the elementary (3–5) level and middle grades (6–8).

Endnotes

(i) Under Section §8101(21)(A) of the Every Student Succeeds Act of 2015 (ESSA), “the term ‘evidence-based’, when used with respect to a State, local educational agency, or school activity, means an activity, strategy, or intervention that — (i) demonstrates a statistically significant effect on improving student outcomes or other relevant outcomes based on — ‘‘(I) strong evidence from at least 1 well-designed and well-implemented experimental study; (II) moderate evidence from at least 1 well-designed and well-implemented quasi-experimental study; or ‘‘(III) promising evidence from at least 1 well-designed and well-implemented correlational study with statistical controls for selection bias; or ‘‘(ii)(I) demonstrates a rationale based on high-quality research findings or positive evaluation that such activity, strategy, or intervention is likely to improve student outcomes or other relevant outcomes.”

(ii) In their 2016 Non-Regulatory Guidance document “Using Evidence to Strengthen Education Investments”, the Office of Elementary and Secondary Education provides the following definition and example for the term: “A quasi-experimental study (as known as a quasi-experimental design study or QED)… means a study using a design that attempts to approximate an experimental design by identifying a comparison group that is similar to the treatment group in important respects. These studies, depending on design and implementation, can meet What Works Clearinghouse Evidence Standards [for high-quality research]. An example of a QED is a study comparing outcomes for two groups of classrooms matched closely on the basis of student demographics and prior mathematics achievement, half of which are served by teachers who participated in a new mathematics professional development (PD) program, and half of which are served by other teachers. This study uses a nonequivalent group design by attempting to match or statistically control differences between the two groups.” (OESE, 2016, pg. 11)

(iii) In their Procedures Handbook, the What Works Clearinghouse provides the following rationale and definition: “In general, to improve the comparability of effect size estimates across studies, the WWC uses student-level standard deviations when computing effect sizes, regardless of the unit of assignment or the unit of intervention. … For continuous outcomes, the WWC has adopted the most commonly used effect size index, the standardized mean difference known as Hedges’ g, with an adjustment for small samples. It is defined as the difference between the mean outcome for the intervention group and the mean outcome for the comparison group, divided by the pooled within-group standard deviation of the outcome measure.” (IES, 2017b, pg. 14)

(iv) Relying on the 2013 technical report on the STAAR scale scores from the Texas Education Agency, the decision was made to combine results for both English and Spanish into a single grade-level mean scale score. From the 2013 STAAR Vertical Scale Technical Report from the TEA’s Student Assessment Division: “Under Texas Education Code (TEC) §39.036, the Texas Education Agency (TEA) is required to develop a vertical scale for assessing student performance in grades 3–8 for reading and mathematics. A vertical scale is a scale score system that allows for direct comparison of student test scores across grade levels within a content area. Vertical scaling refers to the process of placing test scores that measure similar content areas but at different grade levels onto a common scale. A vertical scale was developed for the following grades and subjects: STAAR English grades 3–8 mathematics, STAAR English grades 3–8 reading, STAAR Spanish grades 3–5 reading. Although there is a Spanish version of STAAR mathematics assessments in grades 3–5, a separate vertical scale was not developed because the same scale is used for both language versions. Use of the same scale is possible because Spanish mathematics items are transadapted from the English items. Spanish reading passages and items are uniquely developed to maintain the authenticity of the Spanish assessment.” (Student Assessment Division, 2013, pg. 3)

(v) From the American Statistical Association (ASA): “Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value… The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions hold.” (Wasserstein & Lazar, 2016)

References

Achievement Performance Indicators, Texas Education Code § 39.053. Retrieved from http://www.statutes.legis.state.tx.us/Docs/ED/htm/ED.39.htm#39.053

Adoption and Administration of Instruments, Texas Education Code § 39.023. Retrieved from http://www.statutes.legis.state.tx.us/Docs/ED/htm/ED.39.htm#39.023

Chang, Y. T., & Stevens, C. J. (2014). *Identification of Homogenous School Clusters* (Memorandum to Chief School Officers). Houston ISD: Department of Research and Accountability.

Every Student Succeeds Act of 2015, Pub. L. No. 114–95. 20 U.S.C. § 6301 *et. seq.* U.S. Government Printing Office. Retrieved from https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.htm

Hanita, M., Ansel, D., & Shakman, K. (2017). Matched-Comparison Group Design: An Evaluation Brief for Educational Stakeholders. Retrieved from www.tifcommunity.org

Houston Independent School District (HISD). (2012, July). PR/Award # S374B120011: Application for Grants under the TIF Competition with a Focus on STEM CFDA # 84.374B. Retrieved from www2.ed.gov/programs/teacherincentive/

HISD Communications. (2012, September 27). HISD Wins $15.9 Million Teacher Incentive Fund Grant. Retrieved from www.houstonisd.org/HISDmedia

Institute of Education Sciences (IES). (2017a). *What Works Clearinghouse™ Procedures Handbook (Version 4.0)*. U.S. Department of Education: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/wwc/handbooks

Institute of Education Sciences (IES). (2017b). *What Works Clearinghouse™ Standards Handbook (Version 4.0)*. U.S. Department of Education: National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/wwc/handbooks

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Review of Educational Research*. https://doi.org/10.3102/0034654318759268

Kraft, M. A., Blazar, D., & Hogan, D. (2016, November). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence [Working Paper]. Brown University. Retrieved from http://scholar.harvard.edu/mkraft/publications/

McCarley, K., Ye, R., Selig, H., & Stevens, C. J. (2014). *STAAR Scale Score Standard Deviation Report 2013–2014* (Internal Report). Houston ISD: Department of Research and Accountability.

McCarley, K., Ye, R., Selig, H., & Stevens, C. J. (2013). *STAAR Scale Score Standard Deviation Report 2012–2013* (Internal Report). Houston ISD: Department of Research and Accountability.

Miller, J., Adrien, R., Harmon, B., Koppich, J., Potemski, A., & Yoder, M. (2015). The Evolution of the Teacher Incentive Fund (TIF) Program. Retrieved from https://www.tifcommunity.org

Murnane, R. J., & Willett, J. B. (2011). Experimental research when participants are clustered within intact groups. In *Methods matter: improving causal inference in educational and social science research*. Oxford: Oxford Univ. Press.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107–110. 20 U.S.C. § 6301 *et. seq.* U.S. Government Printing Office. Retrieved from https://www.gpo.gov/fdsys/pkg/PLAW-107publ110/html/PLAW-107publ110.htm

Office of Elementary and Secondary Education (OESE). (2012a, June). 2012 Application for New Grants under the Teacher Incentive Fund (TIF) Program CFDA 84.374A and 84.374B. U.S. Department of Education. Retrieved from https://www2.ed.gov/programs/teacherincentive/2012.html

Office of Elementary and Secondary Education (OESE). (2012b, September). Teacher Incentive Fund FY 2012 Funded Applications. U.S. Department of Education. Retrieved from https://www2.ed.gov/programs/teacherincentive/awards.html

Office of Elementary and Secondary Education (OESE). (2016). Using Evidence to Strengthen Education Investments (Significant guidance). U.S. Department of Education. Retrieved from https://www2.ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf

Office of Innovation and Improvement (OII). (2015). TIF4 Profile Summaries. Presented at the Teacher Quality Programs Project Directors Meeting (May 18—19, 2015), Vienna, VA: U.S. Department of Education.

Price, L. E., Provencher, S. A., & Stevens, C. J. (2018). Teacher Incentive Fund STEM Grant in Houston ISD: A Descriptive Overview. Houston ISD: Department of Research and Accountability. Retrieved from http://www.houstonisd.org/

Price, L. E., & Stevens, C. J. (2017). Teacher Incentive Fund, Cohort 3. Looking Back, Around, And Ahead: HISD's Viewfinder for Teacher and Leader Effectiveness. Houston ISD: Department of Research and Accountability. Retrieved from http://www.houstonisd.org/

Reeves, R., Bigner, Z., & Stevens, C. J. (2017). *STAAR Scale Score Standard Deviation Report 2016–2017* (Internal Report). Houston ISD: Department of Research and Accountability.

Reeves, R., Bigner, Z., & Stevens, C. J. (2016). *STAAR Scale Score Standard Deviation Report 2015–2016* (Internal Report). Houston ISD: Department of Research and Accountability.

Reeves, R., Carney, D., & Stevens, C. J. (2015). *STAAR Scale Score Standard Deviation Report 2014–2015* (Internal Report). Houston ISD: Department of Research and Accountability.

Reeves, R., McCarley, K., Mosier, V., & Carney, D. (2015). A Better Picture of Poverty - Campus "Risk Loads" (Educational Program Report). Houston ISD: Department of Research and Accountability.

Shakman, K., Wogan, D., Finster, M., & Milanowski, A. (2016). *Teacher Incentive Fund (TIF) Evaluation Summit Summary* (Conference Synthesis). Office of Innovation and Instruction: U.S. Department of Education. Retrieved from https://www.tifcommunity.org/

Sondhi, D., Huang, C.-H., McCarley, K., Sage, E. L., & Stevens, C. J. (2014). State of Texas Assessments of Academic Readiness (STAAR) Performance, Grades 3-8, Spring 2014. Houston ISD: Department of Research and Accountability. Retrieved from www.houstonisd.org/Page/63696

Student Assessment Division. (2013). *STAAR Vertical Scale Technical Report*. Texas Education Agency. Retrieved from https://tea.texas.gov/

Student Assessment Division. (2017). *STAAR Performance Labels and Policy Definitions*. Texas Education Agency. Retrieved from https://tea.texas.gov/student.assessment/staar/performance-standards/

Tannenbaum, D. (2013). *COHEND: Stata module to compute Cohen's d*. Boston College Department of Economics. Retrieved from https://ideas.repec.org/c/boc/bocode/s457235.html

Public Education Information Management System (PEIMS). (2013). 2012–2013 Financial Actual Data. Texas Education Agency: State Funding Division. Retrieved from https://tea.texas.gov/Finance_and_Grants/State_Funding/State_Funding_Reports_and_Data/

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on *p*-Values: Context, Process, and Purpose. The American Statistician, 70(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Weiss, I. (2013, August). *Developing a Theory of Action for Master STEM Teacher Programs*. STEM Grantee Webinar Series. Retrieved from https://tifstemcommunity.org

Zawaiza, T., & Robinson, V. (2014, July). Selecting and Supporting Effective STEM Master Teachers. Presented at the TIF4 Annual Project Directors' Meeting (STEM), Bethesda, MD.

Appendix A: Teacher Incentive Fund

Since established by an Appropriations Act in 2006, the Teacher Incentive Fund (TIF) competitive grant program in the U.S. Department of Education (the Department) has supported human capital strategies for teachers and school leaders, “to ensure that students attending high-poverty schools have better access to effective teachers and principals, especially in hard-to-staff subject areas” such as science and math. While the specific programming supported through the TIF grant program has evolved since 2006 (Miller et al., 2015), TIF projects are supported by the Department to develop and implement sustainable performance-based compensation systems (PBCSs) for teachers, principals, and other personnel in high-need schools in order to increase educator effectiveness and student achievement. HISD was awarded over $43 million as part of the first and third cohorts of TIF grantees – $11.8 million in 2006, and $31.3 million in 2010. A recap of these program activities is available on HISD’s website (Price & Stevens, 2017).

In September 2012, HISD was awarded a TIF grant for $15.9 million over five years (OESE, 2012b) — one of just six STEM projects funded among the fourth cohort of awards (TIF4-STEM): HISD, plus Calcasieu Parish (LA), National Institute for Excellence in Teaching (IA), Orange County (FL), Washoe County (NV), and the South Carolina Department of Education.

These grantees committed to the two Absolute Priorities required of all TIF grantees, as well as a third Priority that was specific to STEM programming:

- **Priority 1 (all grantees):** “An LEA-wide human capital management system (HCMS) with educator evaluation systems at the center that (a) is aligned with the local education agency's (LEA's) vision of instructional improvement and (b) uses information generated by the evaluation system to inform key human capital decisions, such as recruitment, hiring, placement, dismissal, compensation, professional development, tenure, and promotion.”
- **Priority 2 (all grantees):** “An LEA-wide educator evaluation system based, in significant part, on student growth. The frequency of evaluation must be at least annually and the evaluation rubric should include at least three performance levels and (a) two or more observations during each evaluation period, (b) student growth for the evaluation of teachers at the classroom level, and (c) additional factors determined by the LEA. In addition, the evaluation system must generate an overall evaluation rating based, in significant part, on student growth and the evaluation system must be implemented within the timeframe specified in Priority 2.”
- **Priority 3 (STEM grantees):** “Improving STEM achievement by developing a corps of skilled STEM master teachers by providing additional compensation to teachers who (a) receive an overall evaluation effectiveness rating of effective or higher under the evaluation system, (b) are selected based on criteria that are predictive of the ability to lead other teachers, (c) demonstrate effectiveness in one or more STEM subjects, and (d) accept STEM-focused career ladder positions. In addressing this priority, each LEA needs to identify and develop the unique competencies that, based on evaluation information or other evidence, characterize effective STEM teachers. Projects also need to identify hard-to-staff STEM subjects and use the HCMS to attract effective teachers, leverage community support and expertise to inform the implementation of its plan, ensure that financial and non-financial incentives are adequate to attract and retain persons with strong STEM skills in high-need schools, and ensure that students have access to and participate in rigorous and engaging STEM coursework.”

See http://www2.ed.gov/programs/teacherincentive/2012-374ab.pdf for the full text of the application package for TIF4 (OSEA, 2012a).

Appendix B: Identification of Homogenous School Clusters

Excerpt of Analysis by Dr. Yu-Ting Chang of the HISD Research and Accountability Department.

MEMORANDUM April 9, 2014

TO: Chief School Officers

FROM: Carla J. Stevens
Assistant Superintendent, Research and Accountability

SUBJECT: **IDENTIFICATION OF HOMOGENOUS SCHOOL CLUSTERS**

The Department of Research and Accountability was asked to perform a non-hierarchical cluster analysis of elementary, middle, and high schools using demographic data. The purpose of this analysis was to develop clusters, or groups, of comparable schools, for the purpose of comparing student performance on the STAAR reading and mathematics assessments for elementary and middle schools, and on the STAAR EOC assessments for high schools within each cluster.

A non-hierarchical, partitioning model, formally known as "K-Means," was performed using STATA (a data and statistical software program). K-Means is a multivariate learning model that processes and classifies an assortment of fairly homogenous variables into sub populations known as "clusters." Schools were then classified into one of several clusters, developed at each level (elementary, middle, and high), based on the relationships between the schools on each of the variables.

In this analysis, the nine variables used were: enrollment, percent economically disadvantaged, percent at risk, percent zoned, percent mobility, percent ELL, percent African American, percent Hispanic, and percent White.

Due to the algorithmic structure of K-Means, each of the nine variables had to be standardized to prevent unequal weighting. For example, if enrollment was not standardized, it would have a much larger scale compared to the other variables, leading to inaccurate cluster results. […]

A total of 35 middle schools were analyzed in this analysis, resulting in six school clusters. A total of 161 elementary schools were analyzed in this, analysis resulting in eight school clusters. A total of 214 HISD schools and 5 NFISD [North Forest ISD] schools were assigned to a cluster based on the characteristics, or pattern of relationships, each school exhibited on the nine variables.

Some schools were omitted from the analysis for various reasons, which include: no mobility rate, no zoned rate, multi-level grade schools, early childhood centers, and specialized schools. […]

Should you have further questions, please contact my office in the Department of Research and Accountability at (713) 556-6700.

cc: Superintendent's Direct Reports, Chief School Officers
School Support Officers, School Office Directors
Lupita Hinojosa

Appendix C: 148 Schools in Sample, by Homogeneous Cluster and Treatment/Comparison Assignment

	Treatment	Comparison		Treatment	Comparison
56 Elementary Schools in Cluster E1					
Anderson		1	Lewis		1
Barrick		1	Lyons		1
Benavidez		1	Martinez, R.		1
Benbrook		1	McNamara		1
Berry		1	Moreno		1
Bonham		1	Neff		1
Bonner		1	Northline		1
Braeburn	1		Park Place		1
Brookline		1	Patterson		1
Burbank		1	Pilgrim Acad.		1
Coop		1	Piney Point		1
Crespo		1	Port Houston		1
Cunningham		1	Robinson		1
De Chaumes		1	Rodriguez		1
DeAnda		1	Rucker		1
Durkee	1		Sanchez		1
Eliot	1		Scarborough		1
Franklin		1	Scroggins		1
Gallegos		1	Seguin		1
Golfcrest		1	Shearn		1
Harris JR		1	Sherman		1
Harris RP		1	Southmayd	1	
Henderson JP		1	Stevens		1
Herrera	1		Sutton		1
Hines-Caldwell		1	Tijerina		1
Janowski		1	Wainwright		1
Kennedy		1	White		1
Ketelsen		1	Whittier		1
			No. in Cluster E1	**5**	**51**
10 Elementary Schools in Cluster E2					
Almeda		1	Garden Villas		1
Cornelius		1	Lantrip		1
Elrod		1	Law	1	
Emerson		1	Roosevelt		1
Garcia		1	Tinsley		1
			No. in Cluster E2	**1**	**9**
21 Elementary Schools in Cluster E3					
Briscoe		1	Helms		1
Browning		1	Jefferson		1
Burnet		1	Looscan	1	
Cage		1	Love		1
Carrillo		1	Memorial		1
Crockett		1	Pugh	1	
Davila		1	Red		1
DeZavala		1	Rusk		1
Durham		1	Sinclair		1
Field		1	Wharton		1
Gregg		1	**No. in Cluster E3**	**2**	**19**

	Treatment	Comparison		Treatment	Comparison
25 Elementary Schools in Cluster E5					
Askew		1	Isaacs		1
Bastian		1	Kelso		1
Bell		1	Martinez, C.		1
Bruce		1	Milne	1	
Cook		1	Montgomery	1	
Daily		1	Paige		1
Dogan		1	Peck		1
Foerster		1	Shadowbriar		1
Fondren		1	Smith, K.		1
Grissom	1		Valley West		1
Gross		1	Walnut Bend		1
Highland Heights		1	Windsor Village		1
Hobby		1	**No. in Cluster E5**	**3**	**22**
24 Elementary Schools in Cluster E6					
Alcott		1	MacGregor		1
Atherton		1	Mading	1	
Blackshear	1		McGowen	1	
Burrus	1		Osborne		1
Codwell	1		Pleasantville		1
Foster	1		Reynolds		1
Frost		1	Ross	1	
Hartsfield		1	Thompson		1
Henderson NQ		1	Wesley		1
Kashmere Gardens		1	Whidby		1
Lockhart		1	Woodson PK-8		1
Longfellow		1	Young		1
			No. in Cluster E6	**7**	**17**
7 Middle Schools in Cluster M1					
Attucks		1	Thomas		1
Cullen		1	Welch		1
Fleming	1		Williams		1
Key		1	**No. in Cluster M1**	**1**	**6**
5 Middle Schools in Cluster M6					
Deady		1	Long		1
Fondren	1		Sugar Grove	1	
Henry		1	**No. in Cluster M6**	**2**	**3**

Number of Schools, by Level and by Group

	Treatment	Comparison	*Level Total*
Elementary	18	114	*132*
Middle	3	13	*16*
Group Total	**21**	**127**	***148***

As outlined in Step 2 of the section “Research Design”, three schools that participated in the TIF4 grant programming were excluded from the analytic sample in this study: Garden Oaks Montessori and Wilson Montessori (K–8) were both dropped from the analytic sample because they did not have comparable schools in HISD. Dodson Elementary was dropped from the sample because it did not have three years of student data: it was closed after 2013, and its zoned students incorporated into the nearby TIF4 school Blackshear Elementary.

Appendix D: Using "A Better Picture of Poverty" to Assess Sample Balance

"Leaders at every level of the school system are being challenged to think and act differently to address the effects of income inequality on academic performance. The majority of schools within Houston ISD are located in high-poverty areas, so it is important to understand which may need the most help – and what kind of help would be most useful. However, simple proxies for poverty, like the proportion of students who receive free and reduced lunch, fail to capture the volume and nature of the challenges that many Houston schools face. Inspired by the November 2014 research report, *A Better Picture of Poverty*, by the Center for New York City Affairs, we identified 23 school and neighborhood risk factors that contribute to chronic absenteeism and low student performance. When the factors are displayed using [color-coding] there emerges a very clear picture of both the kinds of and the volume of educational disadvantage associated with that location; a "heat map" of educational disadvantage."

Excerpt, Campus Risk Load Profiles Fall 2015 (Reeves, McCarley, Mosier, & Carney, 2015)

Risk Factors for Chronic Absenteeism at the TIF4 Project Schools

Overall, the 2015 *Risk Load* report showed two things – that HISD schools are facing complex issues, but that some schools are showing success even with a heavy "risk load." The same is true of the TIF4 project schools. Figure B-1 shows the "heat map" of each school's total risk factors, chronic absenteeism, and the 22 factors associated with it. The sources and definitions of these variables are found in the rest of this Appendix. The median number of Risk Factors facing a TIF4 school is 11, compared to a median of just 8 for all other HISD schools serving grades K–8.

Figure D-1. 2015 Family, School, and Neighborhood Risk Factors for Chronic Absenteeism

Total Number of Risk Factors	2015 Rating	School Name	Chronic Absenteeism	Free Lunch Eligible	Students in Temp. Housing	Special Ed. Students	Black or Hispanic	ELL Students	Students Not Ident. as GT	Immigrant Students	Refugee Students	Principal Turnover	Teacher Turnover	Teacher Vacancy	Student Mobility	Student Transfers	Student Suspensions	Student Safety Issues	CPS Involvement	Child Poverty	Poverty Rate	HS Education or Below	Lack of Adult Professionals	Male Unemployment	Public Housing	Homeless Shelter
18	IR	Sugar Grove MS	-0.100	0.100	1.270	-0.566	-1.089	-1.268	-1.458	-0.758	-1.829	-2.000	-1.079	-2.000	-0.459	-2.000	-0.100	-1.169	-1.143	-0.431	-0.313	-0.308	0.568	0.696	2.000	-2.000
	MET	Fleming MS	0.420	-1.557	-1.105	-0.882	-1.907	0.629	-1.648	1.656	-0.145	-0.111	-0.210	-1.219	-0.392	-2.000	-1.605	0.100	2.000	-1.814	-2.000	-1.119	-0.607	-0.906	2.000	-2.000
17	IR	Blackshear ES	-0.567	-0.390	0.100	-0.143	-1.626	1.152	-1.160	1.793	-0.173	-2.000	-2.000	2.000	-0.748	-2.000	-2.000		2.000	-1.407	-1.428	-0.100	-1.526	-0.823	-2.000	-2.000
		Fondren MS	-0.469	0.313	0.312	-0.286	-0.308	-0.920	-1.690	0.335	-1.890	-0.111	-1.778	-0.810	-0.596	-2.000	-1.239	-0.179	2.000	-0.769	-0.568	0.201	-0.377	-0.314	2.000	-2.000
15	MET	Grissom ES	0.100	-1.467	-2.000	0.321	-1.528	-0.415	-0.573	1.186	2.000	-1.216	-1.507	2.000	0.251	-2.000	-0.100		-2.000	-0.289	-0.238	-0.490	-1.070	-0.606	2.000	2.000
		McGowen ES	0.112	-0.735	-0.657	-0.977	-1.921	1.856	-0.857	1.867	2.000	-0.100	-0.738	2.000	0.111	-2.000	-1.258		-2.000	-0.200	-0.100	-0.838	-2.000	-2.000	2.000	2.000
14	IR	Mading ES	-1.061	-0.100	-0.305	-1.534	-1.784	1.530	-1.509	1.906	2.000	-1.216	-1.138	2.000	-0.100	2.000	-1.611		2.000	-0.500	-0.143	-0.649	-2.000	-2.000	2.000	2.000
13	IR	Foster ES	-1.963	-1.592	0.709	-1.024	-1.764	1.973	-1.775	2.000	2.000	-0.100	0.515	2.000	-0.232	-2.000	-2.000		2.000	-1.054	-0.939	-0.421	-1.539	-0.668	2.000	2.000
	MET	Looscan ES	-0.460	-0.986	2.000	-1.163	-1.843	-0.237	-0.596	1.633	2.000	-0.100	-2.000	2.000	0.109	-2.000	-0.296		2.000	-1.607	-0.921	-0.956	-0.882	0.273	2.000	2.000
		Durkee ES	0.678	-0.641	2.000	-1.441	-1.685	-0.439	-0.612	1.671	2.000	-0.100	-0.151	2.000	0.361	-2.000	0.263		2.000	-1.319	-1.653	-1.726	-0.149	-0.370	2.000	2.000
		Pugh ES	0.684	-1.446	2.000	-1.116	-1.784	-0.285	-0.787	0.346	2.000	-0.100	-0.873	2.000	0.510	2.000	-0.242		2.000	-0.542	-0.407	-1.422	-0.502	-0.565	2.000	2.000
11	IR	Burrus ES	-0.100	-0.850	-0.101	0.738	-1.587	1.867	-1.266	1.630	2.000	2.000	-0.668	2.000	-0.148	2.000	0.773		2.000	-0.764	-0.802	-0.459	-0.100	-0.509	2.000	2.000
	MET	Eliot ES	0.875	-1.415	1.512	-1.395	-1.843	-0.189	-0.435	0.685	2.000	-0.100	-0.873	2.000	-0.310	-2.000	0.846		2.000	0.100	-0.167	-1.808	0.815	1.014	2.000	2.000
10	IR	Montgomery ES	-0.770	-0.349	-2.000	-0.097	-1.626	0.143	-0.647	0.499	1.383	2.000	0.284	2.000	0.317	-2.000	-0.272		2.000	0.175	0.303	-0.489	-0.478	-0.248	2.000	2.000
		Ross ES	0.739	-0.464	0.100	0.599	-1.626	1.072	-0.751	2.000	2.000	-1.216	-0.100	2.000	0.100	2.000	0.623		2.000	-0.926	-0.825	-0.985	-0.302	-0.249	2.000	2.000
	MET	Southmayd ES	0.177	-1.153	1.607	-0.375	-1.862	-0.487	0.318	1.837	2.000	-0.100	0.529	2.000	0.347	-2.000	1.765		2.000	-0.100	-0.252	-1.051	0.167	0.157	-2.000	2.000
		Herrera ES	0.703	-1.760	2.000	-0.560	-1.508	-0.424	-0.488	1.799	2.000	2.000	-0.181	-1.195	0.872	2.000	1.935		2.000	-0.201	-0.514	-1.645	0.969	1.051	2.000	2.000
		Braeburn ES	1.426	-1.728	1.091	0.182	-1.705	-1.593	-0.477	-0.966	-0.339	2.000	0.812	2.000	0.624	-2.000	0.318		2.000	-1.224	-1.151	-0.362	0.299	0.680	2.000	2.000
9	IR	Milne ES	-1.084	-1.070	2.000	-0.606	-1.390	0.831	-0.855	1.670	2.000	-0.100	-0.380	2.000	-0.197	-2.000	-1.211		2.000	0.338	0.152	0.762	0.160	0.325	2.000	2.000
	MET	Law ES	-0.234	-0.600	1.617	-0.560	-1.331	0.420	-0.392	1.921	2.000	2.000	-0.104	2.000	-0.308	-2.000	1.103		2.000	1.045	1.162	0.591	-0.100	-0.523	2.000	2.000
7	IR	Codwell ES	-1.450	-0.328	0.906	0.135	-1.725	1.856	-1.218	2.000	2.000	-0.100	0.830	2.000	0.163	2.000	-1.790		2.000	0.116	0.446	0.100	-1.060	-0.322	2.000	2.000

Sorted by count of school's factors below HISD average (0.00). Showing bounded Z-Scores. Elementary and secondary schools not compared to each other. White cell indicates data unavailable.

-2.000 2.000

Risk Factors Balance between Treatment and Comparison Schools

Appendix Table 1 shows the descriptive statistics of these risk factors for both the TIF4 and Comparison schools: both group means and standard deviations, and the standardized mean difference (Hedges' **g**, or effect size).

Appendix D Table 1. 2014–2015 Risk Load Factors for Treatment and Comparison Schools

Demographic Variable	Treatment (T)		Comparison (C)		|g|
	Mean	SD (Pts)	**Mean**	SD (Pts)	
Student Variables					
Free/Reduced Lunch Eligible	**88.2**	7.7	**85.8**	9.4	0.27
Black or Hispanic	**97.9**	1.9	**95.1**	6.4	0.47
English Language Learner	**28.7**	19.5	**40.0**	19.9	0.57
Immigrant	**1.9**	3.3	**3.3**	3.7	0.39
Asylee/Refugee	**0.70**	1.7	**0.60**	1.9	0.05
Special Education	**7.9**	2.3	**7.1**	3.1	0.27
Gifted/Talented	**7.9**	3.9	**12.5**	7.3	0.67
Family Variables					
Child Protective Services	**0.06**	0.2	**0.01**	0.0	0.57
Homeless/Housing Insecure	**1.3**	3.2	**0.7**	0.8	0.44
Student Mobility	**27.3**	5.3	**25.9**	6.7	0.21
School Environment Variables					
Chronically Absent	**8.1**	4.6	**5.6**	3.8	0.65
Suspended Once or More	**8.9**	11.1	**5.2**	8.2	0.43
If Ss left > Ss transferred in (1/0)	**0.71**	0.5	**0.66**	0.5	0.11
Student Safety Score †	**64.3**	9.8	**64.2**	17.3	0.00
Teacher Turnover, 2014 to 2015	**33.9**	12.7	**26.6**	13.2	0.56
Mid-Year Teacher Vacancies	**0.01**	0.0	**0.01**	0.0	0.18
Principals (Count), 2011 to 2015	**2.1**	0.9	**2.0**	0.9	0.10
Neighorhood Variables					
Children in Poverty	**46.2**	11.4	**41.1**	14.7	0.35
HS Grad or Less	**64.7**	11.9	**60.4**	19.5	0.23
Neighborhood Poverty	**31.6**	8.1	**28.6**	10.0	0.31
Adults in Workforce	**87.2**	4.0	**89.4**	4.3	0.51
Unemployed Men, Age 20-64	**12.6**	5.3	**10.3**	4.9	0.45
If Public Housing in Zone	**0.10**	0.3	**0.15**	0.4	0.15
If Homeless Shelter In Zone	**0.19**	0.4	**0.23**	0.4	0.09
Number of Schools Per Group					
Elementary	18		114		132
† Secondary	3		13		16
Total	21		127		148

Data Source Abbreviations in “A Better Picture of Poverty”

- ACS: American Community Survey 5 Year Estimates, 2010–2014, from the US Census Bureau (Tract)
- City: The City of Houston’s Housing and Community Development Department.
- HRIS: Houston ISD’s Human Resource Information Systems.
- PEIMS Snapshot: The Public Education Information Management System (PEIMS) encompasses all data requested and received by TEA about public education, including student demographic and academic performance, personnel, financial, and organizational information. Data from the October 31, 2014 “PEIMS Snapshot”.
- TAPR: Texas Academic Performance Report (TAPR) 2014–2015.
- SIS: Student Information System, called Chancery. SIS “At Risk” Report from HISD Federal and State Compliance Department.
- YourVoice: A customer satisfaction survey conducted by HISD vendor RDA (2013, 2014, 2015). Student survey items must have a 50% response rate to be included and reported.

Student Variables in "A Better Picture of Poverty"

1. Free/Reduced Lunch Eligible. Percentage of school's students enrolled at the PEIMS snapshot who received free or reduced-price lunch subsidies under the Richard B. Russell National School Lunch Act, or are considered to be economically disadvantaged by the Texas Education Agency. Source: TAPR 2014–2015, from PEIMS Snapshot.
2. Black or Hispanic. Percentage of school's students enrolled at the PEIMS snapshot who are identified as belonging to one of the following groups: African American, or Hispanic. Source: TAPR 2014–2015, from PEIMS Snapshot.
3. English Language Learner (ELL). Percentage of school's students enrolled at the PEIMS snapshot identified as participating in programs for English language learners (ELL). Students are identified as ELL by the Language Proficiency Assessment Committee (LPAC). Source: TAPR 2014–2015, from PEIMS Snapshot.
4. Immigrant. Percentage of school's students enrolled at the PEIMS snapshot identified as Immigrants. Source: PEIMS Snapshot.
5. Asylee/Refugee (Secondary only). Percentage of school's students enrolled at the PEIMS snapshot whose initial enrollment in a school in the United States in grades 7 through 12 was as an unschooled asylee or refugee per Texas Education Code (TEC) Section 39.027(a-1). Source: PEIMS Snapshot.
6. Special Education. Percentage of school's students enrolled at the PEIMS snapshot identified as students with disabilities. Students are placed in special education by their school's Admission, Review, and Dismissal (ARD) committee. Source: TAPR 2014–2015, from PEIMS Snapshot.
7. Students NOT identified as Gifted/Talented: Percentage of school's students enrolled at the PEIMS snapshot who are NOT identified and served in state-approved gifted and talented programs. Source: TAPR 2014–2015, from PEIMS Snapshot.

Family Variables in "A Better Picture of Poverty"

8. Child Protective Services. Percentage of students removed from the school by Department of Family and Protective Services (a.k.a. Child Protective Services) during the school year. Source: SIS "At Risk" Report from HISD Federal and State Compliance Department.
9. Homeless/Housing Insecure. Percentage of school's students enrolled at the PEIMS snapshot who are qualified for at-risk status due to either being flagged as homeless or having residential placement. Source: SIS "At Risk" Report from HISD Federal and State Compliance Department.
10. Student Mobility. Percent of school's students who have been in membership at a school for less than 83% of the school year (missed six or more weeks). Source: TAPR 2014–2015.
11. Chronically Absent. Percentage of school's students enrolled at the PEIMS snapshot who missed 18 or more days of school. Source: Barbara Bush Foundation for Family Literacy, 2014–2015 Data.
12. Suspended Once or More. Percentage of school's students enrolled at the PEIMS snapshot who attend at least one day in a school who received at least one In-School Suspension or Out-of-School Suspension during the school year. Source: SIS "At Risk" Report from HISD Federal and State Compliance Department.
13. If Ss left > Ss transferred in. A binary variable (1/0) capturing whether (1) or not (0) more students left the school than joined the school throughout the year. Source: HISD Demographer in Student Support Services.

14. Student Safety Score (*Secondary only*). Percentage of student respondents who "agree" or "strongly agree" with the statement, *"Overall, I am satisfied that my school is safe and secure".* Source: YourVoice Survey.
15. Teacher Turnover, 2014 to 2015. Percentage of teachers *not* retained at the same campus from the 2013–2014 school year to the 2014–2015 school year. Source: HRIS.
16. Mid-Year Teacher Vacancies. Percentage of teaching positions vacant at the campus on December 1, 2015, as a percentage of total possible teacher population for that campus. Source: HRIS.
17. Principals (Count), 2011 to 2015. Number of unique principals at the school over the previous five years. Source: HRIS.

Neighborhood Variables in "A Better Picture of Poverty"

18. Children in Poverty. Percentage of school's zoned census tract residents ages 18 and younger who live in households below the federal poverty level. Source: ACS.
19. HS Grad or Less. Percentage of school's zoned census tract residents ages 25 and older who attained less than or equal to high school graduation (i.e., no additional formal education after high school). Source: ACS.
20. Neighborhood Poverty. Percentage of school's zoned census tract residents (all ages) who live in households below the federal poverty level. Source: ACS.
21. Adults in Workforce. Percentage of school's zoned census tract residents ages 16 and older who are employed in the civilian labor force. Source: ACS.
22. Unemployed Men, Age 20-64. Percentage of school's zoned census tract male residents ages 20 to 64 who are not employed. Source: ACS.
23. If Public Housing in Zone. Binary variable capturing whether (1) or not (0) a school has Public Housing zoned for attendance. Source: City.
24. If Homeless Shelter in Zone. Binary variable capturing whether (1) or not (0) a school has a homeless shelter zoned for attendance. Source: City.

Appendix E: More on the Methods

Limitations

As illustrated in **Table 1, Table 2,** and **Appendix D Table 1,** the two groups of schools (TIF4, and Comparison) are unequal at baseline along several variables that could affect student outcomes. This does somewhat constrain the generalizability of the findings. Some of these variables were included as controls in the model assessing causal impact (see below). The small sample size for schools serving grades 6–8 (*n*=21) and the resulting degrees of freedom limited the possibilities of adding covariates to the regression model to better account for these baseline differences.

STAAR Performance Levels and STAAR Scale Scores

The first analysis in this report addresses the trends in students' performance levels over the grant period. The cut scores for these performance levels are determined annually by the Texas Education Agency (TEA), and reflect a student's mastery of the content for their current grade level. Under the category definitions revised for 2016–2017 and published in April 2017, the TEA's definitions indicate the following for STAAR in grades 3–8:

- Masters Grade Level (previously Level III: Advanced): "Performance in this category indicates that students are expected to succeed in the next grade or course with little or no academic intervention. Students in this category demonstrate the ability to think critically and apply the assessed knowledge and skills in varied contexts, both familiar and unfamiliar."
- Meets Grade Level (previously Level II: Satisfactory at Final Standard). "Performance in this category indicates that students have a high likelihood of success in the next grade or course but may still need some short-term, targeted academic intervention. Students in this category generally demonstrate the ability to think critically and apply the assessed knowledge and skills in familiar contexts."
- Approaches Grade Level (previously Level II: Satisfactory Phase-In 1 and Level II: Satisfactory 2016). "Performance in this category indicates that students are likely to succeed in the next grade or course with targeted academic intervention. Students in this category generally demonstrate the ability to apply the assessed knowledge and skills in familiar contexts."
- Did Not Meet Grade Level (previously Level I: Unsatisfactory): "Performance in this category indicates that students are unlikely to succeed in the next grade or course without significant, ongoing academic intervention. Students in this category do not demonstrate a sufficient understanding of the assessed knowledge and skills." (Student Assessment Division, 2017)

In consultation with technical assistance providers, HISD's TIF4 project staff determined that the STAAR performance levels were insufficiently rigorous for an investigation of the causal impact of TIF4 because these cut scores changed each year (Shakman, Wogan, Finster, & Milanowski, 2016). Nevertheless, the per-school category counts (or percentages) of students were important to the TIF4 programming for specific purposes: in addition to being used in each school's annual accountability measures from TEA, they were used in the project measures reported annually to USDE.

After considering two other possible dependent variables (Index 2 Student Progress scores from campus-level TEA accountability, and TEKS-level analysis of student achievement), the decision was made to examine the scale scores that underpin the TEA's annual cut scores for performance levels. Consequently the findings of the causal impact analyses were not affected by the TEA's changes in cut scores. For more information on scale scores, see the *STAAR Vertical Scale Technical Report* (Student Assessment Division, 2013).

Modeling the Causal Impact of TIF4 on Math and Science

The model used to evaluate the causal impact of the TIF4 program can be expressed as follows:

$$y_{jt} = \beta_{0j} + \beta_{1t} + \beta_{2t}TIF_j + \beta_{3t}X_{jt} + \varepsilon_{jt}$$

In this model,

- y_{jt} is the average STAAR score in science or mathematics at school *j* in year *t*;
- β_{0j} is a fixed effect for school *j*;
- β_{1t} is a fixed effect for year *t*;
- TIF_j is an indicator variable that equals 1 if school *j* is a participant in the TIF4 program and 0 if school *j* is a comparison school; and
- X_{jt} is a vector of characteristics of school *j* in year *t*.

Note that the coefficients β_{2t} and β_{3t} are year-specific. Of particular interest are the coefficients β_{2t}, which measure the impact of participation in TIF4 in year *t*. Since the TIF4 program had not been implemented in the baseline year (2013), we constrain β_{2t} to equal zero in that year (i.e., $\beta_{2,2013}$ = 0). Consequently, the interpretation of β_{2t} in years after the baseline year (i.e., the interpretation of $\beta_{2,2014}$, $\beta_{2,2015}$, $\beta_{2,2016}$, and $\beta_{2,2017}$) is the cumulative impact of the TIF4 program over the course of its having been implemented for (*t* - 2013) years. For example, the coefficient $\beta_{2,2016}$ is the impact on student achievement of a school having participated in the TIF4 program for three years.

The model is estimated by regressing y_{jt} on a full set of school dummies; a set of year dummies with the baseline year (2013) omitted; interactions between TIF4 status and year dummies (with baseline year omitted); and interactions between school characteristics and year dummies (with baseline year included if the school characteristic is time-variant, omitted if time-invariant). This approach produces estimates of the cumulative impact of TIF4 one year ($\beta_{2,2014}$), two years ($\beta_{2,2015}$), three years ($\beta_{2,2016}$), and four years ($\beta_{2,2017}$) after baseline. The significance of these can be tested individually ($\beta_{2,2014} = 0$, $\beta_{2,2015} = 0$, etc.) or jointly ($\beta_{2,2014} = \beta_{2,2015} = \beta_{2,2016} = \beta_{2,2017} = 0$).

The regression is estimated by ordinary least squares over data sets that are separate by grade and subject but pooled across years. A total of eight regressions are estimated: two in which the outcome variable y_{jt} is average science STAAR score (one each for grades five and eight, the grades in which science is tested); and six in which y_{jt} is average mathematics STAAR score (one each for grades three through eight). The data sets over which each of these eight regressions are estimated include a separate observation for each school for each year from 2013 to 2017. Coefficient standard errors are estimated with clustering by school.

When the outcome variable y_{jt} is average STAAR science or mathematics scores in grades three through five, the school characteristics in X_{jt} include, by school and year:

- average STAAR reading scores by school for that grade and year;
- percent African-American by school and year,
- percent limited English proficient by school and year,
- percent students with disabilities by school and year,
- percent economically disadvantaged by school and year; and,
- percent of students immigrant by school.

Of these, all but the percentage of immigrant students are measured yearly and are time-variant. (See **Appendix D** for details on the sources of these variables.)

When the outcome variable y_{jt} is average STAAR science or mathematics scores in grades six through eight, the school characteristics vector X_{jt} is made up of a more parsimonious set of variables:

- average STAAR reading score by school and year,
- percent African-American by school and year, and

- percent limited English proficient by school and year.

In these grades, the data set is substantially smaller, both in terms of the number of TIF4 schools (3) and the number of comparison schools (13). Including the full set of control variables in these grades substantially reduced the precision of the estimated impacts of TIF4, usually without substantively changing the point estimates.

Average STAAR scores in science, mathematics, and reading are normalized using the mean and standard deviation of STAAR scores across students in Texas by subject, grade, and year. This improved the comparability of the outcome variable y_{jt} from one year to the next. It also produced more easily interpreted estimates of the β_{2t} coefficients that are measured in standard deviations of student-level achievement.

When the outcome variable is STAAR mathematics, the year 2015, which was the first year of a transition to new state mathematics standards, is omitted from the data set. As a result, we do not estimate the impact of TIF4 on mathematics outcomes in 2015, two years out. This does not affect the ability to measure the impact of TIF4 one year (2014) or three or four years (2016, 2017) after implementation.

Technical Details on Specific Grade/Subject Models

Fifth Grade Science

In fifth-grade science, the improvement in science STAAR scores among students in TIF4 schools is statistically significant. We can reject at the .003 level the hypothesis that there is no impact from TIF4 over the four years of implementation. The fifth-grade result is robust to changing the specification of the model to include no variables at all, to including only average STAAR reading scores, and to only including school characteristics other than STAAR reading scores in X_{jt}. In all of these specifications, we can reject the hypothesis of no impact from TIF4 at the .025 level or better.

Eighth Grade Science

The evidence in eighth-grade science is less compelling, even given the substantive point estimate of the impact of the TIF4 program. This is because the sample of schools is sufficiently small that even a substantive measured impact is not necessarily statistically significant. The *p*-value of an *F*-test of the hypothesis that there is no effect from TIF4 on eighth-grade science achievement is $p=0.09$. This means that, if there were no effect from the TIF4 program at all, the probability that there would be a difference in achievement between students in TIF4 schools and in non-TIF schools of the size that we observe in the data is about nine percent. This does not meet the conventional significance threshold of $p \leq 0.05$, although it does meet the more permissive threshold of $p \leq 0.10$. While this level of statistical significance is not as compelling, these results are nonetheless suggestive that the TIF4 program had an impact on eighth-grade science achievement.

As mentioned above, the eighth-grade model includes a more parsimonious set of school characteristics than the fifth-grade model. More specifically, the fifth-grade model includes percent free and reduced-price lunch, percent students with disabilities, and percent immigrant, while the eighth-grade model does not. Adding these variables to the eighth-grade model yields point estimates of the impact of TIF4 similar to those from the more parsimonious model presented in Figure 1. However, it also increases the *p*-value of the hypothesis of no impact from TIF4 to $p=0.37$, which is not statistically significant at any conventional level. The combination of a substantially lower *p*-value but not substantively different point estimates suggests that the estimated eighth-grade science model with additional school characteristics is too imprecise to yield useful information about the robustness of the more parsimonious model's estimate of the impact of the TIF4 program.

In contrast, simplifying the specification of the eighth-grade science model to include only average STAAR reading scores produces similar point estimates with a p-value of .04, and removing all variables also produces similar point estimates with a p-value of 0.02. Both of these p-values are sufficiently low to reject the hypothesis of no impact from TIF4 at conventional levels, although both results also do not control for any improvements over time among schools with specific characteristics relative to other schools, or for the effects of any changes over time in the characteristics of TIF4 schools relative to non-TIF4 schools.

Sixth Grade Mathematics

As shown in **Figure 16**, the point estimates suggest a substantive impact in sixth-grade mathematics — a cumulative impact over the four years of about a fifth of a standard deviation. However, the estimates are not sufficiently precise to be statistically significant at conventional levels; an F-test of the hypothesis that the impact of TIF4 in all four years is zero has a p-value of 0.42.

Seventh Grade Mathematics

The TIF4 program in seventh grade mathematics has an immediate effect of about one-fifth of a standard deviation of student achievement starting in its first year, which increases slightly to about a quarter of a standard deviation in the third year of TIF4. (Recall that we do not measure the effect in the second year, given that we do not include 2015 mathematics scores as an outcome due to the change in mathematics standards at that time.) In the fourth year, the cumulative impact of the TIF4 program ticks upward to about half of a standard deviation of student achievement. This would be a very large impact: a half-standard-deviation increase would improve the achievement of a student at the 25th percentile to the 43rd percentile; that of a student at the 50th percentile to the 69th percentile; and that of a student at the 75th percentile to the 88th percentile. An impact this great may in part be the result of randomness, which is evidenced by the wide two-standard-error confidence intervals around the point estimates of the impact. One way to check this is to see if this uptick persists into the following year; however, given that test scores and statewide documentation for 2018 were not available, we cannot know if this is the case or not.

Regardless, it is unlikely that the impact of TIF4 on seventh-grade mathematics achievement is zero. We can reject the hypothesis that all of the TIF4 program effects across years are zero at the 0.001 level. This result is robust to four different specifications:

- to removing all school characteristics from the model;
- to including only average STAAR reading scores;
- to including all school characteristics described above other than average STAAR reading scores; and
- to adding the school characteristics included in the elementary school models but not in the middle school models (percent students with disabilities, percent free- and reduced-price lunch, and percent immigrant).

In all these specifications, the point estimates of the impact of TIF4 are substantially positive and statistically significant (i.e., we can reject the hypothesis that the TIF4 programs had no effect at the $p \leq 0.005$ level).

Eighth Grade Mathematics

In eighth grade mathematics, we see achievement dip among TIF4 schools relative to non-TIF schools in the first year, only to recover in the third year to a level of about one-quarter of a standard deviation higher among TIF4 schools than among non-TIF schools, and to further improve to about four-tenths of a standard deviation higher in the fourth year.

An additional variable, equal to the percentage of eighth-grade students attempting the algebra assessment in lieu of the eighth-grade mathematics assessment, is added to X_{jt} when the outcome variable y_{jt} is average STAAR mathematics scores in grade eight. This is to adjust for the distortionary effect on measured eighth-grade mathematics scores that takes place when a disproportionately high proportion of students do not take the eighth-grade assessment in favor of algebra. The percentage of students taking the algebra exam enters the regression linearly. Entering this percentage into the regression as a quadratic or cubic rather than solely as a linear term does not have a substantive effect on the estimate.

We cannot reject the hypothesis that TIF4 had no effect (p=0.01). This test attributes to TIF4 not only the higher achievement among TIF4 schools in the third and fourth years, but also the lower achievement among TIF4 schools in the first year (2014); this is because it is not a test that TIF4 has a *positive* effect, but more broadly a test that TIF4 has a *nonzero* effect over the four years.

Notably, this 2014 result may also have been impacted by a policy change in testing from 2013 to 2014. In 2013, advanced students in grade 7 who took the Pre-AP math courses were tested in the grade 8 math STAAR. However, in 2014, policy was changed to have them take their grade-level assessment (grade 7 math). This policy change had a positive impact on the grade 7 mathematics results and an adverse impact on the grade 8 results in 2014. (Sondhi, Huang, McCarley, Sage, & Stevens, 2014) It is possible that the TIF4 schools were affected more by this policy change than the Comparison schools, contributing to the 2014 effect.

However, the fourth-year (2017) effect, which has a point estimate of 0.39 and measures the cumulative impact of the TIF4 program over all four years, is statistically significant at conventional levels; testing its significance using a t-test yields a p-value of 0.04. This suggests that, while we do not measure any positive immediate effect in the first year of TIF4, we do measure a substantive and significant cumulative effect by the end of its fourth year. It is useful to note that this result, while suggestive, is not especially robust. In particular, adding percent students with disabilities, percent free- and reduced-price lunch, and percent immigrant reduces the fourth-year effect of TIF4 from a statistically significant point estimate of 0.39 to a statistically non-significant point estimate of 0.07.

Appendix F: Tables

Appendix F Table 1. STAAR Math, Grades 3–5: Mean Scale Score, Std. Deviation, Student Count

	Grade 3 Math		Grade 4 Math		Grade 5 Math	
	TIF4	Comp.	TIF4	Comp.	TIF4	Comp.
2013	1398.0	1438.3	1456.8	1514.8	1514.7	1554.7
	(46.7)	(50.8)	(45.7)	(53.6)	(37.8)	(50.5)
	1,407	10,646	1,498	10,087	1,377	9,675
2014	1405.8	1445.6	1484.2	1531.0	1545.9	1580.6
	(49.1)	(58.1)	(50.1)	(57.2)	(43.7)	(54.9)
	1,406	11,179	1,449	10,225	1,322	9,648
2016	1390.1	1424.0	1491.0	1524.4	1533.9	1567.2
	(34.3)	(51.3)	(48.5)	(54.1)	(46.1)	(54.0)
	1,561	12,059	1,558	10,922	1,498	10,919
2017	1407.6	1442.4	1504.3	1542.5	1562.5	1596.8
	(49.0)	(54.9)	(62.9)	(56.8)	(44.3)	(54.2)
	1,459	11,693	1,581	11,302	1,477	10,724

- Mean campus scale scores were calculated by year and grade for the STAAR 3–5 mathematics tests. Campus, subject, and grade-level results with fewer than five testers were excluded. Results from first administration English and Spanish test versions were used to calculate the mean campus scale scores. Prior to 2016, the following test versions were excluded from mean campus scale scores: STAAR-L, M, Accommodated, Alternate, and Alternate 2. The scale scores of all students with "S" codes were used. In 2016, the test versions STAAR-L, Accommodated, and Alternate 2 were excluded from mean campus scale scores. In 2017, the STAAR Alt. 2 test version was excluded from mean campus scale scores. (McCarley, Ye, Selig, & Stevens, 2013, 2014; Reeves, Bigner, & Stevens, 2016, 2017; Reeves, Carney, & Stevens, 2015)
- 2015 STAAR Math scores are not shown since they were not used in this analysis

Appendix F Table 2. STAAR Science, Grades 5 and 8: Mean Scale Score, Std. Deviation, Student Count

	Grade 5 Science		Grade 8 Science	
	TIF4	Comp.	TIF4	Comp.
2013	3506.4	3671.3	3547.0	3718.9
	(104.6)	(160.9)	(159.9)	(278.5)
	1,414	9,773	570	1,972
2014	3555.1	3662.7	3501.3	3763.2
	(130.3)	(181.0)	(122.8)	(424.6)
	1,355	9,747	561	2,159
2015	3533	3622	3480	3632
	(159.4)	(155.7)	(90.1)	(343.8)
	1,430	10,121	577	2,183
2016	3625	3676	3630	3657
	(110.3)	(167.7)	(101.7)	(392.4)
	1,495	10,897	672	2,242
2017	3664	3735	3612	3663
	(145.4)	(183.6)	(140.5)	(411.4)
	1,475	10,737	674	2,169

- Mean campus scale scores were calculated by year and grade for the STAAR science tests for grades 5 and 8. Campus, subject, and grade-level results with fewer than five testers were excluded. Results from first administration English and Spanish test versions were used to calculate the mean campus scale scores. Prior to 2016, the following test versions were excluded from mean campus scale scores: STAAR-L, M, Accommodated, Alternate, and Alternate 2. The scale scores of all students with "S" codes were used. In 2016, the test versions STAAR-L, Accommodated, and Alternate 2 were excluded from mean campus scale scores. In 2017, the STAAR Alt. 2 test version was excluded from mean campus scale scores. (McCarley, Ye, Selig, & Stevens, 2013, 2014; Reeves, Bigner, & Stevens, 2016, 2017; Reeves, Carney, & Stevens, 2015)

Appendix Table 3. Math, Grades 6–8: Mean Scale Score, Std. Deviation, Student Count

	Grade 6 Math		Grade 7 Math		Grade 8 Math	
	TIF4	Comp.	TIF4	Comp.	TIF4	Comp.
2013	1533.3	1566.4	1516.1	1559.4	1620.0	1635.3
	(31.9)	(61.9)	(8.7)	(27.5)	(9.9)	(37.8)
	568	2,107	482	1,610	584	1,964
2014	1551.6	1578.4	1556.4	1570.1	1607.5	1633.5
	(22.2)	(75.4)	(22.7)	(32.0)	(27.0)	(36.6)
	588	2,026	586	2,116	534	1,797
2016	1584.5	1582.0	1580.8	1586.8	1630.0	1597.0
	(44.9)	(72.2)	(24.9)	(58.5)	(21.3)	(39.3)
	773	2,122	720	2,151	627	1,924
2017	1567.2	1557.5	1623.4	1587.6	1647.3	1599.2
	(44.4)	(67.8)	(39.3)	(56.9)	(64.5)	(36.8)
	741	2,270	760	2,028	617	1,905

- Mean campus scale scores were calculated by year and grade for the STAAR 6–8 mathematics tests. Campus, subject, and grade-level results with fewer than five testers were excluded. Results from first administration English and Spanish test versions were used to calculate the mean campus scale scores. Prior to 2016, the following test versions were excluded from mean campus scale scores: STAAR-L, M, Accommodated, Alternate, and Alternate 2. The scale scores of all students with "S" codes were used. In 2016, the test versions STAAR-L, Accommodated, and Alternate 2 were excluded from mean campus scale scores. In 2017, the STAAR Alt. 2 test version was excluded from mean campus scale scores. (McCarley, Ye, Selig, & Stevens, 2013, 2014; Reeves, Bigner, & Stevens, 2016, 2017; Reeves, Carney, & Stevens, 2015)
- 2015 STAAR Math scores are not shown since they were not used in this analysis