# Measuring the Implementation and Impact of Aspire's Transforming Teacher Talent

FINAL REPORT ON THE EVALUATION OF AN I3 DEVELOPMENT PROJECT

Andrew P. Jaciw
Jenna Zacamy
Li Lin
Kristen Koue
Boya Ma

Empirical Education Inc.

*February 8, 2016*

Empirical Education

EMPOWERING EDUCATORS FOR EVIDENCE-BASED DECISIONS

Empirical Education Inc. · 425 Sherman Avenue, Ste 210, Palo Alto, CA 94306 · empiricaleducation.com · 650.328.1734

## ACKNOWLEDGEMENTS

We are grateful to the teachers and administrators in Aspire Public Schools for their assistance and cooperation in conducting this research. The research reported here is the independent evaluation of an Investing in Innovation (i3) grant #U411C110424 from the U.S. Department of Education to Aspire Public Schools.

## ABOUT EMPIRICAL EDUCATION INC.

Empirical Education Inc. is a Silicon Valley-based research company that provides tools and services to help K-12 school systems make evidence-based decisions. The company brings its expertise in research, data analysis, engineering, and project management to customers that include the U.S. Department of Education, educational publishers, foundations, leading research organizations, and state and local education agencies.

TABLE OF CONTENTS

Reference this report: Jaciw, A. P., Zacamy, J., Lin, L., Koue, K., & Ma, B. (2016). *Measuring the Implementation and Impact of Aspire's Transforming Teacher Talent: Final Report on the Evaluation of an i3 Development Project.* (Empirical Education Rep. No. Empirical_Aspire_7025-FR1-Y3-O.1). Palo Alto, CA: Empirical Education Inc. Retrievable from http://empiricaleducation.com/pdfs/AspireFR.pdf

## Introduction

In this report we provide an evaluation of Aspire Public Schools' (Aspire) innovative technology-supported professional development (PD) system, Transforming Teacher Talent (t3). In 2011, Aspire was awarded a four-year grant to develop, implement, and evaluate t3.

The t3 system includes a set of tools, PD opportunities, and support. Specifically, the intervention uses a train-the-trainer model in which experienced Aspire teachers receive training on the use of three t3 tools and then provide PD and coaching to other Aspire teachers. Tools include a PD content library, a personalized online PD tool, and a virtual professional learning community.

The fundamental goal of the work supported by the grant was to increase the number of highly effective teachers by 2015. Empirical Education Inc. (Empirical) partnered with Aspire to conduct the evaluation of t3 in 35 of Aspire's California schools. The grant that Aspire received was from the U.S. Department of Education's highly competitive Investing in Innovation (i3) program. The project was funded at the "development" level, meaning that the evaluation was to provide both formative insight—looking to improve the program and its implementation—and a level of evidence of the program's impact or potential impact. The i3 program works with a tiered concept of evidence, an approach also adopted by the recently passed U.S. education law, Every Student Succeeds Act, where specific definitions of levels of evidence are included in the legislative language. The largest grants at the highest "scale-up" tier require strong evidence for funding and are expected to use rigorous experimental designs in their independent evaluations. At the mid-level, "validation" tier, a moderate level of evidence is required for funding. At the "development" level, the proposal must provide at least evidence that the innovation shows promise of impact on the intended outcomes. At all levels, the funding agencies prefer the strongest possible evidence but it is understood that the constraints of developing an innovation during the grant period may limit the evaluator's research design options. As we outline in the impact section of this report, the design that was feasible provides evidence of t3's promise of impact.

Throughout the evaluation, the National Evaluation of i3 (NEi3) Technical Assistance Analysis and Reporting team provided us with support and feedback to "ensure that the government's investment in educational interventions through i3 generates high-quality evidence to inform educational policy decisions and to boost the field's capacity to design and conduct scientifically rigorous education research" (Abt Associates, 2014). The study plan for the t3 evaluation that we provided to the NEi3 Analysis and Reporting team has a parallel structure to this report, with separate sections for reporting on implementation and impact. In addition to this report, Empirical has provided the main results of the implementation and impact studies

to NEi3 for reporting as part of their assessment of the strength of evidence generated by the i3 evaluations and to provide summaries of evaluation findings to i3 and the U.S. Congress.

The evaluation was divided into two parts. The first, reported to Aspire in fall 2014 (Koue, Jaciw, & Zacamy, 2014), evaluated the implementation of the t3 system and provided formative feedback from the t3 leaders and school administrators. The second examined changes associated with the implementation of t3 in teacher performance, as measured through the Aspire Instructional Rubric (AIR) scores, and in student academic achievement, as measured through the California Standards Tests (CSTs) in English language arts (ELA) and mathematics. Results from the second part of the evaluation allow an assessment of the evidence of promise of t3 to achieve impact on important outcomes.

This is the final report of our evaluation of Aspire's t3 system. It summarizes the main findings from both the implementation and impact components of Empirical's evaluation. We start with the background to Aspire's t3 system and the context for the study and evaluation. Next, we summarize the results of the implementation study. Following that, we report the findings from the impact study. The results from the implementation study provide context for the impact study and vice versa. Therefore, while each of the two main components of the evaluation adopt different methods and are informative individually, the full picture of the evaluation results from considering both components as they relate to one another, which we do in the conclusion of this report.

# Background

## ASPIRE PUBLIC SCHOOLS

Aspire is a Charter Management Organization, with 38 schools in 11 cities throughout California and in Memphis, Tennessee. Aspire serves more than 14,600 students in grades K- 12. The Aspire Home Office is located in Oakland, California and includes a leadership team and staff who provide oversight, support, and staff development to all Aspire schools.

The t3 study sample of principals, teachers, and students represents 34 of Aspire's 35 California schools.[1] Distinguishing characteristics of the study sample are that it consists of primarily low-income, high-minority, high-achievement charter schools. The demographics of Aspire's schools in California are displayed in Table 1.

TABLE 1. DEMOGRAPHICS OF ASPIRE PUBLIC SCHOOLS IN CALIFORNIA

| Demographics | |
|---|---|
| **Total schools** | 35 |
| **Total full-time equivalent teachers** | 466 |
| **Student to teacher ratio** | 24:1 |
| **Student population** | 11182 |
| **English Language Learners** | 27%[a] |
| **White** | 9% |
| **Black** | 13% |
| **Hispanic** | 68% |
| **Asian** | 5% |
| **Pacific Islander** | 0% |
| **American Indian/Native Alaskan** | 0% |
| **Multi-racial/No response** | 3% |

[a] Average across all districts in which there is one or more Aspire schools

Note. Percentages may not add up to 100% due to rounding of decimals

Source. http://aspirepublicschools.org/about/aspire-overview/

[1] Not all 34 schools were included in every analysis. Twenty-eight schools were included in the analysis of impact on math outcomes, while eight schools were included in the analysis of impact on ELA.

## THE INTERVENTION: TRANSFORMING TEACHER TALENT

Aspire set the goal to double the number of highly effective teachers by the end of the 2014-15 school year, as measured by classroom observations and walk-throughs that were structured by a protocol, the AIR. Increasing the number of highly effective teachers supports the overarching mission of Aspire's work: to send every single Aspire student to college. To reach its teacher and student achievement goals, Aspire took on the task of building and implementing t3.

### Components of Transforming Teacher Talent

T3 provides PD opportunities around three tools: (1) an expanded PD content library, (2) data collection from informal observations and walkthroughs, and (3) Virtual Collaborations (VCs). The technology underlying t3 was based on a tool set from their vendor/partner, BloomBoard, which provided a customized version of their observation data collection tools and online resource library, which Aspire called "Purple Planet." VCs were supported by Google Hangout.

Three key leadership positions helped teachers utilize each of these opportunities.

1. **Purple Planet Drivers** are experienced teachers who provide teachers at their school sites with hands-on training on how to utilize the expanding components of Purple Planet, the **PD content library**.
2. **Peer Observers** are experienced teachers who provide more frequent targeted informal classroom observations and walkthroughs for teachers with low AIR scores, teachers who are new to the school site, and/or who teach the same grade or content area.
3. **Virtual Collaboration Leaders** (VCLs) are highly effective teachers who facilitate online professional learning communities, or VCs, in Google Hangout for colleagues who teach the same grade level or a similar content across Aspire schools.

The t3 intervention includes trainings for Purple Planet Drivers, Peer Observers, and VCLs—collectively known as t3 leaders. These trainings prepared the t3 leaders to work with teachers at their school sites or across school sites. Training began in spring 2013, and continued throughout the 2013-14 academic year. For each tool, the treatment occurred at two levels: (1) Staff from the Aspire Home Office train t3 leaders, and (2) t3 leaders train, coach, or collaborate with school personnel (e.g., teachers).

### Overview of the Transforming Teacher Talent System Logic Model

The t3 system logic model illustrates the program components (inputs), the activities/outputs that represent the implementation of those components, as well as the intended outcomes that support the t3 system's goals and Aspire's overarching mission (Figure 1). The logic model also includes assumptions and external factors that may facilitate or impede the program's

implementation and, ultimately, its effects. These are formally defined in the logic model narrative in Appendix A.

The t3 system logic model was originally developed in fall 2012. It was updated in fall 2013 and spring 2014 to include changes to the t3 system and more details about the system's inputs, activities, and intended impacts. All iterations of the logic model were developed collaboratively by the Aspire Home Office team and the Empirical Education team.

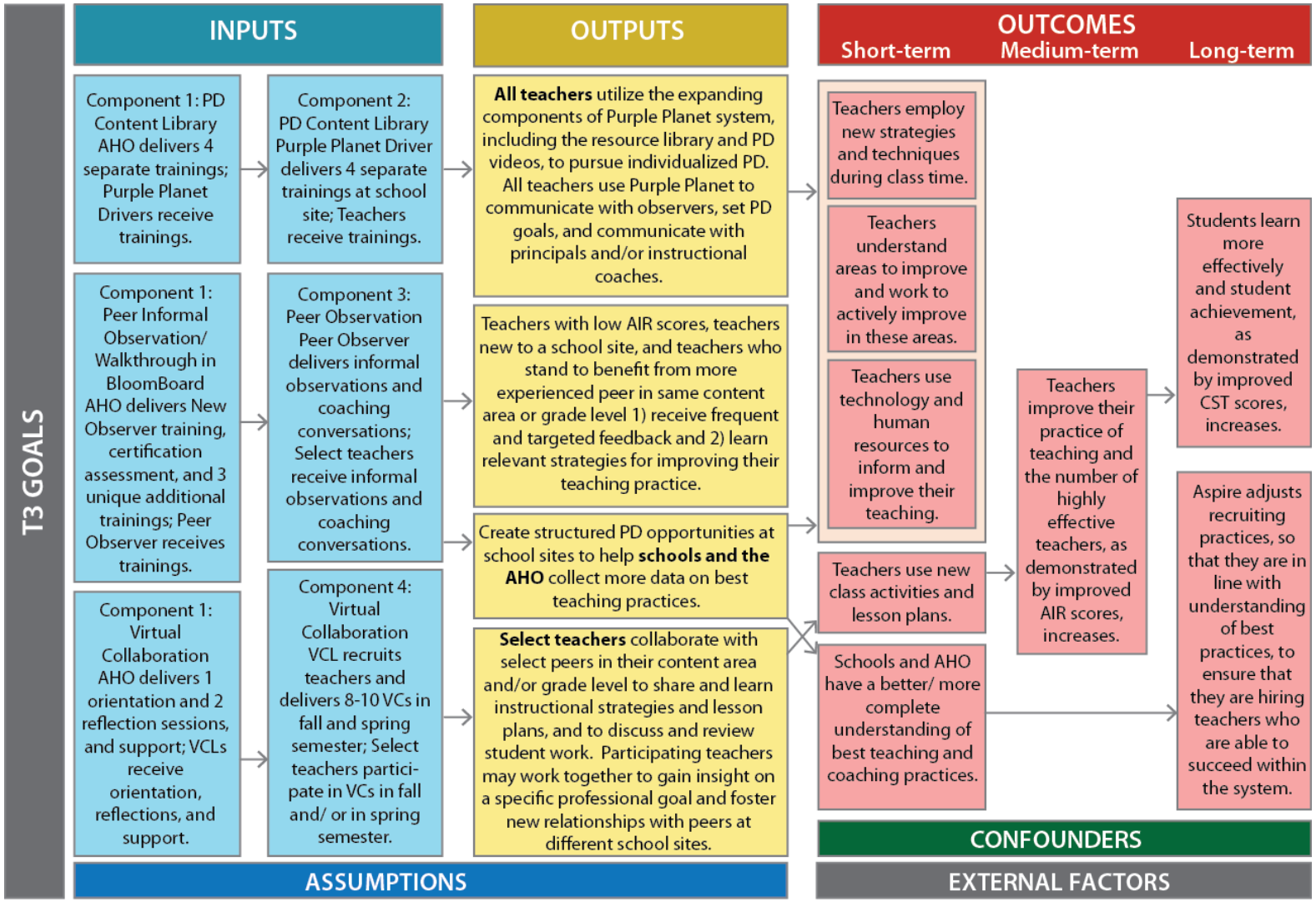


FIGURE 1. TRANSFORMING TEACHER TALENT LOGIC MODEL

Note. AHO stands for Aspire Home Office. VCL stands for Virtual Collaboration Leader. VC stands for Virtual Collaboration. PD stands for Professional Development. AIR stands for Aspire Instructional Rubric. CST stands for California Standards Tests.

## Implementation Study

Providing formative feedback to the developers and measuring the success of their implementation is an important role for the i3 evaluator. A "Fidelity of Implementation" (FOI) metric was mandated by the NEi3 as a way of systematically measuring whether the innovation was faithfully implemented as planned by the developer. The important design features of the innovation are expected to correspond with components of the FOI metric, and these are reflected specifically in the inputs of the logic model (Figure 1). With the logic model, we can look for associations between FOI components and outcome measures to begin considering the components that are critical. In this section, we provide a description of the FOI measurements for four components of the t3 system (corresponding to the logic model inputs), the data collection and measures used to assess FOI, and a summary of FOI findings.

### MEASURING FIDELITY OF IMPLEMENTATION

Over the course of the 2013-14 school year we worked closely with the Aspire Home Office team to develop indicators to assess FOI of the t3 system. A major consideration when developing measures of fidelity was the following tiered process of program delivery and participation.

1. t3 leaders' level of preparedness to support teachers at their school sites or across school sites, measured through t3 leaders participation at Aspire Home Office-led (Aspire-led) trainings
2. t3 leaders' completion of expected work with teachers, measured through delivery of Purple Planet trainings, peer observations and coaching conversations, and VC meetings
3. the extent to which t3 system components reached intended teacher beneficiaries, measured through teachers' participation

The final assessment of fidelity involved four components, with between four and seven indicators measured per component. The four components are as follows.

#### Component 1: Aspire Home Office Provides Adequate Training and Support

This component consists of Aspire providing Purple Planet Drivers, Peer Observers, and VCLs with the support and training they need to effectively interact with teachers at their school sites and across school sites through five indicators.

#### Component 2: Professional Development Content Library

This component consists of activities related to Purple Planet Drivers' participation in Aspire-led trainings, their delivery of trainings at school sites, and teacher participation in school site trainings. Component 2 includes five indicators.

### Component 3: Peer Observation/Walkthrough in BloomBoard

This component consists of activities related to Peer Observers' participation in Aspire-led trainings, their interactions with teachers at school sites, and teachers' participation in Peer Observer-led informal observations and coaching conversations. Component 3 includes four indicators.

### Component 4: Virtual Collaborations

This component consists of activities related to VCLs' participation in Aspire-led trainings, their facilitation of VCs, and teacher participation. Component 4 has seven indicators.

The complete fidelity matrix, with program components, operational definitions, and fidelity indicators is included in Koue, Jaciw, and Zacamy (2014).

These are the research questions for the implementation study.

1. To what extent were key components of the t3 system implemented with fidelity?
2. What are the characteristics of the distributions of the school-level fidelity scores for each component?

## DATA COLLECTION AND MEASURES OF PROGRAM IMPLEMENTATION

We collected implementation data from May 2013 through June 2014 to assess implementation fidelity across school sites for the four components. Data collected through teacher background forms, attendance records, training observations, multiple t3 leader surveys, principal surveys, the BloomBoard device log, and training observations are used to provide evidence of the implementation. See Koue, Jaciw, and Zacamy (2014) for a description of how each component score was calculated and the data collection source and schedule, by FOI indicator. Surveys from t3 leaders and administrators were also used to provide formative feedback to Aspire.

### Aspire Training Attendance Records

We collected attendance records from Aspire-led trainings for t3 leaders to assess the extent to which they participated in Aspire-led trainings as intended. Attendance records also confirmed that the training was delivered. The date Aspire conducted the training was included with the record to determine if the training was delivered on the scheduled date. We used this data to assess the extent to which Aspire delivered trainings to t3 leaders as intended.

### Aspire Training Reports

We collected reports from Aspire trainers to assess the extent to which Aspire-led trainings were delivered as intended.

### Transforming Teacher Talent Leader Activity Logs/Attendance Records

We collected activity logs and attendance records from trainings as well as from the teacher interactions led by t3 leaders at their school sites or in their virtual learning communities. Activity logs and attendance records also provided information about the frequency of t3 leader interactions with participants and/or the dates of interactions. We used these data to assess the extent to which t3 teachers delivered trainings and interacted with teachers as intended.

### Informal Training Observations

We conducted informal observations of a random sample of Aspire-led trainings for each of the t3 tools in the summer of 2013 and over the course of the 2013-14 school year. We compared these observations to Aspire's implementation guide, which describes their expectations for these trainings. The information we gleaned from observations informed the development of survey questions regarding FOI and t3 leaders' attitudes towards the t3 system.

### Transforming Teacher Talent Leader Surveys

We deployed three online surveys to all Purple Planet Drivers, Peer Observers, and VCLs in fall 2013, winter 2013-14, and spring 2014. These surveys collected t3 leaders' self-reported data on their attendance at Aspire-led trainings and the attendance of teachers at their school sites or in their virtual learning communities. The surveys also gathered formative data related to t3 leaders' feelings of preparedness based on training and experience implementing the t3system at their school sites or in their VC. The surveys included questions about t3 leaders' attributes and attitudes to inform Aspire about the qualities of teachers assuming the t3 leadership roles. While we asked all t3 leaders to complete surveys, we did not offer an additional stipend to complete data collection activities. Survey response rates varied by survey and included data from 87% of the Purple Planet Drivers, 67% of the Peer Observers, and 50% of the VCLs.

### Principal Surveys

We deployed one online survey to all principals in spring 2014. These surveys collected principals' self-reported data on the implementation of the t3 system at their school. We collected formative data regarding principals' attitudes toward the t3 system, along with their perceptions of how the t3 system is being implemented at their school sites. We also asked principals to verify the t3 leaders' delivery of school site trainings and other interactions with teachers to assess the extent to which the intervention was implemented at school sites. We received survey data from 45% of the Aspire principals.

## IMPLEMENTATION STUDY FINDINGS

In this section, we review the implementation findings by component. Where appropriate, we include data from t3 leaders' and administrators' surveys to help us better interpret FOI results.

Table 2 below presents an overview of the FOI matrix and final scores by component at the project level. Overall, Aspire reached the threshold for adequate FOI at the project level for one of the four FOI components—Component 1: Aspire Home Office provides adequate training and support. Notwithstanding this limited achievement of FOI thresholds at the project level, the extent to which schools implemented the different program components varied greatly. More simply put, several schools and individuals implemented various components of the t3 system with great success, while others did not.

TABLE 2. COMPONENT-LEVEL PROJECT-LEVEL FIDELITY OF IMPLEMENTATION (FOI) MATRIX

| | Component | Component range | FOI threshold | Project-level score | Met adequate FOI threshold? |
|---|---|---|---|---|---|
| 1 | Aspire Home Office provides adequate training and support | 0 - 59 | 48 or higher | 59 | Yes |
| 2 | Professional Development Content Library (Purple Planet Driver) | 0 - 15 | 75% or more of schools must achieve a score of 12 or higher | 71.4% | No |
| 3 | Informal Observation/ Walkthrough in BloomBoard (Peer Observer) | 0 -25 | 50% or more of schools must achieve a score of 18 or higher | 48.6% | No |
| 4 | Virtual Collaboration (VCLs) | 0 - 7 | 7 | 4 | No |

### Component 1: Aspire Home Office Provides Adequate Support

Component 1 measures the extent to which the Aspire Home Office provided t3 leaders with adequate training and support. This component includes five indicators. Table 3 presents a summary of Aspire's performance on this component.

TABLE 3. ASPIRE HOME OFFICE PROVIDES ADEQUATE SUPPORT SUMMARY

| | Indicator | Indicator score range [a] | FOI threshold | Indicator score | Met adequate FOI threshold? |
|---|---|---|---|---|---|
| **1.1** | Aspire delivers initial training to Purple Planet Drivers (for PD Content Library) | 0 - 6 | 6 | 6 | Yes |
| **1.2** | Aspire delivers ongoing trainings to Purple Planet Drivers (for PD Content Library) | 0 - 9 | 6 | 9 | Yes |
| **1.3** | Aspire delivers initial training to Peer Observers (For Informal Observation/ Walkthrough in BloomBoard | 0 - 9 | 9 | 9 | Yes |
| **1.4** | Aspire delivers ongoing trainings to Peer Observers (Informal Observation/ Walkthrough in BloomBoard) | 0 - 27 | 21 | 27 | Yes |
| **1.5** | Aspire delivers trainings to Virtual Collaboration Leaders | 0 - 8 | 6 | 8 | Yes |
| | Totals | 0 - 59 | 48 | 59 | Yes |

[a] Descriptions of how indicator scores are determined (e.g., the activities that are counted in the indicator score range) are included in the fidelity matrix in Koue, Jaciw, and Zacamy (2014).

Note. PD stands for professional development. FOI stands for Fidelity of Implementation.

Of the components included in the Fidelity Matrix, Aspire scored highest on Component 1, receiving the maximum possible component- and project-level score of 59. Indicators 1.1 and 1.2 measure the support the Aspire Home Office provided to Purple Planet Drivers through regional trainings in the Bay Area, Central Valley, and Los Angeles. The aim of the trainings was to prepare Purple Planet Drivers to deliver information and training on how to utilize the PD Content Library to teachers and staff at their school sites, as well as to help Purple Planet Drivers become experts on the platform. Indicators 1.3 and 1.4 measure the support the Aspire Home Office provided to Peer Observers through the New Observer Orientation training, four ongoing trainings, and a certification exam in the Bay Area, Central Valley, and Los Angeles. The objective of the New Observer Orientation and trainings was to provide Peer Observers with information and resources to effectively observe and coach teachers and identify and collect evidence in BloomBoard. The objective of the certification exam was to calibrate Peer Observers and ensure that their observation scoring was on par with Aspire's master observers. Indicator 1.5 measures the support Aspire provided to VCLs by leading three orientation

sessions and a fall and spring reflection. The three orientation sessions took place in the fall and covered topics including the role and expectations of VCLs, VC tools and hardware, and recommendations for VCLs on beginning their first VC series. Reflections held at the end of the fall and spring semesters were loosely structured and provided VCLs an opportunity to discuss the highlights and challenges they faced during the semester. The reflections also provided an opportunity for Aspire Home Office leaders to discuss mid-course adjustments to the VC and get feedback from VCLs. Aspire delivered all scheduled trainings and received the maximum scores for each of these indicators.

Aspire's high performance on this component shows that 1) the Aspire Home Office successfully provided t3 leaders training opportunities and support and 2) the Aspire Home Office had the resources (including space—both physical and virtual—in which to deliver training and staff with the content knowledge to train t3 leaders) to carry out all scheduled trainings. Moreover, it suggests that Aspire was committed to providing t3 leaders with an opportunity to access the training and tools they needed to fulfill their leadership duties at their school sites.

In addition to delivering scheduled trainings and support to t3 leaders, we learned through survey responses that the Aspire Home Office provided make-up trainings for some t3 leaders, particularly Purple Planet Drivers, when they were unable to attend scheduled regional trainings. This suggests that Aspire Home Office trainers made an effort beyond the minimum implementation requirement to ensure that t3 leaders had access to the information they needed. Survey results showed that overall, t3 leaders felt Aspire-led trainings were useful. T3 leaders felt sufficiently supported by the Aspire Home Office trainers. Nearly all Purple Planet Drivers and all VCLs felt that all trainings and orientations moderately, more than moderately, or completely prepared them—and enhanced their ability—to carry out their roles and responsibilities. Only a few Peer Observers reported that the meeting left them feeling "less than moderately prepared" or "less than enhanced" their ability to serve in their role. The majority reported leaving trainings feeling sufficiently prepared to carry out their roles.

### Component 2: Professional Development Content Library (Purple Planet)

Component 2 measures the extent to which the Aspire Home Office trained the Purple Planet Drivers and the extent to which the Purple Planet Drivers shared the information they learned with teachers at their school sites through trainings. Each of Aspire's 35 schools received a score on each of five indicators and an overall component score. The five indicators included: (2.1) Purple Planet Drivers' participation in initial training; (2.2) Purple Planet Driver initial teacher training delivery; (2.3) Purple Planet Driver participation in ongoing trainings; (2.4) Purple Planet Driver ongoing teacher training delivery; and (2.5) Teacher participation (reach). Table 4

provides a summary of the score range, threshold, mean score, and number and percent of schools reaching the threshold for each indicator for this component.

TABLE 4. PROFESSIONAL DEVELOPMENT (PD) CONTENT LIBRARY SUMMARY BY INDICATOR

| | Indicator | Indicator Score Range | FOI threshold | Mean Indicator Score | Total schools reaching adequate FOI threshold | % of schools reaching adequate FOI threshold |
|---|---|---|---|---|---|---|
| **2.1** | Purple Planet Driver participation in initial Aspire-led training | 0 - 2 | 2 | 1.6 | 28 | 80.0% |
| **2.2** | Purple Planet Driver delivers initial training to teachers | 0 - 1 | 1 | 0.83 | 29 | 82.9% |
| **2.3** | Purple Planet Driver participation in three ongoing Aspire-led trainings | 0 - 6 | 4 | 4.28 | 27 | 77.1% |
| **2.4** | Purple Planet Driver delivers three ongoing trainings to teachers | 0 - 3 | 2 | 2.34 | 29 | 82.8% |
| **2.5** | Teacher participation in PD Content Library Trainings (reach)[a] | 0 - 3 | 3 | 2.1 | 18 | 51.4% |
| | Component total | 0 - 15 | 12 | 11.15 | 22 | 62.8% |

[a] A school received one point if at least 66% of teachers attended the initial school site training. A school received an additional point if at least 66% of teachers attended either the second or third training (whichever had higher attendance) at their school site, and another point if at least 33% of teachers attended the remaining training.

Note. FOI stands for Fidelity of Implementation.

To reach adequate fidelity on this component, schools needed to receive at least 12 points. We determined the project-level score for this component by measuring the percentage of schools that achieved or exceeded the 12-point adequate FOI threshold. Through conversations and thought exercises, the Aspire Home Office team determined that 75% of schools had to achieve the 12-point adequate threshold in order to reach FOI at the project level. At the project level, Aspire did not reach the FOI threshold for the PD Content Library component. Twenty-two Aspire schools (62.8%) reached or exceeded the FOI criteria for this component.

Although Aspire did not meet the project-level FOI threshold for this component, several schools received high scores on indicators measuring Purple Planet Drivers' participation in Aspire-led trainings and their delivery of trainings to teachers at their school sites. There were, however, two primary reasons Aspire did not meet the project-level FOI threshold for this component: 1) five schools did not have Purple Planet Drivers and, therefore, received a score of zero on all indicators, and 2) only 18 schools (51%) met adequate FOI on indicator 2.5: teacher participation. The five Aspire schools that did not have a Purple Planet Driver received scores of zero for each indicator and for the component overall. This, in turn, lowered the project-level score by more than 10%. More importantly, this finding tells us that five schools in the Aspire system did not receive any training on or support using the PD Content Library during the 2013-14 school year. In the absence of a Purple Planet Driver, teachers at these five schools may not be using the PD Content Library, or may be using it only to a limited extent.

### Component 3: Peer Observation/ Walkthrough Using BloomBoard

Component 3 measures the extent to which Peer Observers were trained by the Aspire Home Office and the extent to which Peer Observers worked with teachers at their school sites. Each of Aspire's 35 schools received a score on each of four indicators and an overall component score. The four indicators included: (3.1) Peer Observer participation in initial Aspire-led training (New Observer Orientation); (3.2) Peer Observer participation in ongoing Aspire-led trainings; (3.3) Peer Observer delivery of two observation cycles to teachers at school site; (3.4) Teacher participation in peer observations (reach). Table 5 provides a summary of the score range, threshold, mean score, and number and percent of schools reaching the threshold for each indicator for this component.

TABLE 5. PEER OBSERVATION/WALKTHROUGH IN BLOOMBOARD SUMMARY BY INDICATOR

| | Indicator | Indicator score range | FOI threshold | Mean school indicator score | Total schools reaching adequate FOI threshold | % of schools reaching adequate FOI threshold |
|---|---|---|---|---|---|---|
| **3.1** | Peer Observer participation in initial Aspire-led training | 0 - 3 | 3 | 2.39 | 27 | 77.1% |
| **3.2** | Peer Observer participation in ongoing Aspire-led trainings and certification assessment | 0 - 9 | 8 | 5.76 | 16 | 45.7% |

TABLE 5. PEER OBSERVATION/WALKTHROUGH IN BLOOMBOARD SUMMARY BY INDICATOR

| | Indicator | Indicator score range | FOI threshold | Mean school indicator score | Total schools reaching adequate FOI threshold | % of schools reaching adequate FOI threshold |
|---|---|---|---|---|---|---|
| **3.3** | Peer Observer delivery of two observation cycles (A cycle includes one observation and two coaching conversations) | 0 - 12 | 6 | 7.34 | 23 | 65.7% |
| **3.4** | Eligible teachers receive some or all of an observation cycle[a] | 0 - 1 | 1 | 0.2 | 7 | 20% |
| | Component total | 0 - 25 | 18 | 16.28 | 17 | 48.6% |

[a] Eligible teachers were defined as teachers with low Aspire Instructional Rubric scores (i.e., teachers who received an overall scores of "1" (Entering) or a "2" (Emerging)) and teachers who were new to Aspire. A Peer Observer received one point if he/she conducted at least one observation or one coaching conversation with 30 percent or more eligible teachers at her school. If there was more than one Peer Observer at a school, Peer Observers' work with eligible teachers was assessed collectively (i.e., if at least 30 percent of eligible teachers were reached by the sum of all Peer Observers at a school, then all Peer Observers at the school received one point on this indicator).

Note. FOI stands for Fidelity of Implementation.

To reach adequate fidelity on this component, schools needed to receive at least 18 points. To determine the project-level score for this component, we measured the percentage of schools that achieved or exceeded the 18 point adequate FOI threshold. Fifty percent of schools had to achieve the 18 point adequate threshold in order to reach adequate FOI at the project level. For schools with multiple Peer Observers, we used the average score to calculate indicator- and component-level scores. Figure 2 shows the distribution of schools by FOI score and that 17 Aspire schools (48.6%) reached or exceeded the FOI criteria for this component.

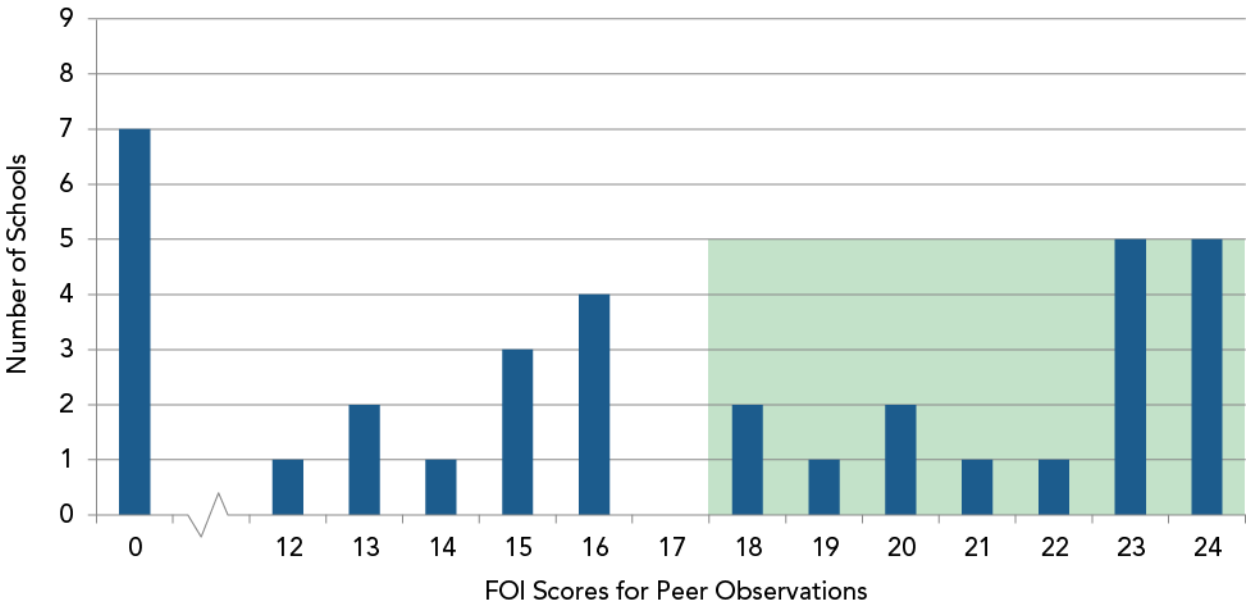


FIGURE 2. DISTRIBUTION OF SCHOOLS BY FIDELITY OF IMPLEMENTATION (FOI) SCORES

Aspire did not meet the project-level threshold for Peer Observations for four reasons: 1) few Bay Area schools met the FOI threshold (only 3 of the 10 Bay Area schools [30%] met or exceeded the threshold. For most of the year, the region did not have a liaison from the group—The College Readiness Promise—that supported the teacher effectiveness and observation process and led regional meetings and trainings); 2) Peer Observers' attendance at ongoing Aspire-led trainings was low, and make-up trainings were not offered; 3) in some instances, schools had two or more Peer Observers that completed role activities to varying degrees, resulting in relatively low school averages of implementations; and 4) Peer Observers' limited reporting on key activities, especially in BloomBoard (our primary data source); plausible values were imputed for missing scores using the best available methods, however, the imputed values are—at best—approximation of achieved but unreported values.

In several instances, Peer Observers worked with multiple teachers, but only a few met the eligibility criteria and/or had complete cycles recorded in BloomBoard. Nearly half of Peer Observers (42%) felt that their presence was completely valuable at their schools sites. Most administrators (67%) also felt that having a Peer Observer at their schools was completely valuable. This suggests that despite challenges Peer Observers faced in meeting with teachers and finding substitutes for their own classes, the program is important to various Aspire stakeholders.

### Component 4: Virtual Collaborations

Component 4 measured VCLs' participation in the Aspire-led orientation and reflections, VCLs' facilitation of VC meetings, and teachers' participation in VCs. The component includes seven indicators: (4.1) VCL participation in an orientation session, (4.2) VCL participation in fall reflection, (4.3) VCL participation in spring reflection, (4.4) Fall VC session delivery, (4.5) Spring VC session delivery, (4.6) Teacher participation in fall VC sessions, and (4.7) Teacher participation in spring VC sessions. Since VCLs worked with teachers across school sites, the component was assessed at the project-level only. Table 6 presents a project-level summary of Aspire's performance on this component.

TABLE 6. VIRTUAL COLLABORATION SUMMARY BY INDICATOR

| | Indicator | Indicator score range | FOI threshold | No. of participants meeting threshold | Indicator score | Met adequate FOI threshold |
|---|---|---|---|---|---|---|
| **4.1** | VCL participation in orientation session | 0 – 1 | 1= 90% of VCLs attend | 12/12 VCLs (100%) | 1 | Yes |
| **4.2** | VCL participation in fall reflection | 0 – 1 | 1= 70% of VCLs attend | 7/10 VCLs (70%) | 1 | Yes |
| **4.3** | VCL participation in spring reflection | 0 – 1 | 1=70% of VCLs attend | 6/6 VCLs (100%) | 1 | Yes |
| **4.4** | Fall VC session delivery | 0 – 1 | 1=90% of VCLs facilitate six VCs | 6/10 VCLs (60%) | 0 | No |
| **4.5** | Spring VC session delivery | 0 – 1 | 1=90% of VCLs facilitate six VCs | 5/6 VCLs (83.3%) | 0 | No |
| **4.6** | Teacher participation in fall VC sessions | 0 – 1 | 1=50% of teachers attend six VCs | 6/23 teachers (26.1%) | 0 | No |
| **4.7** | Teacher participation in spring VC sessions | 0 -1 | 1=50% of teachers attend 75% or more meetings in series | 29/54 teachers (53.7%) | 1 | Yes |
| | Totals | 0 – 7 | 7 | | 4 | No |

Note. VC stands for Virtual Collaboration. VCL stands for Virtual Collaboration Leaders. FOI stands for Fidelity of Implementation.

Over the course of the 2013-14 school year, Aspire made significant changes to the VC program. To measure these changes accurately, we assessed FOI by semester, and the measures and fidelity thresholds for each indicator reflect the structure of the program during the semester of implementation. Indicators 4.1, 4.2, 4.4, and 4.6 measured VCL's activity during the fall semester. During the fall semester, the expectations for VCLs were to 1) attend an orientation

session, 2) recruit a group of teachers in the same grade/ content area as the VCL to participate in VC meetings, 3) lead a VC meeting once weekly for 8 to 10 weeks during the fall semester, and 4) to attend a fall reflection session. The fall VCs intended to engage small groups of teachers for the full 8 to 10 week session. Since VC meetings used Google Hangout, there was a technological capacity limit of 10 (nine participants and one VCL). Given this constraint, a maximum of 90 teachers could have participated in VCs during the fall semester (based on 10 VCLs holding one series each) and in the spring semester, a maximum of 108 teachers could have been reached (based on 12 mini-series held by six VCLs). Indicators 4.3, 4.5, and 4.7 measured the VCL's activity in the spring semester. In the spring semester, significant changes were made to the VC program. Instead of facilitating 8 to 10 sessions for one group of teachers over the course of the semester, VCLs had the opportunity to hold mini sessions, with two to four meetings each. Different teachers could be recruited to participate in each of the mini sessions. VCLs were expected to facilitate at least six meetings, using the mini-series structure or the semester-long session structure from the fall semester. Teachers were offered a stipend of $30 per meeting to participate in VCs.

From survey responses and anecdotes, it appears that the changes made between the fall and spring semesters improved the program. Despite the significant changes to the program, Aspire did not meet the project-level criteria for FOI for this component. There are two important factors that kept Aspire from reaching the FOI threshold for this component: 1) teacher participation and 2) the number of active VCLs. From survey responses, we learned that recruiting and retaining teachers was a major sticking point. While VCLs reported that recruitment efforts in the spring were more successful than in the fall, they also said that many teachers were not interested in participating in VCs because the meetings took place after normal school hours, they could not commit to a weekly meeting, or because they had access to a professional learning community at their school site. Teachers who did join VCs, did not attend regularly. The second major issue that kept Aspire from reaching the FOI threshold for this indicator was the number of active VCLs. While a total of 10 VCLs were included in the sample during the fall semester (two left the program after attending the orientation), only six completed at least six VC meetings, and only eight held any meetings at all. In the spring, there were only six active VCLs. These small sample sizes made it difficult for Aspire to reach the fidelity thresholds on some indicators. It was determined that ideally there would be at least one VCL per grade level cluster (K-1, 2-3, 4-5), one per content area for secondary teachers (math, science, ELA, and social science), and one per specialized K-12 content area (art, music, and special education). All of these grade and content areas were not covered in either the fall or spring semesters. This, in turn, could mean that some teachers, who may have been interested, did not have an opportunity to access a VC group.

## STRENGTHS AND WEAKNESSES IN IMPLEMENTATION

This section covered findings related to the implementation of the t3 system in Aspire's 35 California schools during the 2013-14 school year. In general, we found that the Aspire Home Office team successfully provided the t3 leaders the training opportunities and support consistent with their model. The t3 leaders attended the initial Aspire-led trainings provided to review roles and expectations, and the tools and site-based training objectives. The t3 leaders struggled, however, to reach the expected numbers of teachers at their school sites. Additionally, there was variation in t3 implementation levels across the 35 schools, with some schools fully implementing the t3 systems and others without assigned t3 leaders.

While Purple Planet Drivers generally perceived their role to be valuable to teachers at their site, scheduling a time to conduct longer PD Content Library trainings was a challenge for many, and some reported that they were only able to pass information about the PD Content Library on to a select group of teachers (e.g., team leads). Training attendance records were missing for several schools, and while we were able impute the missing data in the FOI calculation, it is possible that attendance was underreported. As with Purple Planet Drivers, we found variation in Peer Observers' activity at a school site, including limited reporting of observation cycles in the BloomBoard tool and Peer Observers' absence at regional trainings. Despite these challenges, survey results indicate that both administrators and Peer Observers valued this t3 leader role, which required the most face-to-face, direct, and targeted contact with teachers and links to "high stakes" teacher evaluation through AIR.

Aspire made significant adjustments to the VC program during the 2013-14 school year. While VCLs reported that these changes improved the program, recruiting and retaining teachers to participate in the VC sessions continued to be a challenge. VCLs reported that recruitment efforts in the spring were more successful than in the fall. They also said that many teachers were not interested in participating in VCs because they already had sufficient opportunities to collaborate with colleagues at their school sites and did not need the support of a virtual community. Teachers, who did join VCs, did not attend regularly. These findings illustrate the implementation study's primary goal, which was to support Aspire's continuous improvement of t3 during its first year of implementation.

## Impact Study

The second component of our evaluation of Aspire's t3 was a measurement of the impact the program had, and could expect to have in the future, on outcomes of importance to the goals of the program.

### RESEARCH QUESTIONS AND DESIGN

The logic model provided in Figure 1, shows "Teachers improve their practice of teaching, and the number of highly effective teachers, as demonstrated by improved AIR scores, increases " as the key outcome on which improvements in student outcomes depend. We expect certain student-level outcomes to stem from this, for instance "Students learn more effectively, and student achievement, as demonstrated by improved CST scores, increases." Thus, this impact study complements the implementation study and adheres to the logic of the intervention, by examining first the connection between t3 and quality of instruction, and subsequently, the connection between t3 and student achievement.

Working with our advisors at the NEi3, we developed a research design that was the most rigorous feasible approach. Our original design was an experimental comparison of teachers who had access to t3 with those who did not. However, a phased roll out of t3 was difficult, because it was not technically practical to have two versions of the online technology running simultaneously—versions with and without the t3 enhancements. An alternative would be to provide training to only half the teachers, but Aspire determined that such an intervention would provide a weak contrast at best and would be susceptible to contamination (teachers in the control condition finding other ways to learn to use the tools). Thus, we settled on a pre-post design, comparing performance before and after usage of the intervention (the t3 technology tools and training). This is a weaker design and does not establish causality. There may be a confounding factor: something other than the introduction of t3 that could account for an improvement. For example, the introduction of Common Core State Standards may have resulted, by itself, in improved teaching practices. Nevertheless, a measured improvement in the year following the introduction provides promising evidence of a positive effect of t3, if not direct evidence of impact, understood in terms of causation. With the goal of looking for promising evidence of the effects of t3, we proceeded to address the following questions.

We first addressed the question about quality of instruction.

- Were there positive gains in quality of teaching practice as measured through the total score on the AIR when compared to scores in the schools prior to the t3 intervention?

Second, we compared student achievement scores in mathematics and ELA, as measured by the CST, from the year prior to implementing the t3 system (2012-13 school year) to the year when

the t3 system was implemented (2013-14 school year). These analyses address the two main student-level questions of the t3 impact study.

- Were there positive gains in mathematics achievement, as measured by the CST for 3rd-7th grade students in Aspire schools when compared to scores in those schools prior to the t3 intervention?
- Were there positive gains in ELA achievement as measured by the CST for 3rd-11th grade students in Aspire schools when compared to scores in those schools prior to the t3 intervention?

Third, we explored whether the changes in CST scores vary by specific student subgroups.

- Does the change in student achievement in math or ELA between pre-treatment and post-treatment , as measured by CST, vary depending on student (a) incoming achievement, (b) socioeconomic status as measured through eligibility for free or reduced-price lunch, (c) gender, and (d) grade level?

Fourth, we considered two exploratory analyses to get additional insight into the program's effects.

- Is there an association between school-level fidelity in the use of the PD Content Library (Program Component 2) and student achievement outcomes assessed at the school level?
- Is there an association between the number of peer observation cycles per school, one indicator for the Informal Observation / Walkthrough component of the t3 intervention (Program Component 3), and achievement? (We singled out this indicator and component because Aspire indicated specific interest in assessing this association.)

## METHODS

We divide our discussion of methods—including instruments used, sample, and analysis methods—into two sections pertaining to the impact on teachers and impact on students.

### Association between Use of Transforming Teacher Talent and Performance on Aspire Instructional Rubric

#### The AIR Instrument

The AIR includes five domains: (1) Data-Driven Planning and Assessment; (2) Classroom Learning Environment; (3) Instruction; (4) Professional Responsibilities; and (5) Partnerships, Family, and Community. Within each domain are three to five components. Teachers receive scores for indicators measuring elements of each component. There are between one and three indicators per component and 42 indicators in total. Indicator scores are determined both through classroom observations and an observers' assessment of teachers' practices and

interactions with various stakeholders outside the classroom. Domain 2 and Domain 3 indicator scores are assessed during one formal classroom observation and up to three mini observations during a given school year. Most indicator-level AIR scores are determined during the formal observation. If a teacher receives higher scores on some or all of the indicators during mini observations, up to three formal observation indicator-level scores can be replaced by mini observation scores. Observations are conducted by trained and certified Aspire principals, deans, and instructional effectiveness coaches. As part of their training, observers calibrate their observation ratings with the ratings of other observers and with the scoring criteria established by Aspire in the AIR for each indicator. The scoring criteria established by Aspire includes descriptions of specific practices that an observer should observe during a lesson in order to score the teacher at a level one, two, three, or four on a given indicator. Certified observers generally conduct observations independently. AIR score protocol did not change between the 2012-13 and 2013-14 school years, making the scores comparable over time.

**Domains of AIR Used in the Analysis of the Effects of t3**

The outcome measures for the analysis were the overall AIR scores for Domain 2 (Classroom Learning Environment) and Domain 3 (Instruction). We display them in terms of components and indicators in Table 7. This level of detail is necessary to understand the constructs represented through the AIR scores that are used in the analysis.

TABLE 7. ASPIRE INSTRUCTIONAL RUBRIC SCORE DOMAINS 2 AND 3

| Domain | Component | Indicator |
|---|---|---|
| **Domain 2 measures aspects of the classroom learning environment** | 2.1: Create a classroom/culture of learning | 2.1.A: Value of effort and challenge |
| | 2.2: Manage student behavior through clear expectations and a balance of positive reinforcement, feedback, and redirection | 2.2.A: Behavioral expectations<br>2.2.B: Response to behavior |
| | 2.3: Establish a culture of respect and rapport which supports students' emotional safety | 2.3.A: Interactions between teachers and students<br>2.3.B: Student interactions with one another |
| | 2.4: Use smooth and efficient transitions, routines, and procedures to maintain instructional momentum | 2.4.A: Routines, procedures, and transitions |

TABLE 7. ASPIRE INSTRUCTIONAL RUBRIC SCORE DOMAINS 2 AND 3

| Domain | Component | Indicator |
|---|---|---|
| **Domain 3 measures aspects of teachers' instruction** | 3.1: Communicate learning objectives to students | 3.1.A: Communication of the learning objectives of the lesson |
| | | 3.1.B: Connections to prior and future learning experiences |
| | | 3.1.C: Criteria for success |
| | 3.2: Facilitates instructional cycle | 3.2.A: Executes lesson cycle |
| | | 3.2.B: Cognitive level of student learning experience |
| | 3.3: Implementation of instructional strategies | 3.3.A: Questioning |
| | | 3.3.B: Academic discourse |
| | | 3.3.C: Group structures |
| | | 3.3.D: Resources and instructional materials |
| | 3.4: During lesson, teacher makes effective instructional decisions based on formative assessments | 3.4.A: Checking for students' understanding and adjusting instruction |
| | | 3.4.B: Feedback to students |
| | | 3.4.C: Self-monitoring |

Teachers receive a score of one, two, three, or four on each indicator, where a score of one signals that the teacher needs significant improvement in the element of practice measured by the indicator, and a four signals that the teacher has a strong command of the element of practice measured by the indicator.

For overall scoring purposes, Aspire is most concerned with indicator scores and does not generally calculate domain-level scores. For the purposes of analysis (in which we used domain-level scores to assess teacher gains), we aggregated the indicator scores to the domain level two ways.

1. *Metric 1: Average domain scores.* For each of the two domains, we calculated the average (mean) score of all indicators for each teacher. We then examined the average, across teachers, in gains between 2012-13 and 2013-14 on this outcome.
2. *Metric 2: Aspire's score conversion at the domain level.* Aspire uses a five-point rating schema based on the number of 1's, 2's, 3's, and 4's a teacher receives on all indicators to determine the overall score (i.e., across all domains). The five ratings are: 1 = Entering; 2 = Emerging; 3 = Effective; 4 = Highly Effective; 5 = Master. Next, they produce an overall score for the teacher using the following set of rules.

- Master: At least 50% 4s, AND no 1s or 2s
- Highly Effective: 100% 3s and 4s OR over 80% 3s and 4s and at least 30% 4s
- Effective: 61% - 80% 3s and 4s OR over 80% 3s and 4s and less than 30% 4s
- Emerging: 36% - 60% 3s and 4s; Entering: 0 - 35% 3s and 4s

We applied this set of rules to the indicator scores separately for Domains 2 and 3. Then we coded Entering = 1, Emerging = 2, Effective = 3, Highly Effective = 4, and Master = 5 and calculated the gain score for each teacher.

The correlations between the scores on the two metrics within each domain were .92 for Domain 2 and .93 for Domain 3.

**Analysis Model used to Measure the Effects of t3 on AIR Scores**

For each of the metrics above, we computed AIR score gains between 2012-13 and 2013-14 for each teacher. Then we estimated average score gains using (1) an unconditional model (i.e., with no covariates) and (2) an adjusted (or conditional) model, that parses out gains associated with three teacher attributes: years of teaching at Aspire, years of teaching outside of Aspire, and degree level.

The two approaches yield estimates of the average gain in AIR scores over the course of the implementation of t3; that is, the average change in AIR scores between 2012-13 and 2013-14 for the teachers and schools in the sample. With the conditional (adjusted) model, in the context of a pre-post analysis with gain scores as outcomes, the effect of including covariates in the analysis is to make the average gain score dependent on the specific values of the covariates. In this case, it is for the teacher with (a) average years of experience at Aspire, (b) average years of experience excluding years at Aspire, and (c) unknown degree level.

**Sample for Analysis Involving AIR Teacher Outcomes**

We undertook a multistep process to determine the sample of teachers for AIR score analysis. First, we received data for 766 unique teachers who were affiliated with Aspire during at least one of the years of interest (2012-13 or 2013-14). This included 635 teachers in 2012-13 and 693 teachers in 2013-14. The total counts included teachers affiliated with any Aspire school in California (in 2012-13 and 2013-14) and one teacher affiliated with an Aspire school in Tennessee (in 2013-14). During the 2012-13 school year, 582 teachers had indicator and overall AIR scores. During the 2013-14 school year, 655 teachers had indicator and overall AIR scores. Next, we identified 471 teachers who had valid AIR scores in both the 2012-13 and 2013-14 school years in any Aspire school. One of the 471 teachers was affiliated with Hanley Elementary #1 (a school located in Tennessee) in 2013-14 and was removed from the sample.

Next, we checked to see which of the 470 teachers in the sample had valid overall scores[2] as determined by Aspire. Of the 470 teachers, 15 did not have valid overall scores and were eliminated. The final sample includes a total of 455 teachers. Table 8 summarizes the criteria applied to arrive at the analysis.

TABLE 8. STEPS LEADING TO ANALYSIS SAMPLE

| Step | Inclusion criteria | Total teachers in sample | AY 2012-13 teachers in sample | AY 2013-14 teachers in sample |
|---|---|---|---|---|
| 1 | All teachers affiliated with an Aspire school during AY 2012-13 AND/OR AY 2013-14 | 766 | 635 | 693 |
| 2 | All teachers with AIR scores in AY 2012-13 OR AY 2013-14 OR both years | 766 | 582 | 655 |
| 3 | All teachers with AIR scores in AY 2012-13 AND AY 2013-14 | 471 | 471 | 471 |
| 4 | All teachers with any AIR scores in AY 2012-13 AND AY 2013-14 affiliated with a California Aspire school in AY 2012-13 AND AY 2013-14 | 470 | 470 | 470 |
| 5 | All teachers with valid AIR scores in AY 2012-13 AND AY 2013-14 affiliated with a California Aspire schools in 2012-13 and AY 2013-14 | 455 | 455 | 455 |

Note. AY stands for Academic Year. AIR stands for Aspire Instructional Rubric.

## Measuring Impacts on Student Outcomes

### Class Rosters and Demographic Data

We requested and collected class rosters and student demographics from Aspire at the beginning of the study in fall 2012, and collected updated rosters and demographics in fall 2013 and fall 2014. All data were warehoused and firewalled from the primary research personnel—including the data analysts—to avoid leading their judgements and possibly biasing effect estimates, in the event that the primary research questions had to be reconsidered. The data obtained from Aspire were used to link students to teachers and schools, and for analyzing pre-to-post changes in AIR score and CST outcomes. Specifically, in the analysis stage,

[2] A teacher has a valid overall score if she received indicator-level scores on enough indicators to accurately determine an overall score (generally 80% or 90% of all indicator level scores).

demographics were used as covariates to improve the precision of results and to identify subgroups of interest when examining if the results varied across subgroups. Among the data collected from Aspire were the following.

- Names of students and teachers
- Unique identifiers for students, teachers and schools
- Student gender
- Student ethnicity
- Student English proficiency status
- Student disability status (whether or not student has a disability or is in special education, but not the specific condition)
- Student date of birth
- Student grade level
- Classroom teacher name
- Course name and section
- School name

All student and teacher data having individually identifying characteristics were stripped of such identifiers for analysis, and the data were stored using security procedures consistent with the provisions of the Family Educational Rights and Privacy Act. This study falls within the protocol approved by Empirical's Institutional Review Board: Ethical and Independent Review Services. Under this protocol and following Family Educational Rights and Privacy Act guidelines, student or parental permission was not necessary, nor was it was required by Aspire.

**The CST Instrument**

At the start of the study, Aspire identified the CST as the primary outcome measure for this project. According to California's Standardized Testing and Reporting website, "The CSTs are a major component of the [Standardized Testing and Reporting] program. The CSTs are developed by California educators and test developers specifically for California. They measure students' progress toward achieving California's state-adopted academic content standards in ELA, mathematics, science, and history–social science, which describe what students should know and be able to do in each grade and subject tested." (California Department of Education, 2009). Because the test is linked to California's standards, there is no national comparison. Students receive a scale score between 150 and 600. Based on this scale score, California uses five performance levels to report student achievement on the CSTs: Advanced, Proficient, Basic, Below Basic, and Far Below Basic (California Department of Education, 2011). In this study, we

used spring 2014 scores from grades 3 – 7 in math, and from grades 3 – 10 in ELA as outcomes, and spring 2013 scores from grades 2 – 6 in math, and from grades 2 – 9 in ELA as pretests.

Because the CST is not a vertically scaled assessment (we cannot compare scores across grade levels), we z-transformed the outcome scores within each grade level before running the impact analyses. This is done by subtracting the mean for the "lagged" comparison group (i.e., the average of the spring 2013 scores at a given grade level) from each individual score at that grade level, and dividing this difference by the standard deviation of the distribution of posttest scores for this "lagged" comparison group. Following this transformation, the outcome score for each student (i.e., from spring 2014) is expressed as the number of standard deviations away from the mean of the comparison group in his/her grade level. Carrying out the transformation within each grade level essentially puts all scores on a common scale (i.e., in terms of standard deviations away from the mean of the respective lagged comparison group), allowing outcomes to be analyzed together across the grade levels.

**Analysis Models for Student Outcomes**

Pre-post changes are estimated in two ways. With the first approach, we report the average difference in z-transformed scores from pre (spring 2013) to post (spring 2014) periods. This is the "unadjusted" result. With the second approach, we use statistical models to obtain estimates of the changes in achievement scores from pre (spring 2013) to post (spring 2014). The estimates of pre-to-post score changes are "covariate adjusted" which means that they are adjusted for differences, between the groups being compared, in the attributes of individuals that may affect achievement. A series of covariates are included in the estimation equation used to obtain the results. The estimates of change in achievement scores are then expressed as net of, or de-confounded from, the effects of these covariates. This is the "adjusted" result. We used two industry-standard software packages (SAS and HLM) to obtain the estimates.

Both the unadjusted and adjusted estimates of pre-to-post gains are reported as standardized effect sizes; that is, they are divided by the pooled (across spring 2013 and spring 2014) standard deviations of the outcome variables. We also report the *p* value, which is the probability of observing an estimate with a magnitude as large as or larger than the one observed when the actual effect is zero. Conventionally, an estimate is described as "significant" if the *p* value is less than .05. We can also think of having levels of confidence for concluding that the effect is different than zero. A *p* value of less than .05 give us strong confidence, a *p* value spanning .05

and .15 give us some confidence, and *p* values higher than .15 but not exceeding .20 give us limited confidence. Values above .20 give us no confidence of there being a real effect.[3]

In addition to obtaining the adjusted and unadjusted estimates of pre-to-post gains in student achievement, we ran a series of additional analyses using alternative model specifications to assess how robust the results are. The goal is to demonstrate that the results are not sensitive to small differences in the specification of the impact model.

**Differences between Subgroups**

We also examined whether the association between score gains and t3 was different across categories of students. We examined the difference in score gains across levels of: (a) incoming achievement, (b) socioeconomic status, (c) gender, and (d) grade level. To do this, in the model used to obtain the "adjusted" estimate of impact, we included a term that multiplies the variable that indicates the subgroups being compared (e.g., male versus female), by the indicator of whether an outcome is from spring 2013 or from spring 2014 (technically, the product of these two is called an interaction term). The estimate of the coefficient associated with this term tells us if the gains vary across the subgroups being compared.

**Relating Achievement to Implementation**

As part of our exploratory analyses, we considered also the association between CST achievement in spring 2014 and levels of specific measures of implementation. We limited our analysis to math, given the larger sample of schools for which we had math outcomes (compared to ELA). For Component 2, we assessed the correlation between the school-level implementation score for the PD Content Library and student CST math achievement. We also examined the correlation between the average number of peer observation cycles per school, an indicator of fidelity for the Informal Observation / Walkthrough component (Component 3) of the t3 intervention, and student achievement gains.

---

[33] Technically, a two-level hierarchical linear model was used, with students at level-1 and school-specific grade-level blocks at level-2 (e.g., for MATH, with grades 3-7 in the analysis, a school would have 5 such blocks, one for each participating grade.) The same schools are represented in the treatment and comparison samples. Each school-specific grade block contain treatment and control students for a given grade-level in that school. Because students may contribute results for both time points at different grade-levels (spring 2013 and spring 2014), a random effect was initially modeled to account for dependencies in observations arising from repeated measures for individual students; however, the within-student variance was trivially different from zero and therefore the error term was eliminated. Cases with missing outcomes data or pretests were excluded from analysis. For the model used to get the "adjusted" result, for all other covariates, a dummy variable approach is used. With this approach, missing values for a given covariate are set to zero, and another variable is created that takes on a value of 1 to indicate missing values, and 0 otherwise. This is done for each covariate that has at least one missing value.

### Sample for Analysis Involving CST Student Outcomes

The intervention was implemented in 34 of Aspire's 35 California schools. One school opened in fall 2013 and was excluded from analysis because student scores on the CST were not available for the 2012-13 school year. The impact study sample is a subsample of 33 schools (an additional school did not yield any scores for 2013-14 and was therefore excluded from analysis). Schools included in the impact study sample administered the CST in math or ELA in 2013-14. California House Bill 487 suspended mandatory state standardized testing in California public schools during the 2013-14 academic year. Aspire required elementary schools to take either the math or ELA CST in 2013-14 and allowed high schools to opt out of one CST (math, ELA, or a another subject). Principals decided which tests were administered at their school site. A total of 28 schools administered the CST in math in 2013-14 for Grades 3–7, and a total of 8 schools administered the CST in ELA for Grades 3–10 (5 of those schools administered ELA only). All teachers in the eligible grade levels could participate in some or all of the program components. We included students in Grades 3–7 from schools that administered the math CST in 2013-14 in analyses of student math outcomes. Students in Grades 3–10 from schools that administered the ELA CST in 2013-14 are included in analyses of student ELA outcomes. To be included in analysis, students also had to have both a pretest and a posttest—a student contributing a posttest in spring 2014 (i.e., in the treatment group) also had to have a non-missing pretest from spring 2013; a student contributing a posttest in spring 2013 (control) also had to have a non-missing pretest from spring 2012. Analysis was also limited to students who did not change schools. Students who took a modified version of the math or ELA CST or who were given extra time with the regular version of the CST assessment were excluded from analysis.

For the analysis of math, 7800 students in 28 schools fit the criteria described above. For the analysis of ELA, 4066 students in 8 schools met the criteria.

## RESULTS

### Analysis Involving Aspire Instructional Rubric Teacher Outcomes

#### Findings for Domain 2: Aspects of the Classroom Learning Environment

In terms of the Domain 2 scores, on average, teachers' AIR scores positively and statistically significantly increased with exposure to the t3 system.

*Metric 1 (Average Domain Scores).* Figure 3 presents the distribution of Domain 2 scores based on Metric 1 (the average domain score).

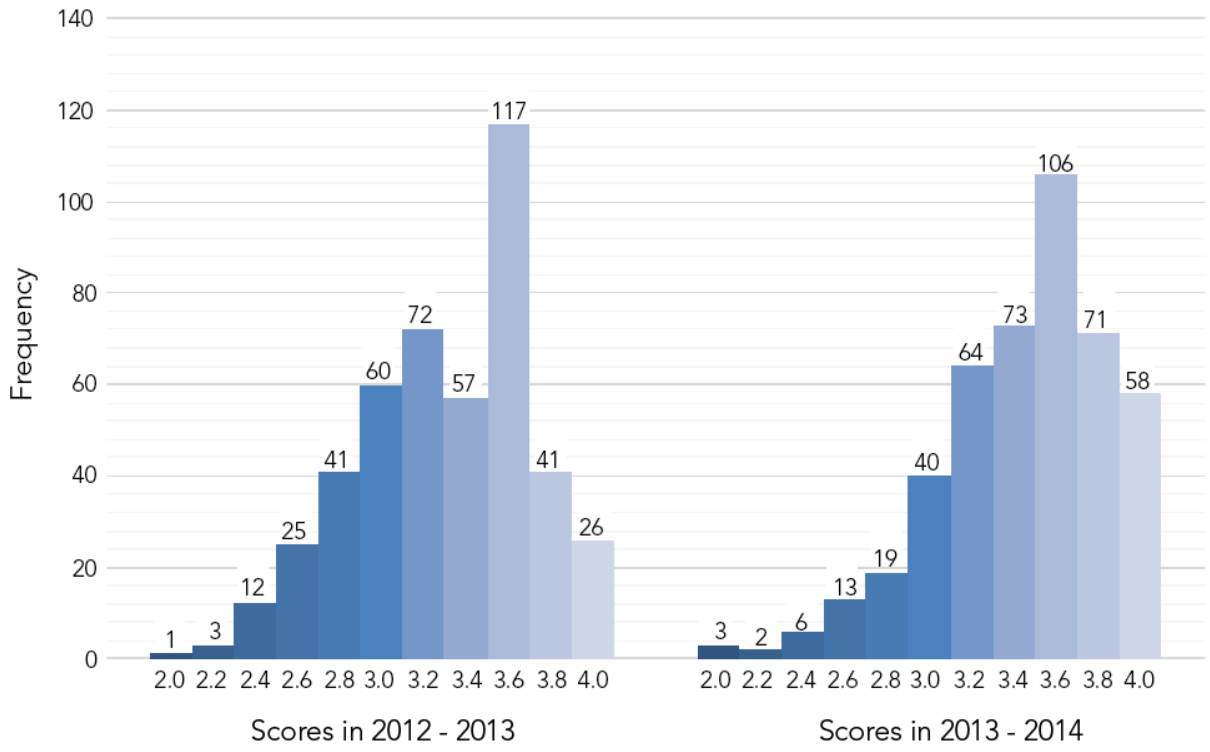


FIGURE 3. DOMAIN 2 ASPIRE INSTRUCTIONAL RUBRIC SCORE DISTRIBUTION FOR METRIC 1 (*N* = 455 TEACHERS)

Table 9 presents descriptive statistics for Domain 2 outcomes using Metric 1. We observed an increase of 0.150 in the mean Domain 2 AIR score from 2012-13 to 2013-14. Similarly, we observed an increase of 0.167 in the median AIR score from 2012-13 to 2013-14.

TABLE 9. DESCRIPTIVE STATISTICS EFFECTS FOR DOMAIN 2 (ASPECTS OF THE CLASSROOM LEARNING ENVIRONMENT) AND METRIC 1 (AVERAGE DOMAIN SCORES)

| | Mean | Median | Standard deviation |
|---|---|---|---|
| **AY 12/13 scores** | 3.284 | 3.333 | 0.413 |
| **AY 13/14 scores** | 3.434 | 3.500 | 0.414 |
| **Gain scores** | 0.150 | 0.167 | 0.412 |

Note. AY stands for Academic Year

In addition to the raw results reported in Table 9, we analyzed the data using an estimation equation, which can give more precise and accurate results, and which allows us to quantify the levels of uncertainty in the results. The findings from the unadjusted and covariate adjusted approaches to analysis for Domain 2 Metric 1 are displayed in Table 10. With both the unadjusted (without covariates) and the covariate adjusted (with covariates) analyses, there was a statistically significant, positive average increase in the mean AIR score from 2012-13 to 2013-14.

The results reach high levels of statistical significance; differences this large result from chance less than 1 out of 1000 times. Therefore, we have a high level of confidence that the AIR score gains are not just due to chance. Expressed as effect sizes, the estimates of the average gains in AIR scores are .357 and .476 standard deviation units. In a bell-shaped distribution of scores, 95% of observations lie within 4 standard deviations. Therefore, the gains observed represent a substantively large shift in the distribution of the outcome.

TABLE 10. AVERAGE GAIN AND COVARIATE EFFECTS FOR DOMAIN 2 (ASPECTS OF THE CLASSROOM LEARNING ENVIRONMENT) AND METRIC 1 (AVERAGE DOMAIN SCORES) (*N* = 455 TEACHERS, J = 35 SCHOOLS)

| | **Unadjusted result** | **Covariate adjusted result** |
|---|---|---|
| **Effect** | Estimate (Standard error) | Estimate (Standard error) |
| **Fixed effects** | | |
| **Average gain in AIR [a] scores** | .148 (.031) *** | .197 (.058)*** |
| **Years at Aspire** | | - .011 (.008) |
| **Years teaching besides Aspire** | | - .005 (.005) |
| **Added gain associated with Bachelor's degree** | | - .063 (.061) |
| **Added gain associated with Master's degree** | | - .053 (.060) |
| **Random effects** | | |
| **School** | .020 (.008) | .020 (.008) |
| **Teacher** | .150 (.010) | .149 (.010) |

[a] AIR stands for Aspire Instructional Rubric

*effect is significant at α=.05; **effects is significant at α=.01; ***effects is significant at α=.001

Note. The random effects estimates are included for technical review. They indicate that the estimation process figures in random school-to-school and teacher-to-teacher differences in average gains. Therefore, the results are robust even with these random fluctuations figured into the estimation.

*Metric 2 (Aspire's Score Conversion).* Figure 4 presents the distribution of Domain 2 scores based on Metric 2 (Aspire's conversion domain score).

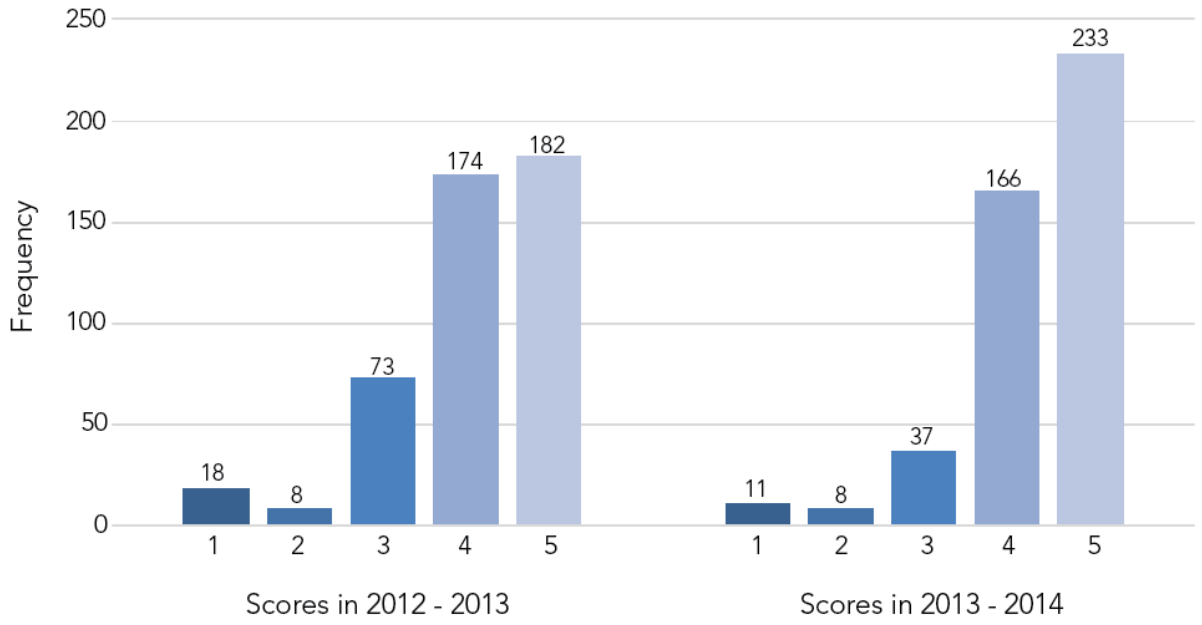


FIGURE 4. DOMAIN 2 ASPIRE INSTRUCTIONAL RUBRIC SCORE DISTRIBUTION FOR METRIC 2 (*N* = 455 TEACHERS)

Table 11 presents descriptive statistics for Domain 2 scores using Metric 2. We observed an increase of 0.237 in the mean Aspire conversion Domain 2 AIR score from 2012-13 to 2013-14. Similarly, we observed an increase of 1 in the median AIR score from 2012-13 to 2013-14. The standard deviation for gain scores was 1.003.

TABLE 11. DESCRIPTIVE STATISTICS EFFECTS FOR DOMAIN 2 (*ASPECTS OF THE CLASSROOM LEARNING ENVIRONMENT*) AND METRIC 2 (ASPIRE'S SCORE CONVERSION)

| | Mean | Median | Standard deviation |
|---|---|---|---|
| **AY 12/13 scores** | 4.086 | 4 | 0.991 |
| **AY 13/14 scores** | 4.323 | 5 | 0.882 |
| **Gain scores** | 0.237 | 1 | 1.003 |

Note. AY stands for Academic Year

The results from the unadjusted and covariate-adjusted models for Domain 2 Metric 2 are displayed in Table 12. With both estimation approaches, there was a statistically significant, positive average increase in AIR scores from 2012-13 to 2013-14.

The results reach high levels of statistical significance; differences this large result from chance less than 1 out of 100 times for the unadjusted result and less than 1 out of 1000 times for the adjusted result. Therefore, we have a high level of confidence that the AIR score gains are not just due to chance. Expressed as effect sizes, the estimates of the average gains in AIR scores are .227 and .418 standard deviation units, for the unadjusted and adjusted results respectively.

TABLE 12. AVERAGE GAIN AND COVARIATE EFFECTS FOR DOMAIN 2 (ASPECTS OF THE CLASSROOM LEARNING ENVIRONMENT) AND METRIC 2 (ASPIRE'S SCORE CONVERSION) (*N* = 455 TEACHERS, J = 35 SCHOOLS)

| | Unadjusted result | Covariate adjusted result |
|---|---|---|
| **Effect** | Estimate (Standard error) | Estimate (Standard error) |
| **Fixed effects** | | |
| **Average gain in AIR [a] scores** | .225 (.070) ** | .414 (.141)*** |
| **Years at Aspire** | | - .024 (.019) |
| **Years teaching besides Aspire** | | - .019 (.013) |
| **Added gain associated with Bachelor's degree** | | - .222 (.150) |
| **Added gain associated with Master's degree** | | - .224 (.150) |
| **Random effects** | | |
| **School** | .093 [.041] | .100 [.043] |
| **Teacher** | .916 [.063] | .900 [.062] |

[a] AIR stands for Aspire Instructional Rubric

*effect is significant at α=.05; **effects is significant at α=.01; ***effects is significant at α=.001

Note. The random effects estimates are included for technical review. They indicate that the estimation process figures in random school-to-school and teacher-to-teacher differences in average gains. Therefore, the results are robust even with these random fluctuations figured into the estimation.

### Findings for Domain 3: Aspects of Teacher Instruction

In terms of Domain 3 scores, on average, teachers' AIR scores positively and statistically significantly increased after exposure to the t3 system.

*Metric 1 (Average Domain Scores).* Figure 5 presents the distribution of Domain 3 scores based on Metric 1 (the average domain score).

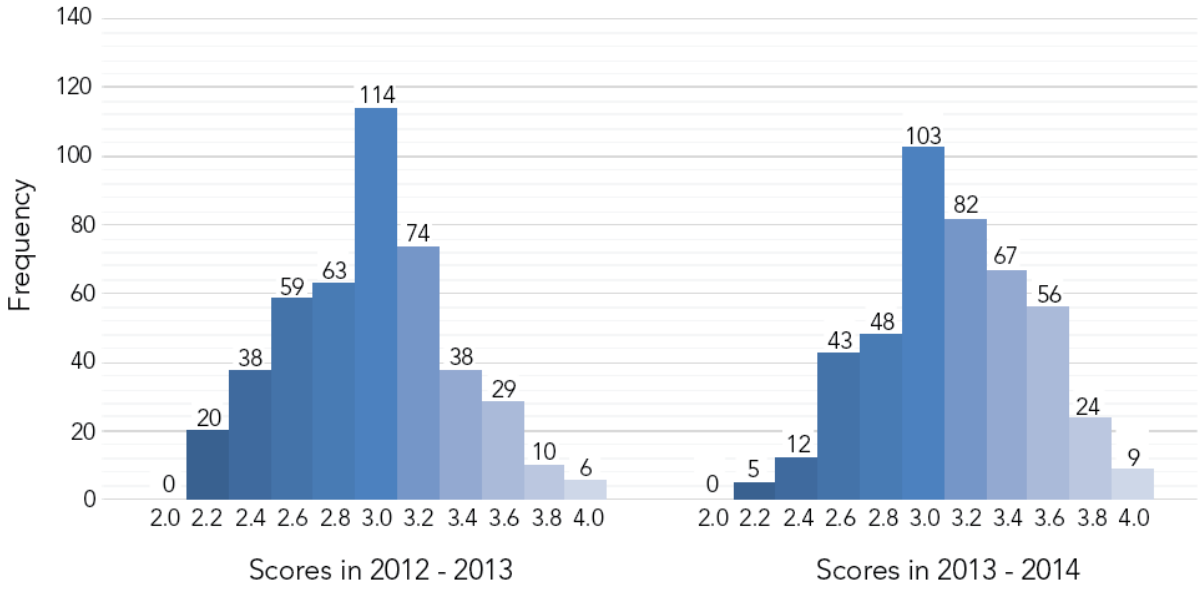


FIGURE 5. DOMAIN 3 ASPIRE INSTRUCTIONAL RUBRIC SCORE DISTRIBUTION FOR METRIC 1 (*N* = 455 TEACHERS)

Table 13 presents descriptive statistics for Domain 3 score using Metric 1. We observed an increase of 0.175 in the mean Domain 3 AIR score from 2012-13 to 2013-14. Similarly, we observed an increase of 0.167 in the median AIR score from 2012-13 to 2013-14.

TABLE 13. DESCRIPTIVE STATISTICS EFFECTS FOR DOMAIN 3 (ASPECTS OF TEACHER INSTRUCTION) AND METRIC 1 (AVERAGE DOMAIN SCORES)

| | Mean | Median | Standard deviation |
|---|---|---|---|
| **AY 12/13 scores** | 2.953 | 3 | 0.399 |
| **AY 13/14 scores** | 3.128 | 3.167 | 0.395 |
| **Gain scores** | 0.175 | 0.167 | 0.365 |

Note. AY stands for Academic Year

The results from the unadjusted and covariate-adjusted estimation approaches for Domain 3 Metric 1 are displayed in Table 14. For both the unadjusted model (with no covariate adjustments) and the adjusted model (with covariate adjustments), there was a statistically significant, positive average increase in mean AIR scores from 2012-13 to 2013-14.

The results reach high levels of statistical significance; differences this large result from chance less than 1 out of 1000 times. Therefore, we have a high level of confidence that the AIR score gains are not just due to chance. Expressed as effect sizes, the estimates of the average gains in AIR scores are .439 and .581 standard deviation units for the unadjusted and adjusted results, respectively.

TABLE 14. AVERAGE GAIN AND COVARIATE EFFECTS FOR DOMAIN 2 (ASPECTS OF THE CLASSROOM LEARNING ENVIRONMENT) AND METRIC 2 (ASPIRE'S SCORE CONVERSION) (*N* = 455 TEACHERS, J = 35 SCHOOLS)

| | Unadjusted result | Adjusted result |
|---|---|---|
| **Effect** | Estimate (Standard error) | Estimate (Standard error) |
| **Fixed effects** | | |
| **Average gain in AIR [a] scores** | .175 (.028) *** | .232 (.051)*** |
| **Years at Aspire** | | - .019 (.007)** |
| **Years teaching besides Aspire** | | - .004 (.005) |
| **Added gain associated with Bachelor's degree** | | - .074 (.054) |
| **Added gain associated with Master's degree** | | - .061 (.054) |
| **Random effects** | | |
| **School** | .018 [.007] | .015 [.006] |
| **Teacher** | .118 [.008] | .115 [.008] |

[a] AIR stands for Aspire Instructional Rubric

*effect is significant at α=.05; **effects is significant at α=.01; ***effects is significant at α=.001

Note. The random effects estimates are included for technical review. They indicate that the estimation process figures in random school-to-school and teacher-to-teacher differences in average gains. Therefore, the results are robust even with these random fluctuations figured into the estimation.

*Metric 2 (Aspire's Score Conversion).* Figure 6 presents the distribution of Domain 3 scores based on Metric 2 (Aspire score conversion domain score).

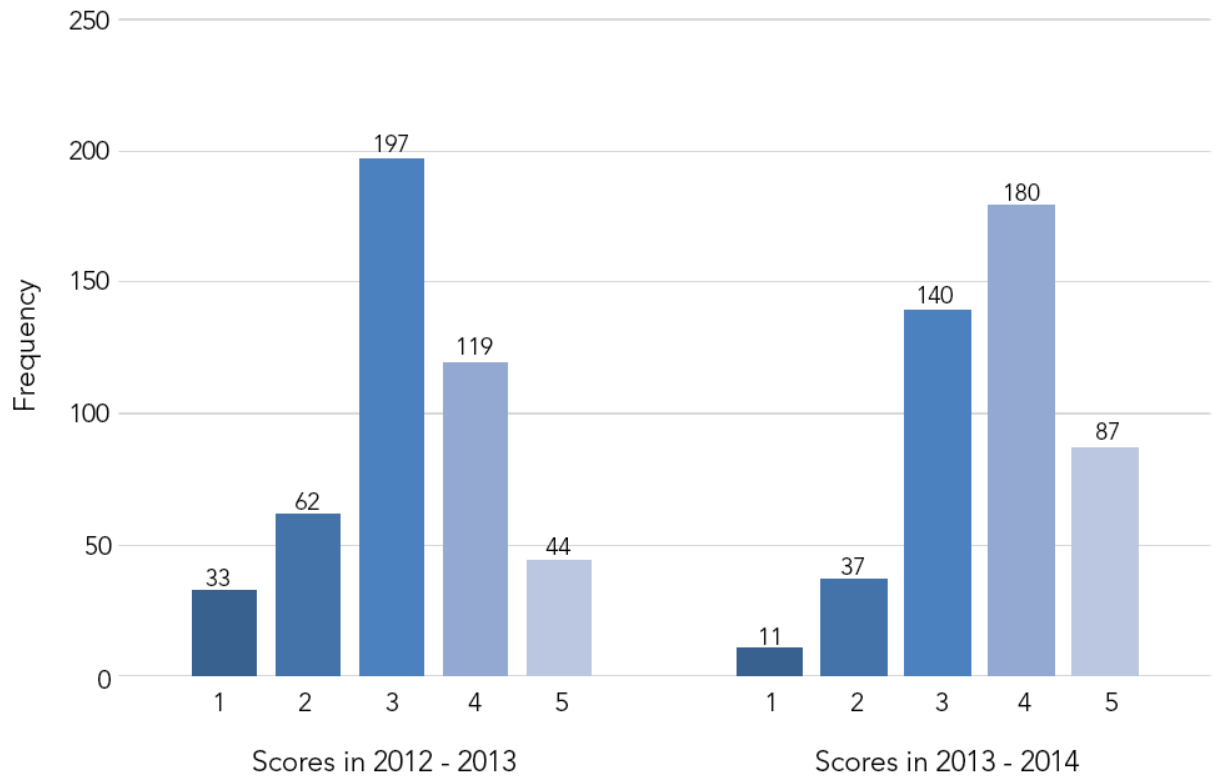


FIGURE 6. DOMAIN 3 ASPIRE INSTRUCTIONAL RUBRIC SCORE DISTRIBUTION FOR METRIC 2 (*N* = 455)

Table 15 presents descriptive statistics effects for Domain 3 score using Metric 2. We observed an increase of 0.475 in the mean Domain 3 AIR score from 2012-13 to 2013-14. Similarly, we observed an increase of 1 in the median AIR score from 2012-13 to 2013-14.

TABLE 15. DESCRIPTIVE STATISTICS EFFECTS FOR DOMAIN 3 (ASPECTS OF TEACHER INSTRUCTION) AND METRIC 2 (ASPIRE'S SCORE CONVERSION)

| | Mean | Median | Standard deviation |
|---|---|---|---|
| **AY 12/13 scores** | 3.174 | 3 | 1.023 |
| **AY 13/14 scores** | 3.648 | 4 | 0.959 |
| **Gain scores** | 0.475 | 1 | 1.026 |

Note. AY stands for Academic Year

The results from the unadjusted and covariates-adjusted models Domain 3 Metric 2 are displayed in Table 16. With both estimation approaches, there was a statistically significant, positive average increase in the average AIR score from 2012-13 to 2013-14.

The results reach high levels of statistical significance; differences this large result from chance less than 1 out of 1000 times. Therefore, we have a high level of confidence that the AIR score gains are not just due to chance. Expressed as effect sizes, the estimates of the average gains in AIR scores are .467 and .537 standard deviation units for the unadjusted and adjusted results, respectively.

TABLE 16. AVERAGE GAIN AND COVARIATE EFFECTS FOR DOMAIN 3 (ASPECTS OF TEACHERS' INSTRUCTION) AND METRIC 2 (ASPIRE'S SCORE CONVERSION) (*N* =455 TEACHERS, J=35 SCHOOLS)

| | Unconditional model | Conditional model |
|---|---|---|
| **Effects** | Estimate (Standard error) | Estimate (Standard error) |
| **Fixed effects** | | |
| **Average gain in AIR [a] scores** | .478 (.070) *** | .549 (.142)*** |
| **Years at Aspire** | | - .045 (.019)* |
| **Years teaching besides Aspire** | | - .011 (.014) |
| **Added gain associated with Bachelor's degree** | | - .113 (.154) |
| **Added gain associated with Master's degree** | | - .054 (.153) |
| **Random effects** | | |
| **School** | .089 [0.417] | .079 [.039] |
| **Teacher** | .969 [0.067] | .959 [.066] |

[a] AIR stands for Aspire Instructional Rubric

*effect is significant at α=.05; **effects is significant at α=.01; ***effects is significant at α=.001

Note. The random effects estimates are included for technical review. They indicate that the estimation process figures in random school-to-school and teacher-to-teacher differences in average gains. Therefore, the results are robust even with these random fluctuations figured into the estimation.

## Analysis Involving California Standards Test Student Outcomes: Impacts

### Mathematics

Results of the analysis of changes in CST math performance associated with the use of t3 are shown in Table 17. The unadjusted result includes the raw means and standard deviations of

the z-transformed scores, as well as counts for students and schools for the analytical sample. The last two columns provide the effect size, that is, the size of the difference between the means for the sample receiving t3, and the comparison group from the year prior, in standard deviation units of the outcome measure, and relative percentile ranking (which tells us where a student at the 50th percentile of performance in the treatment distribution falls with respect to the student at the 50th percentile of the control distribution.) Also provided is the *p* value, indicating the probability of arriving at a difference with a magnitude as large as—or larger than—the magnitude of the one observed, when there truly is no difference. The adjusted row is based on the same sample of students. The mean difference—and therefore the effect size—is adjusted for the effects of covariates, and is therefore potentially more accurate.

TABLE 17. EFFECT SIZES FOR THE END OF THE COURSE ASSESSMENT (MATH *N STUDENTS* = 7800)

| | Timing of outcomes | Means | Standard deviations | No. of students | No. of schools | Effect size | *p* value | Relative percentile standing |
|---|---|---|---|---|---|---|---|---|
| **Unadjusted effect size[a]** | **Spring 2013 outcomes** | 0.020 | .989 | 3914 | 28 | - 0.184 | <.001 | - 7.30% |
| | **Spring 2014 outcomes** | - 0.161 | .983 | 3886 | 28 | | | |
| **Adjusted effect size[b]** | **Spring 2013 outcomes** | 0.020 | .989 | 3914 | 28 | - 0.156 | <.001 | - 6.20% |
| | **Spring 2014 outcomes** | - .134 | .983 | 3886 | 28 | | | |

[a] The unadjusted effect size is Hedges' g.

[b] The adjusted effect size was computed by dividing the regression-adjusted effect estimate by the standard deviation of the outcome distribution. Between-grade differences in the posttest were factored out of the standard deviation in the denominator of the effect size. The *p* value corresponds to the significance test for the estimate of pre-to-post gains in the regression model. The program mean was obtained by adding the regression-adjusted estimate of the pre-to-post gains to the unadjusted control mean.

We observe a negative association between CST math achievement and t3. The change scores are -.184 and -.156 effect size units, for the unadjusted and adjusted results, respectively. In both cases we have a high degree of confidence that the changes are not simply a reflection of sampling variation at the student level, given $p < .001$.

In addition to the unadjusted and adjusted estimates of pre-to-post gains in student achievement, we ran a series of additional analyses using alternative model specifications to assess how robust the results are. The findings were consistent with the above benchmark

results. Estimates of pre-to-post gains ranged between -.199 and -.130 with *p* values all less than .01.

**English Language Arts**

Results of the analysis of changes in CST ELA performance associated with the use of t3 are shown in Table 17. (The interpretation of the effects in the table is the same as in Table 17.)

TABLE 18. EFFECT SIZES FOR THE END OF THE COURSE ASSESSMENT (*N* = 4066 ENGLISH LANGUAGE ARTS STUDENTS)

| | Timing of outcomes | Means | Standard deviations | No. of students | No. of schools | Effect size | *p* value | Relative percentile standing |
|---|---|---|---|---|---|---|---|---|
| Unadjusted effect size[a] | Spring 2013 outcomes | - 0.026 | .998 | 1977 | 8 | - 0.113 | <.001 | - 4.50% |
| | Spring 2014 outcomes | - 0.089 | 1.027 | 2089 | 8 | | | |
| Adjusted effect size[b] | Spring 2013 outcomes | - 0.026 | .998 | 1977 | 8 | - .071 | .004 | - 2.83% |
| | Spring 2014 outcomes | - 0.098 | 1.027 | 2089 | 8 | | | |

[a] The unadjusted effect size is Hedges' g.

[b] The adjusted effect size was computed by dividing the regression-adjusted effect estimate by the standard deviation of the outcome distribution. Between-grade differences in the posttest were factored out of the standard deviation in the denominator of the effect size. The *p* value corresponds to the significance test for the estimate of pre-to-post gains in the regression model. The program mean was obtained by adding the regression-adjusted estimate of the pre-to-post gains to the unadjusted control mean.

We observe a negative association between CST ELA achievement and t3. The change scores are -.113 and -.071 effect size units, for the unadjusted and adjusted results, respectively. In both cases, we have a high degree of confidence that the changes are not simply a reflection of sampling variation at the student level. The *p* values associated with the two outcomes are < .001 and .004, respectively.

In addition to the unadjusted and adjusted estimates of pre-to-post gains in student achievement, we ran a series of additional analyses using alternative model specifications to assess how robust the results are. The findings were consistent with the benchmark results reported in Table 18. Estimates of pre-to-post gains ranged between -.100 and -.043 with *p* values all less than or equal to .05.

**Differences between Subgroups in the Association between t3 Usage and Student Achievement Outcomes**

We examined also whether the association between CST achievement and t3 was different by subgroups.

In math, we found the following.

- A stronger negative association between CST and t3 with higher incoming achievement: there was a reduction in pre-to-post score change of 0.03 effect size units for each one standard deviation gain in incoming achievement ($p$ = .040)
- No difference in association between CST and t3 with student socioeconomic status ($p$ = .307), as measured by eligibility for free or reduced price lunch
- No difference in association between CST and t3 with student gender ($p$ = .476)
- A difference across grades in the association between CST and t3 ($p$ < .001), with regression adjusted estimates of score changes of: -0.100 in 3rd grade, -0.214 in 4th grade, 0.011 in 5th grade, -0.067 in 6th grade, and -0.467 in 7th grade (all in standard deviation units)

In ELA, we found the following.

- A stronger positive association between CST and t3 with higher incoming achievement: there was an increase in pre-to-post score change of 0.036 effect size units for each one standard deviation gain in incoming achievement ($p$ = .061)
- A weak association between CST and t3 with student socioeconomic status ($p$ = .090), with effect sizes of: -0.121 for student not eligible for reduced or free price lunch, 0.020 for students eligible for reduced lunch, and -0.054 for students eligible for free lunch
- No difference in association between CST and t3 with student gender ($p$ = .394)
- A difference across grades in the association between CST and t3 ($p$ = .020), with effect sizes of: -0.068 in 3rd grade, -0.194 in 4th grade, -0.328 in 5th grade, 0.007 in 6th grade, -0.090 in 7th grade, 0.010 in 8th grade, -0.052 in 9th grade, -0.150 in 10th grade

**Relating Achievement to Implementation**

As part of our exploratory analyses, we considered also the association between CST achievement in spring 2014 and levels of implementation. We limited our analysis to math, given the larger sample of schools that provided math outcomes (compared to ELA).

For Component 2, we found no association between the school-level implementation score for the PD Content Library and student CST math achievement (r = .110, $p$ = .578).

We also examined the correlation between the number of peer observation cycles, an indicator of fidelity for the Informal Observation / Walkthrough component of the t3 intervention, and

student achievement gains. We found no correlation between the school average number of observation cycles and CST math achievement (r = .004, $p$ = .984)

## Discussion

In this report, we provided formative and summative results from our evaluation of Aspire Public Schools' (Aspire's) Transforming Teacher Talent (t3) system. We examined (1) levels of attained program implementation, including whether the study reached thresholds for fidelity of implementation on four key components, and (2) changes in Aspire Instructional Rubric (AIR) and California Standards Test (CST) scores that occurred over the course of the implementation period.

Concerning program implementation, thresholds indicating fidelity of implementation were achieved on only one of the four components. Fidelity was achieved with the component corresponding to Aspire's delivery of t3 supports. An earnest effort was made to supply trainers and schools with training necessary to disseminate and use the t3 tools. However, the uptake and delivery of these services in individual schools was lower than the thresholds identified as being important in order to observe impact of t3 on student achievement. This, in part, may explain why implementation did not translate into positive score gains on the CST.

Impacts on AIR scores were positive: on average, there was a positive and statistically significant gain in teachers' AIR scores from 2012-13 (before the t3 system was introduced) to 2013-14 (after the t3 system was introduced). The effect sizes were 0.476 and 0.418 on two metrics for Domain 2, and 0.581 and 0.537 on two metrics for Domain 3. These are substantively important gains (impacts as small as 0.05 – .20 standard deviations have been found to translate into important educational gains). Our interpretation is that impacts on AIR scores are a cause for optimism with regard to the success of t3. The t3 program was chiefly designed to increase instructional quality and effectiveness, and we observe this improvement with the implementation of t3. A small caution is required in the interpretation of these results. One other initiative that was clearly separate from t3 was introduced during the 2013-14 school year—the use of *Common Core Drivers.* These are leaders who work with colleagues from across Aspire regions to deeply understand the Common Core State Standards and share information concerning the standards with fellow teachers. Since we cannot differentiate the effects of *Common Core Drivers* from the components of t3, changes in AIR scores should be considered reflective of both initiatives considered as a bundle.

On the other hand, score changes on the CST were in the negative direction for both math and English language arts (ELA). Lack of impact on achievement may be partially explained by the sub-threshold implementation of several key program components. In addition, the drop in CST scores could reflect more serious confounds that offset the effects of t3. Notably, the transition to the Common Core State Standards had several unintended consequences that very plausibly had a direct influence on the results. First, it likely affected teachers' instructional decisions

related to content coverage—including timing—and which classroom practices were utilized. Second, it greatly changed the testing situation in the state. California was pilot testing the Smarter Balanced Assessment System (an assessment aligned with Common Core State Standards) in mathematics and ELA during the 2013-14 year. The state suspended the requirement of administering the CSTs to give teachers time to focus on refining instruction and to give schools time to prepare to administer the Smarter Balanced computer-based tests effectively. Many districts and schools, including Aspire, chose to administer the CSTs during this transition period to have some measure of academic achievement. Aspire elementary schools were able to administer *either* the mathematics or ELA CST and high schools were able to opt out of one CSTs (mathematics, ELA, or another subject). Principals decided which test(s) would be administered at their school site. The potential misalignment between instruction and assessment, the non-mandatory nature of the assessment, and principals' selectivity in which subjects to test very plausibly had an influence on CST performance in spring 2014; and greatly ambiguates the interpretation of the role of t3 in the changes in CST achievement scores from 2012-13 to 2013-14.

While facing some challenges in implementing this multi-faceted program during a period of policy change in California, we conclude that t3 shows promise of a positive impact on its primary goal, quality of teaching practice, as measured by the AIR.

## References

Abt Associates. (2014). *Investing in Innovation: Evaluating the i3 Program*. Retrieved from http://www.abtassociates.com/projects/2010/investing-in-innovation--evaluating-the-i3-program.aspx

California Department of Education. (2011). *Standardized Testing and Reporting (STAR) Results*. Retrieved on April 22, 2011 from http://star.cde.ca.gov/

California Department of Education. (2009). *Standardized Testing and Reporting (STAR) Program: Information for Parents.* Evanston, IL: Northwestern University. Retrieved from http://starsamplequestions.org/grades_9_11_math.pdf

Koue, K., Jaciw, A.P., & Zacamy, J. (2014). *Aspire Transforming Teacher Talent (t3) System: Fidelity of Implementation and Formative Findings*. (Empirical Education Rep. No. Empirical_Aspire-7025-IR1-Y2-O.2). Palo Alto, CA: Empirical Education Inc.

# Appendix A: Narrative for Aspire's Transforming Teacher Talent System Logic Model

The following narrative describes the inputs, outputs, outcomes, assumptions, confounders, and external factors presented in the logic model (Figure 1).

### INPUTS: TRAINING FOR T3 LEADERS AND T3 LEADERS' INTERACTIONS WITH TEACHERS

Inputs are resources and contributions that are made to support a program. The t3 system logic model has two tiers of inputs. The first tier includes trainings and orientations for t3 leaders led by Aspire Home Office staff. The second tier includes t3 leaders' interactions with teachers.

Trainings and orientations led by the Aspire Home Office for t3 leaders (Component 1) prepare leaders to interact with teachers at their school site or across schools. As such, t3 leaders who receive training and/or orientation from Aspire apply the lessons learned when interacting with teachers.

The second tier of inputs for the t3 system span three additional components (Components 2-4): the PD content library (also known as Purple Planet), informal observations/walkthrough in BloomBoard, and VCs. Interactions between t3 leaders and teachers consist of trainings on Purple Planet, informal observations and coaching conversations for Peer Observations, and facilitation of VCs, respectively.

### OUTPUTS: ACTIVITIES FOR ASPIRE TEACHERS, SCHOOLS, AND THE CHARTER MANAGEMENT ORGANIZATION

The outputs for the t3 system include activities that enhance Aspire teachers', schools', and the Aspire Home Office's professional or organizational learning and development. Outputs most directly benefit Aspire teachers by providing them with new PD resources and techniques for how to best utilize available resources. Outputs result most directly from t3 leaders' interactions with teachers.

#### PD Content Library Outputs

Outputs are processes, actions, and events that use program resources to achieve intended outcomes. Teachers who receive training from Purple Planet Drivers learn about new components of the PD content library. They receive guidance on how to navigate the platform and link to resources (such as the resource library and PD videos) that can support teachers' specific PD goals. This component is intended to reach and benefit all Aspire teachers.

#### Informal Observation/Walkthrough in BloomBoard Outputs

Teachers who work with Peer Observers receive targeted, frequent feedback. It is expected that Peer Observers expose teachers to new, relevant teaching strategies. This output is intended to

reach and benefit teachers with low AIR scores[4] and teachers who are new to Aspire during the 2013-14 school year. Experienced teachers and those with average AIR scores who teach the same grade or content area as a Peer Observer are also eligible to receive the treatment.

Peer Observers are expected to utilize BloomBoard, an online personalized PD tool used by Aspire teachers, during observation cycles. By doing this, Peer Observers could collect additional data on best teaching and coaching practices. By tracking Peer Observers' work and ensuring that they interact with teachers, schools and the Aspire Home Office can create structured PD opportunities at school sites to help schools and the Aspire Home Office staff collect more data on best teaching practices. Outputs and outcomes with direct benefits at the school and Aspire Home Office levels are not measured in this evaluation and included in gray in the logic model.

### VC Outputs

Teachers who participate in VCs build communities of practice beyond their schools sites. They work with peers in their content area and/or grade level to share and learn instructional strategies, share lesson plans, discuss student work, and/or work collaboratively to gain insight on a specific professional goal, like classroom management.

## OUTCOMES: BENEFITS OF THE T3 SYSTEM FOR ASPIRE TEACHERS AND STUDENTS

Outcomes are the direct results or benefits for individuals, families, groups, etc. that result from participation in, or exposure to, a program. It is hypothesized that all outputs will contribute to four major short-term outcomes.

### Short-term Outcomes

1. *Teachers employ new strategies and techniques during class time.* Teachers were exposed to new strategies by utilizing the PD content library; working with, and learning from, Peer Observers; and virtually collaborating with peers. It is hypothesized that teachers will employ new strategies and techniques learned from these resources in their classrooms.
2. *Teachers understand areas for improvement and work to improve in these areas.* All components of the t3 system aim to help teachers improve their practice. Teachers receive training from Purple Planet Drivers on how to set PD goals and review relevant resources to help them achieve goals facilitated through the PD content library. Peer Observers work with teachers to identify and improve specific aspects of their teaching. By working with their peers during VCs, teachers could identify weak areas and seek guidance from others concerning those particular areas. VCLs facilitate professional learning communities to help teachers in the same grade or content area share strategies

[4] Teachers with low AIR scores are defined as those teachers whose AIR scores are below Aspire's system-wide average AIR score (2.97 on a scale of 1- 4) for 2012-13.

and learn from one another. It is hypothesized that by giving teachers opportunities to identify and examine their areas of weakness, they can better target areas for improvement. By providing access to virtual professional learning communities, peers, and online resources, and by receiving encouragement and guidance from peers, it is hypothesized that teachers would work actively to improve their teaching.

3. *Teachers utilize technology and human resources to inform and improve their teaching.* All components provide teachers access to more and new PD resources. Resources include 1) online videos, articles, and reflection questions on the PD content library, 2) highly-effective or master teachers serving as Peer Observers, and 3) grade- and content-specific virtual professional learning communities. It is hypothesized that because teachers have access to a wider range of PD resources, they will utilize these resources to inform and improve their teaching.
4. *Teachers utilize new class activities and lesson plans.* New activities and lesson plans are available in the PD content library, shared with teachers during coaching conversations with Peer Observers, or shared with teachers by peers during VCs. It is hypothesized that teachers will utilize the new class activities and lesson plans they are exposed to.

While the implementation evaluation did not assess organizational changes directly, schools and the Aspire Home Office are also expected to benefit from t3 system inputs and outputs. By creating structured PD opportunities for teachers, the Aspire Home Office and schools will be able to collect more data on best teaching practices. In turn, schools and the Aspire Home Office are expected to have a better and more complete understanding of best teaching and coaching practices.

### Medium-term Outcomes

We anticipate that when teachers use new strategies, technology, human resources, and class activities and lesson plans—along with their increased understanding of, and efforts to address, areas for improvement—their teaching will improve. These efforts are expected to increase the number of highly effective teachers, as demonstrated by improved AIR scores.

The four short-term outcomes are expected to help teachers improve their practice and increase the number of highly effective teachers, as demonstrated through an increase in AIR scores. The change in teacher overall scores on the AIR from the pre- /comparison-period to the post-/treatment-period has been assessed as part of exploratory impact analyses.

One short-term outcome in the logic model describes changes in school practices as a result of t3 outputs: Schools and Aspire Home Office have a better and more complete understanding of best teaching and coaching practices. This short-term outcome is expected to lead to changes in Aspire's recruitment practices, so that they are in line with the Aspire's understanding of best practices for recruiting teachers. In turn, aims ensure that Aspire hires effective teachers, who

are able to succeed within their system. Data on changes in school practices, such as recruitment strategies, have not been collected as part of this evaluation.

### Long-term Outcomes

It is expected that better teaching practices would lead to improved student achievement, as measured through increased CST scores. It is also hypothesized that because Aspire schools and the Aspire Home Office have a better and more complete understanding of best teaching and coaching practices, Aspire will adjust its recruitment practices to ensure that the system hires only teachers who are able to succeed within it. This long-term outcome, nor its related outputs and short-term outcomes, have been evaluated.

#### ASSUMPTIONS

In this logic model (Figure 1), we make the following assumptions about the conditions under which the t3 system will be successful.

- The Aspire Home Office, t3 leaders, and Aspire teachers embrace Aspire's core values, which include an academic culture focused on rigor, the belief that effective teaching is at the heart of student success, and commitment to sharing practices to help catalyze education reform in public schools.
- There are sufficient human resources to implement the intervention.
- Teachers are aware of, and understand, the AIR framework.
- T3 leaders are experienced teachers who exhibit qualities that make them particularly well suited to assume specific leadership roles (i.e. Purple Planet Drivers have an affinity to technology and a willingness to learn new systems).
- T3 leaders assumed their positions voluntarily and are, in turn, motivated to carry out their duties.
- Administrators support the implementation of the t3 system at their school sites.
- Teachers have the prerequisite skills to access online tools associated with the t3 system.

#### CONFOUNDERS[5]

In addition to the t3 system, during the 2013-14 academic year, Aspire introduced several new initiatives. Given the pre/posttest evaluation design, it is not possible to determine if a change in teacher or student outcomes is a result of exposure to the t3 system or other confounding initiatives. The confounding initiatives include the following.

[5] A confounder is a variable that might influence a dependent and independent variable. In the context of the t3 system, a confounder is a competing initiative that is introduced at the same time as the t3 system and could affect the extent to which we can attribute any impacts on teachers or students to their exposure to the t3 system.

- *Aspire Video Library*. This collection includes short video clips of highly effective and master Aspire teachers demonstrating specific teaching strategies from the AIR.
- *New Observation Platform for Coaches and Principals on BloomBoard*. This new observation platform simplifies the process of completing an observation for coaches and principals.
- *Common Core Drivers*. These leaders work with colleagues from across Aspire regions to deeply understand the Common Core State Standards and share information concerning the standards with fellow teachers.

### EXTERNAL FACTORS

In addition to confounders, there are several external factors[6] that may contribute to changes in 2013-14 teacher and student outcomes. These external factors include the following.

- Teachers' access to technology at their school site
- Adoption of Common Core State Standards
- Shifts in teacher demographics
- Changes of school leadership

[6] External factors are existing conditions and occurrences outside of a program that could affect its implementation or its impact on some or all beneficiaries.