CADRE

FEBRUARY 2021

Evaluation of Colorado School Turnaround Network and School Turnaround Leadership Development Grants: Descriptive Analysis of 2015-2020 Cohorts

Benjamin R. Shear, Elena Diaz-Bilello,
Sanford R. Student, and Medjy Pierre-Louis

A report prepared by the Center for Assessment, Design, Research and Evaluation (CADRE) at the CU Boulder School of Education.

School of Education
UNIVERSITY OF COLORADO **BOULDER**

Acknowledgements

We would like to thank Nate Goss, Lindsey Jaeckel, Kate Bartlett, and Marie Huchton at CDE for valuable thought partnership and feedback on drafts of this report. We thank Josh Purdue at CDE for assistance in preparing the necessary data files.

About CADRE

The Center for Assessment, Design, Research and Evaluation (CADRE) is housed in the School of Education at the University of Colorado Boulder. The mission of CADRE is to produce generalizable knowledge that improves the ability to assess student learning and to evaluate programs and methods that may have an effect on this learning. Projects undertaken by CADRE staff represent a collaboration with the ongoing activities in the School of Education, the University, and the broader national and international community of scholars and stakeholders involved in educational assessment and evaluation.

Suggested Citation

Shear, B.R., Diaz-Bilello, E., Student, S.R., and Pierre-Louis, M. (2021). Evaluation of Colorado School Turnaround Network and School Turnaround Leadership Development Grants: Descriptive analysis of 2015-2020 cohorts. Boulder, CO: The Center for Assessment, Design, Research and Evaluation (CADRE).

Please direct any questions about this project to:
benjamin.shear@colorado.edu

Table of Contents

Executive Summary

As part of a broader strategy to help support the lowest performing schools in the state, the Colorado Department of Education (CDE) launched the School Turnaround Network (STN) program in 2014-15 and subsequently launched the School Turnaround Leadership Development (STLD) program the following year. The STN program approaches supports using a comprehensive, holistic perspective by ensuring that entities across three levels (state, district, and school) work together to coordinate school improvement strategies for each school participating in the network. STN participants receive these supports for a three-year period. The STLD program focuses specifically on school leadership development and building the capacity of school leaders to lead and sustain difficult changes at participating schools. The theory of action for these programs and related supports, provided to the lowest performing schools through the Empowering Action for School Improvement (EASI) grants, posits that supports provided by CDE will improve student academic performance and other indicators of school quality over time. To date, 151 schools have received state support through these two programs.

The Center for Assessment, Design, Research and Evaluation (CADRE) at the University of Colorado Boulder was asked by the CDE to form a partnership and carry out studies that would address two areas of interest: 1) build on prior analyses to describe the impact of the STN and STLD programs on student academic performance as evaluated by the state summative assessments; and 2) carry out case studies of select turnaround schools to identify and learn about the work these schools have done as participants in these programs. This report focuses on the first area of interest and provides descriptive analyses of student academic performance at participating schools since the inception of both programs. For this report, we analyzed longitudinal administrative data collected by the state to track the academic performance of schools. The analyses contained in this report focus on the set of schools receiving STN and STLD supports. We also compare the academic performance of these schools to the schools that were eligible to receive these supports but did not participate in these or other EASI programs. Our analyses describe performance trends and patterns for each cohort of STN and STLD schools. We describe trends for each cohort separately rather than in a combined group due to factors such as: differences in STN and STLD program eligibility rules each year, distinct performance trends in each cohort, and the differential impacts of state testing changes on each cohort of schools during this time period.

Key findings from our analyses include:

- On average, we found trends in student achievement consistent with small positive effects for each cohort of STN and STLD schools for years during and after participating in the programs. To put these findings into context, these small positive changes are consistent with the average magnitude of positive effects found in other recent studies of school turnaround interventions.

- Based on examining historical academic performance data, the schools participating in the STLD and STN programs tended to experience a downward trend in performance over time prior to starting these programs. However, for reasons that need to be studied further, these lowest performing schools often tended to experience an increase in performance, as indicated by state accountability ratings, the year prior to receiving supports. This performance increase was consistently observed across each cohort of schools participating in these programs and was also observed in comparison schools that were eligible to participate but did not receive any EASI supports.

- When comparing student achievement in each cohort of STN and STLD schools with schools eligible for but not participating in these programs, no clear pattern of trends surface. That is, differences in average achievement performance trends between the two groups of schools varied across cohorts and programs.

Limitations

As a descriptive study conducted outside of an experimental context, the results shared in this study should be interpreted with considerable caution. The results are useful for identifying performance patterns and trends in the data but using these results to support strong claims about the efficacy (or not) of these programs is limited. A descriptive approach was taken due to the great difficulty of disentangling the effects of these two programs on student achievement from the effects of other factors. Without knowing about the broader context of structural or organizational reforms taking place at both low performing participating and eligible schools, it is unclear whether changes in observed academic performance can be directly attributed to these programs.

Another limitation of this study is that the administrative data analyzed only reflect state summative assessments that may not be sensitive to the changes taking place in these schools. Among STN schools focused on transforming school culture, for example, school climate measures would be considered to be more proximal to the initiatives taking place at these schools compared to more distal state English Language Arts and math tests. In the long-term, the vision and hope for these and other EASI programs is that the set of supports implemented would lead to improvements in academic performance. As highlighted in the educational reform literature, however, detecting significant and larger positive effects in academic performance associated with implementing sustainable school-wide approaches to system changes will often be lagged (Fullan, 2001; Hargreaves and Fullan, 2012; Schleicher, 2018). The small positive trends detected for each cohort of participating schools holds promise that these initiatives are supporting improvements in student academic achievement rather than exacerbating low performance as observed during the years prior to joining these programs.

Conclusion

The limitations we highlight point to the importance of gaining a more comprehensive picture of STN and STLD schools beyond what can be described using standardized test scores. In the fall of 2021, we will conduct the second phase of our study that will include carrying out case studies at a purposeful sample of STN schools that have experienced significant improvement in terms of academic performance. We will undertake these case studies to understand the factors and key practices that have contributed to the success of these schools over time. An important goal for the second phase of this study is to highlight proof-of-concept approaches that appear to have success at these schools, and that could potentially be adapted to meet the needs of other current and future STN or STLD schools.

Introduction

Similar to efforts undertaken by other states, Colorado has invested significant financial resources and personnel into supporting the lowest performing or “turnaround” schools in the state. Over the years, eligible schools received these supports in the form of grants connected to a combination of federal, state, and local sources. During the 2017-18 school year, the state streamlined the application process for these various grants by having all eligible schools apply to one funding application called the Empowering Action for School Improvement (EASI) grant. According to the School and District Transformation Unit at the Colorado Department of Education (CDE), low performing schools eligible to submit an application to EASI can use these funds “to support educator professional development, to implement activities geared toward instructional transformation, or to plan or implement one of the restructuring options that state law requires for schools and districts with persistent low performance” (Jaeckel, Bartlett & Goss, 2020, p. 4).

This report builds on the set of preliminary analyses carried out by CDE staff in the summer of 2020 to examine the progress and performance made by EASI supported schools. In their report, CDE shared findings on student outcomes for schools participating in three support programs that have been in place for several years: The School Turnaround Leadership Development program (STLD), the School Turnaround Network (STN), and the Connect for Success (CFS) program. For this report, CDE requested that we focus our analyses on evaluating outcomes achieved at the school level for the STLD and STN programs. We present a brief overview of these two programs to highlight key differences between the two initiatives.

School Turnaround Leadership Development and School Turnaround Network Programs

Passed in 2014 through S.B. 14-124, the STLD program provides leadership training with the explicit purpose of improving student achievement in the lowest-performing schools and districts in the state. Under this program, schools select a leadership training provider approved by CDE and participate in the training program for a duration of approximately one year. The impetus for this program can be traced to recommendations made by Baker, Hupfeld, Teske and Hill (2013) to CDE that turnaround supports focus on addressing a key challenge found at low performing schools: the lack of school leaders willing to engage in school-wide innovations and transformations while navigating complex social and political systems for implementing difficult reforms. The findings from Baker et al.’s report prompted the creation of the STLD program that focused exclusively on building leadership capacity to carry out a vision for school transformation at low performing schools. As of the 2019-20 academic year (AY), four full cohorts of schools had completed the STLD program. Although the program officially launched in the 2015-16 AY, documentation from school participants is only available for a single Alternative Education Campus (AEC) school. Therefore, 2016-17 represents the first year that a full cohort of schools participated in this program and for which data are available.

The STN program started one year earlier than the STLD program, in the 2014-15 AY. Compared to the one-year duration for the STLD program, STN grant recipients typically spend three years receiving both funding and technical supports from CDE staff. Due to the longer-term nature of this grant, both grant recipients and CDE agree to a set of commitments that require the two institutions to implement targeted interventions by involving individuals situated at the state, district, and school levels. For school recipients, expectations for their engagement include appointing a district point person to participate in all network activities, implementing the improvement strategies as outlined in their memorandum of understanding with the state, and engaging in a performance management process with the district and CDE. For CDE, expectations for their support role include dedicating staff to support and partner with network schools, providing a diagnostic needs assessment to help identify the best interventions, and providing professional development. The underlying theory of change for the STN program is that student performance can only improve if the following four key areas are addressed from a systems standpoint (i.e., the state, district, and schools working together): leadership, instructional transformation, culture and climate, and talent management. The exact nature of supports and interventions provided to each school can vary depending on specific needs identified relative to those four areas. Therefore, a wide range of strategies and programs are implemented by schools participating in the network. One school's approach may entail working with a consultant recommended by CDE to implement programs geared toward improving student engagement and staff morale; in the case of another school, an optional needs assessment conducted by CDE may guide that school to modify the literacy strategies used in the lower grades and to focus on project-based learning opportunities across grades to improve student engagement. As of the 2019-20 AY, six full cohorts have been engaged with the STN program since its inception in 2014-15, although schools in the most recent year had only completed their first year of participation in 2019-20.

STLD and STN Eligibility

CDE identifies schools eligible to participate in either the STLD or STN programs primarily based on annual ratings received from the state's School Performance Framework (SPF) accountability system. Each year since 2009-10, every school receives an accountability rating from the state. In order from lowest to highest, the category ratings are: Turnaround, Priority Improvement, Improvement, or Performance. The exact calculations for SPF scores have changed slightly across years but are based primarily on student performance on state standardized test scores. Each separate Elementary, Middle, or High School receives a rating each year. In Elementary and Middle schools, the rating is based on average student achievement on state test scores as well as student growth as measured by Student Growth Percentiles (Betebenner, 2009). The exact scoring formula is complicated and incorporates both overall student performance and performance reported separately by student subgroups. At the High School level, measures of graduation, dropout rates, and post-secondary enrollment are also factored into the ratings. Each school can earn up to 100% of the possible points (which varies depending on school size and demographics), and each school is then assigned one of the four category ratings based on the percentage of points earned.

The SPF ratings are important to understand in this context because eligibility to receive the STLD or STN support funding depends on a school's SPF rating. In brief, the theory of action is that CDE can use the SPF ratings to identify schools that need additional supports to improve teaching and learning. As a result, the state makes a number of different supports available to schools receiving the lowest two SPF ratings (Turnaround and Priority Improvement). Although the eligibility criteria have changed from year to year, in general schools that receive either of the two lowest SPF ratings can apply for grant funding to support a school's participation in the STLD or STN programs. In recent years, Federal accountability ratings have also been factored into eligibility criteria. Table 1 presents the set of rules used to determine eligibility to apply for the STLD and STN programs from the first year that each program was implemented through the 2019-20 school year. Something to consider when interpreting the data and analyses below is that there is a delay between when a school is eligible to apply for one of the two grants and when participation would actually begin. As an example, schools participating in the most recent (2019-20) cohorts of the STLD or STN programs would have been identified as eligible based on 2018 SPF ratings, which were derived using student achievement data from the 2017-18 academic year.

Table 1. Eligibility Requirements, Eligibility Counts, and Participation Counts for STLD and STN Programs, by Year.

Program	Cohort	SY First Funded	Eligibility Requirement	Eligible Schools	Schools in Cohort	% Selected
School Turnaround Leadership Development Program (STLD)	1	2015-16	PI/T on 2014 SPF final	158	0	0
	2	2016-17	PI/T on 2014 SPF final	155	31	20%
	3	2017-18	PI/T on 2014 SPF or 2nd round with PI/T on 2016 SPF preliminary	260	48	19%
	4	2018-19	PI/T on 2017 SPF and/or federally identified (comp, TS, ATS) for 17-18	267	41	15%
	5	2019-20	PI/T on 2018 SPF and/or federally identified (comp, TS, ATS) for 18-19	327	39	12%
School Turnaround Network (STN)	1	2014-15	PI/T on 2013 SPF final	150	9	6%
	2	2015-16	PI/T on 2014 SPF final	159	14	9%
	3	2016-17	PI/T on 2014 SPF OR less than 10% at benchmark on ELA & Math in 2015 at any grade level	263	9	3%
	4	2017-18	PI/T on 2016 SPF preliminary	164	14	9%
	5	2018-19	PI/T on 2017 SPF final and/or federally identified (comp, TS, ATS) for 17-18	262	11	4%
	6	2019-20	PI/T on 2018 SPF final and/or federally identified* (comp, TS, ATS) for 18-19	320	7	2%

Notes: a) PI=Priority Improvement, T=Turnaround, ELA=English Language Arts, comp=Comprehensive Support, TS=Targeted Support, ATS=Additional Targeted Support. For more information on what these terms mean, visit https://www.cde.state.co.us/accountability. b) The eligibility counts in 2018-19 and 2019-20 differ slightly between STLD and STN because there were a small number of schools that participated in STLD without meeting the recorded eligibility rules, and we counted these schools as "eligible."

Table 1 also reports the total number of schools identified as eligible to participate in each STLD and STN cohort. These counts reflect the number of schools included in the data used for the analyses in this report, as described in more detail below. As a result, there are no schools in the first cohort of the STLD program, because the single school recorded as participating in this cohort by CDE was designated as an Alternative Education Campus (AEC), and hence excluded from our analyses (see details below). Table 1 reveals that, although there were between 150 and 327 schools eligible to participate in these programs each year, only 20 percent or fewer of all eligible schools participated in either STLD or STN programs each year. In this report, we use the eligibility rules specified in Table 1 to identify the entire population of all eligible schools for these programs in the state each year.

Although not evident in Table 1, there are also schools that participated in multiple cohorts of the two programs. As a result, the counts in Table 1 represent 151 unique schools that participated in at least one STN or STLD cohort. Each of the 64 unique schools participating in the STN only participated in a single STN cohort, but over half of these schools also participated in at least one STLD cohort – 20 participated in one STLD cohort and 16 participated in two or more STLD cohorts. Overall, at least three schools in each STN cohort participated in at least one STLD cohort at some point between 2016-2020. Of the 123 unique schools that participated in at least one STLD cohort, 32 participated in two or more STLD cohorts, with one school participating in all four STLD cohorts from 2016-2020, and 36 also participated in the STN at some point in this time period. This is a good reminder that although our analyses focus on schools that participated in the STN or STLD in specific years (cohorts), many of these schools were likely pursuing additional interventions or opportunities to improve student achievement. Further, an STLD school could participate in the program for multiple years if they opt to send different staff members to attend the leadership training program at different timepoints. The 64 schools participating in the STN were spread across 12 districts, while the 123 schools participating in the STLD were spread across 26 districts.

Before moving into the set of questions used to guide the analyses, we first summarize findings from recent studies focused on evaluating the impact of other “turnaround” interventions intended to improve student achievement for the lowest performing schools in other states. We include this review to provide a comparative frame of reference for considering the results and findings highlighted in this report.

Turnaround Studies Summary

We focused on studies produced within the last ten years to consider the magnitude of effects found based on different “turnaround” interventions designed to improve student achievement. In 2015, under the Every Student Succeeds Act (ESSA), the School Improvement Grants (SIG) offered to schools located at the bottom five percent of the performance distribution (based primarily on test scores) in their respective states since 2009, was revamped. Prior to 2015, the SIG grants were criticized as being too prescriptive and the re-authorization of these grants under ESSA aimed to provide states and districts with the latitude to implement evidence-based interventions selected to meet the specific needs of each school (Sun, Penner, & Loeb, 2017). In the set of studies we reviewed, a mix of turnaround interventions are studied. Some of the studies we reviewed focused on evaluating the impact of interventions during the pre-2015 SIG era. At that time, there were four SIG turnaround models implemented: Transformational, Turnaround, Restart, and School Closure. The Transformational Model best resembles the STN and STLD programs evaluated in this report since that model focused broadly on implementing changes in the areas that the STLD and STN programs focus on. The Turnaround Model entailed replacing the principal of a school and dismissing a minimum of 50% of the school's staff. The Restart Model required the district to reopen a school using a charter operator, management organization or an educational management organization. The School Closure Model entailed shutting down a school and re-enrolling their students in other schools in the district.

Since the start of the SIG grants to the present, defining the model or intervention implemented at these lowest performing schools is not clear-cut. In the case of a school that engaged in a Transformational Model before 2015, the district may have decided to dismiss the principal and 50% of their teachers a year or two after implementing the Transformational Model strategies at that school.

For this school, the staff dismissals implemented two years later would mean that this school adopted both Turnaround and Transformational Models. Presently, the blending of interventions that cut across different SIG models used in the past is not uncommon although the hope is that programs such as STN and STLD can provide more targeted interventions.[1] The fact that schools can still employ a variety of strategies that can and do cut across different SIG models, however, makes it challenging to anticipate the magnitude of effects on student achievement that any single turnaround initiative is likely to have. Here we summarize findings from recent studies on a range of turnaround programs to provide a sense for plausible effects we might expect the STLD or STN programs to have on student achievement. Overall, findings reported across studies on the success of turnaround initiatives are mixed, with effect sizes[2] ranging from null (i.e., not statistically significant) to small or medium positive effects.

Schueler et al. (2020) recently conducted a meta-analysis of 67 studies focused on evaluating the efficacy of interventions aimed at improving student achievement in low-performing schools. This recently published meta-analysis includes many of the studies reviewed in this summary. The authors restricted their review of studies to those that evaluated achievement against a comparison group. Schueler et al. found that on average, these programs reported small to medium-sized, positive, statistically significant effects on student achievement in math (0.062 standard deviations on average), and small, positive (though not statistically significant) effects in English Language Arts (ELA) achievement (0.016 standard deviations on average), primarily on state summative tests. A key takeaway from this meta-analysis is that despite mixed findings, interventions implemented for one year can produce positive effects in some cases, although longer-term programs appeared to have a larger effect on outcomes.

While there are not many examples in the literature of turnaround programs that are more similar to the STLD in nature (i.e., programs that provide one-year treatments that focus on developing school leadership), there are several studies that suggest the effects of turnaround programs are best observed and maintained over time, especially when the intervention is provided for several years. In a study of 65 schools that received school improvement grants in Texas, Dickey-Griffith

[1] In Colorado, we see the application of different SIG models being implemented at schools in districts such as Denver Public Schools (DPS). DPS has implemented interventions at Manual High School, for example, that cut across practices found in the Transformational, School Closure, and Turnaround SIG models.

[2] To understand the practical implications of the effect sizes reported within the educational context, the range of effect sizes captured in the studies reviewed would be considered to be small if below 0.05 standard deviations, "medium" if between 0.05 and 0.2, and "large" if above 0.2 when using standardized test scores as an outcome measure (Kraft, 2020). In educational research contexts, even effect sizes classified as small can be considered to be practically significant.

(2013) used a difference-in-differences approach to evaluate the impacts of receiving these grants and found negative impacts on student standardized tests in elementary and middle school. However, for that study, the author acknowledges that effects were only estimated after one year and that more positive effects may be detected after several years of implementation. De la Torre et al. (2012) also used a difference-in-difference approach to study turnaround initiatives implemented for four years in low performing elementary schools in Chicago. In that study, the authors found small effects that were not statistically significant in the first year and reported significant and larger effects in later years of the intervention on reading and math standardized tests. However, some studies find that even three years may not be enough to have effects on standardized tests in reading and math. In Dee and Dizon-Ross (2017) the authors used a regression discontinuity approach to estimate the causal effect of turnaround programs in 1,172 Louisiana schools and found null effects. Additionally, in another large-scale study evaluating the effects of turnaround interventions on 290 schools across 22 states that received federal school improvement grants (SIGs) for three years (between 2010-2013), Dragoset et al. (2017) found that implementing these SIG models in a school had no statistically significant effects on students' math and reading achievement.

In turnaround programs that were longer-term or multi-year, school-wide interventions requiring the involvement of state and district actors (more similar in nature to STN), the effects on student achievement were slightly larger and better sustained over time than in other studies but were still mixed. A study of three cohorts of turnaround programs in the state of Massachusetts found medium to large, statistically significant effects on math and reading achievement after the first, second, and third years of turnaround program implementation using a comparative interrupted time series (CITS) design (LiCalsi, Citkowicz, Friedman & Brown, 2015). More recent studies on these turnaround programs in Massachusetts, completed in 2016 by LiCalsi and García Píriz as well as in 2017 by Kistner, Melchior, Marken, and Stein, continued to find positive, statistically significant effects after each year of program implementation, across primary and secondary grade levels. However, in a study of three different types of turnaround programs (school restart, charter management, or a turnaround model managed by a district) in Tennessee schools, Zimmer, Henry and Kho (2017) found that only the district management turnaround model yielded positive effects on student achievement in math, reading, and science. Both of these studies analyze a state department of education's efforts to support low performing schools through grants and program supports over a multi-year process, yet the results of these two studies were variable.

The research literature reviewed suggests that the effects of turnaround initiatives and interventions are inconclusive. Even though the results are mixed across the literature, key themes emerge that suggest longer-term interventions tend to be more successful and that turnaround programs seem to be more effective in improving math achievement relative to reading achievement. However, the effects of these interventions are likely to vary depending on context, the design of the studies, and the set of interventions used. While in some studies of turnaround programs have found positive effects in some subject areas across multiple grade levels, other studies using similar designs evaluating similar programs have found null effects associated with these turnaround programs. The implications of these findings suggest that the results from our descriptive study may also yield mixed results since STN and STLD schools engaged in a wide range of programs and implementation approaches to improve student achievement outcomes. These variations in programs may also apply to the set of comparison schools eligible for, but not participating in, the STN and STLD programs. At best, we would hope that even if evidence of effects were marginal, the results would trend in a positive direction for the cohorts of schools participating in the STN and STLD programs.

Guiding Questions

This report presents descriptive statistical analyses based on statewide administrative data intended to address three questions:

- Which schools received STLD or STN supports and how do they compare to the population of all schools statewide and to other schools that were eligible to receive supports in terms of student demographic characteristics, prior student achievement, and SPF ratings?

- What type of changes are observed in SPF ratings, student achievement, and student growth percentiles for schools participating in STLD and STN support programs?

- How do the trends in student achievement and growth at schools participating in the STLD and STN programs compare to these same metrics at other eligible, demographically comparable schools that did not participate in either program?

These analyses are intended to address needs identified by CDE staff in multiple ways. Answers to the first question will provide a detailed description of the schools and students being served by the STLD and STN funds. Understanding how these schools may differ from the population of all schools that were eligible might also help reveal why certain schools do or do not participate, and potentially allow CDE to improve the process of matching supports to schools. Answers to the second question provide preliminary data about whether there is evidence that the programs are having their intended effects in terms of improving overall school performance as represented by the SPF, or in terms of improving student achievement as measured by state summative tests and student growth metrics. Finally, the third question is intended to provide preliminary evidence about whether it might be reasonable to attribute any changes observed in student achievement to participation in the programs.

Data and Analytic Samples

The descriptive analyses in this report draw on a number of different data sources provided by CDE. We provide more details about the data construction process in Appendix A. Here we provide a brief overview of the data sources and the primary outcome variables. For the analyses in this report, we focus primarily on school SPF ratings and student achievement test results. These data are available through the 2018-19 academic year, as state assessment data were not collected or reported during the 2019-20 academic year due to the COVID-19 pandemic.

Program Participation

We received files indicating which schools received STLD funds in each year from 2015-16 through 2019-20, STN funds each year from 2014-15 through 2019-20, and other non-STLD and non-STN funds through the EASI application process in 2017-18 through 2019-20 years. These counts are described in Table 1. When presenting results by cohort below, for schools that participated in more than one STLD cohort, we generally assign that school to the first cohort they participated in when tracking trends over time, unless otherwise noted. Because we describe trends for STN and STLD cohorts separately, there are some schools that are represented in both STN and STLD results.

School and Student Demographic Data

School and student demographic data are drawn from CDE's October count enrollment files, which report official school enrollment for each public school in Colorado and include data about student demographic characteristics of students enrolled at each school. These demographic characteristics are based on the major student sub-groups reported for CDE accountability purposes and include the percent of students identified as male or female, the percent of students identified as belonging to a minority racial/ethnic group[3], the percent of students eligible for free or reduced-price lunch (FRL), the percent of students identified as English Language Learners (ELL), and the percent of students with an individualized education plan (IEP). We also received files indicating whether each school was designated as an Alternative Education Campus (AEC)[4] each year and whether each school was designated as "rural" or "small rural" per the CDE guidelines.[5]

Student Performance Framework (SPF) Ratings

Historical SPF ratings for each school from the 2009-10 AY through 2018-19 AY, including whether schools were identified for Comprehensive Support or Targeted Support under Federal ESSA rules in 2017-18 through 2018-19, were provided by the CDE accountability office. School SPF ratings are based primarily on annual student achievement test scores. These tests are administered each spring, while the SPF ratings are reported in the fall. As a result, the SPF rating for a school in any given AY is based on test score data from the prior AY. A school that received funding for the STLD program in the 2018-19 AY, for example, would have been eligible to apply for this funding if the school's SPF rating were Turnaround or Priority Improvement in the 2017-18 AY, but the SPF rating that applied for the 2017-18 AY would have been based on standardized testing from the spring of the 2016-17 AY. There are some missing data in the SPF ratings. First, due to the transition in tests used for state accountability during the 2014-15 AY, there were no SPF ratings assigned in the 2015-16 AY, which would have used the spring 2015 test results. Second, in a small number of cases an official SPF rating may not be assigned to a school due to exceptions such as low participation rates in accountability testing or other exceptions, although these are relatively rare.

Student Achievement Data

We summarize student achievement data from Spring 2009 through Spring 2019 at each school using two different metrics: average test scores and average student growth percentiles (SGP). We include these metrics for tests measuring student achievement in math and English Language Arts (ELA) in grades 3-8 for every year, and in grades 9-11 in years when year-end accountability tests were administered in these grades. We summarize achievement data at the school level because the STLD and STN programs are awarded to an entire school. A more detailed description of the achievement data preparation and calculations is provided in Appendix A.

[3] CDE pupil membership guidelines classify the following groups of students as minorities: American Indian or Alaskan Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or other Pacific Islander, and students falling under two or more races.

[4] Our analysis excludes AECs since these schools are evaluated using a separate accountability framework and are classified as AECs due to serving distinct student populations relative to other public schools in the state. More information on AECs can be located here: https://www.cde.state.co.us/accountability/stateaccountabilityaecs

[5] See: https://www.cde.state.co.us/sites/default/files/documents/ruraledcouncil/download/ruraldefinitionletter12813.pdf. We considered a school "non rural" for a small number of schools where we didn't have a current designation.

The average test score for each school is computed by averaging across all valid test scores at each school in each year. Before calculating average test scores for each school, we first standardized students' test scores by grade, subject, and year at the statewide level. This is to account for the fact that different tests are administered in each grade and subject, and that there have been changes to the tests used within grades over time. These changes to tests across grades and years complicate interpretation of the student achievement trends during the years the STN and STLD programs were implemented, something we discuss further below. As a result of standardizing test results, a value of 0 represents the statewide average score in a given grade, year, and subject; positive values represent averages that are higher than the statewide average and negative values represent averages that are lower than the statewide average.

We also summarize trends in SGP over time. Since 2009, Colorado has used the SGP methodology to measure "growth" in student learning each year. Each student's SGP indicates how a student's current year test score compares to other students in the state who had similar prior test scores in the same subject. We summarize these at the school-level by computing the mean SGP (MGP) at each school, which differs slightly from the median SGP used in the SPF calculations. We use the mean SGP because it has been shown to have lower sampling variability and better statistical properties (e.g., McCaffrey et al., 2015). The MGP is intended to provide an indicator that better represents a school's effect on student learning than does the average test score in a single year, which does not take into account how much progress students have made in the past year. Two limitations of MGPs relative to average test scores are that they cannot be computed for as many students (a student must have appropriate prior year test scores), and they tend to contain more random sampling error and measurement error than do average test scores.

Analytic Samples

The analyses beginning in the next section are based on two different analytic samples. The first, represented by the counts in Table 1, includes the population of all non-AEC schools and is based on records in the October count data files. This analytic sample represents the population of participating schools and is used when presenting descriptive statistics about demographics and SPF ratings (where applicable). The second analytic sample is based on schools for which student achievement data are available. As noted above, there were no achievement tests administered in 2019-20, so the second analytic sample based on achievement data does not include the most recent STN or STLD cohorts (Cohorts 6 and 5, respectively). In addition, there are some years in which a school did not have sufficient achievement test score data to be included in the analyses. As a result, although most schools from the earlier cohorts are included in the second analytic sample, there are a small number of additional school by year observations not included in the second analytic sample (see Appendix A for more details).

Comparison of Participating and Eligible Schools

This section describes the populations of participating and eligible schools in more detail. Table 2 presents the number of participating schools for each cohort of the STLD and STN programs, broken out by school structure indicating whether each school enrolled students in Elementary, Middle, or High School grades, or a combination of multiple grade levels. The table reports these counts for all schools in each STN cohort and also for all schools participating in the STLD program each year; the counts for the STLD program include all schools participating each year, which includes some schools that participated in multiple years, to be consistent with the counts in Table 1. The table also reports the average percent of schools of each type across cohorts, averaged across participating schools ("Avg. Part. %") and averaged across all eligible schools for each cohort ("Avg. Elig. %"). Table 2 indicates, for example, that across STN cohorts, 70% of participating schools were Elementary schools, while only 55% of eligible schools were Elementary schools, indicating that Elementary schools were relatively more likely to receive STN supports. For reference, the table also reports the percentage of schools of each type statewide in the 2018-19 academic year ("% All Schools in 2018-19"; these proportions were similar across academic years from 2015-2020), which indicates that statewide, 52% of schools were Elementary schools and 15% are Middle schools, etc. The distribution of school types among eligible schools is similar to the distribution statewide.

Table 2. Counts of Eligible and Participating Schools in STLD and STN, by Cohort and School EMH Level.

		Cohort						**Averages**		
Program	**Level**	**1**	**2**	**3**	**4**	**5**	**6**	**Avg. Part. %**	**Avg. Elig. %**	**% All Schools in 2018-19**
STN	E	8	9	6	9	7	5	70%	55%	52%
	M	1	2	3	5	1	2	22%	16%	15%
	H	0	0	0	0	2	0	3%	12%	15%
	EM	0	2	0	0	1	0	4%	9%	9%
	MH	0	1	0	0	0	0	1%	6%	5%
	EMH	0	0	0	0	0	0	0%	4%	4%
	Total	**9**	**14**	**9**	**14**	**11**	**7**			
STLD	E		20	32	22	23		61%	54%	52%
	M		6	5	10	4		16%	15%	15%
	H		1	7	4	6		11%	14%	15%
	EM		2	2	1	3		5%	8%	9%
	MH		2	2	4	3		7%	6%	5%
	EMH		0	0	0	0		0%	4%	4%
	Total		**31**	**48**	**41**	**39**				

Table 2 highlights that on average the vast majority (96%) of schools participating in the STN program enrolled only Elementary and/or Middle School students (levels E, M, or EM), while 80% of eligible schools enrolled only E and/or M students across cohorts, and statewide approximately 76% of schools did. For STLD, the mix of EMH levels is more balanced; 82% of participating schools enroll students in only E/M grades across cohorts, while 77% of eligible schools do. These patterns appear to be relatively consistent across cohorts within the two programs. Because the achievement tests administered in the E/M grades versus H grades differ and have changed in different ways over time, these differences are relevant when forming comparison groups and interpreting results in the analyses below.

To further characterize similarities and differences between schools participating in the STLD and STN programs, other eligible schools, and the population of all schools in Colorado, we turn to the data presented in Tables 3a and 3b below.

Table 3a summarizes student demographics for all schools participating in the STLD or STN programs during the first year they participated, and also reports the same summary statistics for all schools that were ever eligible for one of these programs in the first year they were eligible. The first row, for example, reports that for the 64 unique schools that participated in one of the six STN cohorts, on average 76% of the students at these schools were FRL-eligible, which is often used as a proxy for poverty levels in a school or community. In contrast, among all schools that were ever eligible for the STN program (including the schools that eventually participated), on average 64% of enrolled students were FRL-eligible. These figures were nearly identical for the 123 unique schools participating in versus ever eligible for the STLD program. The final two columns report the mean and standard deviation of each variable for all schools statewide in the 2018-19 academic year. Statewide, approximately 45% of students at each school were FRL eligible, with a standard deviation of 27.3 percentage points across all schools.

Table 3a. Average School-Level Demographics, Locale, and Enrollment by Program and Eligibility.

	STN		STLD		All Schools (2018-19)	
Variable	**Participants**	**All Eligible**	**Participants**	**All Eligible**	**Mean**	**SD**
% FRL	75.6%	64.0%	76.5%	63.8%	45.1%	27.3%
% Minority	68.5%	61.8%	75.4%	61.9%	44.9%	26.5%
% ELL	26.7%	27.6%	33.6%	26.7%	16.1%	19.1%
% IEP	13.4%	11.7%	13.3%	12.1%	11.3%	5.0%
Rural	14.1%	25.4%	11.4%	24.3%	27.6%	44.7%
Enrollment	441	460	512	472	489	413
N Schools	64	583	123	503	1744	

Table 3a indicates that, relative to all eligible schools, schools participating in the STN and STLD programs tended to serve slightly higher proportions of students eligible for FRL, higher proportions of minority (non-white) students, and were less likely to be in rural communities.

Schools participating in the STLD program enrolled a slightly higher proportion of ELL students than all eligible schools and all schools statewide on average. In terms of enrollment size, while STN schools were similar in size or slightly smaller relative to eligible schools, on average, STLD schools enrolled slightly more students than eligible schools, although the differences do not appear substantial relative to the statewide distribution of enrollment numbers. The percentage of students with an IEP was relatively similar across participating and eligible schools and were both similar to the statewide average of 11% of students. Taken together the patterns show that schools eligible to participate in the STN and STLD programs tended to serve higher proportions of students from historically disadvantaged groups relative to the statewide population of schools, and that among eligible schools, those participating in the programs served slightly higher percentages of these students.

Table 3b. Prior Year Average Test Scores and MGPs by Program and Eligibility.

		STN		**STLD**		**All Schools (2018-19)**	
Subject	**Variable**	**Participants**	**All Eligible**	**Participants**	**All Eligible**	**Mean**	**SD**
ELA	MGP	45.10	47.40	46.77	47.22	50.23	6.96
ELA	Avg. Score	-0.51	-0.43	-0.56	-0.43	-0.04	0.41
MATH	MGP	44.11	46.13	44.94	46.48	50.23	8.20
MATH	Avg. Score	-0.55	-0.45	-0.59	-0.44	-0.06	0.43
N MGP		**54**	**442**	**97**	**365**	**1604**	
N Avg. Score		**54**	**444**	**97**	**367**	**1613**	

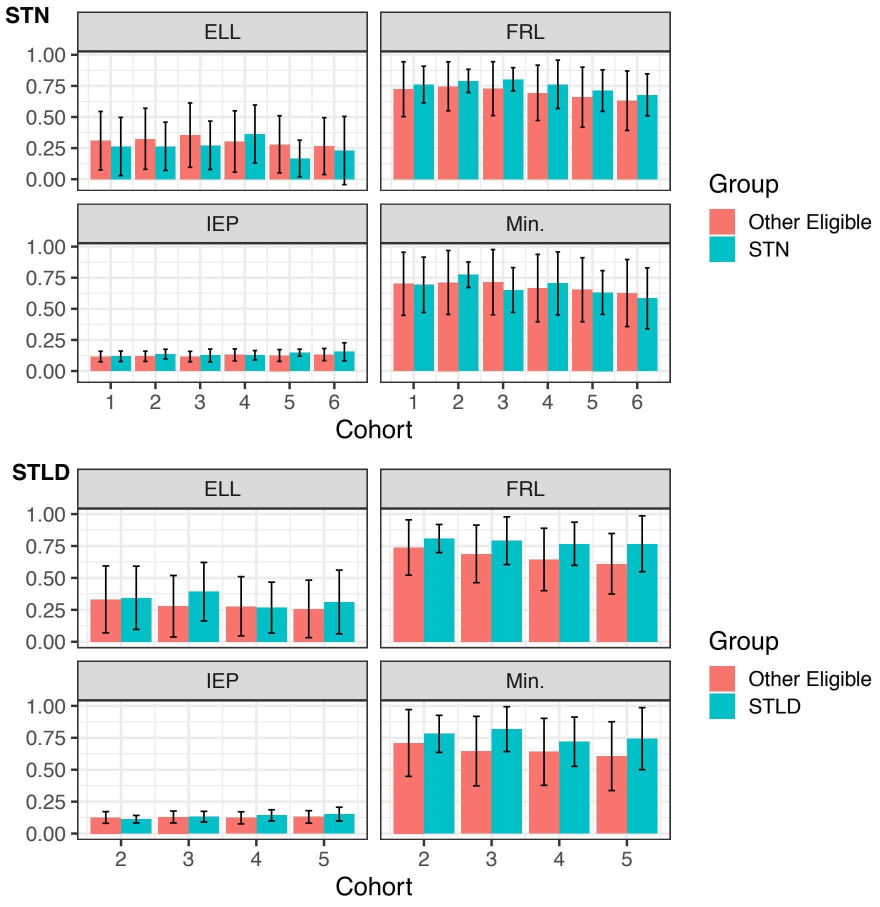
Note: MGP=mean growth percentile; sample size for ELA All Schools was N=1603.

Table 3b summarizes the student achievement variables for the same groups of schools. The table reports average test scores and average MGP for participating and eligible schools in the year prior to starting participation or to being eligible for the first time to participate. Schools in the most recent cohorts are not included in the achievement summary statistics, and hence the number of schools is smaller than in Table 3a. Table 3b also summarizes current year average test scores and MGPs for all schools statewide in the 2018-19 academic year as a reference point. These are helpful for interpreting the achievement variables for participating and eligible schools. In any given year, the average test score is close to 0 across schools (by construction due to standardizing scores within subject, grade, and year), while the standard deviation of average scores across schools is approximately 0.40 in both math and ELA, although the standard deviation of scores at the student-level would be approximately 1.0 due to the standardization. The standard deviation of MGPs across schools is about 7 SGP points in ELA and 8 SGP points in math.

In both math and ELA, schools that were eligible to participate had much lower average test scores and MGPs than the statewide distribution. This is expected, because eligibility to participate is based primarily on SPF ratings, which in turn are based primarily on average test scores and SGP. Average achievement among eligible schools was about 0.4 student-level standard deviations below the statewide average, which is equivalent to about 1 school-level standard deviation below the statewide average; MGPs were approximately 0.5 school-level standard deviations below the statewide average of 50. The participating schools also tended to have lower average test scores and MGPs than all eligible schools overall. However, although these are lower on average, there was substantial variation and overlap in the distribution of achievement for participating and eligible schools, and the exact differences varied across cohorts.

To present some of the differences in Table 3a visually, Figure 1 shows the average demographic characteristics for eligible and participating schools in each cohort of STN (upper plots) and STLD (lower plots). The colored bars represent the average value across schools, and the error bars show plus or minus one standard deviation. The differences described in Table 3a can be seen in Figure 1, but the figure also shows the substantial overlap in the distributions, as indicated by the overlapping error bars. Figure 1 also shows that the relative differences between eligible and participating schools varies somewhat across cohorts. In part this is due to the relatively small sample sizes of schools participating in each cohort (particularly for STN) so that one or two schools can affect the average values. But there is also evidence of some trends in the differences – for example, the percent of FRL-eligible students among eligible and participating schools, trends downward slightly in the more recent cohorts. This could be due in part to changing eligibility rules over time, which have increased the number of schools that are eligible to participate each year. These differences are another reason that motivate our analysis of outcomes for each cohort separately in the subsequent sections.

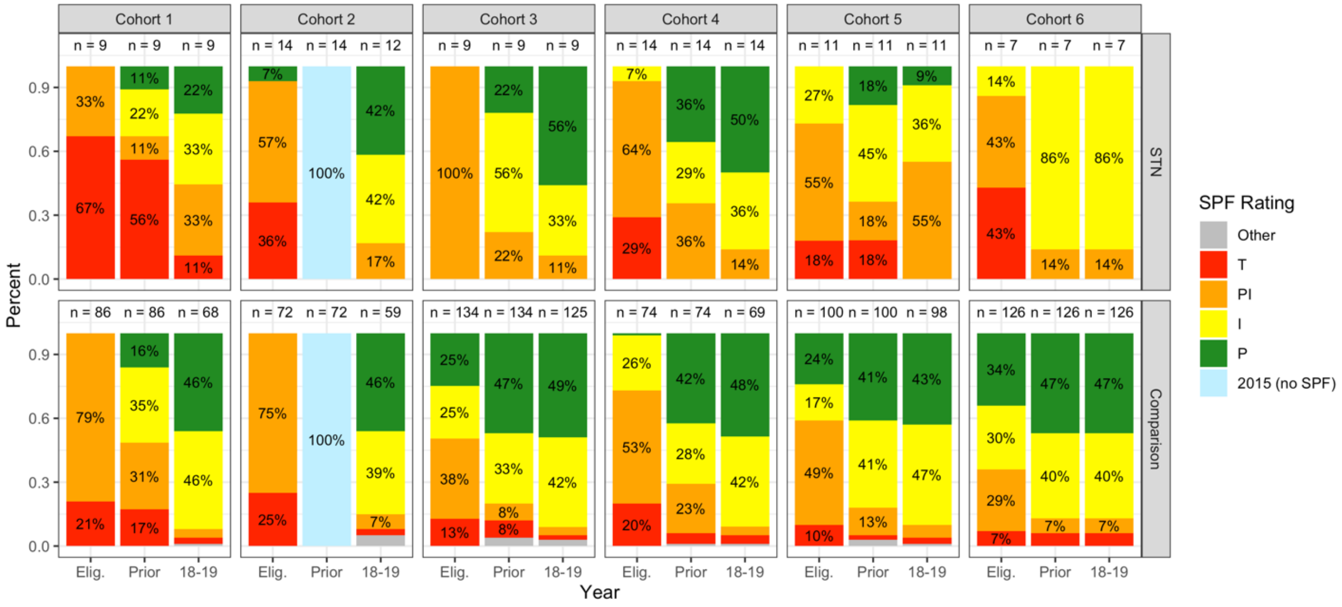
Figure 1. Demographic Characteristics for Eligible and Participating Schools, by Cohort.



SPF Ratings

This section summarizes the SPF ratings for participating schools and a subset of other eligible schools. The SPF ratings are one of the primary metrics the state uses to evaluate schools. In addition, the SPF ratings are the primary metric used to determine which schools are eligible to participate in the STN and STLD programs. Each year from 2010-2019, approximately 10% of schools were given a Turnaround (T) or Priority Improvement (PI) rating, with about 2-3% being given a T rating. Schools with a T/PI rating are considered the lowest performing schools in the state and are the schools the STN and STLD programs are primarily intended to support. As part of the theory of action for these programs, an assumption is that if the STN and STLD programs are effective at supporting schools to improve achievement outcomes for students, then this should be reflected by schools earning higher SPF ratings after participating in the programs. To evaluate whether there is evidence of this happening, we describe trends in the proportion of schools being given the lowest T/PI ratings across years.

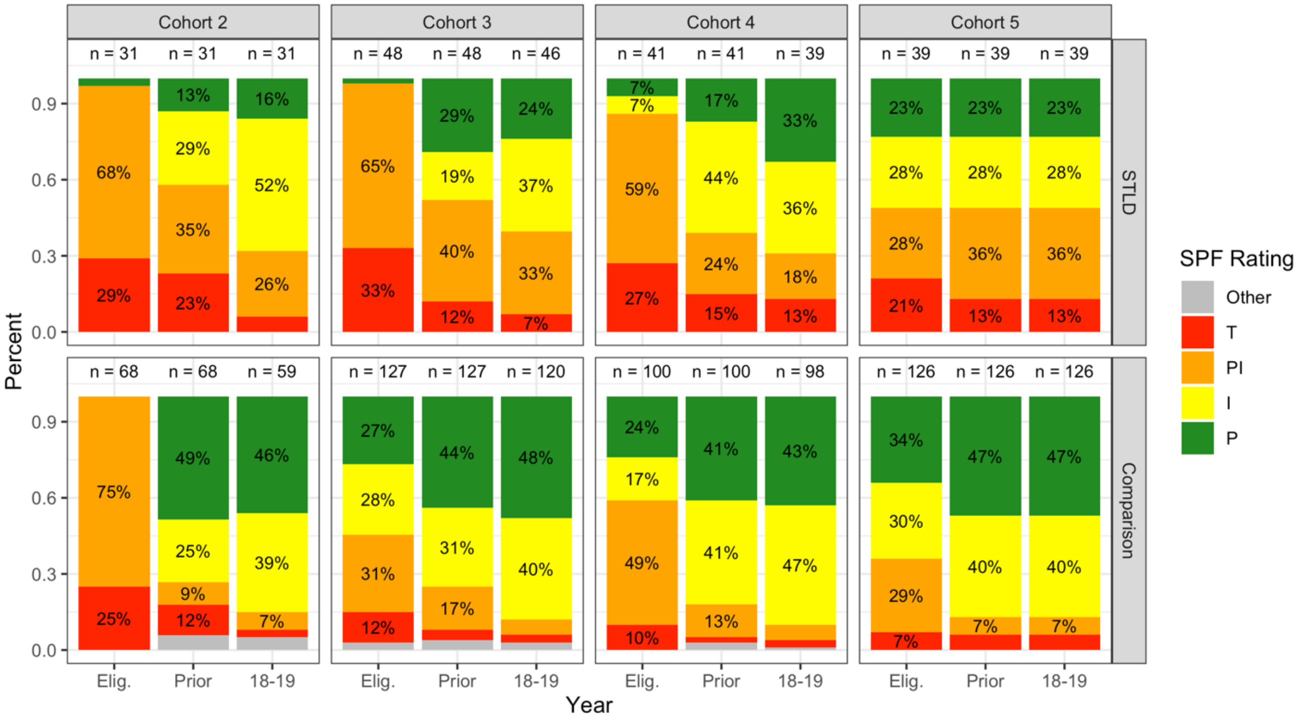
Figure 2a. SPF Ratings by Cohort, Year, and Eligibility for STN Schools.



Figures 2a and 2b show SPF ratings over time for each of the STN (Figure 2a) and STLD (Figure 2b) cohorts, as well as for a set of comparison schools. The figures include all schools participating in each cohort; note that some schools in the STN cohorts also participated in the STLD program, and some schools in each STLD cohort also participated in an STN cohort or in multiple STLD cohorts. The comparison schools are schools that were eligible for each cohort, but that never received funding for STN or STLD, nor did they receive any other form of EASI funding from 2017-2020. The figures show the distribution of SPF ratings for these schools in three years of interest: 1) the year used to determine eligibility for a particular cohort ("Elig."), 2) the year just prior to starting the STN or STLD program, which was usually the year after the eligibility year ("Prior"), and 3) the most recent year for which SPF ratings were reported, 2018-19. Each of these years helps to better understand the performance of these schools in different ways. As an example, for STN Cohort 1, the "Elig." year SPF ratings were based on test score data in 2012-13 and the "Prior" year SPF ratings were based on test score data in the 2013-14 year, because participation began in the 2014-15 academic year. Note that no SPF data are available for the “Prior” year for STN Cohort 2 due to no SPF ratings being reported in 2014-15, and that in the most recent STN and STLD cohort, the “Prior” year and 2018-19 are the same year.

For the earliest cohorts (STN cohorts 1-4 and STLD cohorts 2-3) starting before the 2018-19 AY, eligibility was based primarily on SPF ratings. For these cohorts, between 93% and 100% of participating schools received a T/PI rating at eligibility, as expected. In the more recent cohorts, however, lower proportions of eligible schools received a T/PI rating in the year used to determine eligibility (anywhere from 49% to 86%). This is true for both participating and non-participating schools, although the difference is larger for non-participating schools. In the most recent STN and STLD cohorts, for example, less than 50% of eligible, non-participating schools were identified as T/PI in the year used to determine eligibility.

Figure 2b. SPF Ratings by Cohort, Year, and Eligibility for STLD Schools.



By 2018-19 (the most recent year for which schools received an SPF rating), the majority of schools participating in the STN and STLD programs moved up from T/PI ratings to P/I ratings. On average across STN cohorts, 91% of schools were T/PI at eligibility while only 26% were T/PI in 2018-19; across STLD cohorts, 83% were T/PI at eligibility while only 38% were T/PI in 2018-19 (see Table B1 in Appendix for exact percentages). At the surface level, this suggests a positive outcome indicating that performance on metrics included in the SPF calculations improved over time. But there are two caveats to this trend. First, the 2018-19 SPF ratings are much higher both for schools that participated in the STN or STLD programs, and also for the comparison schools that were eligible to participate but chose not to (among the comparison schools, approximately 10% were T/PI in 2018-19). Second, the “Prior” column within each panel shows that SPF ratings for participating and non-participating schools had already improved substantially relative to the eligibility year, before these schools officially began program participation. Among STN cohorts only 35% of schools on average still had a T/PI rating in the prior year, while only 50% of STLD schools maintained these lower ratings. Any changes in SPF ratings from the Eligibility year to the Prior year cannot be directly attributed to participation in the STN or STLD programs. This suggests that these schools were likely engaged in various other activities in addition to (or instead of) the STN and STLD programs that may have supported the improvement in SPF ratings over this time period.

In summary, although schools participating in the STN and STLD programs showed substantial improvement in SPF ratings by the 2018-19 AY, we raise two caveats to attributing this improvement directly to participation in the STN or STLD programs. First, academic performance improvements appear to have started before participation began. Second, other eligible schools that did not participate in either program also saw substantial improvement in SPF ratings during the same time period. Both of these caveats suggest that other factors in addition to the STN and STLD programs were likely affecting schools’ ratings, although it is reasonable to assume that receiving the funds also contributed to schools receiving higher SPF ratings in subsequent years. Finally, the differences between participating and comparison schools in SPF ratings in the eligibility year also suggest that direct comparisons to the other eligible schools is not necessarily a valid indicator of what we would have expected in participating schools had they not participated. Appendix B contains a table presenting the detailed values from Figures 2a and 2b.

Achievement and Growth Percentile Trends

Although these achievement indicators are the primary data that contribute to each school’s SPF ratings summarized in the prior section, the complicated rules used to combine indicators to produce an SPF rating make it difficult to make direct inferences about how the achievement indicators are changing over time at these schools. To provide a more detailed description of trends in student achievement outcomes, this section summarizes the average test scores and MGP values across schools directly by each program area beginning with the STN program. We computed the average test score or MGP for each school by calculating the average across all students in each school with valid scores in a given year, averaged across grades. As noted above, because we standardized the test scores within grade, subject, and year, the average standardized test score for each school in a given year indicates how high scores were for students in that school, relative to students taking tests in the same subject and grades across the state.

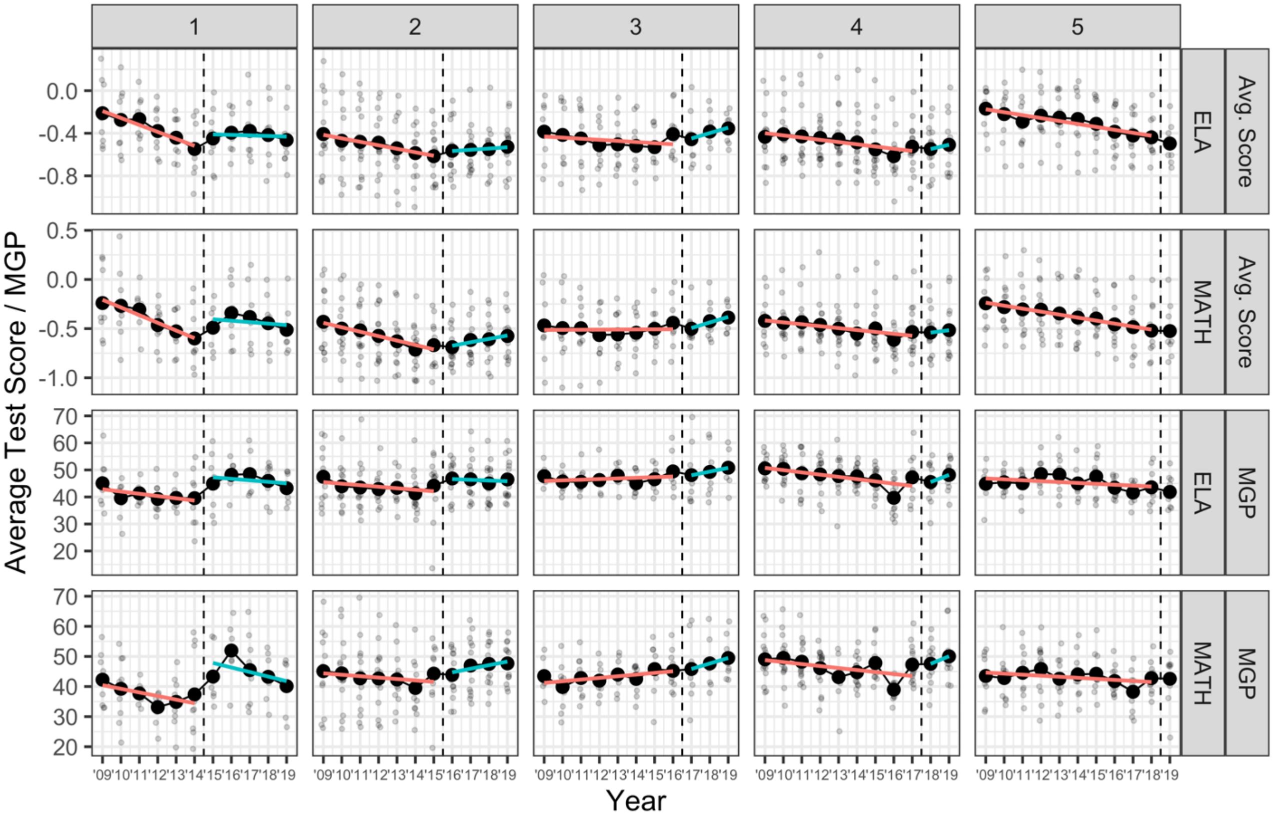
We summarize these trends separately for each cohort for three reasons. First, because the eligibility rules for each cohort varied, there could be different patterns in achievement and growth metrics. Second, because the timing of participation varied, different years constitute the “pre” and “post” participation years, and there are differing numbers of pre- and post-participation data. Finally, the achievement tests administered from 2009-2019 changed substantially beginning in 2015, and these changes correspond to different pre- and post-participation years for different cohorts.

STN Trends

Figure 3 shows the trends in average (standardized) test scores and MGP for all schools participating in each STN cohort, separately by subject for each cohort. Each column represents a single cohort, while each row represents a different outcome variable (average test scores, “Avg. Score”, or MGPs, “MGP”, in either math or ELA for a single cohort of the STN program). Within each panel the light gray dots represent individual schools, the larger black dots show the average outcome across schools, and the vertical dashed lines separate the pre- and post-participation years. Linear trend lines have been added to help visualize trends in the outcomes during pre- and post-participation years.

The 20 plots represented in Figure 3 show that there is considerable variability in outcomes both across schools within each cohort, and in the trends across cohorts. We start by summarizing average test scores, because these are more straightforward to interpret. Test scores at the participating schools were well below the state average and in most cohorts there is evidence that test scores were declining relative to other students in the state prior to participating in the STN. As a reference, on the standardized metric, an average of -0.50 at a school indicates that students scored, on average, 0.5 standard deviations lower than other students taking tests in the same subject, grade, and year. This is a large difference from the state average; as noted in the literature review above, the average effect of turnaround initiatives tended to range from 0 to 0.05 standard deviations. In addition, because the standard deviation in standardized scores across schools is approximately 0.4, these schools tend to have some of the lowest average test scores of any schools across the state. Only about 10% of schools would be expected to have average test scores of –0.5 or lower each year.

Figure 3. Trends in Average Test Scores and MGP for STN Schools, by Cohort and Subject.



Note: there are N=57 unique schools represented in the figure.

Figure 3 suggests there is evidence of trends that are consistent with positive effects on student achievement from participating in the STN, although the changes to the tests beginning in 2015 make some of the trends difficult to interpret directly. In Cohorts 1 and 2, there is evidence that average test scores were steadily declining in the years prior to participation, followed by some increases in the first years after starting participation in the STN. However, the increases did not appear to continue steadily in all years. Moreover, the first two years of STN participation for these schools took place in the 2014-15 and 2015-16 school years, which were also the first two years of the new CMAS tests in grades 3-9, which affected nearly all students in these schools (because they were almost entirely E/M schools). Standardizing scores by grade and year adjusts for some of these changes, but there are multiple possible explanations for why student test scores in these schools might have increased, relative to other students in the state, in the early years of implementation. Test participation rates statewide declined in 2015 and 2016 relative to earlier years, for example, and the tests were intended to measure different content standards.

In Cohort 3, there is a similar pattern although it also appears that average scores increased slightly beginning in 2015 and 2016, which were prior to starting the STN for this cohort of schools. This also suggests that changes to the tests in these years may be at least partly responsible for the change in trends beginning in 2015. It is difficult to describe trends for the most recent cohorts, because there are only 1 or 2 years of data after participation. In Cohort 4, for example, average test scores in the two most recent years do appear to be increasing slightly, although there were also some increases in 2017, the year just before participation in the STN began.

Table 4 provides another way to summarize the trends. Table 4 reports the average year to year change in average test scores and MGPs, separately by program and for pre- and post-participation year observations. Across all observations in Figure 3, the average change in school-level average test scores from year to year was approximately -0.01 (SD=0.12) in both subjects. These changes were systematically different in pre- and post-participation years, however; the average change in all pre-participation years was approximately -0.02 (SD=0.12), while the average change in all post-participation years was approximately between 0.01 and 0.02 (SD=0.14) across subjects. While the absolute magnitude of these changes is small, they are consistent with small positive effects of program participation that reversed negative trends. A caveat, as the figures make apparent, is the considerable variability across schools and cohorts in the exact pattern and magnitude of these changes.

Table 4. Average Year to Year Changes in Average Test Scores and MGPs, by Program.

			All			Pre-Participation			Post-Participation		
Program	Subject	Outcome	N	Mean	SD	N	Mean	SD	N	Mean	SD
STN	ELA	Avg. Score	552	-0.01	0.12	392	-0.02	0.12	160	0.01	0.14
	Math	Avg. Score	552	-0.01	0.13	392	-0.02	0.13	160	0.02	0.14
	ELA	MGP	551	-0.11	7.31	391	-0.34	6.92	160	0.46	8.20
	Math	MGP	552	0.24	8.37	392	-0.03	8.13	160	0.91	8.92
STLD	ELA	Avg. Score	936	0.00	0.14	220	-0.02	0.12	716	0.00	0.15
	Math	Avg. Score	936	0.00	0.15	220	-0.02	0.13	716	0.00	0.16
	ELA	MGP	934	0.05	8.24	219	-0.30	6.69	715	0.16	8.67
	Math	MGP	935	0.09	9.46	220	-0.27	7.68	715	0.20	9.95

Interpreting the trends in the MGPs is more complicated. First, MGPs tend to include more sampling and measurement error from year to year making them more variable from year to year even when there are no true changes in student learning. Second, although the MGPs are useful because they provide more information about patterns of achievement at each school, they are more complex to interpret because they are derived using multiple years of data and a quantile regression model that tracks cohorts of students with similar academic starting points over time. Thus, although schools will tend to have higher MGPs in years where average test scores are higher, it is possible for a school to have high average test scores but low MGPs (if the test scores were high relative to the statewide distribution but low relative to schools where students had similar prior year scores) or vice versa (if test scores were low relative to the statewide distribution but high relative to schools where students had similar prior year scores). This also makes interpreting changes in MGPs across years complicated – it could be that average test scores increased from one year to the next while MGPs do not. The average MGP across schools statewide is 50 each year. Figure 3 shows that across most years, the average MGP of schools in the STN cohorts were below this statewide average, although in many years only by a small amount.

Finally, we note that the 2015 MGPs were not officially reported due to changes in the testing program, and subsequent changes in 2016-2018 have affected MGPs in one or more grades.

In Cohorts 1, 2, and 4 there appears to be a slight negative trend in MGPs prior to participating in the STN, with positive trends during and after participation. However, in Cohort 1 the trend is inconsistent and the large year to year changes, particularly in math, suggest that the test transitions may be related to these patterns. In Cohort 3, MGPs had a slight positive trend prior to joining the STN, and these trends appear to continue. There is a noticeable dip in MGPs in 2016 for Cohort 4 that is difficult to explain, although the 2016 MGPs were used to compute 2016 SPF ratings used to determine eligibility for this cohort, suggesting that the sharp drop in MGPs in 2016 could be related to a combination of changes to the tests as well as the use of preliminary SPF ratings to determine eligibility. Finally, in Cohort 5 there were slight negative trends in MGPs over time, but with only a single year of data after participation began it is difficult to infer whether these trends have changed.

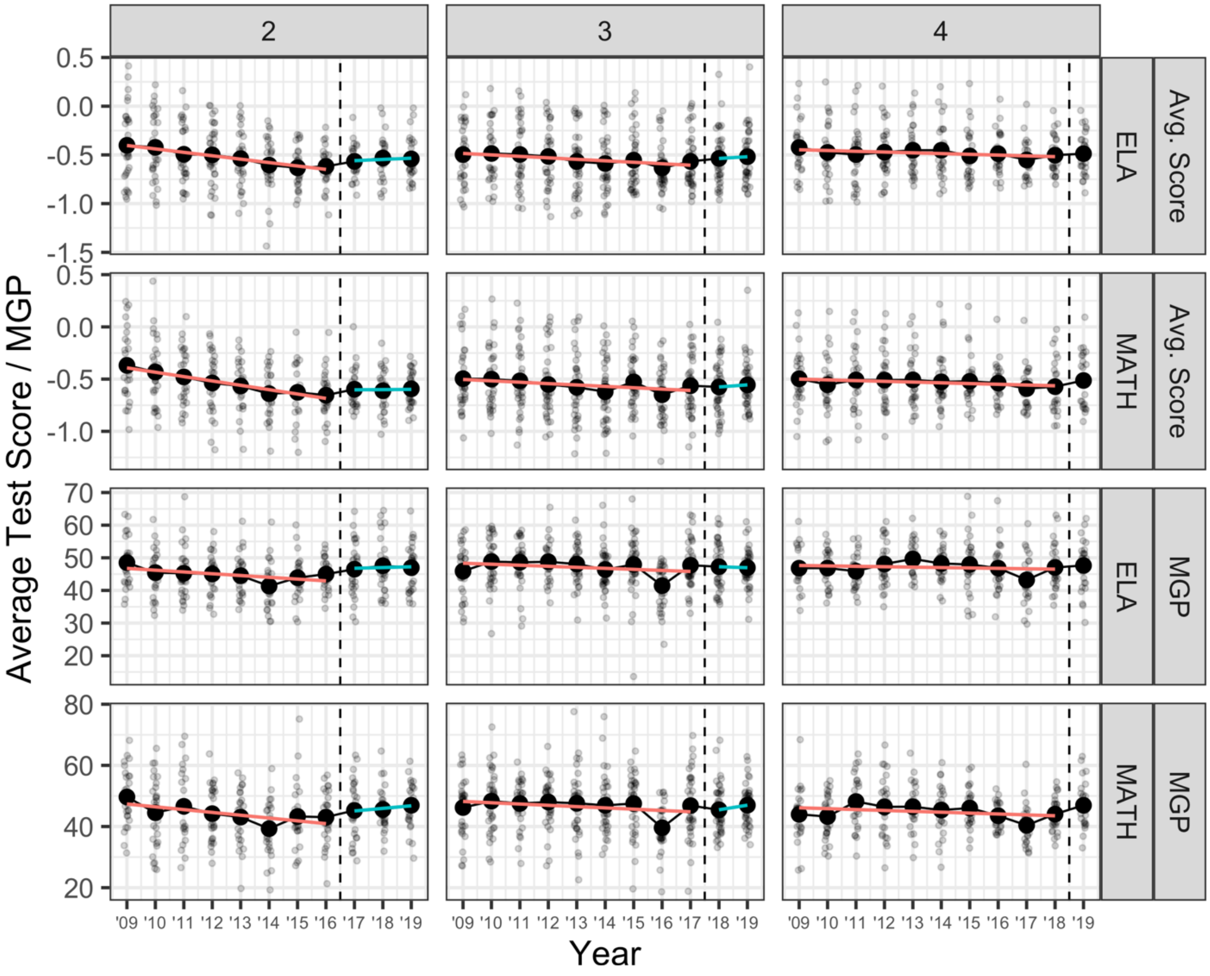
As with the average test scores, we can also summarize year to year changes in MGPs across schools for years before and after participating in the STN. The average year to year changes in MGPs were -0.34 (SD=6.9) and -0.03 (SD=8.1) in ELA and math, respectively for all pre-participation years. These are very small changes relative to the overall distribution of MGPs, although they are both negative. In years after participation began, the average changes were 0.46 (SD=8.2) and 0.91 (SD=8.9) in ELA and math, respectively. If we exclude Cohort 1 and the 2015 MGPs from these calculations, the average pre-participation changes are -0.55 (SD=6.6) and -0.50 (SD=8.0) in ELA and Math, while the average post-participation changes are 0.4 (SD=7.8) and 1.1 (SD=7.7). Again, these are relatively small changes, but they do suggest that negative average changes (i.e., declining) MGPs reversed to positive average changes after participation began.

In sum, we say fairly consistent negative trends in average test scores and MGPs leading up to participation in the STN, suggesting that these were not schools where achievement outcomes were low for a single year of eligibility that led to participation, but rather were schools where achievement was consistently low. While there is some evidence that these negative trends either slowed or reversed after participating, these patterns vary across cohorts and because they coincide with large changes to the tests students were taking, especially in the earliest cohorts, it is difficult to make any strong conclusions about whether these represent true changes in student learning and whether they are due primarily to participation in the STN or to other factors. The patterns in the MGP data are less consistent, and there are some relatively large changes to MGPs both immediately before and after beginning participation that are difficult to explain and should be investigated further, particularly in Cohorts 1 and 4.

STLD Trends

Figure 4 presents the same set of average test score and MGP plots for schools participating in the STLD program. For these plots, we have included each school only in the first cohort they participated in. So, for example, if a school participated in both STLD Cohorts 2 and 3, we only include them in the plots for Cohort 2. We do this so that all of the observations for the pre-participation years (to the left of the vertical dashed line) are truly years before the school participated in the STLD program. Based on Figure 3, the trends in average test scores and MGP are similar for the STLD cohorts, although somewhat less variable from year to year and cohort to cohort.

Figure 4. Trends in Average Test Scores and MGP for STLD Schools, by Cohort and Subject.



Note: in this figure, schools that participated in multiple STLD cohorts are grouped with the first cohort they were a part of. There are N=102 unique schools represented.

In Cohort 2 (plots in the first column), there were negative trends in both average test scores and MGPs leading up to participation, and either leveling off or positive trends during and after participation. There are some exceptions, however, for example MGPs for these schools appear to have begun increasing in 2015, two years before beginning participation. In Cohort 3 (the plots in the middle column), there is a less apparent trend in average test scores and MGPs prior to participation, and visually there appears to be an increase in average test scores in 2015 for these schools. The MGPs seem to remain slightly below the state average of 50 each year, with the exception that there is a noticeable dip in 2016 for Cohort 3, similar to that seen in STN Cohort 4 above (Cohort 3 STLD eligibility was based on a combination of 2014 and 2016 SPF ratings). This would be worth investigating to understand further. The two years of data during and after participation in STLD are consistent with small positive effects, but with only two time points it is difficult to discern a clear trend. Finally, for Cohort 4, there is a noticeable dip in 2017 MGPs (2017 SPF data were used for eligibility), but otherwise there is only a slight downward trend in both average test scores and MGPs prior to participation. Although the single year of post-participation data is consistent with an upward trend, we cannot infer a trend from the single data point available.

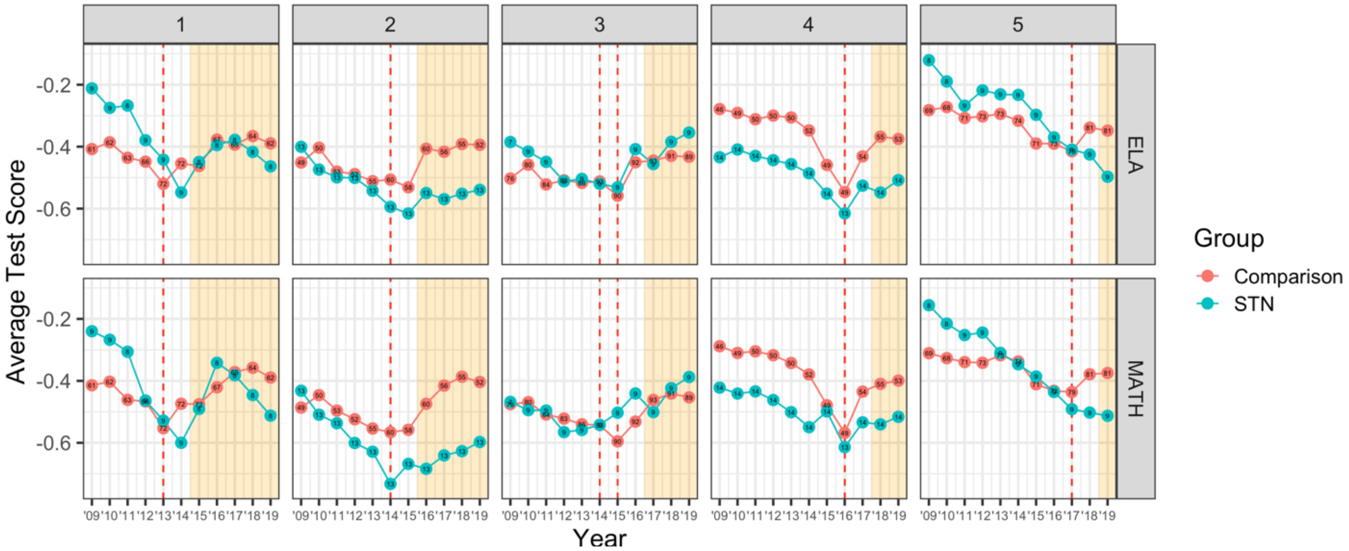
Summarizing average year to year changes in average test scores and MGPs in Table 4 suggests similar patterns observed for STN schools. During pre-participation years, the average changes in test scores each year were approximately -0.02 (SD=0.12) across subjects while the average changes in post-participation years were between 0 and 0.01 (SD=0.15) across subjects. These are consistent with the visual evidence of slight negative average changes prior to participation, and either flat or slightly positive average changes after participation. For MGPs, the average changes in pre-participation years were just under -0.30 in both subjects (SD=6.7 in ELA and 7.7 in math) while the average changes in post-participation years were 0.16 (SD=8.7) and 0.20 (SD=10) in ELA and math. Again, these suggest consistent negative trends in pre-participation years and positive trends in post-participation years, although the absolute magnitudes are very small relative to the distribution of overall changes and MGPs across schools.

Achievement Trends with Comparison Groups

This section compares the trends in achievement from the prior section to a set of similar comparison schools for each STN or STLD cohort. We introduce a comparison group of schools to help contextualize and interpret the trends summarized above. While the trends reported above for schools participating in the STN and STLD programs are consistent with small positive effects of the programs, on average, there is the possibility that some of these changes are due to other factors. These other factors could include the changes to the tests used or to other programs schools identified by the state as “lower-performing” may have been implementing during this time period. Comparing the trends to those observed for similar schools that did not participate in the programs can help determine whether this might be the case. However, we also caution that comparing the trends across the two groups cannot be used to show the direct effect of participating in these two programs. First, as seen above when comparing demographic variables and SPF trends, although schools that were eligible for the programs but did not participate are similar to participating schools (particularly relative to all schools in the state), the two populations do differ in systematic ways. Second, schools in both groups could be participating in other initiatives that affect student achievement.

To construct a comparison group for each STN cohort, we include all other schools that were eligible to participate in the same STN cohort, but which never participated in either the STN or the STLD program, and also never received other non-STN or non-STLD EASI supports in the 2017-18 through 2019-20 academic years. We further limit the STN comparison group to schools enrolling students in E/M grades (E, M, or EM schools), because there are so few schools enrolling students in high school grades participating in the STN program. For the comparisons in this section, we also do not include the three STN participating schools that enrolled high school grades students. To construct a comparison group for each STLD cohort, we use a similar procedure. We begin with all schools eligible to participate in a given cohort and then exclude any schools that received funds to participate in the STN, STLD, or EASI programs at any point. The only restriction we make on school structure is to exclude a small number of eligible schools enrolling students across all EMH levels, because there were no such schools that participated in the STLD program. As with the prior section, we only include schools in a single STLD cohort; for schools that participated in multiple STLD cohorts we include them only in the first cohort they participated in. Based on this process there is a separate comparison group for each cohort of STN or STLD, but there are also some schools that may appear in multiple comparison groups (and schools that may be in both an STN comparison group and an STLD comparison group).

Figure 5. Trends in Average Test Scores for STN Schools and Comparison Schools, by Cohort and Subject.

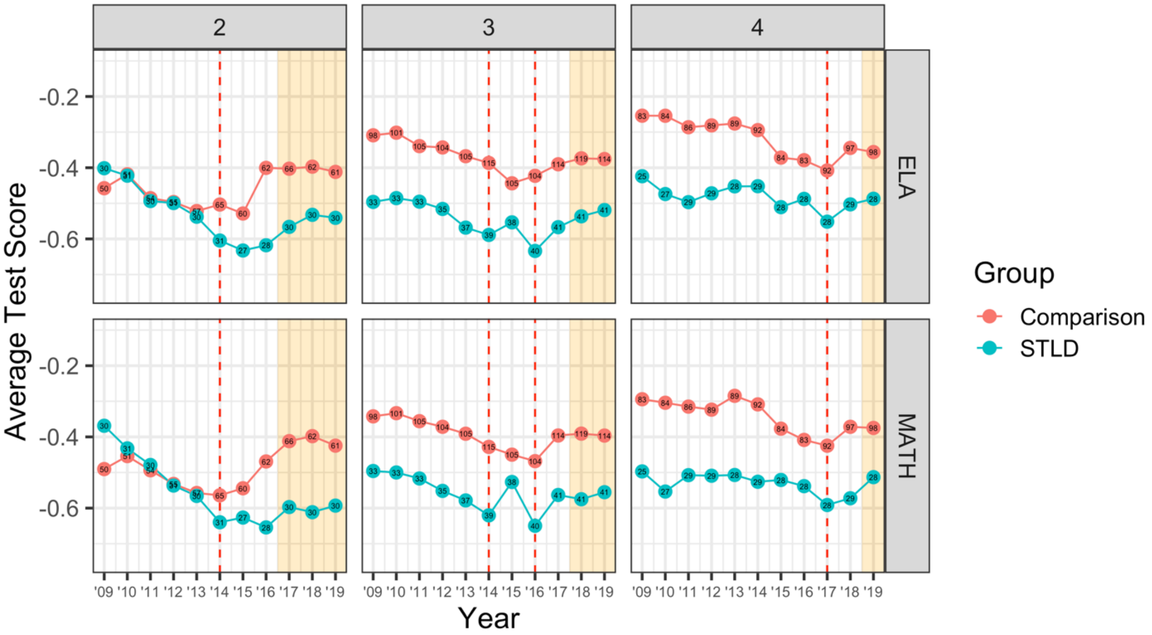


Notes: The numbers within circles indicate the sample size at each time point. The yellow shaded region shows post-participation years. The red dashed vertical lines show the years of SPF data used to determine eligibility.

Figures 5 and 6 show the trends in average test scores across participating schools and the comparison group for each cohort. Each figure includes a single panel for average test scores in either math or ELA for each cohort; the rows represent subjects and the columns represent cohorts. Within each panel, the yellow shaded region represents years during and after participation began for each cohort, while the dashed vertical line indicates the year(s) of test score data used in SPF ratings that determined eligibility. In each panel the orange dots represent average test scores for comparison schools, while the blue dots represent average test scores for the participating schools. The trends for the participating schools are nearly identical to those in the plots above, although there are slight differences in the samples (e.g., the 3 high schools participating in STN are not included). Although it would be ideal from a statistical standpoint to combine data across cohorts, the varying eligibility rules and timing of the programs required us to create separate comparison groups for each cohort and to compare trends separately cohort by cohort. Because the figures in the prior section show considerable variability in the MGP trends and because these are more complex to interpret over time, this section focuses only on comparisons based on average test scores.

Comparing the trends across the STN schools and comparison schools in Figure 5 suggests a few insights. First, average test scores for both groups were generally similar, and substantially lower than the statewide average. Second, the trends in scores in both groups follow similar broad patterns, which suggests that some of the changes in trends summarized above may be due to additional factors beyond STN participation. In multiple cohorts, for example, average test scores in the comparison schools appear to trend upwards at about the same time that some upward trends are seen in the participating schools. Third, however, is evidence of variability in the comparisons across cohorts. If trends in scores were driven primarily by changes made to the tests, we would expect them to be similar patterns across cohorts. Instead, the relative trends across cohorts varies. In Cohort 1 test scores in the comparison schools began to increase in 2014, whereas in the Cohort 2 and 3 comparison group scores appear to begin increasing most noticeably in 2016. In the Cohort 4 comparison group the same drop in average test scores in 2016 appears and is unusual given the other data points; this again suggests that the way scores and SPF ratings were used to identify eligibility for this cohort, possibly in conjunction to changes in the tests may be contributing to these trends. Other variability is also apparent, for example in some cohorts the participating schools appear to have slightly higher average test scores in years prior to participation (Cohorts 1 and 5); in other cohorts, prior achievement is very similar (Cohorts 2 and 3); and in one cohort the comparison schools had higher average scores (Cohort 4). Relative to the distribution of average test scores across schools in the entire state, however, these differences are relatively small.

Figure 6. Trends in Average Test Scores for STLD Schools and Comparison Schools, by Cohort and Subject.



Notes: The numbers within circles indicate the sample size at each time point. The yellow shaded region shows post-participation years. The red dashed vertical lines show the years of SPF data used to determine eligibility.

Figure 6 depicting trends for STLD schools shows less variability in the trends both across cohorts and across the two groups of schools. In Cohort 2, test scores in pre-participation years were similar in participating and comparison schools, and in both groups scores appear to begin increasing prior to 2017, when schools would have first participated in the STLD program, although the increase is more dramatic for the comparison schools. In Cohorts 3 and 4 the participating schools had consistently lower average test scores prior to beginning the STLD program than did the comparison schools, and there is evidence of scores increasing on average leading up to and continuing into the STLD participation years. Note that for Cohort 3, eligibility was based both on 2014 and 2016 SPF ratings and test score data, and there is a noticeable increase in test scores in 2015 for participating schools between these years. This again suggests that the changes to the tests and SPF rating calculations may be contributing to the trends seen during and around 2015 and 2016.

Because it can be difficult to visually discern and summarize comparisons among so many trends, we also fit a series of multiple linear regression models to these data. While multiple regression models are often used to estimate causal effects or to make inferences from samples to populations, we use multiple regression as a descriptive tool to summarize the relative trends across groups shown in Figures 5 and 6. For each cohort and subject separately, we used a regression model to estimate unique linear pre- and post-participation trends for the participating and comparison schools. Details about the models and results from these analyses are located in Appendix C. The results of these analyses were similar to the visual summaries above. There was evidence that even among the comparison schools, achievement tended to increase in the years following the start of the STN and STLD programs, but the increases were slightly larger, on average, for participating schools. Again, we caution that because of evidence that the comparison and participating schools differ in systematic ways this does not definitively indicate a causal effect.

Taken together, there is no single pattern of trends that is consistent across all cohorts. The summary analyses suggest that the data are consistent with small positive effects of participation in the STN and STLD programs, even when compared to similar schools that did not participate in the programs. At the same time, the evidence that achievement tended to increase for the comparison schools that did not participate in the programs make it difficult to determine how much of the increases in achievement should be attributed to participation in the STN or STLD programs, and how much to other factors, including changes to tests or other initiatives the schools were pursuing during these years.

Discussion

One clear pattern that surfaces from our descriptive analyses is that, on average, the schools participating in the STN and STLD programs tend to be concentrated in urban areas that disproportionately serve students of color in lower-income households. It is also clear, based on examining student academic performance prior to joining these programs, that the STN and STLD programs tend to attract schools that have sustained consistently low performance over time. Through these two support programs, CDE is helping to fulfill their stated mission to ensure equity and opportunity for every student by meeting a key initiative around expanding access and opportunities for historically underserved students. In providing the financial resources to sustain these programs, the state is helping to fulfill a key commitment to equity, diversity, and inclusion by channeling resources to schools that tend to serve students coming from historically underserved groups and communities. Although our analyses show that these programs are not associated with large positive gains on state standardized assessments, the analyses point to the promise of these programs to slowly begin the process of reversing negative performance trends observed on average in these schools prior to joining. Additionally, since these programs are designed to facilitate school-wide transformational and cultural changes that take time to implement, future studies may also require including a broader range of indicators that may be more sensitive to organizational and cultural shifts taking place at these schools.

A key question we would like to be able to answer is: how does receiving funding to participate in the STLD or STN programs affect student achievement and other outcomes at a school? More precisely, we would like to know, for schools that received STLD or STN funding, how would achievement (or other outcomes) have been different if the school had not participated in these CDE-sponsored programs? Unfortunately, this is a question that cannot be answered directly, because we cannot know what would have happened in these schools if they had not received funding and customized supports through these programs. Because schools choose whether or not to participate in these programs, there could be other unique characteristics of participating schools that explain any observed changes. At the same time, if the programs are successful at supporting schools in improving student achievement, we would expect to see certain patterns such as evidence of improved student achievement in years after starting participation.

In an effort to provide initial data to help answer this question, and to provide data that could be used to strengthen these programs in the future, this report described trends in student achievement and SPF ratings for schools participating in the STN and STLD cohorts from 2014-15 through 2018-19. This report is the first of two studies that will be carried out. While this report focused on presenting descriptive statistical analyses, the second report will conduct detailed case studies for a small number of schools participating in the STN program. The primary aim of this report was to summarize broad descriptive trends in achievement outcomes at participating schools that can provide context for the case studies to be carried out beginning in the Fall of 2021 and as a starting point for subsequent research on the STN and STLD programs.

Although the analyses cannot support strong causal claims, the trends observed in student achievement outcomes were consistent with participation in the STN and STLD programs having small, positive effects on student achievement in math and ELA. The trends were consistent with the magnitude of effects we would expect based on the prior studies reviewed above and summarized in a recent meta-analysis of turnaround programs (Schueler et al., 2020). Although the systematic changes in achievement were small, many factors affect student test scores, making it unusual for any single program or intervention to have particularly large effects on student test scores. In addition, many of the changes implemented as part of the STN and STLD programs may not be expected to improve student achievement immediately. As a result, the small positive trends observed in the data are still consistent with positive, educationally meaningful improvements in student achievement for schools receiving supports to participate in the STN and STLD programs. In addition, by 2018-19, the most recent year for which SPF ratings were reported, the majority of schools that participated in either the STN or STLD earned Performance or Improvement SPF ratings, the two highest ratings. Taken together, we believe the results present promising, though inconclusive, evidence about the efficacy of these supports.

We also compared the achievement trends and student demographic profiles of participating schools to schools that were eligible to receive these supports but opted not to apply for them. Although these comparisons did not change our primary interpretations, they identified two systematic patterns that could potentially be studied in future research. First, although both the participating and non-participating schools enroll substantially higher proportions of students from historically disadvantaged groups than the state overall, it appears that participating schools enroll consistently higher proportions of these students, including higher proportions of minority (non-white) students and higher proportions of FRL-eligible students. This finding has mixed implications. On the one hand, because participating schools are those identified by the state as the lowest performing, it is consistent with prior research suggesting that students of color and students living in poverty tend to have access to fewer educational opportunities. On the other hand, it means the resources provided through the STN and STLD programs are being provided to schools and students who stand to benefit most from them, consistent with CDE goals. The systematic difference between schools that do and to not choose to participate in these programs suggests that future research could investigate how and why schools decide to participate in the STN or STLD programs. Understanding the decision-making process and goals of participating schools could help CDE to recruit additional schools when funds are available and to better tailor supports to match the goals of participating schools.

The second trend that emerged when looking at outcomes for the comparison schools pertains to SPF ratings. As noted above, although SPF ratings improved substantially for participating schools by 2018-19, the SPF ratings of these schools also improved noticeably prior to the beginning of participation in the STN and STLD programs. In addition, the same trend was observed among the comparison schools that did not participate in the focal programs. It is not clear whether this is common or specific to the cohorts of schools included in the current analyses. Further research about year-to-year trends in school SPF ratings could be useful for understanding the SPF rating system and understanding additional interventions or programs that schools might be undertaking.

Limitations and Future Directions

There are some important caveats and limitations that should be taken into consideration when interpreting these results or planning and carrying out future evaluations, specifically of the STN and STLD programs or in similar contexts.

One challenge specific to this context are the numerous changes that have been made to the standardized tests used to measure student achievement in Colorado since 2015. These changes occurred during the same time span that the majority of schools included in this study were participating in the STN and STLD programs. The changes made to the tests were unrelated to the two programs and were part of broader revisions made to state content standards. The largest changes were made in the 2014-15 AY, but there have been subsequent changes to the tests administered at nearly every grade level, as well as variability in test participation rates in some years. Table A3 in the Appendix summarizes changes made to the tests from 2014 – 2019. Although the testing changes likely improved the quality of the assessments and alignment to new content standards, they also pose methodological challenges when attempting to study change in student achievement over time. While we used standardization to adjust for the changes to the test score scales, this cannot fully account for differences that may have been observed if consistent tests had been used over time. Most recently, disruptions to schooling and state accountability testing due to the COVID-19 pandemic will likely limit the use of state accountability tests for describing trends in achievement moving forward in the next few years. Future evaluations of the STN and STLD programs in regard to student achievement will need to address these challenges.

A more general challenge, noted above, is that schools self-select to participate and may differ systematically from non-participating schools. As a result, it is difficult to draw conclusions about the direct effect of participation on student achievement outcomes or on other outcomes or practices. This is a common challenge faced by studies attempting to evaluate the effects of school turnaround initiatives and many other interventions in education carried out in observational contexts. It may be possible to use additional quasi-experimental statistical methods such as a matching approach to make stronger causal inferences, but at least two challenges described above would need to be addressed. First, the eligibility rules for participation changed a number of times, making it complicated to identify a consistent group of comparison schools across cohorts. While it is often preferable to combine data across cohorts to make statistical results more reliable, the variability we found across cohorts suggests this should be done with caution. Second, as noted in prior research on school turnaround initiatives, the specific activities or practices that schools implement as part of the programs can vary considerably. In addition, some schools participated in both the STN and STLD programs or multiple STLD cohorts and there was relatively little data about other initiatives or supports schools might have been receiving. The variability and overlap are complicating factors that need to be taken into account in future studies and when interpreting the results in this report.

Finally, although large-scale administrative data allow us to document important trends over time and help to situate the STN and STLD school supports in a broader context, there are questions these data are not well-suited to answer. Data on student learning outcomes in these datasets are often based on state standardized tests, as was the case in this analysis. Because state standardized tests are intended to assess students' learning relative to a broad range of content, they may not be ideally suited to assess learning related to the specific content or skills that schools are focused on in regards to specific supports. Additionally, the initiatives or interventions undertaken by schools as part of the turnaround programs may focus on non-academic outcomes that are not well represented by standardized test scores. Finally, despite the variability in the practices schools adopt as part of school turnaround efforts and in the context in which these initiatives are undertaken, many of these factors are not represented in large-scale administrative datasets. The case study analyses to be carried out in the next phase of this study will attempt to address some of these concerns by providing more detailed descriptions of practices observed in schools participating in the STN program and the school contexts. Going back to an earlier point about the inclusion of a broader set of indicators in future studies, CDE may want to develop data collection protocols that can gather some of this information, and which could be included in future large-scale descriptive analyses such as this report to better understand the context and effects of school turnaround efforts.

References

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). https://doi.org/10.18637/jss.v067.i01

Baker, R., Hupfeld, K., Teske, P., & Hill, P. (2013). *Turnarounds in Colorado: Partnering for innovative reform in a local control state*. Denver, CO: University of Colorado Denver, Buechner Institute for Governance School of Public Affairs. Retrieved from https://www.cde.state.co.us/sites/default/files/documents/turnaround/download/schoolturnaroundreport.pdf

Betebenner, D. W. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4), 42–51. https://doi.org/10.1111/j.1745-3992.2009.00161.x

Dragoset, L., Thomas, J., Herrmann, M., Deke, J., James-Burdumy, S., Graczewski, Cheryl., Boyle, A., Upton, R., Tanenbaum, C., & Giffin, J. (2017). *School improvement grants: implementation and effectiveness*. (NCEE 2017-4013). Washington DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from: https://files.eric.ed.gov/fulltext/ED572215.pdf

de la Torre, M., Allensworth, E., Jagesic, S., Sebastian, J., Salmonowicz, M., Meyers, C., & Gerdeman, R.D. (2012). *Turning around low-performing schools in Chicago*. Research 36 Report. Chicago, IL: The University of Chicago Consortium on Chicago School Research. Retrieved from https://consortium.uchicago.edu/sites/default/files/2018-10/12CCSRTurnAround-3.pdf

Dickey-Griffith, D. (2013). Preliminary effects of the school improvement grant program on student achievement in Texas. *Georgetown Public Policy Review*, 21-39. Retrieved from http://gppreview.com/wp-content/uploads/2014/02/Dickey-Griffith-D.pdf

Fullan, M. (2001). *The new meaning of educational change*, 3rd Edition. New York: Teachers College Press.

Hallberg, K., Williams, R., Swanlund, A., & Eno, J. (2018). Short comparative interrupted time series using aggregate school-level data in education research. *Educational Researcher*, 47(5), 295–306. https://doi.org/10.3102/0013189X18769302

Hargreaves, A. & Fullan, M. (2012). *Professional capital: Transforming teaching in every school*. New York, NY: Teachers College Press.

Jaeckel, L., Bartlett, K., & Goss, N. (2020). *School transformation grant report*. Denver, CO: Colorado Department of Education, School and District Transformation Unit. Retrieved from https://www.cde.state.co.us/cdedepcom/schooltransformationgrantreport

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kistner, A.M., Melchior, K., Marken, A.A., & Stein, L.B. (2017). *Lessons learned in Massachusetts high school turnaround: A resource for high school leaders*. Washington, DC: American Institutes for Research. Retrieved from https://www.air.org/sites/default/files/downloads/report/Massachusetts-High-School-Implementation-Report-Oct-2017_0.pdf

LiCalsi, C., Citkowicz, M., Friedman, L. B., & Brown, M. (2015). *Evaluation of Massachusetts office of district and school turnaround assistance to commissioner's districts and schools: Impact of school redesign grants*. Washington, DC: American Institutes for Research. Retrieved from https://www.air.org/sites/default/files/downloads/report/15-2687_SRG_Impact-Report_ed_FINAL.pdf

LiCalsi, C., & García Píriz, D. (2016). *Evaluation of level 4 school turnaround efforts in Massachusetts. Part 2: Impact of school redesign grants*. Washington, DC: American Institutes for Research. Retrieved from https://www.doe.mass.edu/turnaround/howitworks/impact-study.pdf

R Core Team. (2020). R: *A language and environment for statistical computing. R Foundation for Statistical Computing*. http://www.R-project.org/

Schleicher, A. (2018). Making education reform happen. In *World class: How to build a 21st-century school system*. Paris, France: OECD Publishing.

Schueler, B.E., Asher, C.A., Larned, K.E., Mehrotra, S., and Pollard, C. (2020). *Improving low-performing schools: A meta-analysis of impact evaluation studies*. (EdWorkingPaper: 20-274). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/qxjk-yq91

Sun, M., Penner, E. K., & Loeb, S. (2017). Resource- and approach-driven multidimensional change: Three-Year effects of school improvement grants. *American Educational Research Journal*, 54(4), 607–643. https://doi.org/10.3102/0002831217695790

Zimmer, R., Henry, G. T., & Kho, A. (2017). The effects of school turnaround in Tennessee's Achievement School District and Innovation Zones. *Educational Evaluation and Policy Analysis*, 39(4), 670-696. https://doi.org/10.3102/0162373717705729

Appendices

Appendix A: Achievement and Other Data Preparation

Sample restrictions and adjustments for characterizing population of schools in all years:

- Begin with October count file list.
- Exclude any schools that did not enroll any students (N=0).
- Exclude any schools that had an AEC designation in any year from 2009-10 through 2018-19.
- There were a very small number of schools that received STLD or STN funding, but were not considered "eligible" based on the rules in Table 1. We coded these schools to be "eligible" to receive funding when tabulating counts of schools.
- When determining "rural" designation, we used the 2019 designation when available. For schools not included in this list we imputed "non rural."

Sample restrictions for test score data samples. We began with the sample of school-by-year observations based on the rules above and then further restricted as follows.

- We use state accountability test score data from spring 2009 to spring 2019. Although neither of the STLD/STN programs were active prior to 2014-15, the earlier years of data allow us to compare pre-treatment trends.
- We limit only to valid test scores and SGPs used in accountability ratings as coded in the provided data files from CDE.
- We standardized student test scores by grade, subject, and year. We do not include student test scores if fewer than 10 students took a particular test in a given grade (e.g., if only three 7th graders took Geometry tests in a given year). We include 11th grade ACT data in 2009-2016, using a Math score and a composite Reading/ELA score that we use as the ELA test score. Table A3 reports the different tests administered in different grades each year from 2009-2019 and highlights the substantial number of test changes that occurred from 2015-2019 at different grade levels.
- We remove any school-by-year observation that does not have at least 5 valid test scores in both Math and ELA so that the sample of schools in both subjects remains constant. We also remove any MGP observations based on fewer than 5 valid test scores in a given subject; the sample of schools in the MGP analyses can thus differ slightly from the sample of schools with average test scores and can differ across subjects. This occurs even without our sample restriction, because in some cases a student can have a valid current year test score but not have a valid SGP.
- We remove any school-by-year observations where there were valid test scores for fewer than 75% of eligible students who could have participated in testing. We do this because if achievement tests are administered to a smaller proportion of students, we worry that they may not accurately reflect the achievement level of students overall at the school.

Table A1 summarizes the achievement variables and associated sample size per school across all school by year by subject observations in the final dataset. This dataset includes all schools across all years, not only the participating STN and STLD schools. Table A2 summarizes achievement variables (combined across subjects) and demographic variables for all school by year observations for schools that ever participated in the STN or STLD programs.

Table A1. Summary Statistics for Achievement Variables by Subject Across all Schools and Years.

	ELA					Math				
Variable	Avg.	SD	Min.	Max.	N	Avg.	SD	Min.	Max.	N
MGP	50.44	7.49	4.33	90.67	16994	50.28	8.82	10.44	98.90	16999
MGP N	220.63	216.30	5	1835.00	16994	225.90	225.97	5	1909	16999
Avg. Score	-0.03	0.41	-1.47	1.65	17141	-0.05	0.43	-1.53	1.67	17141
Avg. Score N	300.10	279.78	5	2669.00	17141	300.64	280.28	5	2672	17141

Table A2. Summary Statistics for Achievement and Demographic Variables for STN and STLD Schools.

Program	Variable	N Schools	N Obs.	Mean	SD	Min.	Max.
STN	MGP	57	1225	44.86	7.86	13.62	69.56
	Avg. Score	57	1226	-0.47	0.27	-1.10	0.44
	% ELL	57	1226	0.28	0.21	0.00	0.80
	% Female	57	1226	0.48	0.03	0.40	0.56
	% FRL	57	1226	0.76	0.16	0.16	0.98
	% IEP	57	1226	0.13	0.04	0.04	0.30
	% Minority	57	1226	0.68	0.20	0.03	0.98
	Avg. Score N	57	1226	273.36	182.04	56.00	896.00
STLD	MGP	102	2091	46.01	7.92	13.62	77.59
	Avg. Score	102	2094	-0.54	0.27	-1.44	0.44
	% ELL	102	2094	0.35	0.23	0.00	0.84
	% Female	102	2094	0.48	0.03	0.16	0.68
	% FRL	102	2094	0.77	0.17	0.00	0.99
	% IEP	102	2094	0.12	0.04	0.02	0.37
	% Minority	102	2094	0.75	0.19	0.07	0.99
	Avg. Score N	102	2094	322.81	263.86	14.00	1619.00

Achievement Test Changes Over Time

Table A3 summarizes the tests administered at different EMH and grade levels from 2014 through 2019 and how these have changed over time.

Table A3. Summary of tests administered in each grade and subject from 2014-2019.

Year	Elementary	Middle	High School
2014	G3: TCAP Math and ELA (with Spanish version)	G6: TCAP Math and ELA	G9: TCAP Math and ELA
	G4: TCAP Math and ELA (with Spanish version)	G7: TCAP Math and ELA	G10: TCAP Math and ELA
	G5: TCAP Math and ELA	G8: TCAP Math and ELA	G11: ACT Composite
2015	G3: CMAS (PARCC) Math and ELA (CSLA field test year)	G6: CMAS (PARCC) Math* and ELA	G9: CMAS (PARCC) Math* and ELA
	G4: CMAS (PARCC) Math and ELA (CSLA field test year)	G7: CMAS (PARCC) Math* and ELA	G10: CMAS (PARCC) Math* and ELA
	G5: CMAS (PARCC) Math and ELA	G8: CMAS (PARCC) Math* and ELA	G11: ACT Composite
2016	G3: CMAS (PARCC) Math and ELA**	G6: CMAS (PARCC) Math* and ELA	G9: CMAS (PARCC) Math* and ELA
	G4: CMAS (PARCC) Math and ELA**	G7: CMAS (PARCC) Math* and ELA	G10: PSAT Math and EBRW
	G5: CMAS (PARCC) Math and ELA	G8: CMAS (PARCC) Math* and ELA	G11: ACT Composite
2017	G3: CMAS (PARCC) Math and ELA**	G6: CMAS (PARCC) Math* and ELA	G9: CMAS (PARCC) Math* and ELA
	G4: CMAS (PARCC) Math and ELA**	G7: CMAS (PARCC) Math* and ELA	G10: PSAT Math and EBRW
	G5: CMAS (PARCC) Math and ELA	G8: CMAS (PARCC) Math* and ELA	G11: SAT Math and EBRW
2018	G3: CMAS (PARCC) Math and ELA**	G6: CMAS (PARCC) Math* and ELA	G9: PSAT Math and EBRW
	G4: CMAS (PARCC) Math and ELA**	G7: CMAS (PARCC) Math* and ELA	G10: PSAT Math and EBRW
	G5: CMAS (PARCC) Math and ELA	G8: CMAS (PARCC) Math* and ELA	G11: SAT Math and EBRW
2019	G3: CMAS (PARCC) Math and ELA**	G6: CMAS (PARCC) Math and ELA	G9: PSAT Math and EBRW
	G4: CMAS (PARCC) Math and ELA**	G7: CMAS (PARCC) Math and ELA	G10: PSAT Math and EBRW
	G5: CMAS (PARCC) Math and ELA	G8: CMAS (PARCC) Math and ELA	G11: SAT Math and EBRW

** indicates that students could take specific end-of-course assessment aligned with class enrollment (Algebra I, Geometry, Algebra II, Integrated Math 1, Integrated Math 2, or Integrated Math 3).*

*** indicates students could take the CSLA instead of the standard ELA test.*

Appendix B

Table B1 summarizes the SPF ratings for participating and comparison schools, by cohort, as summarized in Figures 2a and 2b in the text. The table also indicates the years that were used for eligibility and prior year SPF rating reports for each cohort.

Table B1. Distribution of SPF Ratings for Participating and Other Eligible Schools at Eligibility Year, Prior Year and in 2018-19, by Program.

			STN					**Other Eligible**				
	Years Used				**Elig.**	**Prior**	**2018-19**			**Elig.**	**Prior**	**18-19**
Cohort	Elig.	Prior	N1	N2	% PI/T	% PI/T	% PI/T	N1	N2	% PI/T	% PI/T	% PI/T
1	2013	2014	9	9	100%	67%	44%	86	68	100%	48%	7%
2	2014	2015	14	12	93%	--	17%	72	59	100%	--	10%
3	2014+	2016	9	9	100%	22%	11%	134	125	51%	16%	6%
4	2016	2017	14	14	93%	36%	14%	74	69	73%	28%	8%
5	2017	2018	11	11	73%	36%	55%	100	98	59%	15%	9%
6	2018	2019	7	7	86%	14%	14%	126	126	36%	13%	13%
Average:					**91%**	**35%**	**26%**			**70%**	**24%**	**9%**
			STLD					**Other Eligible**				
	Years Used				**Elig.**	**Prior**	**2018-19**			**Elig.**	**Prior**	**18-19**
Cohort	Elig.	Prior	N1	N2	% PI/T	% PI/T	% PI/T	N1	N2	% PI/T	% PI/T	% PI/T
2	2014	2016	31	31	97%	58%	32%	68	59	100%	21%	10%
3	2014+	2017	48	46	98%	52%	40%	127	120	43%	21%	9%
4	2017	2018	41	39	86%	39%	31%	100	98	59%	15%	9%
5	2018	2019	39	39	49%	49%	49%	126	126	36%	13%	13%
Average:					**83%**	**50%**	**38%**			**60%**	**18%**	**10%**

Notes: N1=sample size at eligibility year; N2=sample size in 2018-19, Other Eligible are schools that were eligible to participate in the indicated STN/STLD cohort, but that never participated in either of these programs nor other EASI programs in 2017-2020. The "Elig." column summarizes SPF ratings in the year ratings were used to determine eligibility; the "Prior" column summarizes SPF ratings in the year just prior to beginning participation (and after eligibility year); the "18-19" column summarizes SPF ratings based on the 2018-19 AY, the most recent year of SPF ratings available. For STLD Cohort 3, 2016 SPF ratings are summarized for the eligibility year, although SPF ratings from both 2014 and 2016 were used to determine eligibility. For STN Cohort 4, the final 2016 SPF ratings are summarized for eligibility year despite using the preliminary 2016 SPF ratings for eligibility determinations.

Appendix C: Regression Analysis Details

This section briefly describes the regression models used to summarize trends in average test scores for the participating versus comparison schools. We fit a separate multiple regression model for each program by cohort by subject combination for cohorts with at least two years of post-participation data. Specifically, the regression model for each cohort and subject is a mixed-effects regression model with the following form:

$$y_{st} = \beta_0 + \beta_1 yearC_t + \beta_2 post_t + \beta_3 part_s + \beta_4 post_t * part_s + \beta_5 part_s * yearC_t + \beta_6 post_t * yearC_t + \beta_7 part_s * post_t * yearC_t + \boldsymbol{\gamma X}_{st} + u_s + e_{st}$$

In this model y_{st} is the average observed test score in school *s* in year *t*, $yearC_t$ is the year centered relative to the first year of participation (so that for example *yearC* = 0 corresponds to 2015 for the first STN cohort), $post_t$ is an indicator equal to 0 in pre-participation years and 1 during/after participation ($post_t$ equals 1 if $yearC_t \geq 0$), $part_s$ is an indicator equal to 1 if a school participated in the STN/STLD program and 0 otherwise, $\boldsymbol{X}_{st}$ is a vector of time-varying demographic covariates, u_s is a random school-level error term assumed to be normally distributed, and e_{st} is a year-specific error term assumed to be normally distributed. These models include demographic variables representing the percent of students in each school who are eligible for FRL, the percent of students identified as minority students, and the percent of students identified as ELL. We include the time-varying demographic variables to adjust for systematic differences in the demographic characteristics of students in the participating and comparison schools. These models are estimated via restricted maximum likelihood via the lme4 package (Bates et al. 2020) in the R Software package (R Core Team, 2020).

This is a common form of regression model used to compare trends in outcome variables in comparative interrupted time series (CITS) analyses when the trends for the treatment and comparison groups may differ prior to treatment, as they appear to in some cohorts above (e.g., Hallberg et al., 2018). We estimate 12 separate regression models of this form – one for each subject for each of the first 4 cohorts of STN and the first 2 of STLD schools. We restrict to these cohorts so that we have at least 2 years of test score data after participation in the programs began and can estimate a linear trend both pre- and post-participation. The model uses all available observations so that we summarize the trends in all available data, although the general conclusions are similar when restricting the sample to a constant sample of schools that have data across all years.

The primary coefficients of most interest are β_4 and β_7. The coefficient β_4 quantifies whether there are differences in average test scores after the first year participating in the STN or STLD program, above and beyond changes observed in this year for demographically similar comparison schools. Similarly, the coefficient β_7 quantifies whether there are changes in the linear trend in test scores in participating schools, above and beyond those observed in the trend for demographically similar schools in the comparison group. We can combine these two coefficients to calculate the total average anticipated difference in scores after K years as:

$$\delta = \beta_4 + K * \beta_7$$

The term δ represents how much higher (or lower) average test scores were in participating schools, relative to what would have been expected if the pre-participation trend for participating schools continued for *K* additional years, with changes in post-participation years equivalent to those observed in demographically similar comparison schools. If we calculate these values using *K* = 2 for STN (i.e., an average difference after 3 years of participation based on the model parameters) the estimated values range from 0.242 student-level SDs (Math, Cohort 1) to -0.052 student-level SDs (Math, Cohort 3). The estimated difference for Cohort 1 is extremely large, and the large deviations in the trend plots suggest there were likely effects due to the test changes that coincided with the start of this cohort that should be investigated. The average estimated 3-year difference across subjects and the 4 cohorts was 0.07. The 1-year average difference was -0.002, and the 2-year average difference was 0.035. For the STLD cohorts, the 1, 2, and 3-year average differences across the two cohorts and subjects were 0.01, 0.04, and 0.07, respectively. The differences for the STLD cohorts tend to be less variable; for example, the smallest 1-year estimated difference was -0.016 (Math, cohort 2) and the largest was 0.031 (ELA, cohort 3), while the 3-year differences ranged from 0.12 (ELA, cohort 2) to 0.04 (Math, cohort 3).[7]

We should be cautious about interpreting these differences as causal effects that can be attributed directly to the STN and STLD programs. If any differences between the comparison and participating schools were due only to participation and the trends were linear, then these differences could be interpreted as causal effects of participating in the programs. Unfortunately, both of these assumptions are likely false. The initial distributions of SPF ratings and the demographic characteristics of students at the participating and comparison schools suggest there are small but systematic differences between the two groups of schools, although the exact nature of these differences varies across cohorts. We also have limited information about what other programs schools in either group might have been participating in – as an example, some schools participated in both the STN and STLD programs, and it is possible that schools in the comparison groups received funding to participate in other programs not recorded in the data provided to us. The figures above also suggest that while linear trends are a reasonable summary of the general patterns over time, there are cases where changes from year to year appear non-linear; the first cohort of STN is an example of this.

Finally, two additional points should be noted about the regression results. First, although we are focused on describing trends rather than statistical inference, we note that the two coefficients of interest, β_4 and β_7, would not be considered “statistically significant” in the majority of these models. Second, for most of the cohorts, the regression models are based on a very small number of years of post-participation data. As a result, the post-participation trends are being estimated with relatively little data, and care should be taken when inferring the trends that might be observed in subsequent years.

Despite these caveats, the observed differences are consistent with the magnitude of effects reported in prior studies reviewed above, suggesting that they are consistent with the small, positive effects we might expect, as well as the evidence that the exact effects are likely to vary across subjects and contexts. Table C1 summarizes demographics for the participating and comparison schools used in the regression models. Tables C2 and C3 summarize the regression model estimates for each cohort and subject.

[7] The pattern of these results was similar when using a less parametric regression model that estimated separate 1, 2, and 3-year post-participation differences rather than estimating a post-participation linear trend difference for participating schools.

Table C1. Sample Sizes and Prior Year Demographics for Participating and Comparison Schools.

STN									
Cohort	**Group**	**N Total**	**N Balanced**	**FRL**	**ELL**	**Minority**	**Rural**	**MNSS**	**MGP**
1	Comparison	73	53	0.73	0.33	0.69	0.28	-0.46	46.8
2	Comparison	60	41	0.74	0.35	0.72	0.29	-0.55	44.4
3	Comparison	96	70	0.78	0.42	0.76	0.24	-0.49	48.8
4	Comparison	57	37	0.68	0.33	0.62	0.26	-0.43	48.0
1	STN	9	8	0.75	0.28	0.70	0.44	-0.57	38.5
2	STN	13	13	0.83	0.28	0.79	0.00	-0.64	44.2
3	STN	9	7	0.79	0.29	0.66	0.00	-0.42	47.3
4	STN	14	13	0.77	0.40	0.70	0.07	-0.53	47.3
STLD									
Cohort	**Group**	**N Total**	**N Balanced**	**FRL**	**ELL**	**Minority**	**Rural**	**MNSS**	**MGP**
2	Comparison	66	44	0.74	0.35	0.70	0.31	-0.44	48.7
3	Comparison	120	77	0.68	0.32	0.65	0.30	-0.39	48.5
2	STLD	31	25	0.83	0.38	0.80	0.21	-0.64	44.0
3	STLD	41	32	0.79	0.44	0.82	0.07	-0.57	47.3

Note: N Total represents sample size used in models. N balanced represents sample of schools that have data for all 11 years. The demographics and achievement variables are for the full sample used in the models, and are calculated based on the year prior to beginning participation.

Table C2. STN Regression Models.

	C1 ELA	C1 Math	C2 ELA	C2 Math	C3 ELA	C3 Math	C4 ELA	C4 Math
Part	-0.066 (0.08)	-0.116 (0.09)	-0.061 (0.07)	-0.118 (0.07)	-0.05 (0.08)	0.044 (0.09)	-0.026 (0.06)	-0.025 (0.06)
Post	0.070** (0.02)	0.074** (0.03)	0.128** (0.03)	0.121** (0.03)	0.065** (0.02)	0.123** (0.02)	0.084** (0.03)	0.064* (0.03)
yearC	-0.009 (0.01)	-0.01 (0.01)	-0.011* (0.00)	-0.010* (0.01)	0.003 (0.00)	-0.008* (0.00)	-0.014** (0.00)	-0.014** (0.00)
Part*Post	0.1 (0.07)	0.198* (0.08)	-0.041 (0.06)	-0.041 (0.06)	-0.02 (0.07)	-0.126 (0.08)	-0.062 (0.06)	-0.026 (0.06)
Part*yearC	-0.047** (0.01)	-0.059** (0.02)	-0.016 (0.01)	-0.028** (0.01)	-0.004 (0.01)	0.018 (0.01)	-0.001 (0.01)	-0.001 (0.01)
Post*yearC	0.023** (0.01)	0.034** (0.01)	0.013 (0.01)	0.034** (0.01)	-0.002 (0.01)	0.013 (0.02)	-0.006 (0.03)	0.023 (0.03)
Part*Post*yearC	0.029 (0.02)	0.022 (0.03)	0.022 (0.02)	0.036 (0.03)	0.052 (0.04)	0.037 (0.05)	0.07 (0.07)	0.025 (0.07)
Constant	-0.487** (0.03)	-0.520** (0.03)	-0.534** (0.03)	-0.582** (0.03)	-0.489** (0.03)	-0.567** (0.03)	-0.464** (0.03)	-0.487** (0.03)
Demographics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Var(res.)	0.023	0.03	0.026	0.031	0.028	0.036	0.024	0.025
Var(int.)	0.032	0.035	0.035	0.031	0.039	0.038	0.021	0.03
Var(y)	0.11	0.101	0.111	0.096	0.101	0.092	0.107	0.1
STN Schools	9	9	13	13	9	9	14	14
Control Schools	73	73	60	60	96	96	57	57
Observations	819	819	743	743	1,054	1,054	710	710

*Note: * p<0.05; ** p<0.01*

Table C3. STLD Regression Models.

	C2 ELA	C2 Math	C3 ELA	C3 Math
Part	-0.115* (0.05)	-0.121* (0.05)	-0.056 (0.04)	-0.057 (0.04)
Post	0.076** (0.03)	0.125** (0.03)	0.024 (0.02)	0.031 (0.02)
yearC	0.001 (0.00)	-0.001 (0.00)	-0.003 (0.00)	-0.0004 (0.00)
Part*Post	0.028 (0.04)	-0.016 (0.05)	0.031 (0.04)	-0.001 (0.04)
Part*yearC	-0.027** (0.01)	-0.033** (0.01)	0.002 (0.00)	-0.001 (0.00)
Post*yearC	-0.014 (0.02)	-0.009 (0.02)	-0.007 (0.02)	-0.005 (0.02)
Part*Post*yearC	0.046 (0.03)	0.043 (0.03)	0.018 (0.04)	0.02 (0.05)
Constant	-0.489** (0.03)	-0.543** (0.03)	-0.422** (0.02)	-0.439** (0.02)
Demographics	Yes	Yes	Yes	Yes
Var(res.)	0.027	0.032	0.027	0.031
Var(int.)	0.035	0.031	0.025	0.028
Var(y)	0.103	0.087	0.107	0.1
STN Schools	31	31	41	41
Control Schools	66	66	120	120
Observations	971	971	1595	1595

*Note: * p<0.05; ** p<0.01*