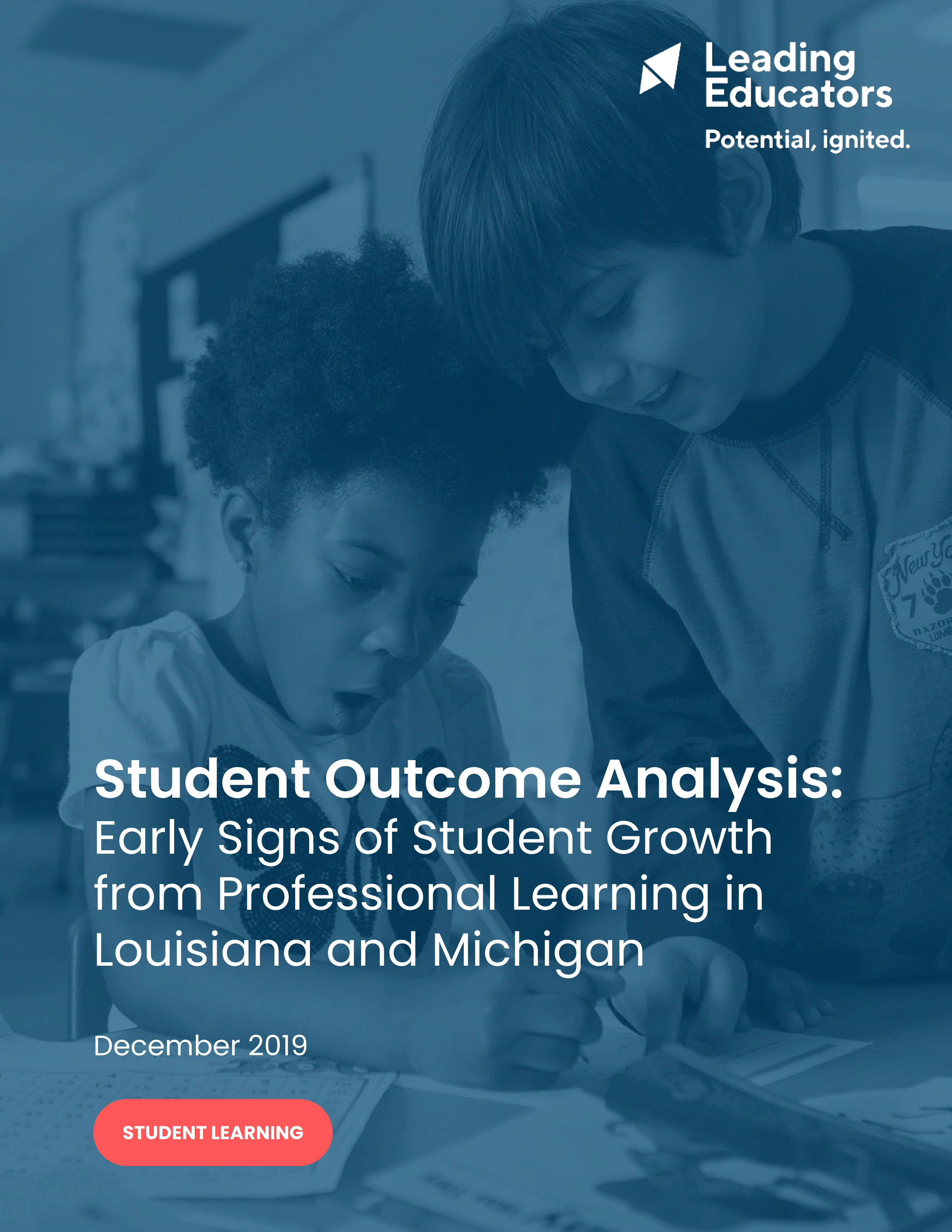


Leading Educators
Potential, ignited.

Student Outcome Analysis:
Early Signs of Student Growth from Professional Learning in Louisiana and Michigan

December 2019

STUDENT LEARNING

Leading Educators

Leading Educators is reinventing professional development for educators, transforming schools into equity-centered environments where teachers have strong opportunities to learn and practice, in collaboration with others, to bring engaging lessons to life. When we invest in educators, and empower them to lead other educators, we ignite exponential impact across entire school systems, ensuring consistently excellent teaching for every student in every classroom, day after day, year after year.

Leading Educators
1824 Oretha Castle Haley Blvd.
New Orleans, LA 70113

leadingeducators.org

Leading Educators
Potential, ignited.

Student Outcome Analysis:
Early Signs of Student Growth from Professional Learning in Louisiana and Michigan

December 2019

Acknowledgements

Many individuals within and beyond Leading Educators were instrumental to this study.

Rebecca Taylor-Perryman led the evaluation effort and our incredible team: Tim Tasker, Ariana Audisio and Dawnell Powell. Ariana Audisio led the writing of this report with Kelley Allenspach-Ruyle and Kami McDonald. Adan Garcia led editing and release, and Jake Ramirez led the report design.

This learning is possible only because of the efforts of our programmatic teams on the ground and the hundreds of partner educators who have committed themselves to lifelong learning in pursuit of the best possible outcomes for every student. We want to recognize the team in Louisiana: Sara Neal, Kesley McLachlan, and Kelley Allenspach-Ruyle; as well as the team in Greater Grand Rapids, Michigan: Mary Kay Murphy, Jennifer Sierra, Tina de la Fe, Kristine Schipper, and Candice Johnson.

We are grateful for the generous support of the Charles and Lynn Schusterman Foundation, the Overdeck Family Foundation, the Bill & Melinda Gates Foundation, the Baptist Community Ministries of New Orleans, the greater New Orleans Foundation, and the Doug & Maria Devos Foundation.

We thank the data team from the Louisiana Department of Education and the data teams from the five districts we served in Michigan for their assistance providing the student data necessary for this analysis. Kata Mihaly and Geoff Grimm from RAND Corporation provided invaluable support with the RATE software and analysis design.

Finally, we are indebted to the staff and families of the school systems we serve in greater New Orleans and Greater Grand Rapids who have trusted us to be their partner in improving outcomes for approximately 9,500 students.

Table of Contents

Executive Summary

More than half of U.S. students are not ready for their next step because instruction is consistently below grade-level, especially for students of color and students from low-income households. Leading Educators aims to erase this gap in students' opportunities by partnering with school systems to build and sustain the conditions, teaching, and leadership that enable a strong culture of content-based professional learning. Equity lives in the choices teachers make every day, so Leading Educators uses professional learning to empower teachers with the deep knowledge of learning standards and pedagogical skills necessary to reach every student, no matter what level they start at. This evaluation attempts to measure Leading Educators' efficacy in reaching these intended outcomes using a rigorous methodology. This evaluation specifically measured the effect of programs in Louisiana and Michigan during the 2017-2018 school year on student standardized scores in mathematics and English language arts (ELA).

During the 2017-2018 school year, Leading Educators supported school systems in Louisiana and Michigan to establish teacher-led professional learning structures. Local teams of teacher leaders who led professional learning for their peers in partner schools received cohort-based professional development to prepare them for effective leadership, as well as monthly coaching. Within schools, the teacher leaders facilitated 90-minute blocks of sequenced professional learning aligned to their curriculum on a weekly or bi-weekly basis. In meetings, content teams analyzed data, set goals, learned individually and collaboratively, applied new learning, and monitored student progress. Meetings were linked to key areas of focus in a cycle so that teachers could build deep capacity in a narrow set of key topics connected to school priorities.

With the support of RAND Corporation, Leading Educators used a quasi-experimental study design and combined two methods for the analysis: propensity score weighting (PSW) and difference in differences (DiD). While the DiD method accounts for changes over time that are not explained by the intervention, the PSW accounts for selection bias. RAND provided extensive guidance to ensure Leading Educators could provide schools and districts with strong information about how to support improvement in student outcomes.

This study found a positive and statistically significant effect on math scores in both sites, a positive and significant effect in ELA scores in Michigan and a positive and not significant effect in ELA in Louisiana. The effects are reported in standard deviation units and can be converted to an improvement index using the difference between the percentile range of an average student in a school served by Leading Educators and a comparable school. The improvement indexes for math were 12% in Louisiana and 4% in Michigan. The improvement index for ELA in Michigan was 6%. We found no discernible effects in ELA in Louisiana during the first year. Finding positive and significant effects across considerably different contexts is strong evidence for the scalability of the Leading Educators model. Lipsey et al. (2012, p. 34) calculated the mean effect size of 89 randomized studies that used broad scope standardized tests results at the elementary level as the outcome measure. Applying this framework, all effects of Leading Educators' intervention are above the mean effect of similar randomized studies.

One possible explanation for the difference in results observed in math and ELA in Louisiana could be differences at the school level that influence math and ELA in different ways, such as the use of high quality curriculum and competing priorities and interventions. The professional development program focused exclusively on math in some schools and focused exclusively on ELA in others. The demographics of supported math students differed from the demographics of supported ELA students in the same region but the bigger differences are observed across the two regions. Compared to the state of Louisiana, a much higher proportion of the supported students in Louisiana are students of color and a slightly higher proportion are English learners. In Michigan, a much greater proportion of supported students are students of color, English language learners, and disadvantaged when compared with state averages.

Introduction

Majorities of students are not ready for their next step beyond high school due to inconsistent access to grade-appropriate teaching and learning. School systems seek effective strategies for improving instruction at scale to strengthen the opportunities students have in the classroom to reach college and career readiness. Many support organizations exist to meet this demand, but few interventions have significantly affected student outcomes when assessed with the most rigorous evaluation methods. The most common analyses focus on changes in percentage of students achieving proficiency, but that metric cannot easily detect movement above or below the cut score, and trends seen can vary widely within the same dataset due to the placement of the cut score.[1] Consequently, trends can appear deceptively large or small dependent on factors that have little to do with actual student growth. Other studies that compare average scores use methods that cannot establish causality. For example, comparing outcomes before and after an intervention fails to take into account other events that may have affected the outcomes. Likewise, studies that compare participants and non-participants fail to account for selection bias. Student analyses that move beyond proficiency and simple correlational methods can provide schools and districts with stronger information about how a particular choice may have impacted student outcomes.

The purpose of this impact evaluation is to answer the question: What is the effect of Leading Educators' model of content-based professional learning on student learning? More specifically, what was the effect of programs in Louisiana and Michigan during the school year 2017-2018 on student standardized scores in mathematics and English language arts (ELA). Leading Educators used a quasi-experimental study design and combined two methods for the analysis: Propensity Score Weighting (PSW) and Difference in Difference (DiD). The study found a positive and significant effect for math scores in both sites, a positive and significant effect for ELA scores in Michigan and a positive and not significant effect for ELA in Louisiana. Leading Educators' logic model anticipates positive and significant effects after one or two years of partnership. These results are especially promising since positive, significant effects were detected at both sites after the first year. Schools supported in Michigan and Louisiana have considerably different conditions, suggesting that this type of support can work in a variety of contexts.

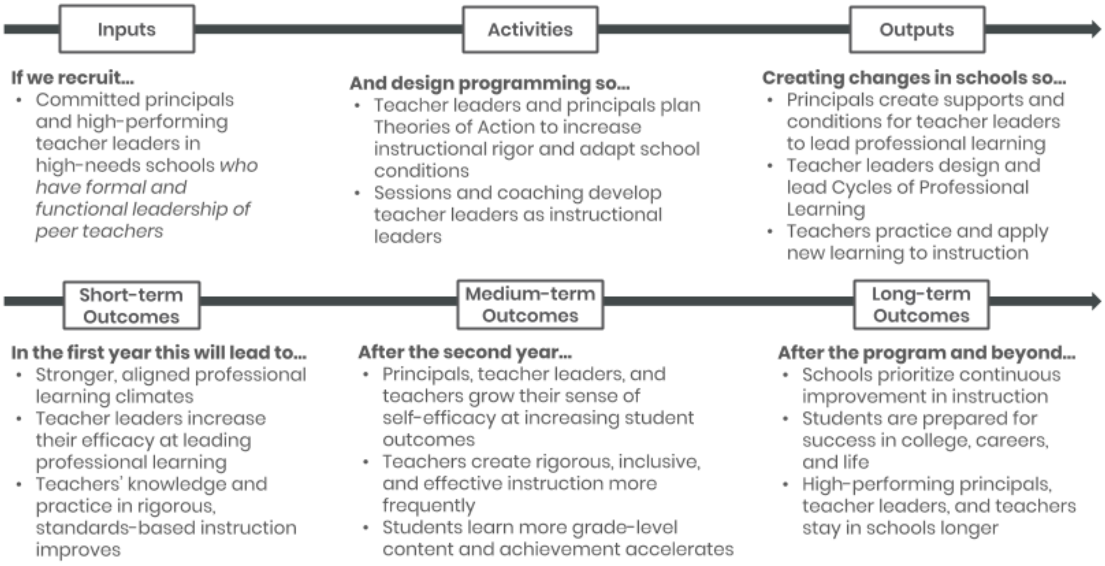
Leading Educators Program Model

During the 2017-2018 school year, Leading Educators partnered with school systems in Louisiana and Michigan to build teacher knowledge and practice to meet the demands of rigorous college and career readiness standards. Using local teams of teacher leaders as key drivers for instructional improvement, Leading Educators supported system leaders to design and implement job-embedded, content-specific professional learning that allows teachers to learn, plan, and practice with grade-appropriate content in collaboration with peers.

Teacher leaders who supported schools received cohort-based professional development to prepare them for effective leadership of professional learning, as well as monthly coaching. Within schools, the teacher leaders facilitated 90-minute blocks of sequenced professional learning aligned to their curriculum on a weekly or bi-weekly basis. In content-focused cycles of professional learning, groups of content-alike teachers met weekly to tailor their instruction using standards-based instructional planning and practice. In meetings, content teams analyzed data, set goals, learned individually and collaboratively, applied new learning, and monitored student progress. Meetings were linked to key areas of focus in a cycle so that teachers could build deep capacity in a narrow set of key topics connected to school priorities. The program comprised about 78 hours of out-of-school sessions (summer and weekend sessions) and 14 to 15 hours of in-school coaching and support. On average, teacher leaders conducted about one and a half content cycles over the course of the year, with the most common topic in ELA being "Building Knowledge through Text Sets" and the most

[1] See Ho (2008) for a discussion of the limitations of using proficiency indicators.

Figure 1: Leading Educators Logic Model for Program Year 2017-2018



Inputs

If we recruit...
- Committed principals and high-performing teacher leaders in high-needs schools *who have formal and functional leadership of peer teachers*

Activities

And design programming so...
- Teacher leaders and principals plan Theories of Action to increase instructional rigor and adapt school conditions
- Sessions and coaching develop teacher leaders as instructional leaders

Outputs

Creating changes in schools so...
- Principals create supports and conditions for teacher leaders to lead professional learning
- Teacher leaders design and lead Cycles of Professional Learning
- Teachers practice and apply new learning to instruction

Short-term Outcomes

In the first year this will lead to...
- Stronger, aligned professional learning climates
- Teacher leaders increase their efficacy at leading professional learning
- Teachers' knowledge and practice in rigorous, standards-based instruction improves

Medium-term Outcomes

After the second year...
- Principals, teacher leaders, and teachers grow their sense of self-efficacy at increasing student outcomes
- Teachers create rigorous, inclusive, and effective instruction more frequently
- Students learn more grade-level content and achievement accelerates

Long-term Outcomes

After the program and beyond...
- Schools prioritize continuous improvement in instruction
- Students are prepared for success in college, careers, and life
- High-performing principals, teacher leaders, and teachers stay in schools longer

common topic in math being "Curricular Shifts in Math."

Figure 1 shows the links between Leading Educators program activities and the expected outcomes. Based on experience and multiple meta-analyses (Timperley & Alton-Lee, 2018; Kraft, 2018; Boulay et al., 2018; Dolfin et al, 2019), Leading Educators expects to see improvements in the short term in school climate, teacher leader beliefs, and leadership skills. In the medium term (1 to 2 years), after teachers have integrated their new knowledge and skills into their practice, a measurable increase in student outcomes will be detected.[2]

Description of the files and analysis

The methodology for this evaluation was developed with the support of the RAND Corporation. It combines two statistical methods for the analysis: Propensity Score Weighting and Difference in Difference. PSW is used to calculate weights that are applied to students in the comparison sample in order to make the group composition similar to the group of students served by the intervention. The weights are calculated by taking into account a combination of demographic variables and their relative influence on student outcomes. DiD is a type of regression analysis that calculates the difference before and after an intervention between participants and non-participants. The difference between the pre/post differences in outcomes of the participants and non-participants was used to isolate the

[2] Timperley & Alton-Lee (2008) synthesize the evidence from 97 empirical studies of professional development models that have demonstrated to have a positive impact on outcomes on diverse learners and conclude among other things that it takes one or two years for teachers to build knowledge and change beliefs and practice. Boulay et al. (2018) evaluated 67 i3 education interventions. Only twelve of them had a statistically significant positive impact on at least one student academic outcome. Nine of the twelve evaluations reported a positive increase on outcomes in ELA and mathematics and only two of those nine interventions had a positive effect at the end of year 1 of the program, with most reporting positive effects in year two and some reporting positive effects after year 3 (Clark, 2015; May, 2016; Parkinson, 2015; Gallagher, 2016; Meyers, 2016; Brandt, 2013; Mokher, 2016; M. Bos, 2019; Fong, 2015; Jones, 2016). Kraft (2018) also uses evidence from a meta-analysis where he finds that smaller effects are seen when cumulative decisions and sustained effort over time is required, as is the case of Leading Educators model which emphasises school system change and supports a balanced combination of math and ELA content knowledge, equity mindsets and leadership skills. Dolfin (2019) evaluated a professional development intervention very similar to Leading Educators and found that students scores on ELA assessments improved only after year two.

effect of the intervention over time.[3] Combining both methods (DID and PSW) integrates the benefits of both model assumptions. While the DiD method accounts for changes over time that are not explained by the intervention, the PSW accounts for selection bias. RATE (RAND Analysis Toolkit in Education) is the software used for the analysis, developed by the RAND Corporation.

The state of Louisiana provided anonymized, student-level results on the Louisiana Educational Assessment Program (LEAP) from the 2014-2015 through 2017-2018 school years. The file contained indicators for all students taught by Leading Educators teacher leaders and their peers with grades 3-8 math and ELA LEAP assessment data and demographic information of students from the entire state of Louisiana. This provided the analysis with a very large comparison group, essential for the effectiveness of the PSW analysis in particular. As opposed to receiving separate datasets directly from the schools, the district level dataset guarantees consistency of the indicators and naming conventions and reduces the chances of errors in data processing and cleaning.

The impact evaluation in Michigan used files from four of the five districts Leading Educators served from two different sources and with some variability in the content and characteristics of the datasets. The merged files with the common variables included 14 schools, ten of them elementary schools and four middle schools. Two high schools were excluded from the dataset because the standardized tests are not comparable with the ones for elementary and middle school and because there was no data for the three years required. Data was not available for four elementary schools and one middle school served. Half of the schools in the dataset were schools supported by Leading Educators and the other half were comparison schools. The file includes anonymized student level demographics and results from the M-Step for 3rd to 8th grade students from 2015-2016 to 2017-2018 school years. All files include indicators for all students taught by Leading Educators teacher leaders and their peers.

The Louisiana and Michigan analysis and dataset have two common limitations. First, teachers had completed nine months of content-focused, job-embedded professional learning. It may take more than one year to produce strong enough effects that can be detected by statistical models. Second, the files do not include teacher level demographics, program attendance or other implementation data that would allow analysis of connections with teacher characteristics and program implementation. Finally, the number of schools served in Louisiana and Michigan is too small to include the school level characteristics in the PSW analysis. This means that the evaluation design could not include the peer effect to balance the characteristics of the treatment and control groups.[4] Nevertheless, the research team was able to include them in the DiD analysis and improve the estimates in that way, which is one of the advantages of combining the two methods.

The Louisiana dataset is also missing Free and Reduced Lunch status, a critical variable as a significant body of research demonstrates individual socioeconomic status (SES) and school composition both affect student achievement. The Louisiana Department of Education will likely be able to add this status in the future. Another limitation was the inability to include an indicator for location in the Louisiana analysis due to data sharing restrictions to protect the privacy of Louisiana students. Future studies will account for maintaining student privacy while including some indicators for location and urbanity. Conditions in rural Louisiana schools are likely quite different from conditions in the urban schools that received the intervention. For Michigan, this variable is not as relevant because the comparison schools are located in the supported school districts and should share most of the neighborhood characteristics of the treatment schools.

In Michigan, six of the 14 schools that Leading Educators supports were excluded from the analysis. That means the impact of Leading Educators' programs could only be evaluated in 57% of the supported schools. Furthermore, the dataset is limited to scores for students in 3rd to

[3] See Appendix 1 and 2 for a more detailed explanation of the PSW and DiD methods.

[4] "Peer effect" is the presumption that student composition in schools has an effect on student outcomes.

8th grade. The Michigan dataset did not include data from the entire state, so the sample of comparison schools was more limited. After applying the weights calculated using the PSW analysis, the treatment and comparison group observed characteristics are very similar.

Student demographics

The professional development program focused exclusively on English language arts (ELA) in some supported schools, while others focused on math. Because the demographics of Leading Educators math students differed from the demographics of ELA students, different groups were used for subject-specific analysis. For each group, a similar comparison sample was estimated using the PSW analysis.

Compared to the students in the ELA sample, a greater proportion of math students in Louisiana were white (13% more), Hispanic (4% more) and English language learners (2% more), and fewer were Black (20% less) and exceptional learners (3.5% less). Students in the ELA sample performed about 0.1 standard deviations further below the state average on the ELA assessment compared to the math sample relation to the math average.

Compared to Louisiana State, a much higher proportion of Leading Educators students are students of color and a slightly higher proportion are English learners.

Compared to the students in the ELA sample in Michigan, a greater proportion of math students were Black (22% more), white (13% more) and Asian (17% more), and fewer were Hispanic (54% less), English learners (11%), students with free and reduced lunch status (14% less), and exceptional learners (3% less). Students in the ELA sample performed 0.1 standard deviations further below the state average on the ELA assessment compared to the math sample relation to the math average. Compared with the state of Michigan, a much greater proportion of Leading Educators supported students are students of color, English language learners, and economically disadvantaged.

Demographic differences between Michigan and Louisiana can have implications for the way the impact of Leading Educators' programs is interpreted. Positive and significant effects under different circumstances suggest that the program can be effectively implemented and scaled in varying contexts.

Table 1: Louisiana Partnership and State Demographic Information

	% Students with exceptionalities	% of Students who are Black	% of Students who are White	% of Students who are Hispanic	% of Students who are Asian	% of students who are English Learners	Mean baseline scale score
Leading Educators Math students	6.70%	62.90%	21.30%	11.20%	2.20%	6.69%	-0.24
Leading Educators ELA students	10.91%	83.42%	8.94%	5.59%	1.16%	4.30%	-0.33
Louisiana State 2016	14.00%	44.00%	45.00%	6.00%	2.00%	3.16%	-

Source: Louisiana Department of Education and this study.

Table 2. Michigan Partnership Demographic Information for Math and ELA Schools

	% Students with exceptionalities	% of Students who are Black	% of Students who are White	% of Students who are Hispanic	% of Students who are Asian	% of students who are English Learners	% of Students with Free and Reduced Lunch Status	Mean baseline scale score
Leading Educators Math students	7.90%	28.70%	16.60%	31.10%	17.10%	33.90%	79.50%	-0.37
Leading Educators ELA students	10.60%	7.00%	3.80%	84.60%	0.00%	45.20%	93.30%	-0.47
Michigan State 2016	12.91%	18.02%	66.62%	7..68%	3.25%	6.20%	45.86%	-

Source: Michigan Department of Education and this study.

Results

The idea behind the PSW technique is to calculate weights that will be used in the regression analysis to balance the comparison and treatment samples according to the observed characteristics of the students and their environment. An important assumption in this method is that all the observed characteristics that are correlated with student outcomes and that are different in the two groups have been included. There are a few important characteristics that meet this criteria and were not included in the analysis. Teacher demographic information and indicators of program participation were not included because they were not part of the data sharing agreement for the school year 2017-2018. School level characteristics were not included in the PSW analysis due to sample size. The student level variables included in the PSW analysis in both regions are race/ethnicity, gender, prior year scores, English language proficiency, and exceptionality status. For Michigan analysis, we were able to also include free and reduced lunch status.

The Difference in Difference analysis was performed including both student and school level variables. In Louisiana school level aggregates for Black and White race, exceptionality status, and English language learner status in both regions were included. In Michigan, school aggregates for Hispanic race and free and reduced lunch status were also added. The student level variables included are the same variables included in the PSW, a strategy known as a doubly robust estimation.

The results of the Difference in Difference analysis for ELA and math are presented in **Table 3**. The probability value of zero means that the margin of error of our effect size estimate is close to zero, and therefore, we can be very confident about the results given that all model assumptions were met. The P value of 0.69 found in the ELA sample in Louisiana means that the margin of error is too high and it is not possible to determine if the program had an effect on student scores or if the effect was too small to be detected with the available data. The effect sizes can then be interpreted as the difference in the change in score after one year in the program between a student in a school served by Leading Educators and a comparison student. In other words, a student in a supported math school in Louisiana scored 0.31 standard deviations better after one year than a comparison student. In Michigan, a student in a Leading Educators school scored 0.15 standard deviations better in ELA and 0.10 standard deviations better in math after one year than a comparison student.

One possible explanation for the difference in results observed in math and ELA in Louisiana could be connected with differences at the school level that influence math and ELA in different ways. Data collected by Leading Educators for the teacher cohort of 2017 shows that 9% more math teachers reported the use of at least one high-quality aligned curriculum compared to ELA teachers. Competing priorities and possibly contradictory interventions that were implemented after Hurricane Katrina could have had different effects on teachers' knowledge and practice of math and ELA. More years of analysis are necessary to understand why the program had a significant large effect in math earlier than expected.

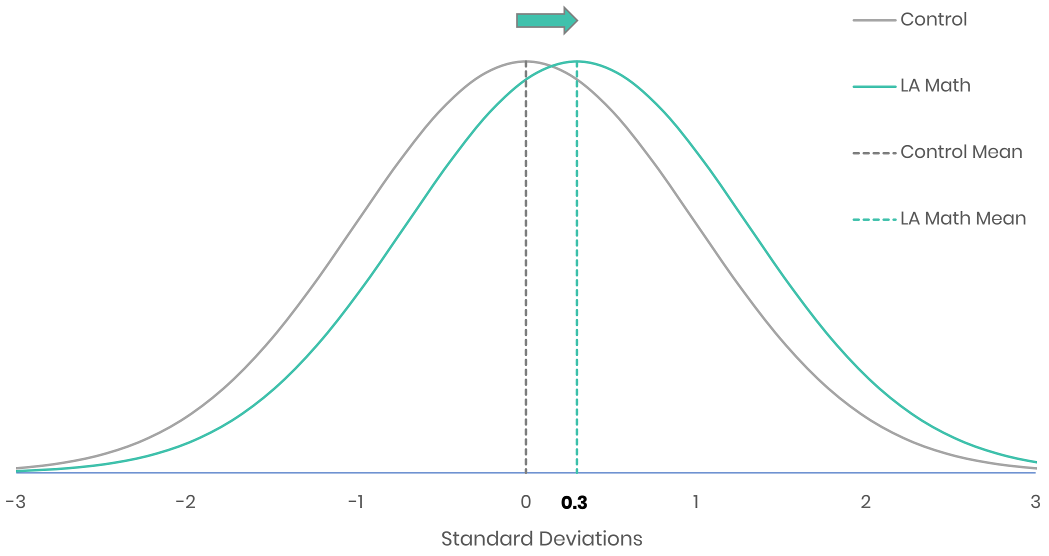
Figure 2 illustrates the distribution of the scores for the math program in Michigan compared to the control sample. The difference between the mean of the Michigan distribution and the mean of the control distribution, represents the magnitude of the effect. We will dedicate the next section to giving context to this magnitude and to understand its practical significance.

Table 3. Results from the Difference in Difference Analysis using Propensity Scores to Weight the Comparison Sample

Subject	Effect Size	Probability Value	N (total observations, across years, treatment and control)
Math Louisiana	0.31	0.00**	8,601
Math Michigan	0.10	0.00**	31,626
ELA Louisiana	0.05	0.69	6,947
ELA Michigan	0.15	0.00**	30,924

**Statistically significant at the 1% significance level (P <.001)

Figure 2: Distribution and Effect Size of Louisiana Math Program, 2017-2018



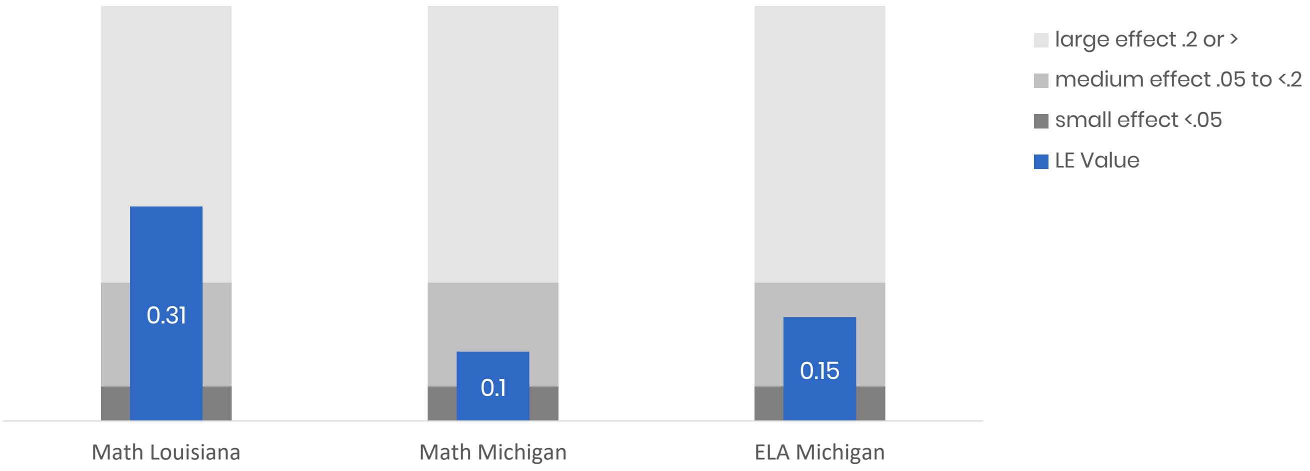
Interpreting the magnitude of the results

Several strategies[5] can be used to understand the magnitude of the effect size, and therefore, the magnitude of Leading Educators' impact on student outcomes. An intuitive and easy to understand strategy involves comparing the effect size with the effects obtained by other reputable evaluations that calculate causal effects of similar professional development programs. Finding and discerning which studies meet the criteria can be challenging, and Kraft's (2018) effect size interpretation framework is a helpful tool. According to his research, an effect size above 0.2 can be considered large when compared with other teacher professional development programs in middle and high school.[6]

Because these programs also vary widely in terms of cost and scalability, these two factors should also be included in the interpretation. The core of Leading Educators' professional development program was designed not only to be scalable, but also to be sustainable and cost effective. The program cost $238 per pupil in Louisiana and $308 per pupil in Michigan, a low cost relative to other professional development programs according to Kraft's interpretation framework.

Lipsey et al. (2012, p.34) calculated the mean effect size of 89 randomized studies that used broad scope standardized tests results at the elementary level as the outcome measure. The mean is compared to Leading Educators' effect sizes in **Figure 2**. All effects of Leading Educators' intervention are above the mean effect of similar randomized studies.[7]

Figure 3. Michigan and Louisiana Effect Sizes Compared to Kraft's Benchmarks



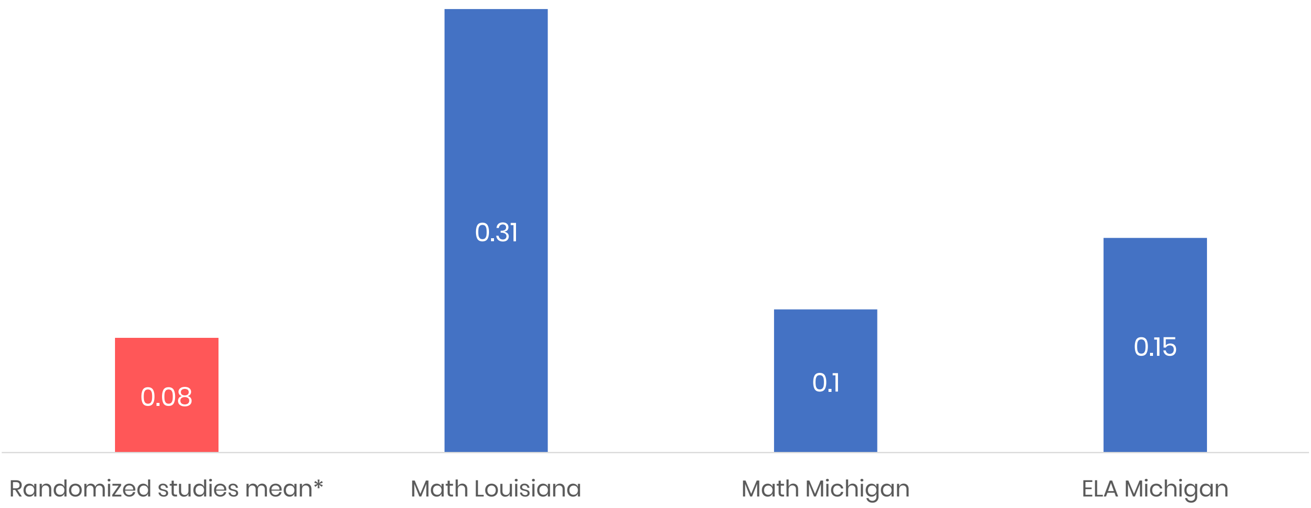
Source: Kraft (2018) and this study.

[5] See appendix 3 for a decision tree to contextualize effect sizes.

[6] Effect sizes in Elementary school are expected to be higher and therefore an effect size that is considered large in High school could be considered medium or less large in Elementary school.

[7] Slavin & Cheung (2015) found that the average effect size of randomized studies was smaller compared to quasi-experimental studies, but also noticed that these results have not been consistently found in other reviews about this comparison.

Figure 4. Michigan and Louisiana Effect Sizes Compared to Randomized Studies



Source: Lipsey et al. (2012) and this study.

Another preferred strategy that can be used to understand the magnitude of the effect is to calculate an improvement index. An improvement index converts the standard deviation units into percentile change of the average student. For the math students in Louisiana, an effect size of 0.31 is equivalent to a 12% improvement index, or a 12 point increase in the percentile range of an average student in the comparison schools, if he had attended a supported school. We could also suggest that 62% (50% + 12%) of Leading Educator students performed better in math than the comparison students.[8] For the ELA students in Michigan, an effect size of 0.15 is equivalent to a 6% improvement index, or a 6 point increase in the percentile range of an average student in the comparison schools if she had attended a supported school. The effect size for the math program in Michigan of 0.10 is equivalent to a 4% improvement index.

The improvement index is the effect size translation chosen by the What Works Clearinghouse (WWC), an initiative of the U.S. Department of Education's Institute of Education Sciences (IES). Improvement indexes reported by WWC for teacher level interventions that measure impact on ELA and math are presented beside Leading Educators' improvement indexes in **Table 4**. None of these teacher level programs with studies that meet WWC standards had discernible effects on ELA achievement, and only Teach For America (TFA) had a positive and significant effect on math, translated to a 4% improvement index. TFA is comparable to Leading Educators in terms of cost and scalability.

[8] The improvement index was calculated using a table that lists the proportion of the area under the standard normal curve, by looking at the value under the z-score value that is equal to the effect size and subtracting 50% from that value. For example, for a .25 effect size, the area under the table equals 60%. Then 60%-50% equals an improvement index of 10 percentile points. "We could then conclude that the intervention would have led to a 10% increase in percentile rank for an average student in the comparison group and that 60% (10% + 50% = 60%) of the students in the intervention group scored above the comparison group mean." (What Works Clearinghouse, 2011).

Table 4. What Works Clearinghouse and Leading Educators improvement index comparison

PD Program	Impact assessed	Focus area	Grades	Improvement Index
Leading Educators - Michigan	ELA	Content knowledge and conditions	3-8	6%
Leading Educators - Louisiana	ELA	Content knowledge and conditions	3-8	No discernible effects
Teach for America (TFA)	ELA	Non traditional trained teachers	K-12	No discernible effects
National Board for Professional Teaching Standards Certification	ELA	Content knowledge and practice	3-8	No discernible effects
Leading Educators - Michigan	Math	Content knowledge and conditions	3-8	4%
Leading Educators - Louisiana	Math	Content knowledge and conditions	3-8	12%
Teach for America (TFA)	Math	Non traditional trained teachers	K-12	4%
National Board for Professional Teaching Standards Certification	Math	Content knowledge and practice	3-8	No discernible effects
TNTP Teaching Fellows	Math	Non traditional trained teachers	6-8	No discernible effects

Source: Institute of Education Sciences: What Works Clearinghouse

The last strategy discussed here involves converting the standard deviation units to other units of reference connected to a more tangible goal or concept like the racial achievement gap, the socioeconomic achievement gap, or units of time like days of schooling. We choose not to convert to units of time because there are many concerns in recent research about the reliability of this type of conversion and its interpretation. The Black/White achievement gap was estimated to be one standard deviation by Miksic (2014). Hannussek (2019) estimates that the SES gap between those at the 90th and 10th percentile of the achievement distribution is about 2 standard deviations. Thus, the effect size of 0.31 standard deviations obtained in math in Louisiana is equivalent to one third (30%) of the Black/White gap and about one sixth (16%) of the SES gap. The corresponding conversions for math and ELA in Michigan are 10% and 15% of the Black/White gap and 5% and 7.5% of the SES gap.

Figures 5 to **7** compare the size of the effect with the achievement gaps at the state level in the corresponding areas of content. **Figure 5** shows that the improvement in Math in Louisiana is almost the same size as the math performance gap between White and Hispanic students and about 40% the size of the Black/White math gap. In Michigan, the largest improvement was found in ELA and can be compared with reducing the Hispanic/White gap by 30% and the Black/White gap by 17% (see **Figure 6**).

Figure 5: Size of FY18 improvement in Math scores in Louisiana Compared to 2013 Louisiana Math Gaps

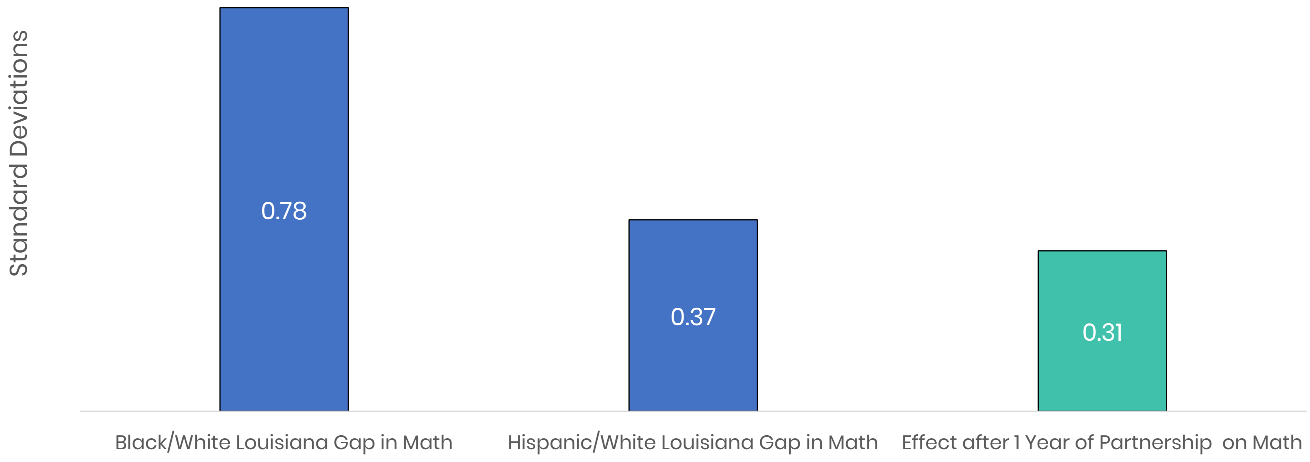


Figure 6: Size of FY18 improvement in ELA scores in Michigan Compared to 2013 Michigan Reading Gaps

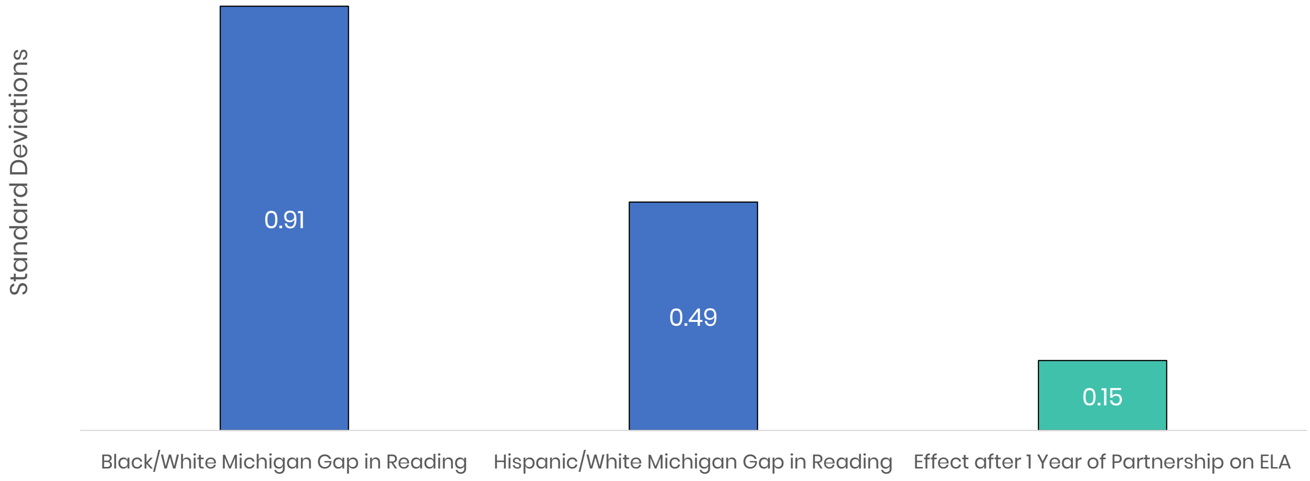
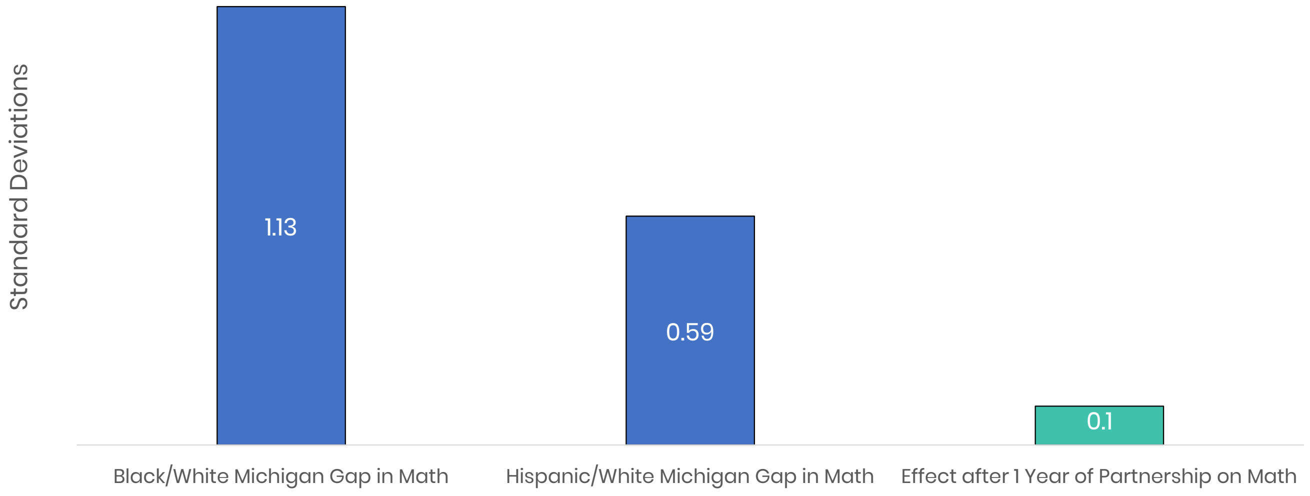


Figure 7: Size of FY18 improvement in Math scores in Michigan Compared to 2013 Michigan Math Gaps



Source Figures 5-7: Racial and Ethnic Achievement Gaps. (n.d.) and this study

Conclusion

Based on research and past experience, Leading Educators expects effects on student achievement to take two years to be detected by statistical methods. Teacher leaders must improve their own content knowledge and develop their leadership skills while simultaneously leading learning and developing their practice and the practice of their colleagues. While some of the learning can be immediately applied in the classroom and transferred to students, other types of learning involve several rounds of practice and feedback. Detecting a positive impact on student standardized scores after only 9 months of support surpasses these expectations. The effects observed range from medium to large when compared with competing professional development programs and their impact is even higher when the comparable low cost per student of the program is taken into account. This study found no discernible effects in the ELA group in Louisiana, which is in alignment with Leading Educators' logic model. One possible explanation for the difference in impact between math and ELA could be found in unobserved characteristics at the school level that could have affected ELA and Math scores in different ways. After the tragedy of Hurricane Katrina, the New Orleans school system experienced a drastic transformation that included the coexistence of numerous programs and organizations working on similar goals. Perhaps in math, Leading Educators was able to leverage the strengths of other interventions and competing priorities. Another possible explanation is connected to a higher proportion of math teachers reporting the use of high quality curriculum when compared with ELA teachers. By calculating the effects in the following school year, and with new evidence collected through the continuous improvement work of the program team, future analysis may be able to identify the factors that helped increase student outcomes at a significant magnitude earlier than expected in math and ELA in Michigan and in math in Louisiana.

Finding a positive and significant impact in two regions with contrasting characteristics in the student populations and school systems suggests that the Leading Educators model can work in a variety of contexts and can be implemented with fidelity at different scales. The results from this study are strong evidence that the Leading Educators model is a highly cost-effective and scalable teacher professional development program, capable of producing sizable increases in students' academic outcomes in partnership with underserved school systems.

References

Baird, Matthew D. and John F. Pane (2018). Translating Standardized Effects of Education Programs into More Interpretable Metrics. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/working_papers/WR1226.html.

Boulay, Beth; Goodson, Barbara; Olsen, Rob; McCormick, Rachel; Darrow, Catherine; Frye, Michael; Gan, Katherine; Harvill, Eleanor; Sarna, Maureen. (2018). The Investing in Innovation Fund: Summary of 67 Evaluations. Final Report. U.S. Department of Education. https://ies.ed.gov/ncee/pubs/20184013/pdf/20184013.pdf

Brandt, C., Meyers, C. & Molefe, A. (2013). The Impact of the enhancing Missouri's Instructional Networked Teaching Strategies (eMINTS) Program on Teacher Instruction and Student Achievement – First Year Results. In R. McBride & M. Searson (Eds.), Proceedings of SITE 2013--Society for Information Technology & Teacher Education International Conference (pp. 2023-2031). New Orleans, Louisiana, United States: Association for the Advancement of Computing in Education (AACE). Retrieved October 3, 2019 from https://www.learntechlib.org/primary/p/48400/.

Clark Tuttle, Christina; Booker, Kevin; Gleason, Philip; Chojnacki, Gregory; Knechtel, Virginia; Coen, Thomas; Nichols-Barrer, Ira; Goble, Lisbeth. (2015). Understanding the Effect of KIPP as it Scales: Volume I, Impacts on Achievement and Other Outcomes. Final Report of KIPP's Investing in Innovation Grant Evaluation. Mathematica Policy Research. http://www.kipp.org/wp-content/uploads/2016/09/kipp_scale-up_vol1-1.pdf

Dolfin, Sarah; Richman, Scott; Choi, Jane; Streke, Andrei; DeSaw, Cheryl; Demers, Alicia and Poznyak, Dmitriy (2019). Evaluation of the Teacher Potential Project. Washington, DC: Mathematica.

Eric A. Hanushek, Paul E. Peterson, Laura M. Talpey, and Ludger Woessmann (2019). The Unwavering SES Achievement Gap: Trends in U.S. Student Performance. NBER Working Paper No. 25648

Fong, Anthony B; Finkelstein, Neal D; Jaeger, Laura M; Diaz, Rebeca; Broek, Marie E. (2015). Evaluation of the Expository Reading and Writing Course. Findings from the Investing in Innovation Development Grant. WestEd. https://www.wested.org/wp-content/uploads/2016/11/1438034849ERWC_Report-3.pdf

Gallagher, H. Alix; Arshan, Nicole; Woodworth, Katrina. (2017). Impact of the National Writing Project's College-Ready Writers Program in High-Need Rural Districts, Journal of Research on Educational Effectiveness, 10:3, 570-595, DOI: 10.1080/19345747.2017.1300361

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. Educational researcher, 37(6), 351-360.

Jones, Curtis J; Christian, Michael; Rice, Andrew. (2016). The Results of a Randomized Control Trial Evaluation of the SPARK Literacy Program. SREE Spring 2016 Conference Abstract Template. https://files.eric.ed.gov/fulltext/ED567484.pdf

Kraft, M.A. (2018). Interpreting Effect Sizes of Education Interventions. Brown University Working Paper.

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). Translating the Statistical Representation of the Effects of Education

Interventions into More Readily Interpretable Forms. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the IES website at http://ies.ed.gov/ncser/.

May, Henry; Sirinides, Philip M; Gray, Abigail; and Goldsworthy, Heather. (2016). Reading Recovery: An Evaluation of the Four-Year i3 Scale-Up. CPRE Research Reports. Retrieved from http://repository.upenn.edu/cpre_researchreports/81

M. Bos, Johannes; Dhillon, Sonica; Borman, Trisha; O'Brien, Brenna; Graczewski, Chery; Park, So Jung; Liu, Feng; Adelman-Sil, Ethan; Hu, Lynn. (2019). Building Assets and Reducing Risks (BARR) Validation Study Final Report. American Institutes for Research. https://www.air.org/sites/default/files/downloads/report/AIR-BARR-Validation-Final-Report-July-2019.pdf

Meyers, Coby V; Molefe, Ayrin; Brandt, W. Christopher; Zhu, Bo; Dhillon, Sonica. (2016). Impact Results of the eMINTS Professional Development Validation Study. Educational Evaluation and Policy Analysis. Volume: 38 issue: 3, page(s): 455-476. https://doi.org/10.3102/0162373716638446.

Miksic, M. (2014). The persistent achievement gaps in American education. CUNY Institute for Education Policy.

Mokher, Christine; Lee, Steve; Sun, Christopher. (2016). Final Findings from Impact and Implementation Analyses of the Northeast Tennessee College and Career Ready Consortium. CNA Education. https://files.eric.ed.gov/fulltext/ED569930.pdf.

Parkinson, Julia; Salinger, Terry; Meakin, John; Smith, Deeza-Mae. (2015). Results From a Three-Year i3 Impact Evaluation of the Children's Literacy Initiative (CLI): Implementation and Impact Findings of an Intensive Professional Development and Coaching Program. American Institutes for Research. https://www.cli.org/wp-content/uploads/2015/09/CLI-i3-Impact-Report-July-2015.pdf

Slaving, Robert; Cheung, Alan. (2015). How Methodological Features Affect Effect Sizes in Education. Best Evidence Encyclopedia. Baltimore: Johns Hopkins University. http://www.bestevidence.org/word/methodological_Sept_21_2015.pdf

Timperley, Helen and Alton-Lee, Adrienne (2008). "Reframing Teacher Professional Learning: An Alternative Policy Approach to strengthening Valued Outcomes for Diverse Learners". REVIEW OF RESEARCH IN EDUCATION 32: 328. DOI: 10.3102/0091732X07308968

What Works Clearinghouse (2016). WWC Summary Of Evidence For This Intervention: Teach for America (TFA). https://ies.ed.gov/ncee/wwc/Intervention/6.

What Works Clearinghouse (2011). Procedures and standards handbook (version 3.0). Washington, DC: U.S. Department of Education (Appendix F). Available at https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf

Appendix 1: Propensity score weighting

The idea behind the propensity score is that while we can't observe the same individual in two worlds at the same time, we can find individuals who have very similar characteristics and assign people to treatment and comparison groups based on their observable characteristics. From a practical perspective, we should identify comparisons using determinants of program participation. If the list of relevant determinants is very large, or if each characteristic takes on many values, it may be hard to identify a comparison group. As the number of characteristics or dimensions increases, evaluators may run into what is called the curse of dimensionality. For example, if two characteristics (age, gender) explain program participation, it may easy to find relevant comparisons, where women are paired with women and then ages can be approximately matched. But if in addition there is a need to consider income, education level, ethnic group, and other characteristics, identifying the comparison groups is much harder. Therefore, these characteristics are combined into a single score called the propensity score, a number between 0 and 1 that represents the predicted probability of participating in the intervention.

Rather than identifying comparisons based on multiple characteristics, participants are instead assigned a propensity score. Intuitively, people with a similar propensity score have a similar probability of being in the treatment group. Maybe these individuals are not exactly the same on the observed characteristics, but overall they are equally likely to participate in the program. Good estimates of the propensity score can be effective at diminishing or eliminating confounders in the estimation of treatment effects.

While propensity scores have been used in the past to directly match treatment and comparison groups, it has been shown that a better use of these scores is to generate weights. These weights are then used in running a regression to estimate the treatment effect, so that a person receiving a higher weight is more similar to the treatment group, and some people receive a weight of 0.

Conditions and Assumptions:

- All characteristics excluded from the matching process (because they are unobserved or unmeasured) are either not correlated with the outcomes or do not differ across participants and non-participants.
- There are data available on a large number of participants and non-participants. Because individuals need to be paired on fundamental characteristics that are unaffected by the intervention, this requires a significant amount of data.
- The data contain a large number of characteristics that are either unaffected by the intervention (e.g., age or gender) or measured before the program was implemented (such as outcomes at baseline, if available).

Limitations

First, to use this method we need a large, detailed dataset of participants and non-participants, ideally before and after the intervention. The dataset must contain sufficient information to adequately estimate propensity scores. A related limitation is that propensity score weighting is not possible unless there is sufficient overlap in the propensity score between the treatment and comparison groups. In addition, note that we can only estimate propensity scores for individuals based on the characteristics that are observed in the data. Thus, the critical assumption is that there are no unobserved or unmeasured differences between the two groups that affect outcomes, which can be a strong assumption in many settings, especially if there is no information on the outcome of interest at baseline.

Appendix 2: Difference in Difference

The key idea of DID is to combine two methods to improve inferences. The first is the difference in outcomes before and after the program for those who participated in the intervention. The main concern with this difference is that any changes in outcomes could be the result of the intervention or another event that occurred in the time between the two data collections. The second difference compares the outcomes of participants and non-participants. Used alone, this is also problematic because these groups could differ in unobservable characteristics that may drive the differences in outcomes.

However, by combining both of the simple differences into a DID framework, we can address some of these concerns. The DID method gives the change in the outcome of the treatment group over time, relative to the change in the outcome of the comparison group. Here is one way of looking at the broad idea: In addition to the difference in outcomes among participants, we can also make use of the pre/post differences among non-participants. Just like the participants, the non-participants are subject to the concurrent events. But unlike the participants, the non-participants are not affected by the education intervention. Thus, by calculating the difference between the pre/post differences in outcomes of the participants and non-participants, we can isolate the effect of the intervention over time.

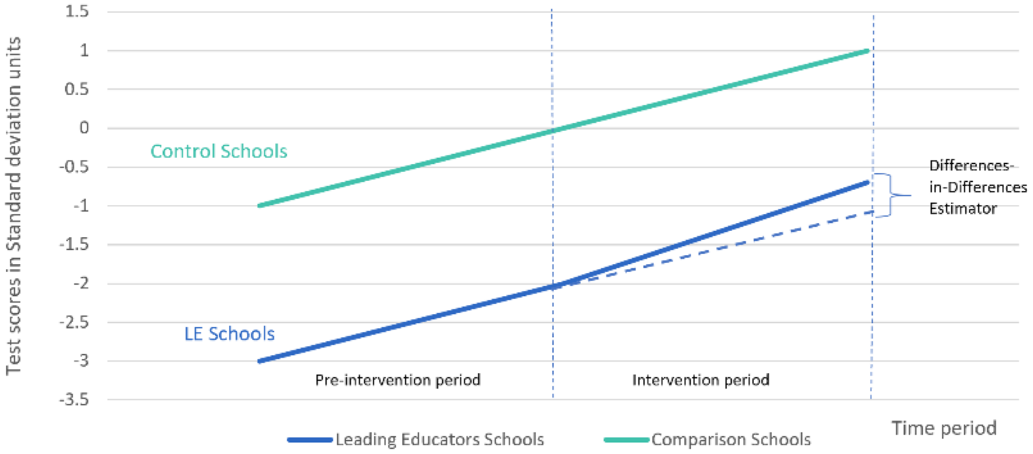
Conditions and Assumptions

- In the absence of the treatment, the change in outcomes for the “treatment” group would have been similar to the change in outcomes for the comparison group. This is often called the “parallel trends” assumption and allows us to calculate the counterfactual for impact evaluation. The two groups do not need to be exactly comparable at baseline. We merely require that their trajectory over time would have been similar
- There is reasonable certainty that the participants and non-participants only differ in their exposure to the intervention. If either group has been affected by some other event, the DID method will lead to a biased result.
- There are at least two instances of data collection for participants and non-participants: before and after the intervention. Additional pre-program data are not critical, but they are useful to implement robustness checks (see below).
- The same individuals can be followed over time; while this condition is not necessary, it is the preferred situation.

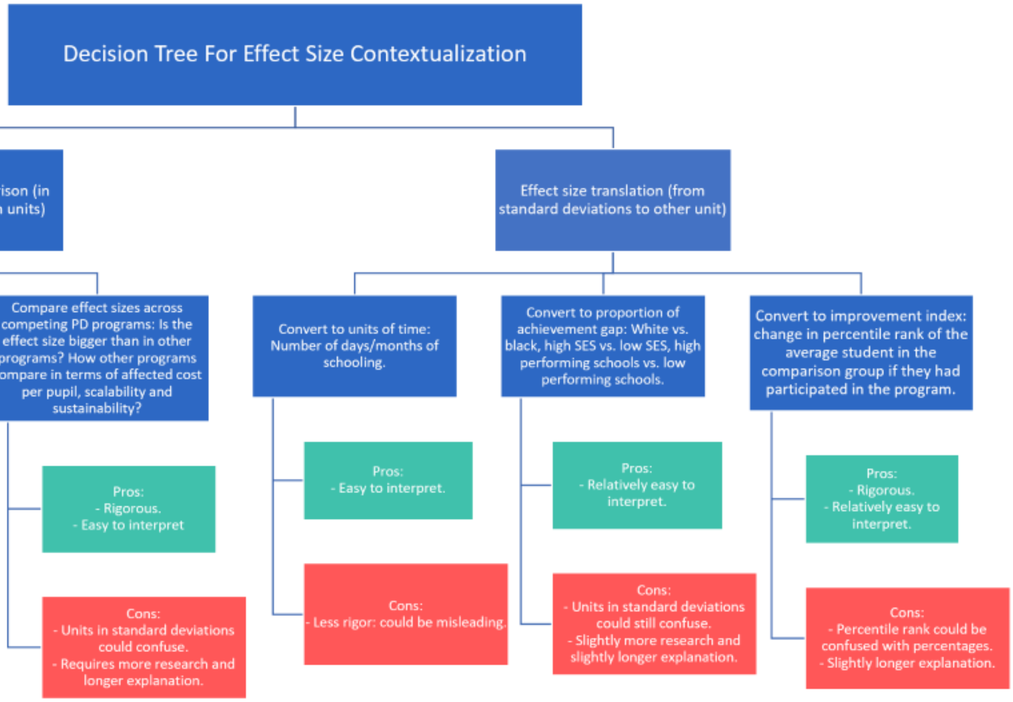
Limitations

We need to assume that there are no other differential time trends between participants and non-participants, because any difference in the changes between the two groups is attributed to the program.

Differences-in-Differences Estimator



Appendix 3: Decision Tree for Effect Size Contextualization



Decision Tree For Effect Size Contextualization

Effect sizes comparison (in standard deviation units)

Compare effect sizes across years and regions: Do they increase when implementation fidelity increase? Does it vary depending on the program archetype?

Pros:
- Rigorous.
- Easy to interpret.

Cons:
- Units in standard deviations could confuse.
- Requires more research and longer explanation.

Compare effect sizes across competing PD programs: Is the effect size bigger than in other programs? How other programs compare in terms of affected cost per pupil, scalability and sustainability?

Pros:
- Rigorous.
- Easy to interpret

Cons:
- Units in standard deviations could confuse.
- Requires more research and longer explanation.

Effect size translation (from standard deviations to other unit)

Convert to units of time: Number of days/months of schooling.

Pros:
- Easy to interpret.

Cons:
- Less rigor: could be misleading.

Convert to proportion of achievement gap: White vs. black, high SES vs. low SES, high performing schools vs. low performing schools.

Pros:
- Relatively easy to interpret.

Cons:
- Units in standard deviations could still confuse.
- Slightly more research and slightly longer explanation.

Convert to improvement index: change in percentile rank of the average student in the comparison group if they had participated in the program.

Pros:
- Rigorous.
- Relatively easy to interpret.

Cons:
- Percentile rank could be confused with percentages.
- Slightly longer explanation.

Leading Educators

Potential, ignited.