Wisconsin Evaluation Collaborative

Supporting PreK-12

July 2021

Evaluation of the Achievement Gap Reduction Program

2015-16 through 2019-20

for the Wisconsin Department of Public Instruction

Wisconsin University of Wisconsin–Madison

Wisconsin Center for Education Research
University of Wisconsin–Madison

About the Authors

Jed Richardson

Jed Richardson is an Assistant Scientist with the Wisconsin Evaluation Collaborative. He holds a Ph.D. in Economics from the University of California, Davis. Questions about this report can be directed to jed.richardson@wisc.edu.

Grant Sim

Grant Sim is a Researcher with the Wisconsin Evaluation Collaborative. He holds a Master's Degree in Public Affairs from the University of Wisconsin–Madison.

About the Wisconsin Evaluation Collaborative

The Wisconsin Evaluation Collaborative (WEC) is housed at the Wisconsin Center for Education Research at the University of Wisconsin-Madison. WEC's team of evaluators supports youth-serving organizations and initiatives through culturally responsive and rigorous program evaluation. Learn more at http://www.wec.wceruw.org.

Acknowledgements

The authors gratefully acknowledge Alison Bowman and Brie Chapa for their many contributions to this report's design and data visualization.

Copyright Information

Contents

Section 1

Executive Summary

Achievement gaps by socioeconomic status have been a persistent feature of the United States' education landscape for at least the past fifty years.[1] The Achievement Gap Reduction (AGR) program is an initiative of the Wisconsin Department of Public Instruction (DPI), as specified by 2015 Wisconsin Acts 53 and 71, that aims to improve the academic performance of students in schools with high concentrations of low-income students. AGR functions as a revision and continuation of the Student Achievement Guarantee in Education (SAGE) program. Similar to SAGE, AGR spans kindergarten to third grade and provides funds to participating Wisconsin schools based on their numbers of economically disadvantaged students. To receive AGR funding, schools must implement one or more strategies in each participating grade:

- Provide professional development related to small group instruction and reduce class size to one of the following:
  - No more than 18
  - No more than 30 in a combined classroom having at least 2 regular classroom teachers
- Provide data-driven instructional coaching for one or more teachers of one or more participating grades. The instruction shall be provided by licensed teachers who possess appropriate content knowledge to assist classroom teachers in improving instruction in math or reading and possess expertise in reducing the achievement gap.
- Provide data-informed, one-to-one tutoring to pupils in the class who are struggling with reading or mathematics or both subjects. Tutoring shall be provided during regular school hours by a licensed teacher using an instructional program to be found effective by the What Works Clearinghouse of the Institute of Education Sciences.[2]

This report presents the results of the annual AGR evaluation completed by the Wisconsin Evaluation Collaborative (WEC) within the Wisconsin Center for Education Research at the University of Wisconsin–Madison. The goal of this year's evaluation was to examine the following questions:

1. How are AGR schools implementing the AGR program as specified by 2015 Wisconsin Acts 53 and 71?
   a. What is the breakdown of strategy usage across the state?
   b. How does implementation of the three strategies differ across schools?

1 Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessman, L. (2020). Long-run trends in the U.S. SES-Achievement Gap. NBER Working Paper No. 26764. Retrieved May 14, 2021 from https://www.nber.org/papers/w26764.

2 2015 Wisconsin Act 53. Wisconsin Senate. Section 118.44.

2. To what extent is AGR meeting intended outcomes, including impacts on standardized test scores, attendance, and disciplinary events?
    a. How does AGR impact vary by student characteristics?
    b. How does AGR's impact on outcomes compare to impacts associated with the SAGE program?
    c. Are there differences between the three AGR strategies' relationships to intended outcomes?

Because AGR targets higher poverty schools where outcomes are typically lower and demographic profiles differ from Wisconsin averages, simple comparisons of outcomes between AGR schools and other, unfunded Wisconsin schools would produce biased results. To address this selection bias, WEC uses a two-part statistical method to better understand how AGR impacts student achievement, attendance, and discipline outcomes, and to compare AGR's impact to those of its predecessor, SAGE. The first part of the analysis uses propensity score matching to identify non-AGR Wisconsin schools that are similar to those receiving AGR funding. These observationally similar schools function as a comparison group for the second step of the analysis, estimating the impact of AGR through multivariate regression techniques.

Due to the COVID-19 pandemic, the 2020-21 evaluation differs slightly from previous evaluations. Data limitations resulting from the lack of PALS, MAP, and STAR testing in the spring of 2020 necessitated that we omit 2019-20 PALS from the analysis and adjust the statistical methodology to plausibly include 2019-20 MAP and STAR scores in the analysis. To provide the most complete evaluation of AGR possible, we present kindergarten PALS results through 2018-19 and Grades 1-3 MAP/STAR results through 2019-20. Both attendance and discipline data for 2019-20 were of sufficient quality to include in the analysis without adjustment.

How are AGR schools implementing the program?

In 2019-20, the most recent year of data, 412 schools implemented the AGR program, serving over 73,000 students in kindergarten through third grades. As previously noted, to fulfill AGR obligations schools could implement any combination of three strategies: reduced class size, instructional coaching, and/or tutoring.

- Over 70 percent of schools utilized multiple strategies – 39.7 percent of schools implemented reduced class size and instructional coaching together and 22.9 percent of schools implemented all three.
- Single strategies were employed less frequently, although 11.4 percent of schools implemented instructional coaching alone and 15.6 percent of schools implemented reduced class size alone.
- Comparatively few schools used tutoring as a strategy, either on its own or in combination with other strategies.

To what extent is AGR meeting intended outcomes?

The impact analysis examined how AGR students performed compared to non-AGR students in similar schools, while controlling for student characteristics. The impacts described in this report and in previous evaluations of the SAGE program are consistent with the school finance literature that finds mixed evidence of school funding impacts on test scores but substantial impacts on long-term student outcomes such as high school graduation. In this report, and in previous SAGE evaluations, test score impacts are large in kindergarten but otherwise indistinguishable from zero.[3] Evaluation of SAGE, with the benefit of 15 years of program data, found large impacts of K-3 SAGE on eventual high school persistence and completion. Results from this analysis included:

- A positive and significant impact of the AGR program on statewide reading growth in kindergarten through 2018-19, as measured by the PALS assessment. AGR is associated with a 0.11 standard deviation increase in PALS scores relative to similar, non-AGR schools. Increased growth on PALS is associated with a 0.05 standard deviation narrowing of the statewide kindergarten achievement gap in 2018-19.
- No estimated impact of the AGR program on statewide reading or math growth in Grades 1-3, as measured by the MAP and STAR assessments.
- No estimated impact of the AGR program on statewide attendance or out-of-school suspension rates.

The evaluation also examined the impact of the program by various subgroup populations and found:

- Large, positive, and significant impacts of the AGR program on kindergarten reading growth for low-income students.
- Large, positive, and significant impacts of AGR on kindergarten reading for English learners, Hispanic students, Asian students, and students in urban settings.
- Students in special education in AGR schools experienced less growth (0.06 standard deviations) than special education students in non-AGR schools.
- Positive and significant impacts of the AGR program on behavior, as measured through a reduction in suspensions, for English learners and Hispanic students.
- AGR is associated with a 0.7 – 1.4 day increase in absences per year for urban students, Black students, and students in grades two and three.

The evaluation also examined AGR program impacts compared to its predecessor, the SAGE program. Results showed that AGR likely has a larger, positive impact on kindergarten reading.

Are there differences in outcomes depending on the AGR strategies schools use?

The evaluation provided preliminary evidence of associations (not causal impacts) between outcomes and the AGR strategies schools choose. Results included:

- Increased reading growth, relative to class size reduction, associated with coaching in Grade 1 and increased math growth associated with tutoring, relative to class size reduction, in Grade 3.
- Reduced suspensions associated with reduced class size relative to other strategies.
- Analyses of coaching and tutoring frequency and intensity found few associated differences in outcomes.

3 Meyer, R. Dokumaci, E., Sim, G., Steele, C., Suchor, K., & Vadas, J. (2015). SAGE Program Evaluation Final Report. Value-Added Research Center. Retrieved May 15, 2021 from https://dpi.wi.gov/sites/default/files/imce/sage/pdf/sage_2015_evaluation.pdf.

Section 2

Introduction

The Achievement Gap Reduction (AGR) program is an initiative of the Wisconsin Department of Public Instruction (DPI) that provides funding to improve the academic performance of students in schools with high concentrations of low-income, or economically disadvantaged, students. AGR functions as a revision and continuation of the Student Achievement Guarantee in Education (SAGE) program, which the Wisconsin legislature and DPI initiated in 1995 to address the need for additional resources for economically disadvantaged students, particularly in urban areas. Beginning with the 1996-97 school year, the SAGE program administered state aid to schools that implemented reduced class sizes in kindergarten through third grade. A school typically qualified for the SAGE program if at least 30 percent of the student population was economically disadvantaged and its school district included one or more schools with at least 50 percent of the student population qualifying as economically disadvantaged.

In 2015, Wisconsin recognized the need to add flexibility to SAGE, reorganizing and renaming the program with the enactment of Wisconsin Acts 53 and 71. Wisconsin began a gradual phase-in of AGR in 2015-16 by transitioning schools from SAGE to AGR, with the final phase out of previous SAGE programs by the end of the 2017-18 school year. Like SAGE, AGR targets funding to schools with economically disadvantaged students through contracts to implement the program in kindergarten through third grade. Each year, the state provides approximately $110,000,000 distributed across participating schools. For 2019-20, AGR schools received approximately $2,621 for each economically disadvantaged student in grades K-3. In order to receive funding under AGR contracts, schools must implement at least one of three prescribed strategies in each participating grade. Each school, and each grade within a school, may implement different strategies. The three strategies include:

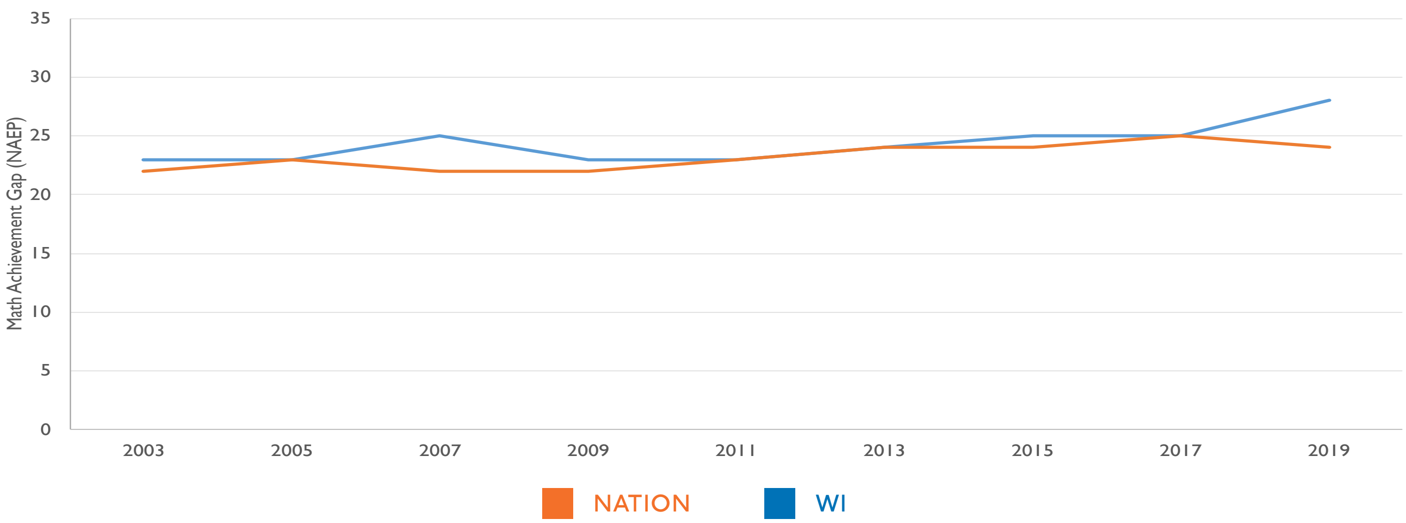
- Provide professional development related to small group instruction and reduce the class size to one of the following:
    - No more than 18.
    - No more than 30 in a combined classroom having at least 2 regular classroom teachers.
- Provide data-driven instructional coaching for one or more teachers of one or more participating grades. The instruction shall be provided by licensed teachers who possess appropriate content knowledge to assist classroom teachers in improving instruction in math or reading and possess expertise in reducing the achievement gap.
- Provide data-informed, one-to-one tutoring to pupils in the class who are struggling with reading or mathematics or both subjects. Tutoring shall be provided during regular school hours by a licensed teacher using an instructional program to be found effective by the What Works Clearinghouse of the Institute of Education Sciences.[4]

4 2015 Wisconsin Act 53. Wisconsin Senate. Section 118.44.

Context

The AGR program seeks to reduce the achievement gap for economically disadvantaged students. Going back to the influential 1966 Coleman Report, researchers have noted achievement gaps between students from different socioeconomic backgrounds.[5] Over the past fifty years, however, nationwide achievement gaps by socioeconomic status have been stagnant.[6] Wisconsin is no exception. As shown in Figures 1 and 2, during the 2000s neither Wisconsin nor the nation made any progress reducing gaps for economically disadvantaged 4th graders on NAEP math and reading, respectively. Researchers and policymakers have hypothesized dozens of causes for the socioeconomic achievement gap. These causes include neighborhood factors such as exposure to entrenched poverty and violent crime,[7] differences in summer opportunities,[8] differences in the amount of time parents are able to spend with their children,[9] and differences in neural development owing to exposure to high-poverty, and potentially traumatic, environments.[10]

Figure 1: Grade 4 Math Socioeconomic Achievement Gaps, 2003-2019



5 Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). Equality of Educational Opportunity. Washington, D.C: U.S. Government Printing Office.

6 Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessman, L. (2020). Long-run Trends in the U.S. SES-Achievement Gap. NBER Working Paper No. 26764. Retrieved May 14, 2021 from https://www.nber.org/papers/w26764.

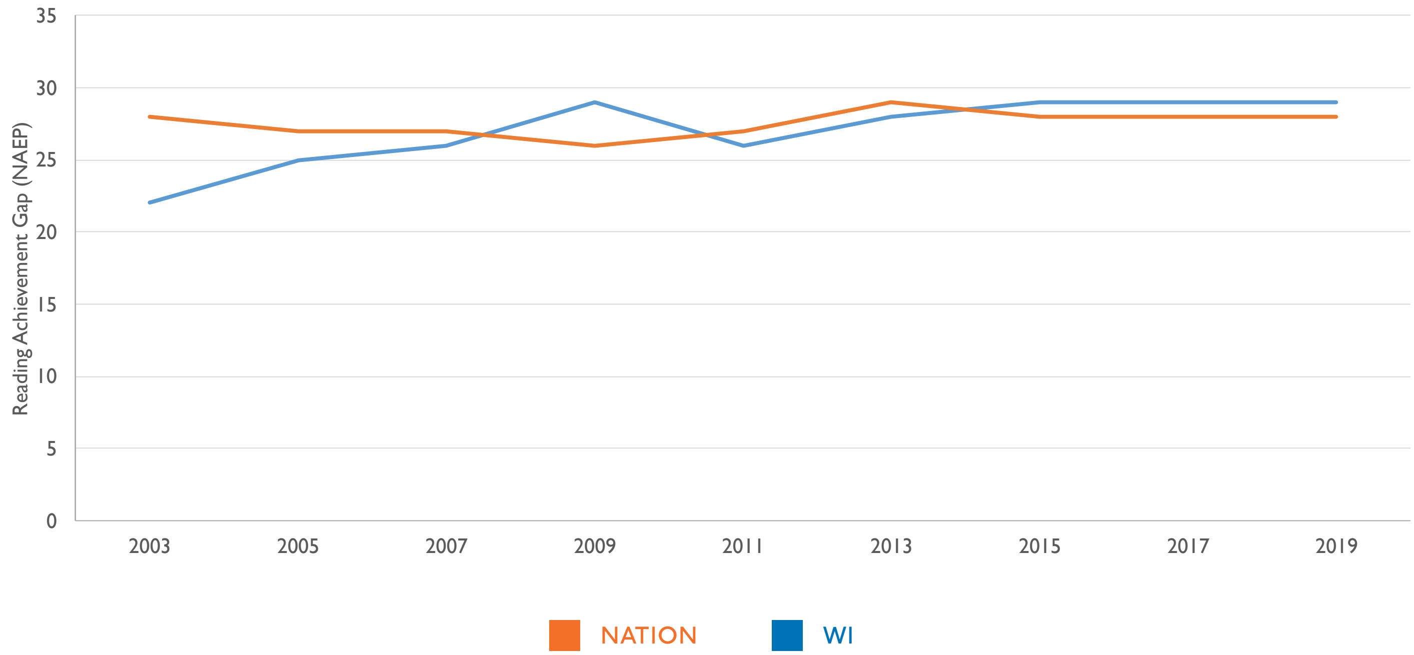
7 Burdick-Will, Julia, Jens Ludwig, Stephen W. Raudenbush, Robert J. Sampson, Lisa Sanbonmatsu, and Patrick Sharkey. 2011. "Converging Evidence for Neighborhood Effects on Children's Test Scores: An Experimental, Quasi-Experimental, and Observational Comparison." In Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances, edited by Greg J. Duncan and Richard J. Murnane. New York: Russell Sage Foundation: 255-276.

8 Leefat, S. (2015). The Key to Equality: Why We Must Prioritize Summer Learning to Narrow the Socioeconomic Achievement Gap. Brigham Young University Education and Law Journal, 2015(2), 549-84.

9 Guryan, Jonathan, Erik Hurst, and Melissa Kearney. 2008. "Symposia: Investment in Children: Parental Education and Parental Time with Children." Journal of Economic Perspectives 22, no. 3 (Summer): 23-46.

10 Nelson, Charles A., and Margaret A. Sheridan. (2011). "Lessons from Neuroscience Research for Understanding Causal Links Between Family and Neighborhood Characteristics and Educational Outcomes." In Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances, edited by Greg J. Duncan and Richard J. Murnane. New York: Russell Sage Foundation: 27-46.

Figure 2: Grade 4 Reading Socioeconomic Achievement Gaps, 2003-2019



AGR's strategy for closing the socioeconomic achievement gap is to provide additional funding to districts with large proportions of economically disadvantaged students. This strategy is consistent with the school finance literature on how school funding impacts student outcomes. In recent research that is especially pertinent to AGR, Jackson et al. (2021) use plausibly exogenous variation in school funding associated with the Great Recession to show that decreases in school resources led to increases in the socioeconomic achievement gap.[11] Jackson et al.'s results echo a growing list of studies that draw causal links between increased school funding and improved student outcomes.

A 2021 meta-analysis of credibly causal studies finds that a $1,000 per pupil increase in spending for four years improves test scores by 0.04 standard deviations, graduation rates by 2.1 percentage points, and the likelihood of college enrollment by 3.9 percentage points.[12] Furthermore, an individual study shows that for low-income students, increased spending increases educational attainment, increases adult wages, and lowers the incidence of poverty.[13]

11 Jackson, C. K., Wigger, C., & Xiong, H. (2021). Do school spending cuts matter? Evidence from the Great Recession. American Economic Journal: Economic Policy, 13(2), 304-35.

12 Jackson, C. K., & Mackevicius, C. (2021). The Distribution of School Spending Impacts. NBER Working Paper No. 28517. Retrieved May 15, 2021 from https://www.nber.org/papers/w28517.

13 Jackson, C. K., Johnson, R. C., & Persico, C. (2015). The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms. The Quarterly Journal of Economics, 131(1), 157-218.

This Evaluation

2015 Wisconsin Acts 53 and 71 include a provision for an annual evaluation of the AGR program starting in the 2018-19 school year. DPI contracted with the Wisconsin Evaluation Collaborative (WEC) within the Wisconsin Center for Education Research at the University of Wisconsin–Madison for these evaluation services. This report provides results from the AGR program evaluation from 2015-16 through 2019-20.

To serve as a foundation for the evaluation, WEC worked in collaboration with DPI to develop the following overarching evaluation questions:

1. How are AGR schools implementing the AGR program as specified by 2015 Wisconsin Acts 53 and 71?
    a. What is the breakdown of strategy usage across the state?
    b. How does implementation of the three strategies differ across schools?
2. To what extent is AGR meeting intended outcomes, including impacts on standardized test scores, attendance, and disciplinary events?
    a. How does AGR impact vary by student characteristics?
    b. How does AGR's impact on outcomes compare to impacts associated with the SAGE program?
    c. Are there differences between the three AGR strategies' impacts on intended outcomes?

Due to the COVID-19 pandemic and associated school closures during Spring 2020, administrative data for 2019-20 does not include spring test scores or, in some cases, complete attendance information. As a result, the evaluation methodology and contents of this report differ slightly from previous years.

This report has eight main sections including the introduction. The evaluation data and methodology section includes details on data, analysis designs, and statistical models used to evaluate program impacts, as well as the limitations of this evaluation. The AGR demographics section contains information on the characteristics of AGR students and schools relative to Wisconsin overall to provide context for later findings. This section also contains testing patterns and growth analysis samples, which describe coverage of common assessments in Grades K-3 and the samples chosen for estimating AGR impacts on growth (note that, due to a lack of 2019-20 posttest information, we report kindergarten reading impacts from the 2018-19 evaluation). AGR Implementation describes the distribution of AGR strategies that schools chose for 2019-20. The AGR impacts section provides the results of analyses of AGR impact on math growth and reading growth. This section is further divided to provide overall impacts, impacts of AGR compared to SAGE, impacts by student subgroups, and differences in outcomes by AGR strategy. The section on school board report findings includes results from an examination of 2019-20 reports by AGR districts and schools. The End-of-Year Report findings provide the results from the 2019-20 survey of AGR schools. The final section of the report includes a summary of findings and thoughts on future evaluations. This report also contains two appendices: a technical appendix that provides further details on statistical methodology, and an appendix including the instrument for the 2019-20 End-of-Year Report survey.

Section 3

Evaluation Data and Methodology

In order to understand how AGR impacts student achievement outcomes, and to compare AGR's impacts to those of its predecessor, SAGE, we must identify a plausible comparison group of schools and students. Because AGR targets higher poverty schools where outcomes are lower on average, naïve comparisons of AGR schools' outcomes to those of other Wisconsin schools would produce biased, negative program impacts. To address this selection bias, the evaluation uses Propensity Score Matching (PSM) to identify non-AGR-funded Wisconsin schools that are similar to those receiving AGR funding. These observationally similar schools act as a comparison group for analyses of AGR impacts.

The analysis includes students in Grades K-3 at all schools that received SAGE and AGR funding during the 2012-13 through 2019-20 academic years. In addition, for purposes of comparison, the evaluation includes K-3 students at subsets of non-AGR, non-SAGE schools.

Data

To identify plausibly equivalent, non-AGR schools for a comparison group and to estimate impacts, the evaluation combines several sources of student- and school-level data for the academic years 2012-13 through 2019-20. Student-level achievement test data, student demographics, and enrollment records came from DPI administrative data. DPI also provided school-level data on AGR and SAGE funding by year. School-level teacher average salaries were sourced from DPI Public Staff Reports, and school location information came from school report card files.[14]

- *Demographic characteristics* include gender, race/ethnicity, English learner status, special education status, and low-income or economically disadvantaged status as measured by free or reduced-price lunch eligibility. School- and grade-level measures of demographic characteristics were calculated from student-level data.

- *Achievement test data* include fall, winter, and spring administrations of the MAP and STAR. For Grades 1-3, MAP and STAR scores were equated and combined into a single test measure in order to attain a sufficient student sample. Previous AGR evaluations have used the Phonological Awareness Literacy Screening (PALS) for Kindergarten reading, but no PALS spring data exist for the 2019-20 academic year.

- *Enrollment data* include school attended and grade.

- *School-level data* include SAGE and AGR funding by year, teacher average salaries, and school location (city, suburb, town, rural).

14 Public Staff Reports are available at https://publicstaffreports.dpi.wi.gov/PubStaffReport/Public/PublicReport. School report cards can be found at https://apps2.dpi.wi.gov/reportcards/

Predicting Spring 2020 Assessment Scores

In March of 2020, the COVID-19 pandemic rapidly and dramatically changed K-12 education. With little time to form contingency plans, nearly all Wisconsin schools opted to forgo Spring 2020 assessments, including both required statewide tests, which the US Department of Education waived, and districtwide exams, such as PALS, MAP, and STAR, that schools use for tracking student progress. Cancellation of Spring 2020 assessments complicates this evaluation but does not lessen the value of understanding how AGR impacts student learning.

To address the lack of Spring 2020 assessment scores, we estimate a predictive model that uses student math and reading scores from Fall 2019 and Winter 2020, student demographics, and school characteristics, to predict what Spring 2020 test scores would have been had the 2020 school year proceeded normally. We use both Fall 2019 and Winter 2020 assessments to predict Spring 2020 scores to capture actual test score growth that had occurred before the COVID-19 school closures and use that actual growth to predict the growth trajectory that is most likely to have occurred from Winter to Spring 2020. The primary advantage of this strategy is that, with Spring 2020 scores, the past evaluation methodology can be used to fit 2019-20 into the overall evaluation framework both now and in the future. One drawback is that, because schools often do not administer winter PALS, the assessment used to evaluate Kindergarten reading, the prediction methodology is only possible for MAP and STAR exams in grades 1-3.

To test the validity of using predicted Spring 2020 test scores in the impact evaluation, we replaced 2018-19 actual test scores with predicted scores and re-estimated analyses from the 2018-19 evaluation. We then compared both 2018-19 actual test scores and resulting impact analyses with the newly-calculated, predicted 2018-19 test scores and impact analyses. Both the test scores and estimated impacts matched well. Although we cannot know how predicted Spring 2020 test scores match with actual scores, results from this year's evaluation, using actual test scores for 2012-13 through 2018-19, actual test scores for Fall 2019 and Winter 2020, and predicted test scores from Spring 2020, compare favorably to past results.

Identifying Comparison Schools

Using the data described above, we aggregated each school's K-3 data to find a comparison group of non- AGR schools. Matching followed two separate strategies. For attendance and discipline outcomes, we matched schools based on 2012-13 data. For math and reading testing outcomes, however, wide variation in schools' testing coverage both across time and across grades prevented matching at the school level (see Table 5 - Table 8). Instead, we chose to match at the school-grade-year level using each school's fall data.

During the 2019 evaluation, we tested multiple variations of PSM in order to (1) achieve the best match between AGR and comparison schools, and (2) retain as many AGR observations as possible. To do so, we tested combinations of demographic and academic variables and several matching algorithms. This testing process resulted in a kernel matching procedure, which we continue to use in this report. Kernels place higher weights on untreated observations nearest to a treatment observation and assign successively lower weights to untreated observations as their distance from a treatment observation increases. Table 1 lists the covariates in the matching model that provide the best balance and sample retention.

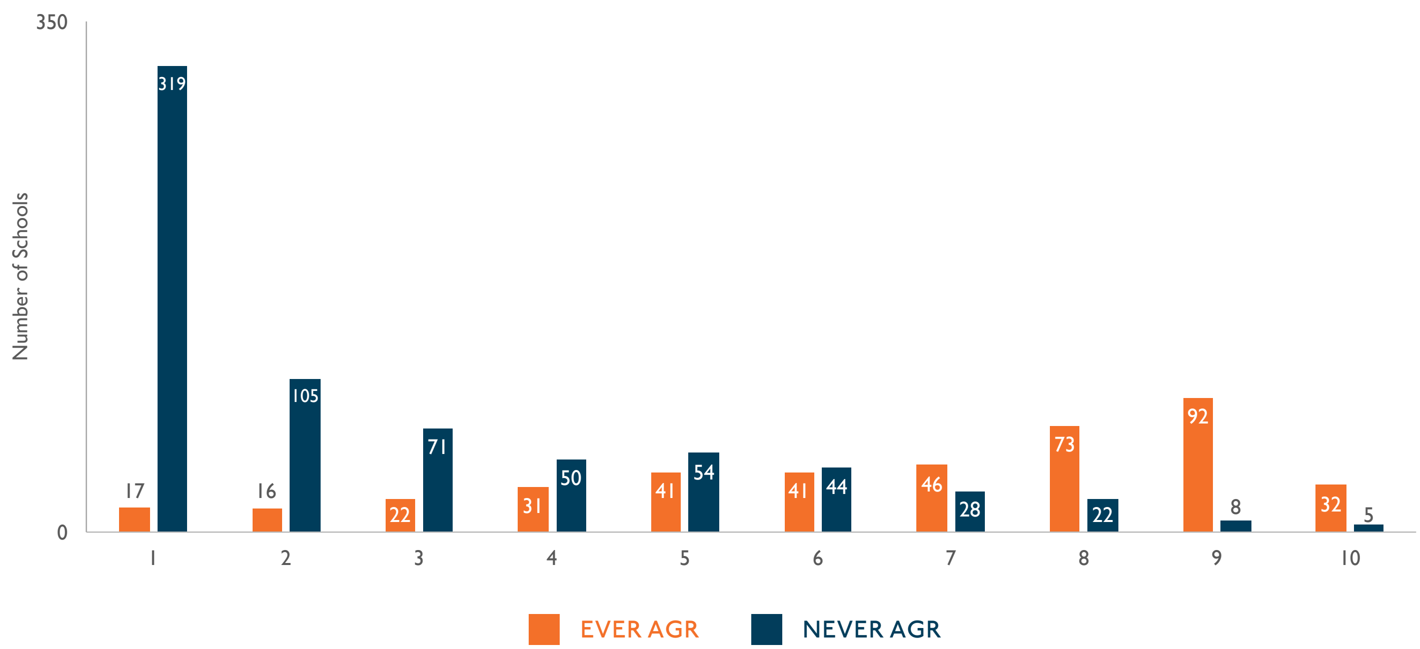
Table 1: Propensity Score Matching Controls by Analysis Type

CONTROL VARIABLE	GROWTH ANALYSIS	ATTENDANCE/ DISCIPLINE ANALYSIS
Student Population	✓	✓
% Black, Hispanic, White, Other Race/Ethnicity*	✓	✓
% Free/Reduced-price Lunch	✓	✓
% English Learner	✓	✓
% Special Education	✓	✓
Average Teacher Salary	✓	✓
Local Description (City, Suburb, Town, Rural)*	✓	
Average Standardized Fall Math Score**	✓	
Average Standardized Fall Reading Score	✓	
Grade Indicators***		✓
Attendance Rate in 2012-13		✓
Suspension Rate in 2012-13		✓

Note: * Due to collinearity, we omitted one Race/Ethnicity category and one Local Description category from the model ** For PALS< only the PALS reading pretest is included, due to low participation in the MAP/STAR exam in kindergarten. *** Indicators equal one if schools include that grade.

When matching is successful, there should be sufficient overlap in propensity scores of treated (AGR) and untreated (non-AGR) schools to ensure that there is a plausible comparison group for analysis. Figure 3 below shows the overlap between AGR and non-AGR schools for MAP/STAR math in 2019-20. In each decile of the propensity score distribution, there are at least five comparison (untreated) schools. Most deciles have more than 10 comparison schools, showing sufficient overlap for the analysis. Overlap is similar across all models.

Figure 3: Common Support for Matching - MAP/STAR Math (2019-20)



Analysis

After matching and computing predicted 2020 MAP and STAR scores, we estimated AGR impacts via multivariate regression models. These models include all school-level matching covariates listed in Table 1 above, as well as student-level demographic variables, student-level pretest scores, and grade-by-year fixed effects. A full listing of analysis variables can be found in Table 2 below. All models include weights generated by the kernel PSM procedure.

Table 2: Analysis Model Controls

CONTROL VARIABLE	GROWTH	ATTENDANCE	DISCIPLINE
Student Demographics Gender, Race/Ethnicity*, Free/Reduced-price Lunch, English Learner, Special Education	✓	✓	✓
School Demographic Percentages Gender, Black, Hispanic, White, Other Race/Ethnicity*, Free/Reduced-price Lunch, English Learner, Special Education	✓	✓	✓
School Population	✓	✓	✓
Local Description (City, Suburb, Town, Rural)*	✓	✓	✓
Student Fall Test Scores**	✓		
School Average Fall Test Scores	✓		
School Attendance Rate in 2012-13		✓	
School Suspension Rate in 2012-13			✓

Note: * Due to collinearity, we omitted one Race/Ethnicity category and one Local Description category from the model. ** For math and reading models, both subject pretests are included.

Limitations

The methodology outlined above provides the most rigorous possible evaluation given the rollout of AGR and available data. There are several limitations, however, that could impact this report's results and conclusions.

The primary limitation stems from PSM's primary assumption that schools matched on observable characteristics such as test scores and demographics are also matched on unobserved characteristics, such as schools' ability to properly implement AGR strategies or instructor quality in the local hiring market. If unobserved characteristics are not balanced between AGR and comparison schools and are related to both outcomes and AGR participation, estimates of AGR impacts will be biased.

The second limitation occurs because all AGR schools previously participated in SAGE, which had been in operation for over 15 years at the beginning of this study's sample period. As a consequence, AGR schools are matched to non-AGR schools based on post-SAGE outcomes. Matching schools on post-program data risks biasing the results toward zero (toward estimating smaller impacts), because schools would be matched on previous-period outcomes that already include the treatment impact (in this case, the SAGE program is similar enough to AGR to raise similar concerns). Omitting these outcomes from the matching model, however, resulted in poor matches and would have caused significant bias.

The third limitation stems from predicting MAP and STAR scores for Grades 1-3 for Spring 2020. The test score predictions assume that test score growth that occurred between Fall 2019 and Winter 2020 would have continued through the spring, had the COVID-19 lockdowns not occurred. While actual growth patterns of students likely changed between Winter and Spring 2020 due to COVID-19, the results in this analysis examine the impact of AGR assuming growth in a typical year.

To the extent that the limitations described above bias impact estimates, the results should not be considered causal. In particular, if AGR schools are systematically more (less) effective than schools in the matched comparison group, impact estimates will be biased upward (downward).

The final limitation occurs due to inconsistent testing patterns (described in detail in Testing Patterns and Growth Analysis Samples below). In general, during the sample period Wisconsin did not require schools to use specific assessments in Grades K-3, which creates difficulties for identifying a consistent, sufficiently sized sample for estimating growth impacts. Although in previous years the tested population of AGR schools used for the evaluation has been observationally similar to the untested sample of AGR schools, in 2019-20 differences between tested and untested AGR populations grew. However, available data cannot support analysis of whether schools' choices of tests are related to outcomes and participation in AGR.

Section 4

AGR Demographics

Table 3: Number of AGR Schools by Grade and Year

GRADE	2015-16	2016-17	2017-18	2018-19	2019-20
Kindergarten	88	393	392	394	396
First	91	398	398	402	400
Second	91	398	399	402	400
Third	88	391	392	394	394
Any (K-3)	96	408	409	412	412

Table 4: Number of AGR Students by Grade and Year

GRADE	2015-16	2016-17	2017-18	2018-19	2019-20
Kindergarten	4,139	18,385	18,292	18,341	18,429
First	4,571	19,290	18,869	18,741	18,533
Second	4,682	20,056	19,219	18,911	18,592
Third	4,544	19,509	19,278	18,339	18,092
Any (K-3)	17,936	77,240	75,658	74,332	73,646

This section and the sections that follow present evaluation results aligned to the evaluation questions listed above. We begin with information on the characteristics of AGR students and schools. Table 3 shows the number of AGR schools for each of the first five years of the program. The first AGR cohort started in 2015-16 with 96 schools, followed by the second cohort in 2016-17, which brought the total to 408 schools. The final cohorts added a small number of schools in 2017-18 and 2018-19. In 2019-20, the overall number of AGR schools remained at its 2018-19 level.

The numbers of students in AGR schools from 2015-16 to 2019-20, overall and by grade, are presented in Table 4. The first cohort of AGR schools included approximately 18,000 students, while the addition of the second cohort in 2016-17 brought the total to over 77,000 students. Since then, student participation has declined to its present level of 73,646.

Figure 4: Race/Ethnicity of AGR and WI Students, 2019-20

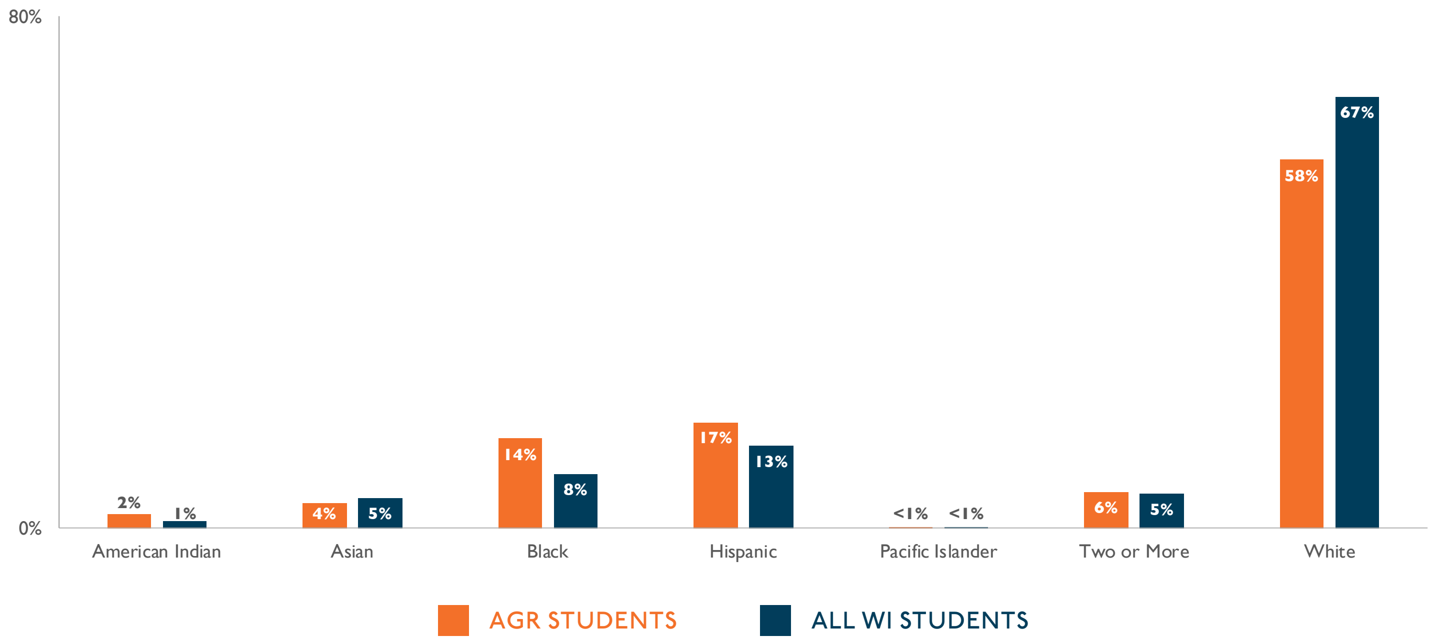


Figure 5: Percentage of AGR and WI Students That Were English Learners, Eligible for Free/Reduced-price Lunch, and in Special Education, 2019-20

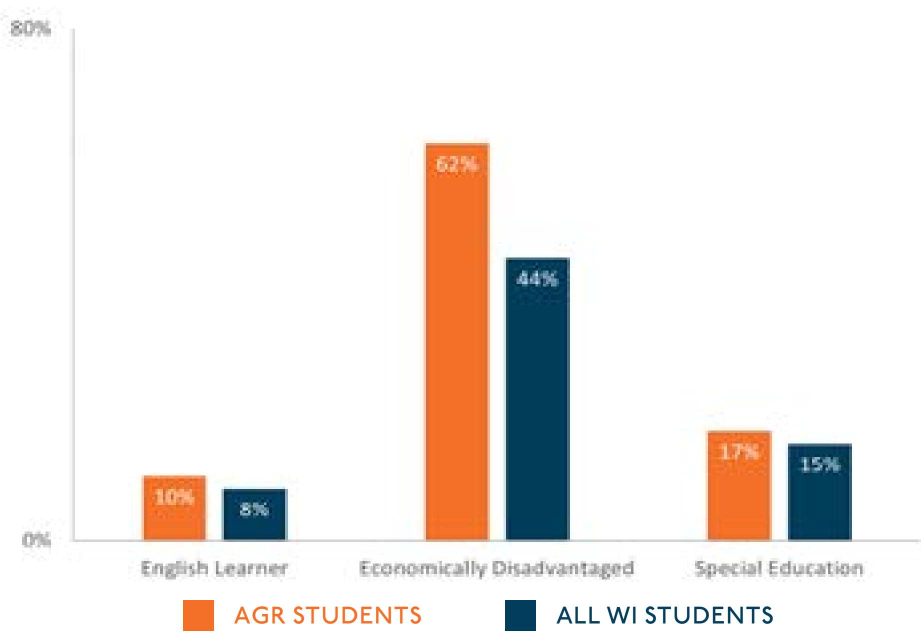
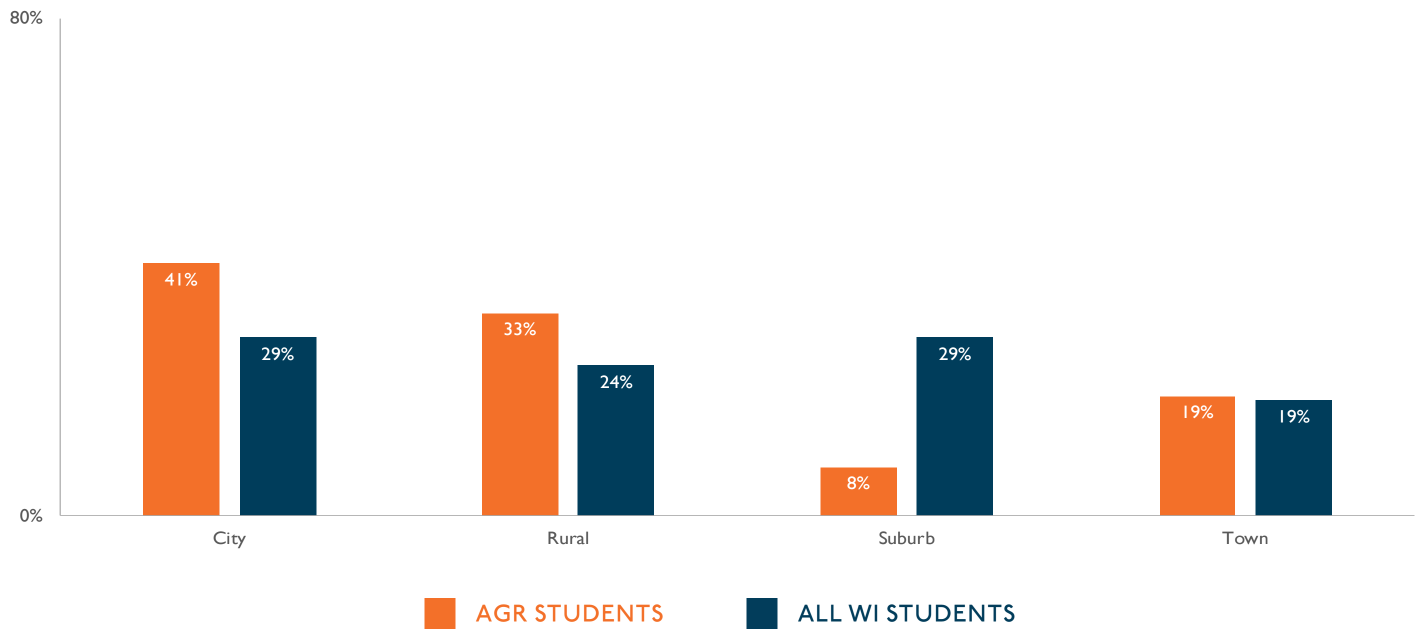


Figure 4 and Figure 5 compare the demographic characteristics of AGR students to all K-3 Wisconsin students (including AGR students) in 2019-20. Relative to Wisconsin as a whole, a higher proportion of AGR students were Black, Hispanic, English learners, and eligible for free/reduced-price lunch, and a lower proportion of students in AGR schools were White.

Figure 6: Locale Description of AGR and WI Students, 2019-20



AGR schools were more likely to be located in urban or rural settings and less likely to be in suburban areas, as shown in Figure 6. This corresponds to the higher proportion of students eligible for free/reduced-price lunch, seen previously, as city and rural areas of the state have larger populations of people living in poverty.[15]

15 See https://nces.ed.gov/programs/maped/ACSMaps/

Testing Patterns and Growth Analysis Samples

Throughout the sample period, shifting testing patterns in Grades K-3 complicated efforts to estimate AGR's impacts on test score growth. Under Wisconsin's current testing policy, the first common, state-mandated accountability test occurs during the spring of third grade. Although schools test students throughout Grades K-3, schools and districts are allowed to choose their own assessments. This policy results in substantial variation

in testing patterns both across and within schools. In addition to variation between schools regarding the assessments they select, many schools began administering a new test and/or quit using a test in the middle of the sample period. Other schools tested some of Grades K-3 but not others, and yet others changed which grades they tested during the sample period. As a result, less than half of the overall population of AGR schools and students are appropriate for use in growth analysis. Given pre-COVID-19 testing patterns, we used two strategies to build sufficient samples. First, we split the growth analysis sample into two samples: Grade K and Grades 1-3. In 2018-19, less than half of all Wisconsin kindergarteners took the PALS, which had been a state-mandated reading assessment for the grade from 2012-13 through 2015-16. For the current evaluation, the lack of Spring 2020 PALS prevented evaluators from having adequate data for growth modeling. Similar to previous years, no combination of assessments provided adequate coverage for kindergarten math.[16]

16 See the Technical Appendix, Figure A5, Table A35, and Table A36, for results from an impact analysis of reading growth on kindergarten MAP and STAR. Results from this impact analysis are strikingly similar to past results from kindergarten PALS impact analysis, providing further support for robust AGR impacts in kindergarten. Note that kindergarten MAP/STAR coverage peaks around 20 percent of Wisconsin and AGR schools. As such, results should be interpreted with caution.

Table 5: Percentage of Wisconsin Schools Using MAP or STAR Math Tests

GRADE	2015-16	2016-17	2017-18	2018-19	2019-20
Kindergarten	13%	11%	9%	8%	7%
First	32%	32%	34%	35%	33%
Second	44%	43%	43%	45%	40%
Third	53%	53%	52%	50%	44%

Note: While 2015-16 through 2018-19 percentages reflect schools that used MAP or STAR in Fall and Spring, 2019-20 percentages reflect schools that used MAP or STAR in Fall and Winter.

Table 6: Percentage of Wisconsin Schools Using MAP or STAR Reading Tests

GRADE	2015-16	2016-17	2017-18	2018-19	2019-20
Kindergarten	9%	7%	16%	21%	21%
First	18%	16%	24%	26%	26%
Second	42%	42%	41%	43%	39%
Third	53%	52%	50%	50%	44%

Note: While 2015-16 through 2018-19 percentages reflect schools that used MAP or STAR in Fall and Spring, 2019-20 percentages reflect schools that used MAP or STAR in Fall and Winter.

The second strategy we used to build sufficient growth analysis samples was to use both the MAP and STAR assessments for Grades 1-3 math and reading.[17] As shown in Table 5 and Table 6, in 2019-20 between 26-44 percent of Wisconsin schools used either the MAP or STAR in first, second, or third grade, with usage rates above 40 percent for second and third graders in both subjects, although these percentages are lower than in previous years. To combine MAP and STAR into a single measure, we equated assessment scores using national norms.[18]

17 In addition to STAR Reading, the analysis also uses STAR Early Literacy results for Grade 1.

18 See Thum, Y. M. & Hauser, C.H. (2015). NWEA 2015 MAP norms for student and school achievement status and growth. NWEA Research Report. Portland, OR: NWEA.

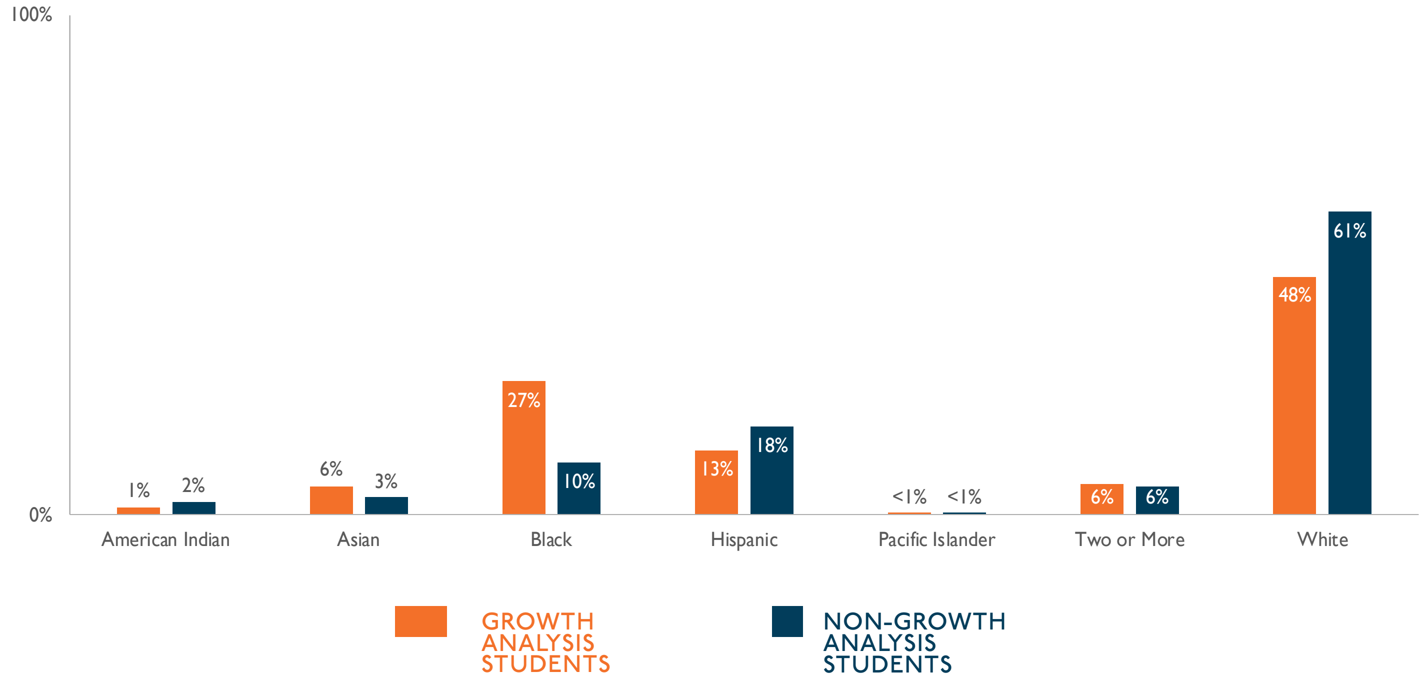
Table 7: Number of Growth Analysis AGR Schools and Percentage of All AGR Schools by Grade, 2019-20

	AGR SCHOOLS		AGR STUDENTS	
	N	%	N	%
Kindergarten	N/A	N/A	N/A	N/A
First	100	25.0	4,016	21.7
Second	144	36.0	5,757	31.0
Third	161	40.9	6,728	37.2
Overall (K-3)	173*	42.0	16,501*	29.9

*For 2019-20, overall (K-3) sample sizes decrease, relative to other years, in large part due to the lack of kindergarten PALS testing in Spring 2020.

Table 7 shows the number of schools overall and by grade that had testing information and were used in the analyses of academic growth. As a reference, the table also shows the percentage of all AGR schools in the tested population. This table displays similar information but for students instead of schools. As seen in Table 7, testing patterns restrict the first grade sample most. In first grade, the growth analysis only includes less than one-quarter of the entire sample of students in 2019-20. This restriction lessens as grade level increases. Technical Appendix Tables A2 and A3 display the sample sizes of AGR schools and students, respectively, for all sample years.

Figure 7: Race/Ethnicity of AGR Growth Analysis and Non-Growth Analysis Students, 2019-20



Because the growth analysis sample of students is smaller than the entire population, as mentioned in the limitations section, the growth analysis results may not apply to all AGR students. To observe whether the tested and untested samples differ, Figure 7 - Figure 9 compare demographic characteristics between AGR students used in the growth analysis and AGR students not used in the growth analysis due to lack of assessment information. The growth analysis AGR sample has higher proportions of Black students and those eligible for free/reduced-price lunch and lower proportions of White students. Although those differences were small in other years, in 2019-20 difference between the tested and untested populations grew. The schools included in the growth analysis are more likely urban and less likely rural.

Figure 8: Percentage of AGR Growth Analysis and Non-Growth Analysis Students that were English Learners, Free/Reduced-price Lunch Eligible, and in Special Education, 2019-20

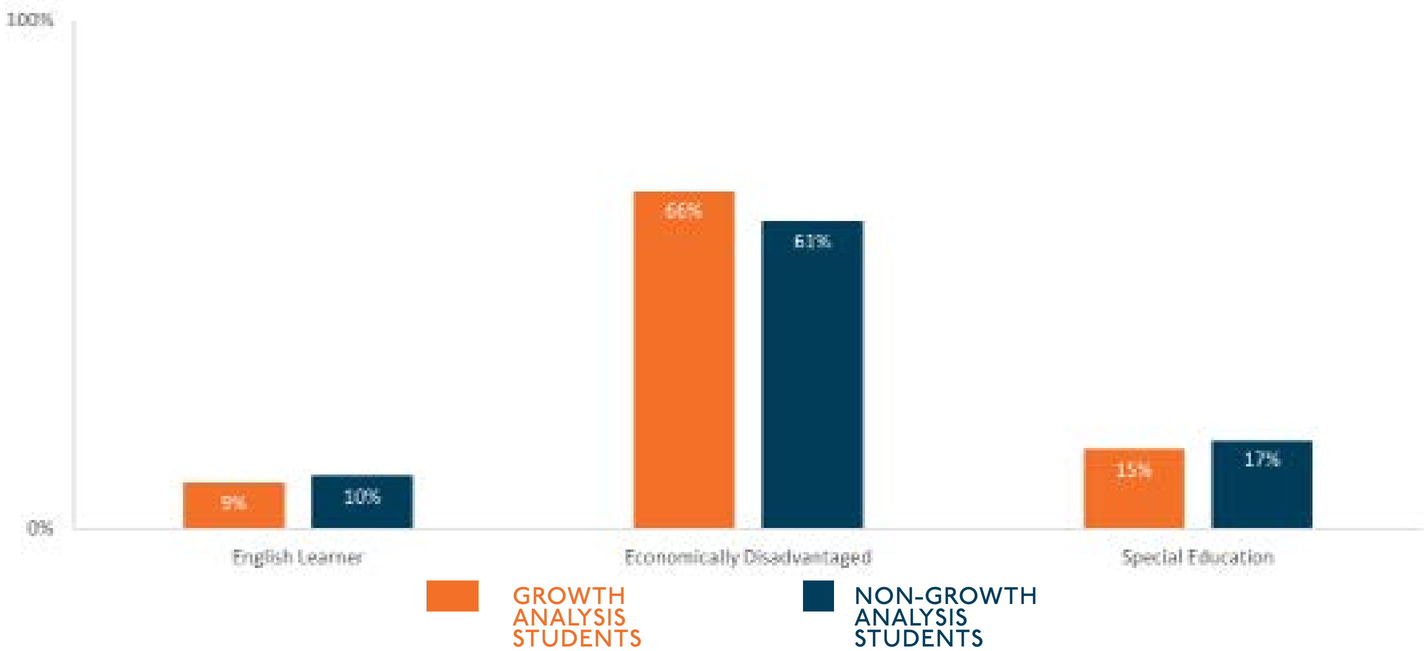
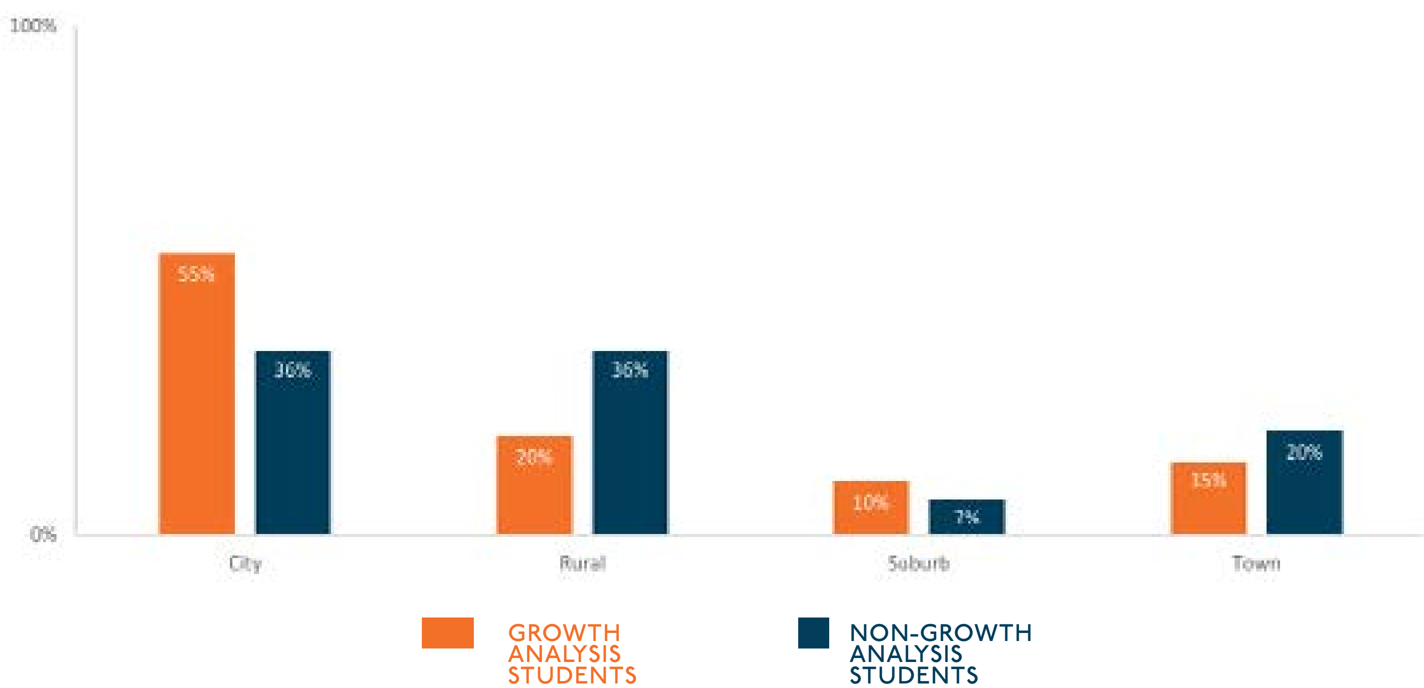


Figure 9: Locale Description of AGR Growth Analysis and Non-Growth Analysis Students, 2019-20



Section 5

AGR Implementation

This section of the report examines the usage of the three possible AGR strategies that schools could use as part of AGR. As noted previously, the three strategies include:

- Provide professional development related to small group instruction and reduce the class size to one of the following:
  - No more than 18.
  - No more than 30 in a combined classroom having at least 2 regular classroom teachers.
- Provide data-driven instructional coaching for the class teachers.
- Provide data-informed, one-to-one tutoring to pupils in the class who are struggling with reading or mathematics or both subjects.

As the AGR program allows schools to use more than one strategy within a school, there are seven possible combinations schools could implement: class size reduction only, coaching only, tutoring only, class size reduction and coaching, class size reduction and tutoring, coaching and tutoring, and all three strategies. Figure 10 provides information on the strategy combinations AGR schools implemented during 2019-20. This figure also provides information on the number and percentage of students affected by each strategy combination. The most frequently used strategies were class size reduction and coaching, all three strategies, class size reduction only, and coaching only. Very few schools used only tutoring as a strategy.

Figure 10: Implementation of AGR Strategies

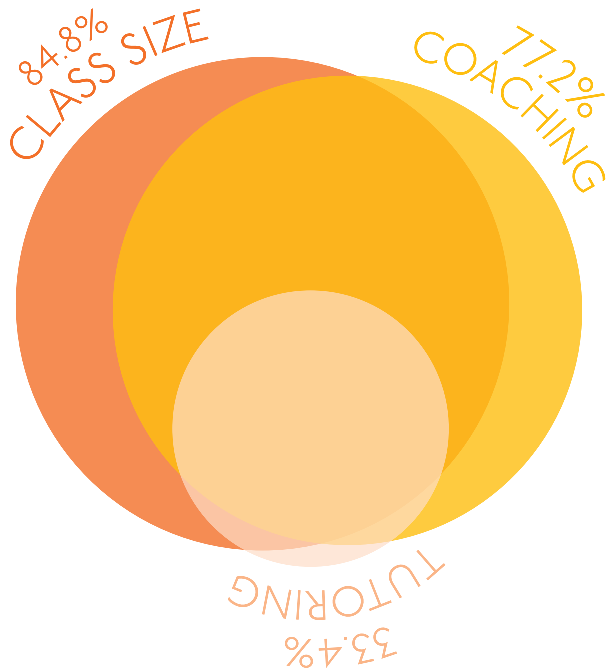
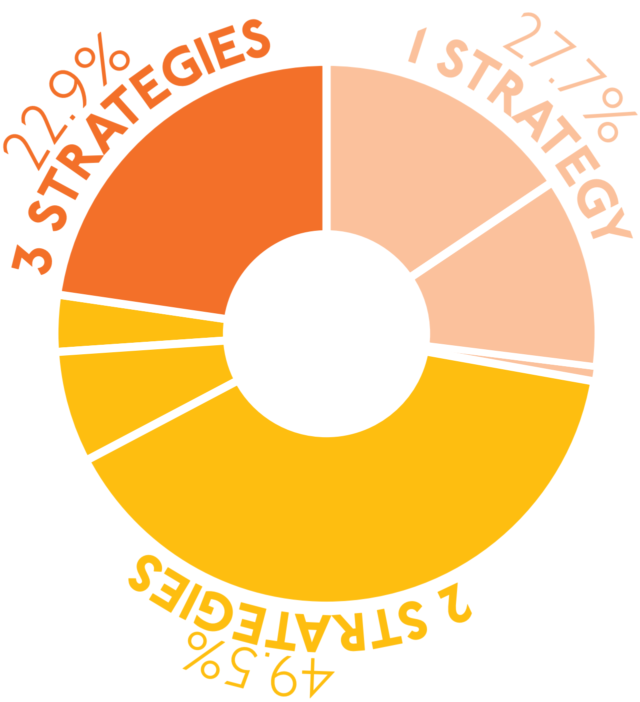


Figure 11: Schools' AGR Strategy Choices



1 STRATEGY

CLASS SIZE
64 SCHOOLS | 15.6%

COACHING
47 SCHOOLS | 11.4%

TUTORING
3 SCHOOLS | 0.7%

2 STRATEGIES

CLASS SIZE + COACHING
163 SCHOOLS | 39.7%

CLASS SIZE + TUTORING
27 SCHOOLS | 6.6%

COACHING + TUTORING
13 SCHOOLS | 3.2%

3 STRATEGIES

CLASS SIZE + TUTORING +COACHING
94 SCHOOLS | 22.9%

Section 6

AGR Impacts

This section of the report examines the results from statistical analyses undertaken to determine the impact of the AGR program. The intent is to answer the second and third guiding evaluation questions:

2. To what extent is AGR meeting intended outcomes, including impacts on standardized test scores, attendance, and disciplinary events?
    d. How does AGR impact vary by student characteristics?
    e. How does AGR's impact on outcomes compare to impacts associated with the SAGE program?
3. Are there differences between the three AGR strategies' impacts on intended outcomes?

To answer these questions, we begin by providing results on the program's overall and by-grade impacts. We then examine the impact of AGR compared to previous SAGE implementation, followed by impacts of AGR by student subgroup populations. Finally, we provide the results of analyses for different AGR strategy combinations.

All AGR impact analyses examine how students performed on outcome measures including math growth and reading growth. For each of these outcomes, this report provides a figure of results at each applicable grade level and overall (across all applicable grades). These figures show a measure or measures of the program impact and whether the impact is statistically significant.[19] More detailed results, including p-values, can be found in the Technical Appendix.

Overall Impacts

The impact analysis examines how AGR students performed compared to non-AGR students in similar schools, while controlling for student characteristics. Impacts are shown for each grade and for all grades combined.

19 Statistical significance is determined by a p-value that indicates the likelihood of observing the reported impact or a more extreme impact assuming that there is no actual impact of the program. Larger p-values indicate weaker evidence of an impact, while smaller p-values indicate stronger evidence of an impact. Throughout the report, the evaluation uses a threshold of 0.05 to determine if a result was statistically significant from zero. All p-values presented in this report are corrected to account for multiple estimates (see the Technical Appendix for details).

Figure 12: Impact of AGR on Math Growth

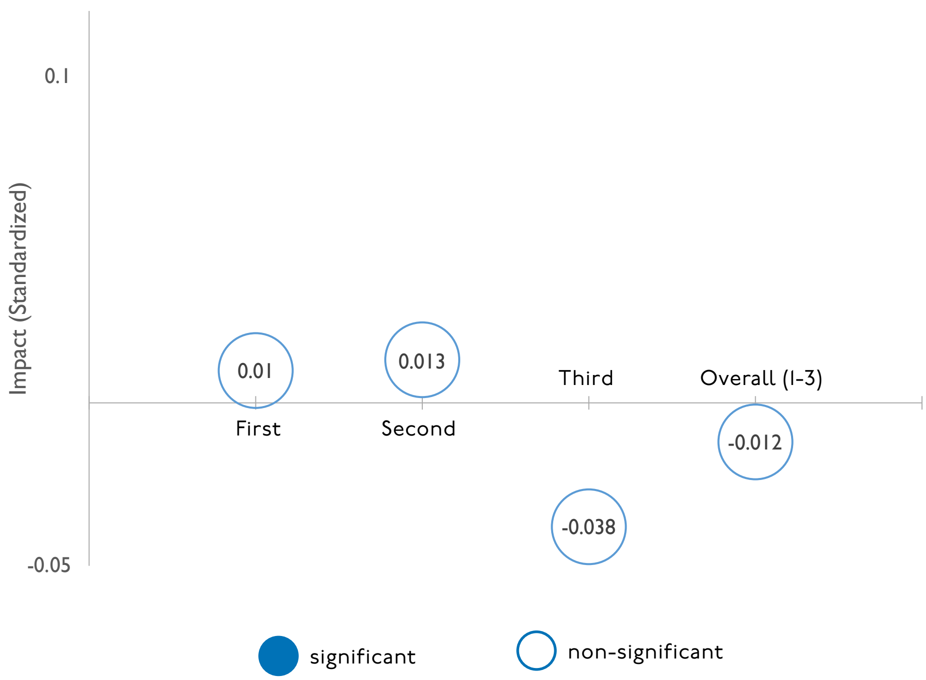


Figure 12 presents the impacts of AGR on math growth on a standardized scale representing the number of standard deviations from zero.[20]

Impacts show the difference between average AGR student growth and non-AGR student growth for students in similar schools. Results across all grades reveal little difference in math growth between AGR students and non-AGR students.

Figure 13: Impact of AGR on Reading Growth

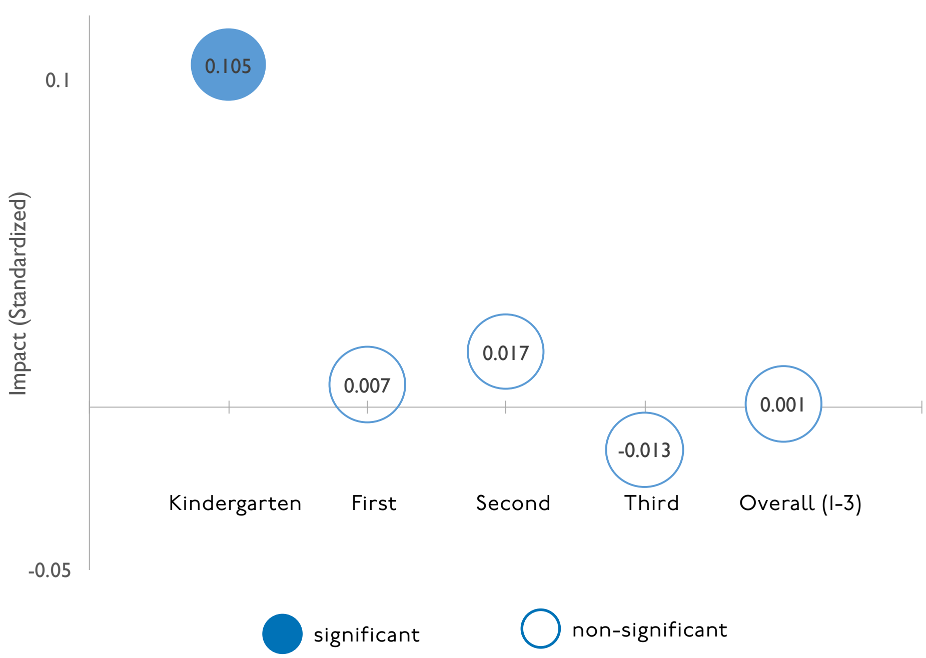


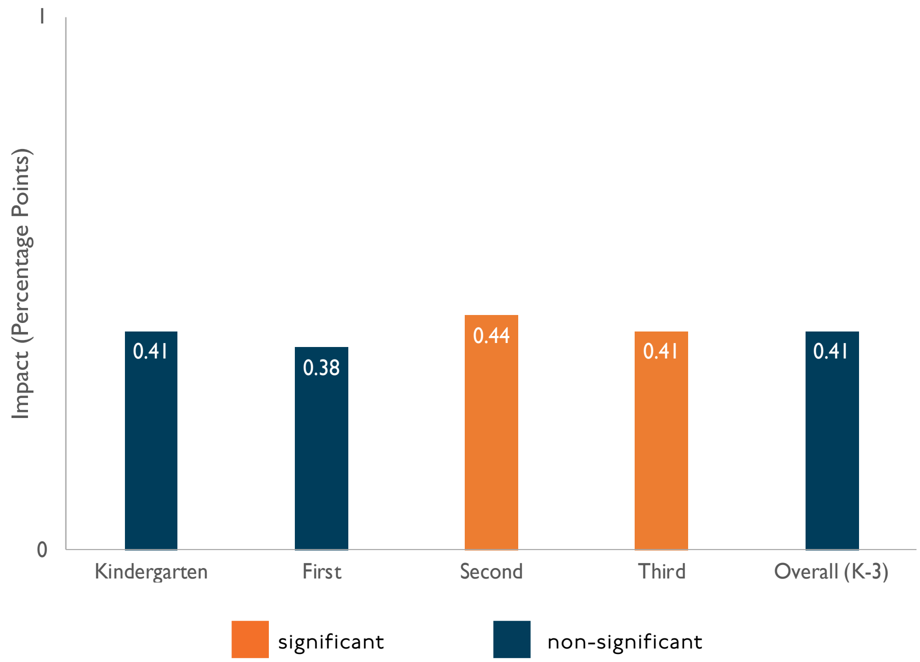
Figure 13 shows impacts of AGR on reading growth. As with math growth, this figure shows average differences in growth on a standardized scale.

Results from MAP/STAR reading assessments indicate little difference in average growth between AGR and non-AGR students in Grades 1-3. Previous years' results show consistently substantive, statistically significant impacts of AGR on kindergarten PALS Reading scores.[21]

20 For detailed impact results, see Technical Appendix Tables A10 – A13.

21 Also see Figure A5, Table A32, and Table A33 in the Technical Appendix for impact analysis of kindergarten MAP/STAR reading growth through 2019-20. While MAP/STAR coverage in kindergarten continues to be low, results show similar impacts to previous analyses of kindergarten PALS reading growth.

Figure 14: Impact of AGR on Absence Rates



The impacts of AGR on absence rates appear in Figure 14. As indicated, while overall there was little impact of AGR on absence rates, there were higher, statistically significant absence rates for second and third grade students in AGR compared to non-AGR students. On average, AGR students had an absence rate 0.4 percentage points higher than their third grade peers in matched non-AGR schools. This translates to approximately 0.7 more absence days, a relatively small impact that is measured with statistical precision due to the large sample size included in the analysis.

Figure 15: Impact of AGR on Suspension Rates

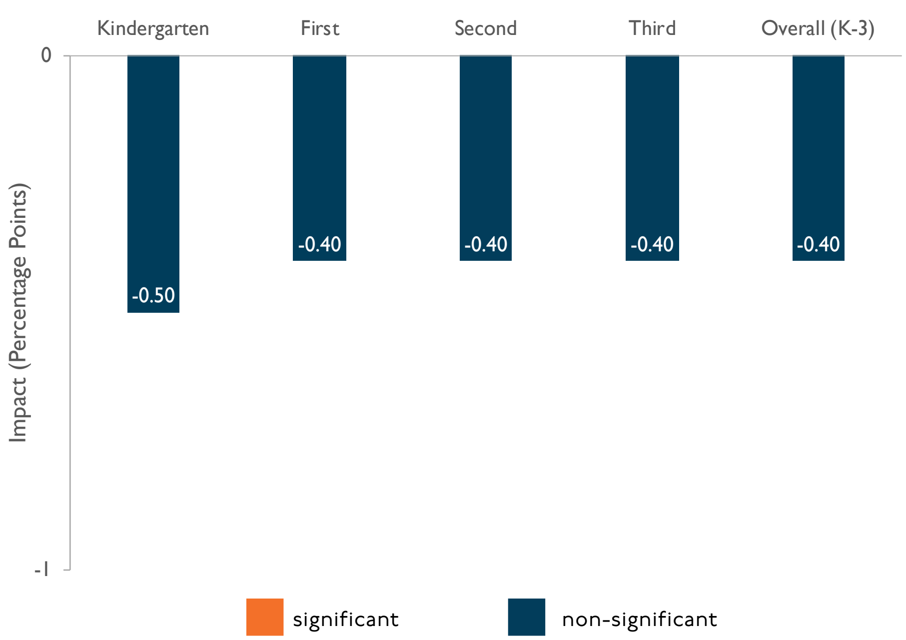


Figure 15 presents the impacts of the AGR program on student discipline as measured by out- of-school suspensions. While the overall impact of the program shows a decrease in the suspension rate, this result is not statistically significant.

Impacts Compared to SAGE

The following results provide information on the impact of the AGR program compared to previous SAGE implementation. This analysis estimates impacts using AGR schools both before and after their transition from SAGE to AGR. Figure 16 shows the impact of AGR compared to SAGE on math growth using both a standardized measure and a measure on the MAP scale. Overall, differences in math growth between AGR and SAGE were small and not statistically significant.

Figure 16: Impact of AGR Compared to SAGE on Math Growth

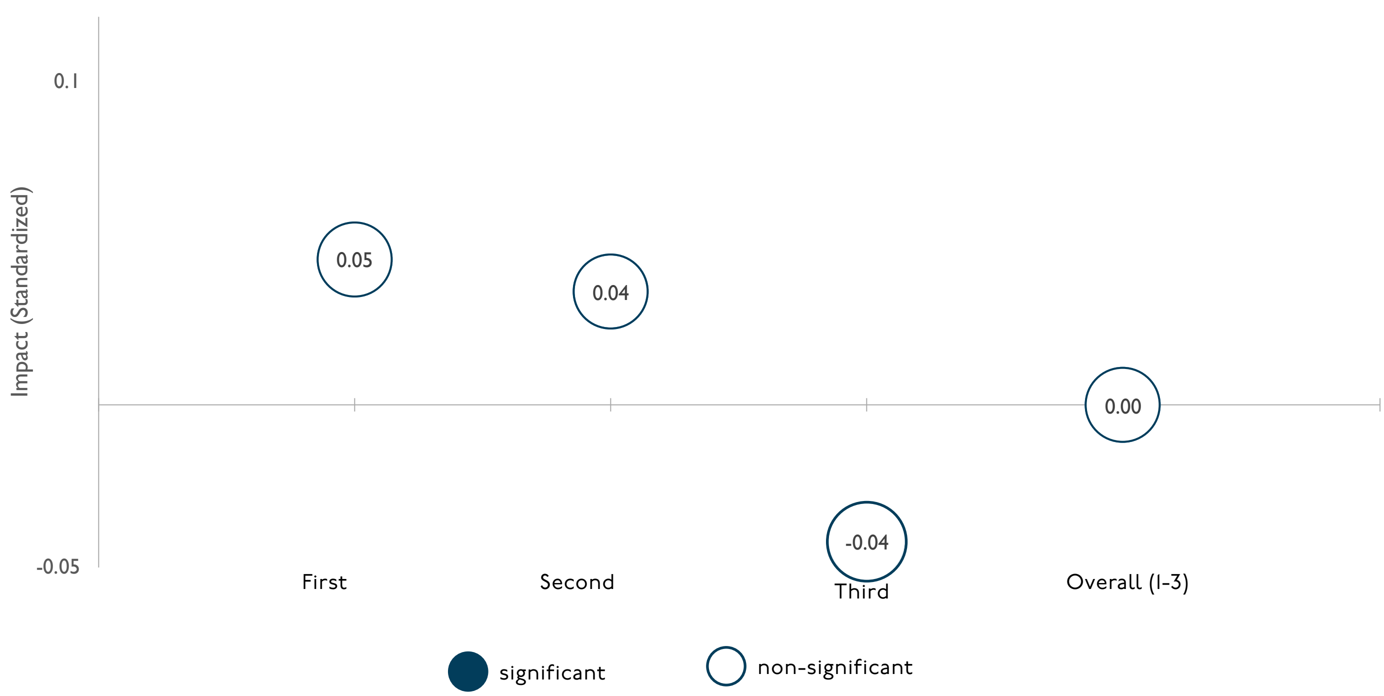


Figure 17: Impact of AGR Compared to SAGE on Reading Growth

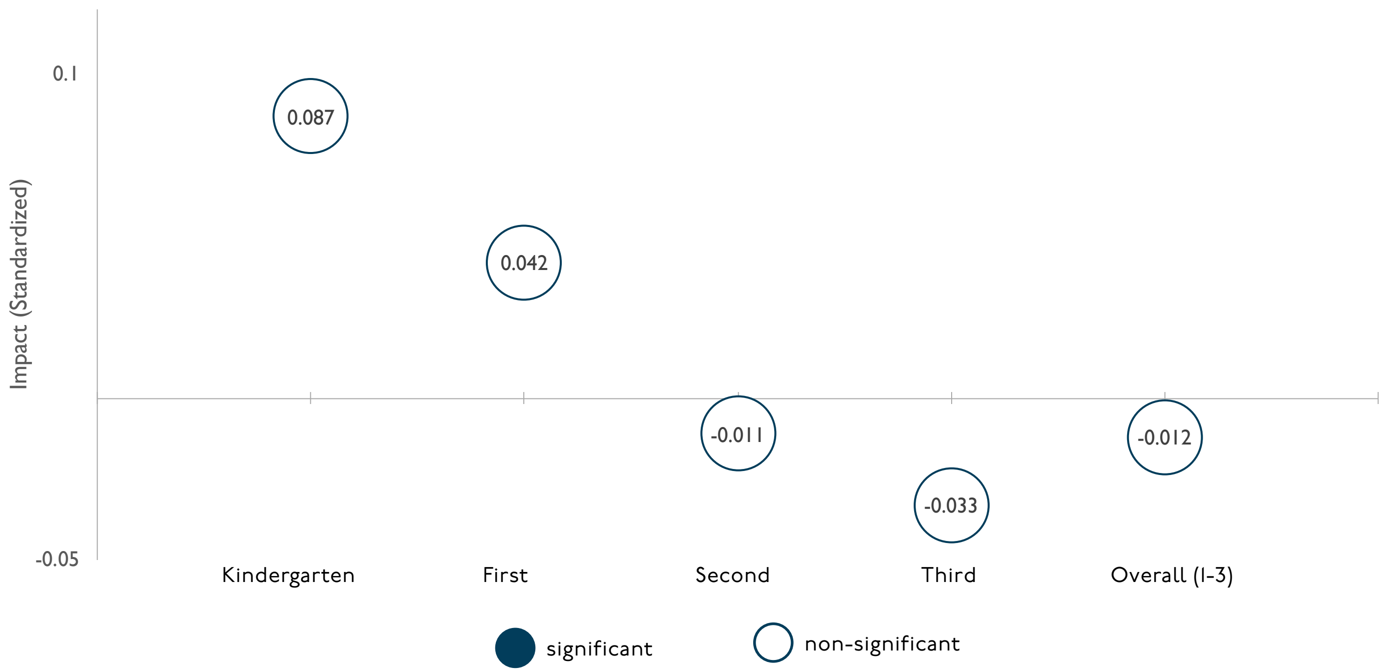
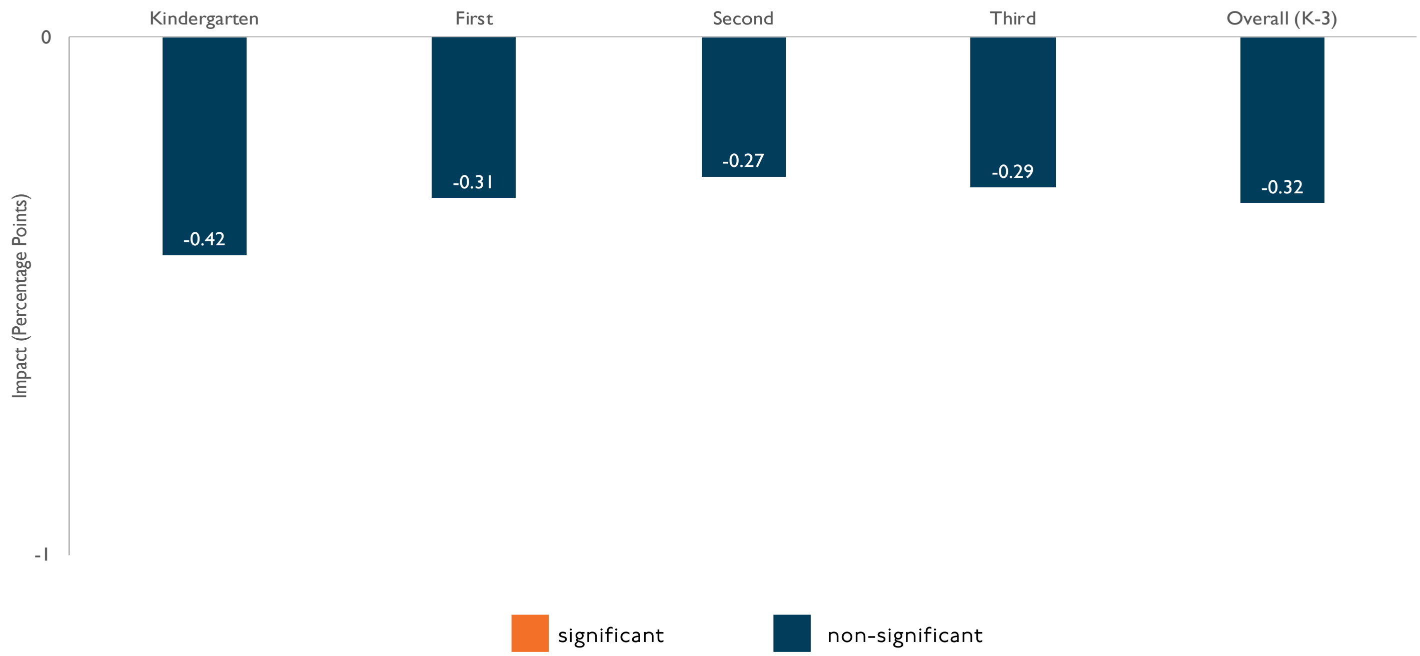


Figure 17 includes differences in reading growth between AGR and SAGE. The evaluation found few differences in reading growth in first through third grade. Previous evaluations have found large and significant impacts of AGR on kindergarten PALS, relative to SAGE (approximately 1.2 PALS scale score points higher growth).

Figure 18: Impact of AGR Compared to SAGE on Absence Rates



Absence rate comparisons between AGR and SAGE reveal few differences, as seen in Figure 18. The estimated average AGR absence rates are slightly lower than SAGE rates, but none of the differences were statistically significant. As with the overall results, all AGR-SAGE differences are too small to be significant for policy.

Figure 19: Impact of AGR Compared to SAGE on Suspension Rates

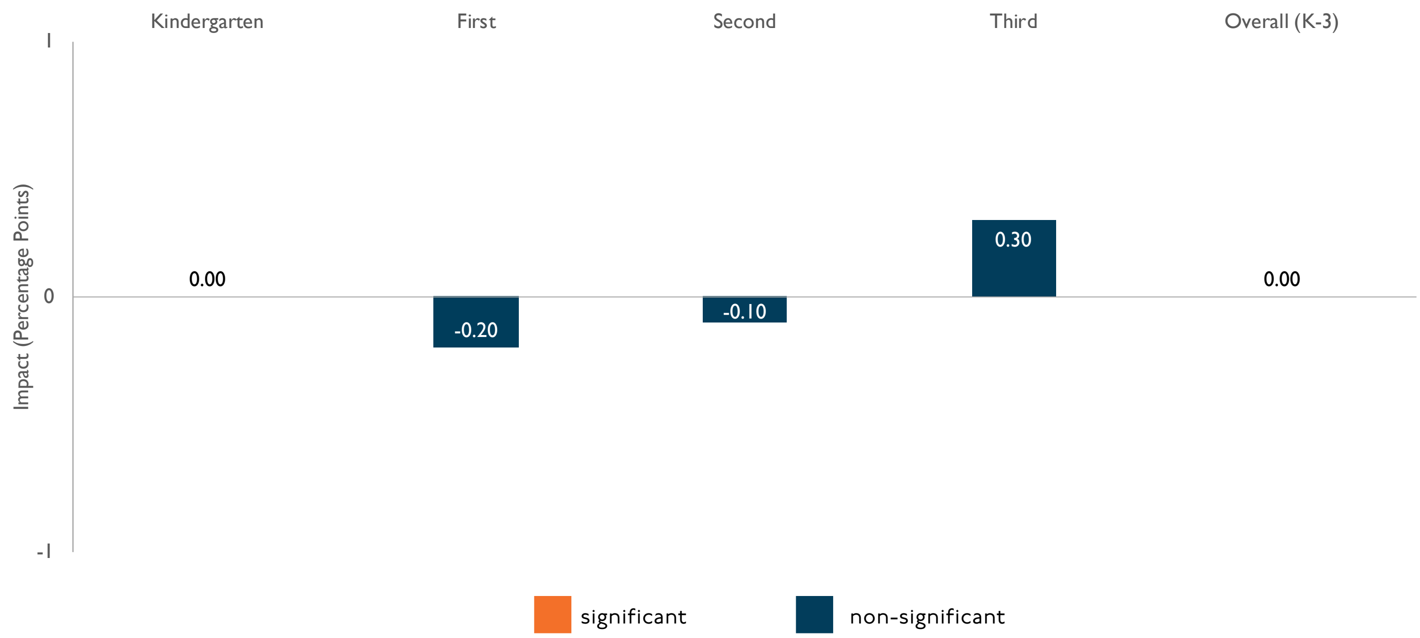


Figure 19 shows the AGR impact compared to the SAGE impact for student discipline. The evaluation found little difference in the suspension rates between AGR students and SAGE students.

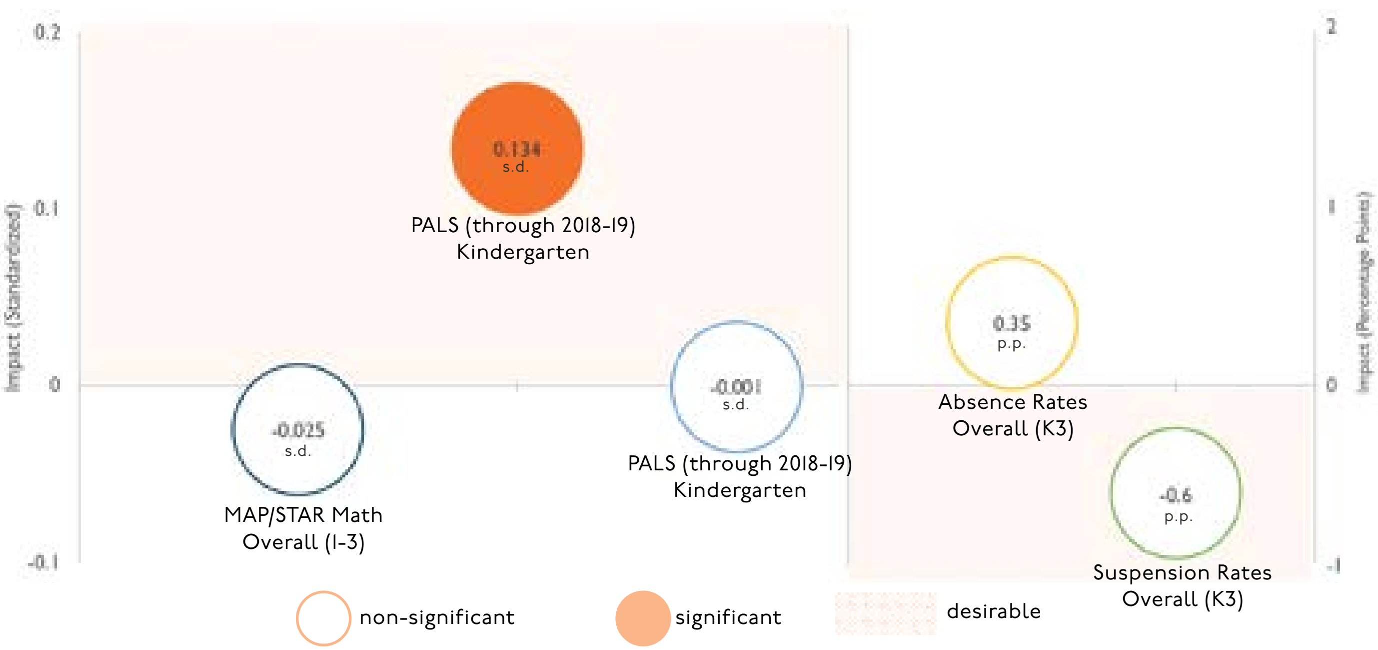
Impacts by Demographics

In addition to examining the overall, statewide impact of the AGR program on student performance, this evaluation also examines whether the AGR program has different impacts for different subgroups of students. Given AGR's focus on closing the achievement gap, examining the program impact for specific subgroups, particularly economically disadvantaged students, is important for determining whether the program is meeting its goals. For each of the subgroups, results compare outcomes of subgroup students that attended AGR schools to outcomes of subgroup students who attended observationally similar, non- AGR schools (e.g., economically disadvantaged students at AGR schools versus economically disadvantaged students at similar, non-AGR schools). Results are pooled across all applicable grade levels to measure the impact of AGR. The Technical Appendix also includes impacts for the following subgroups of students: females, Asian students, Black students, Hispanic students, White students, economically disadvantaged students (as measured by free or reduced- price lunch, or FRL status), English learner (EL) students, special education students, and students in schools within cities.

Impacts for Students Eligible for Free/ Reduced-price Lunch

Given AGR's focus on reducing gaps between economically disadvantaged students and their higher income peers, we present AGR impacts for students who are eligible for free or reduced-price lunch (Figure 20). Outcomes for students eligible for free/reduced-price lunch at AGR schools are compared to outcomes of students eligible for free/reduced-price lunch at observationally similar, non-AGR schools. These results mirror pooled impact results shown above. There are no statistically significant impacts on outcomes in grades 1-3. In past evaluations, students eligible for free/reduced- price lunch had an average PALS reading growth 0.13 standard deviations higher than students eligible for free/reduced-price lunch in similar non-AGR schools (approximately 2 points higher growth on the PALS assessment).

Figure 20: Impacts on Students Eligible for Free/Reduced-price Lunch - All Outcomes



Differences in Outcomes by Strategy

This set of student performance results examines the difference in AGR impacts for each combination of strategies. For the four following tables, the impact is the difference in the outcome between AGR students in schools with the strategy combination listed and AGR students in schools with small class size only. Apart from class size reduction, the strategy combinations examined include coaching only, tutoring only, small class size and coaching, small class size and tutoring, coaching and tutoring, and all three strategies.

Differences in outcomes by strategy should not be interpreted as causal and should be considered only limited evidence of how strategy usage might impact test score growth, absences, and discipline. The analysis includes only AGR schools with no comparison schools. Because Wisconsin law allows AGR schools to select their strategies, differences in outcomes could be biased by omitted variables. For example, more effective schools may systematically choose certain strategies, in which case differences in outcomes could be caused by school effectiveness rather than the strategies themselves.

Table 8 shows the impact on math growth for each strategy combination compared to small class size. Strategy combinations associated with higher average math growth compared to small class size included tutoring only in third grade.

Table 8: Differences in Math Growth by Strategy, Compared to Small Class Size Only

OUTCOME	GRADE	STRATEGY	IMPACT (STANDARIZED)	IMPACT (APPROX. PALS OR MAP SCALE)	P-VALUE
MAP/STAR Math	First	Coaching Only	0.161	2.19	0.069
		Tutoring Only	N/A	N/A	N/A
		Class Size and Coaching	0.099	1.35	0.451
		Class Size and Tutoring	-0.002	-0.03	0.994
		Coaching and Tutoring	0.161	2.19	0.487
		All Three	0.129	1.75	0.170
	Second	Coaching Only	0.070	0.95	0.556
		Tutoring Only	N/A	N/A	N/A
		Class Size and Coaching	0.036	0.49	0.778
		Class Size and Tutoring	0.145	1.96	0.224
		Coaching and Tutoring	0.064	0.87	0.640
		All Three	0.005	0.06	0.998
	Third	Coaching Only	0.027	0.38	0.759
		Tutoring Only	0.335*	4.62	0.000
		Class Size and Coaching	0.028	0.38	0.749
		Class Size and Tutoring	0.023	0.32	0.807
		Coaching and Tutoring	0.077	1.06	0.328
		All Three	0.065	0.90	0.443
	Overall (1-3)	Coaching Only	0.066	0.90	0.433
		Tutoring Only	0.221	3.02	0.152
		Class Size and Coaching	0.043	0.59	0.610
		Class Size and Tutoring	0.066	0.91	0.502
		Coaching and Tutoring	0.090	1.22	0.197
		All Three	0.057	0.78	0.460

Note: P-values corrected to account for multiple estimates. N/A indicates too few schools employing a strategy to accurately estimate results. * Statistically significant at the 0.05 level.

Looking at the differences in kindergarten PALS reading growth for each strategy combination compared to small class size only, the analysis found few differences across strategy combinations, as seen in Table 9. Results on differences in reading growth in Grades 1-3, found in the same table, indicate higher average reading growth for coaching only in first grade when compared to small class size only.

Table 9: Differences in Reading Growth by Strategy, Compared to Small Class Size Only

OUTCOME	GRADE	STRATEGY	IMPACT (STANDARIZED)	IMPACT (APPROX. MAP SCALE)	P-VALUE
PALS (through 2018-19)	Kindergarten	Coaching Only	-0.132	-1.82	0.369
		Tutoring Only	0.001	0.01	0.996
		Class Size and Coaching	-0.104	-1.43	0.252
		Class Size and Tutoring	-0.001	-0.01	1.001
		Coaching and Tutoring	-0.109	-1.50	0.518
		All Three	-0.030	-0.42	0.842
MAP/STAR Reading	First	Coaching Only	0.222*	3.02	0.003
		Tutoring Only	N/A	N/A	N/A
		Class Size and Coaching	0.179	2.44	0.097
		Class Size and Tutoring	-0.068	-0.93	0.706
		Coaching and Tutoring	0.216	2.95	0.230
		All Three	0.099	1.34	0.251
	Second	Coaching Only	0.055	0.74	0.564
		Tutoring Only	N/A	N/A	N/A
		Class Size and Coaching	0.072	0.97	0.477
		Class Size and Tutoring	0.033	0.45	0.813
		Coaching and Tutoring	0.042	0.57	0.731
		All Three	0.010	0.13	0.949
	Third	Coaching Only	0.023	0.32	0.746
		Tutoring Only	0.136	1.88	0.103
		Class Size and Coaching	0.069	0.96	0.145
		Class Size and Tutoring	0.005	0.07	0.999
		Coaching and Tutoring	0.059	0.81	0.512
		All Three	0.018	0.24	0.821
	Overall (1-3)	Coaching Only	0.072	1.07	0.225
		Tutoring Only	0.083	1.24	0.505
		Class Size and Coaching	0.090	1.35	0.087
		Class Size and Tutoring	-0.002	-0.03	0.998
		Coaching and Tutoring	0.083	1.25	0.159
		All Three	0.033	0.49	0.657

Note: P-values corrected to account for multiple estimates. N/A indicates too few schools employing a strategy to accurately estimate results. * Statistically significant at the 0.05 level.

Table 10 displays differences in absence rates for each of the strategy combinations compared to class size reduction only. As this table illustrates, there were no statistically significant differences between absence rates across employed strategies.

Table 10: Differences in Absence Rates by Strategy, Compared to Small Class Size Only

OUTCOME	GRADE	STRATEGY	IMPACT (PERCENTAGE POINTS)	IMPACT (APPROX. DAYS)	P-VALUE
Absence Rate	Kindergarten	Coaching Only	-0.29	-0.5	0.752
		Tutoring Only	0.58	1.0	0.500
		Class Size and Coaching	-0.15	-0.3	0.809
		Class Size and Tutoring	0.46	0.8	0.724
		Coaching and Tutoring	-0.69	-1.2	0.745
		All Three	0.01	0.0	1.002
	First	Coaching Only	-0.02	0.0	0.993
		Tutoring Only	0.34	0.6	0.629
		Class Size and Coaching	0.01	0.0	0.996
		Class Size and Tutoring	-0.40	-0.7	0.787
		Coaching and Tutoring	0.11	0.2	0.895
		All Three	0.43	0.7	0.490
	Second	Coaching Only	-0.03	0.0	0.986
		Tutoring Only	0.56	1.0	0.293
		Class Size and Coaching	0.18	0.3	0.749
		Class Size and Tutoring	-0.20	-0.3	0.899
		Coaching and Tutoring	0.36	0.6	0.552
		All Three	0.22	0.4	0.701
	Third	Coaching Only	0.11	0.2	0.853
		Tutoring Only	0.74	1.3	0.145
		Class Size and Coaching	0.22	0.4	0.743
		Class Size and Tutoring	0.56	1.0	0.442
		Coaching and Tutoring	0.13	0.2	0.914
		All Three	0.08	0.1	0.905
	Overall (1-3)	Coaching Only	-0.04	-0.1	0.961
		Tutoring Only	0.56	1.0	0.222
		Class Size and Coaching	0.07	0.1	0.908
		Class Size and Tutoring	0.11	0.2	0.937
		Coaching and Tutoring	0.03	0.0	0.998
		All Three	0.19	0.3	0.736

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

Table 11 displays differences in absence rates for each of the strategy combinations compared to class size reduction only. As this table illustrates, there were no statistically significant differences between absence rates across employed strategies.

Table 11: Differences in Suspension Rates by Strategy, Compared to Small Class Size Only

OUTCOME	GRADE	STRATEGY	IMPACT (PERCENTAGE POINTS)	P-VALUE
Suspension Rate	Kindergarten	Coaching Only	1.0	0.620
		Tutoring Only	1.0	0.076
		Class Size and Coaching	0.8	0.501
		Class Size and Tutoring	-0.3	0.565
		Coaching and Tutoring	0.7	0.505
		All Three	0.7	0.246
	First	Coaching Only	0.2	0.948
		Tutoring Only	0.2	0.814
		Class Size and Coaching	1.1	0.480
		Class Size and Tutoring	0.4	0.583
		Coaching and Tutoring	0.5	0.654
		All Three	0.2	0.825
	Second	Coaching Only	2.2	0.214
		Tutoring Only	0.8	0.137
		Class Size and Coaching	0.5	0.721
		Class Size and Tutoring	1.1	0.074
		Coaching and Tutoring	0.5	0.705
		All Three	0.3	0.722
	Third	Coaching Only	1.1	0.732
		Tutoring Only	1.6*	0.011
		Class Size and Coaching	0.3	0.862
		Class Size and Tutoring	0.3	0.751
		Coaching and Tutoring	0.1	0.960
		All Three	0.0	1.002
	Overall (1-3)	Coaching Only	0.9*	0.017
		Tutoring Only	0.3	0.563
		Class Size and Coaching	1.1	0.291
		Class Size and Tutoring	0.7	0.477
		Coaching and Tutoring	0.4	0.637
		All Three	0.3	0.559

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

Differences in Outcomes by Intensity and Frequency of Strategy Use

Results from the End-of-Year Report (see below) provide information on the frequency of coaching per semester, the frequency of tutoring per month, the characteristics of instructional coaches, the usage of high-quality instructional coaching practices, and the usage of high-quality tutoring practices. Impacts analyzed by frequency show the effect associated with one additional instructional coaching session or one additional tutoring session. Impacts analyzed by qualification or practice show the effect associated with that qualification or practice regardless of the other qualifications or practices employed. As with the previous section, results shown below should not be considered causal due to selection concerns associated with schools choosing which strategies to use.

Table 12 shows the relationships between frequency and intensity of coaching and tutoring with math growth across grades 1-3. There were no statistically significant relationships between math test score growth and frequency, coach characteristics, coach practices, or tutoring practices.

Table 12: Differences in Math Growth by Frequency and Intensity of Coaching and Tutoring

OUTCOME	FREQUENCY, CHARACTERISTIC, OR PRACTICE	IMPACT (STANDARDIZED)	IMPACT (APPROX. MAP SCALE)	P-VALUE
MAP/STAR Math	Frequency			
	Coaching frequency per semester	-0.001	-0.02	0.745
	Tutoring frequency per month	-0.012	-0.17	0.218
	Coach Characteristics			
	Coach training	0.002	0.03	1.000
	Previous coaching experience	0.007	0.10	0.906
	Content specialist in their subject	-0.053	-0.73	0.226
	Coaching Practices			
	One-to-one teacher coaching	0.038	0.52	0.663
	Team teacher coaching	0.030	0.41	0.664
	Keeps a coaching log	-0.021	-0.29	0.732
	Advises teachers to set goals	0.021	0.29	0.732
	Coaching focuses on teacher goals	-0.011	-0.16	0.853
	Maintains a focus on equity	-0.034	-0.47	0.486
	Encourages reflective practices	0.040	0.54	0.561
	Discusses data with teachers	-0.022	-0.30	0.742
	Observes teacher practices	-0.022	-0.30	0.744
	Tutoring Practices			
	Reviews student data	0.060	0.81	0.729
	Models appropriate learning behavior	0.104	1.43	0.227
	Adapts to student learning styles	0.042	0.58	0.728
	Maintains a focus on equity	0.013	0.18	0.857
	Provides scaffolding	-0.062	-0.85	0.219
	Communicates regularly with teacher	-0.083	-1.14	0.286

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

As with the math outcome, the relationships between PALS growth and frequency of coaching and tutoring, coach characteristics, coach practices, and tutoring practices were not statistically significant, as seen in Table 13.

Table 13: Differences in PALS Growth by Frequency and Intensity of Coaching and Tutoring

OUTCOME	FREQUENCY, CHARACTERISTIC, OR PRACTICE	IMPACT (STANDARDIZED)	IMPACT (APPROX. PALS SCALE)	P-VALUE
PALS (through 2018-19)	Frequency			
	Coaching frequency per semester	-0.006	-0.08	0.617
	Tutoring frequency per month	-0.008	-0.10	0.829
	Coach Characteristics			
	Coach training	0.049	0.68	0.750
	Previous coaching experience	0.022	0.31	0.893
	Content specialist in their subject	-0.150	-2.06	0.100
	Coaching Practices			
	One-to-one teacher coaching	-0.006	-0.08	0.984
	Team teacher coaching	-0.042	-0.58	0.820
	Keeps a coaching log	-0.007	-0.10	0.970
	Advises teachers to set goals	0.045	0.61	0.789
	Coaching focuses on teacher goals	0.015	0.20	0.958
	Maintains a focus on equity	-0.081	-1.12	0.583
	Encourages reflective practices	-0.030	-0.42	0.899
	Discusses data with teachers	0.103	1.42	0.515
	Observes teacher practices	-0.004	-0.06	0.994
	Tutoring Practices			
	Reviews student data	-0.102	-1.40	0.532
	Models appropriate learning behavior	-0.170	-2.34	0.454
	Adapts to student learning styles	-0.027	-0.38	0.887
	Maintains a focus on equity	0.061	0.84	0.702
	Provides scaffolding	0.127	1.75	0.500
	Communicates regularly with teacher	-0.127	-1.75	0.509

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

Table 14 shows the relationships between reading growth and coaching and tutoring frequency and intensity for grades 1-3. Significant results included a negative impact on reading growth for a coach being a content specialist in their subject.

Table 14: Differences in Reading Growth by Frequency and Intensity of Coaching and Tutoring

OUTCOME	FREQUENCY, CHARACTERISTIC, OR PRACTICE	IMPACT (STANDARDIZED)	IMPACT (APPROX. MAP SCALE)	P-VALUE
MAP/STAR Reading	FREQUENCY			
	Coaching frequency per semester	-0.001	-0.02	0.748
	Tutoring frequency per month	-0.008	-0.12	0.247
	COACH CHARACTERISTICS			
	Coach training	0.000	-0.01	1.003
	Previous coaching experience	0.010	0.15	0.816
	Content specialist in their subject	-0.079*	-1.18	0.011
	COACHING PRACTICES			
	One-to-one teacher coaching	0.022	0.34	0.737
	Team teacher coaching	0.018	0.27	0.740
	Keeps a coaching log	-0.049	-0.73	0.226
	Advises teachers to set goals	0.016	0.24	0.739
	Coaching focuses on teacher goals	0.037	0.55	0.390
	Maintains a focus on equity	-0.009	-0.13	0.841
	Encourages reflective practices	0.031	0.47	0.659
	Discusses data with teachers	-0.020	-0.30	0.754
	Observes teacher practices	-0.053	-0.80	0.195
	TUTORING PRACTICES			
	Reviews student data	-0.031	-0.46	0.826
	Models appropriate learning behavior	0.089	1.33	0.254
	Adapts to student learning styles	0.103	1.54	0.105
	Maintains a focus on equity	-0.006	-0.10	0.945
	Provides scaffolding	-0.009	-0.13	0.850
	Communicates regularly with teacher	-0.111	-1.65	0.223

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

Table 15 and Table 16 show the relationships between coaching and tutoring frequency, characteristics, and practices with absence rate and out-of-school suspension rates, respectively. As seen from these two tables, there were no statistically significant relationships.

Table 15: Differences in Absence Rates by Frequency and Intensity of Coaching and Tutoring

OUTCOME	FREQUENCY, CHARACTERISTIC, OR PRACTICE	IMPACT (PERCENTAGE POINTS)	IMPACT (APPROX. DAYS)	P-VALUE
Absence Rate	Frequency			
	Coaching frequency per semester	-0.02	0.0	0.512
	Tutoring frequency per month	0.02	0.0	0.860
	Coach Characteristics			
	Coach training	0.17	0.3	0.819
	Previous coaching experience	-0.27	-0.5	0.553
	Content specialist in their subject	0.40	0.7	0.225
	Coaching Practices			
	One-to-one teacher coaching	-0.21	-0.4	0.752
	Team teacher coaching	-0.16	-0.3	0.747
	Keeps a coaching log	0.11	0.2	0.745
	Advises teachers to set goals	0.00	0.0	1.001
	Coaching focuses on teacher goals	0.19	0.3	0.754
	Maintains a focus on equity	0.18	0.3	0.746
	Encourages reflective practices	-0.23	-0.4	0.723
	Discusses data with teachers	-0.15	-0.3	0.773
	Observes teacher practices	0.34	0.6	0.561
	Tutoring Practices			
	Reviews student data	-0.65	-1.1	0.486
	Models appropriate learning behavior	1.00	1.7	0.496
	Adapts to student learning styles	-0.66	-1.2	0.734
	Maintains a focus on equity	0.31	0.5	0.555
	Provides scaffolding	0.48	0.8	0.732
	Communicates regularly with teacher	-0.71	-1.2	0.560

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

Table 16: Differences in Suspension Rates by Frequency and Intensity of Coaching and Tutoring

OUTCOME	FREQUENCY, CHARACTERISTIC, OR PRACTICE	IMPACT (PERCENTAGE POINTS)	P-VALUE
Suspension Rate	FREQUENCY		
	Coaching frequency per semester	0.0	0.896
	Tutoring frequency per month	0.0	0.843
	COACH CHARACTERISTICS		
	Coach training	0.0	0.999
	Previous coaching experience	0.0	0.996
	Content specialist in their subject	0.2	0.740
	COACHING PRACTICES		
	One-to-one teacher coaching	-0.4	0.558
	Team teacher coaching	-0.2	0.823
	Keeps a coaching log	0.2	0.782
	Advises teachers to set goals	-0.1	0.932
	Coaching focuses on teacher goals	0.1	0.952
	Maintains a focus on equity	-0.7	0.113
	Encourages reflective practices	0.4	0.448
	Discusses data with teachers	-0.3	0.744
	Observes teacher practices	0.1	0.911
	TUTORING PRACTICES		
	Reviews student data	-0.2	0.811
	Models appropriate learning behavior	0.0	0.996
	Adapts to student learning styles	-0.2	0.822
	Maintains a focus on equity	0.1	0.896
	Provides scaffolding	-0.3	0.739
	Communicates regularly with teacher	0.1	0.816

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

Section 7

School Board Report Findings

As part of participation in AGR, schools and districts agree to report to their boards on the strategies they implemented and their success in meeting the performance objectives listed in their AGR contracts. DPI provides a suggested reporting template that the majority of schools use. The impact evaluation uses data from these school board reports, in conjunction with data from the End-of-Year Report, to determine the strategies that schools use in each grade and year. Due to reporting inconsistencies between schools, however, data from school board reports is less reliable and covers fewer schools than the End-of- Year Report. Below, we describe strategies and performance objectives data from the school board reports.

Table 17 below lists all possible combinations of the three strategy types – reduced class sizes, instructional coaching, and one-to-one tutoring – similar to strategies data from the End-of-Year Report (see Table 19). The breakdown of strategies is very similar to those provided from the End-of-Year Report. Schools most commonly reduce class sizes, although instructional coaching was used by over half the reporting schools. Schools are more likely to use combinations of strategies, with 14 percent using all three.

Table 17: 2019-20 School-level AGR Strategies, School Board Report Data

STRATEGIES	%
Coaching Only	15%
Class Size Only	23%
Tutoring Only	2%
Coaching and Class Size	36%
Coaching and Tutoring	4%
Class Size and Tutoring	6%
All Three	14%

Data on the types of performance objectives schools use appears in Table 18. We classify performance objectives into three primary types – achievement (e.g., bringing all students up to proficiency), growth (e.g., improving student scores by 10 points), and gap closing (e.g., improving scores for students receiving free and reduced-price lunch relative to other students' scores). Almost 90 percent of schools set achievement objectives, and 47 percent set growth objectives. Very few set performance objectives to close school achievement gaps, although it should be noted that the AGR program offers supplemental school-based funding to close gaps across schools, not necessarily within schools.

Table 18: 2019-20 School-level AGR Performance Objectives, School Board Report Data

TYPES OF PERFORMANCE OBJECTIVES	%
Achievement only	48%
Growth only	8%
Gap closing only	1%
Achievement and growth	37%
Achievement and gap closing	2%
Growth and gap closing	0%
All 3 goals	2%

Section 8

End-of-Year Report Findings

This section of the report provides the results from the 2019-20 End-of-Year Report survey of AGR schools, fielded during the summer of 2020. The 2019-20 survey was similar to the 2018-19 version, but with one additional item to measure student virtual school attendance from mid-March to June 2020 and another additional, open-ended item that asked which strategies schools have used to ensure that all students were able to regularly participate in virtual schooling. Despite the pandemic, nearly 100 percent of AGR schools responded to the survey, similar to previous years.

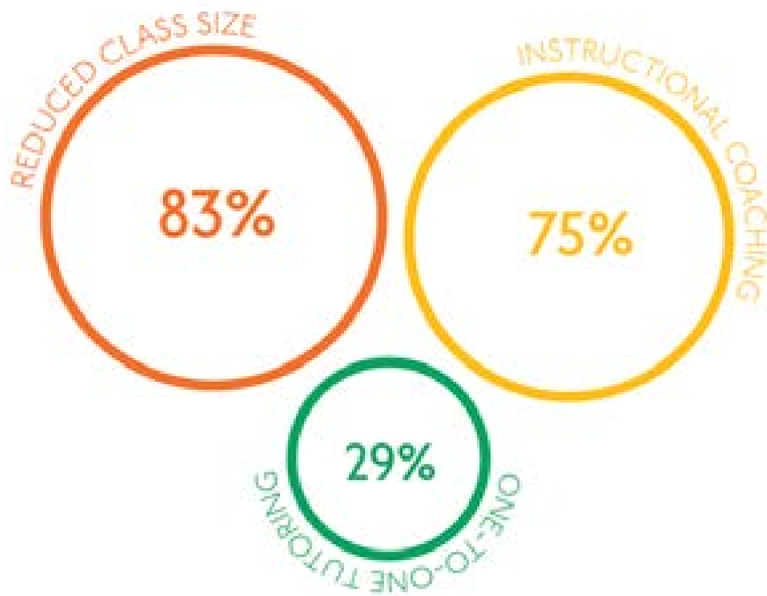
As described in previous reports, schools that had transitioned from SAGE to AGR have taken advantage of AGR's increased flexibility. As shown in Table 19, only 18 percent of responding schools employed only the reduced class size strategy. A majority of schools opted for multiple strategies – 69 percent used more than one strategy, including 18 percent that used all three strategies. Reducing class size remained the most popular strategy, as shown in Table 19. Eighty-three percent of AGR schools used reduced class size (83 percent of schools). Instructional coaching (75 percent of schools) was also common. One-to-one tutoring was used by only 29 percent of sample schools.

Table 19: 2019-20 School-level AGR Strategies from End-of-Year Report

STRATEGIES	%
Coaching Only	12%
Class Size Only	18%
Tutoring Only	1%
Coaching and Class Size	41%
Coaching and Tutoring	4%
Class Size and Tutoring	6%
All Three	18%

Note: 411 respondents.

Figure 21: Percentage of AGR Schools Using Each Strategy, EOY Report



Distribution of AGR Strategies Across Schools

In general, schools chose to use the same strategies across classrooms within each grade (Table 20). This trend was strongest for grades with reduced size classrooms. Within each grade, approximately 90 percent of schools chose to use reduced class sizes in at least three-quarters of classrooms or not at all. For one-to-one tutoring and instructional coaching, approximately 70 to 80 percent of schools used the strategy in at least three-quarters of the classrooms or not at all. The number of grades using instructional coaching in more than 75 percent of classrooms increased noticeably from prior years.

Table 20: Schools' Distributions of Classrooms with Strategies, by Grade

	PERCENT OF CLASSROOMS				
GRADE	NONE	LESS THAN 25%	25%-50%	51-75%	MORE THAN 75%
REDUCED CLASS SIZE, N= 341					
Kindergarten	8%	1%	6%	3%	82%
First	25%	1%	6%	3%	65%
Second	26%	1%	6%	4%	63%
Third	30%	1%	6%	3%	60%
ONE-TO-ONE TUTORING, N= 120					
Kindergarten	12%	13%	11%	3%	62%
First	11%	8%	12%	5%	64%
Second	11%	12%	13%	3%	62%
Third	11%	13%	12%	3%	62%
INSTRUCTIONAL COACHING, N= 309					
Kindergarten	8%	1%	6%	3%	82%
First	25%	1%	6%	3%	65%
Second	26%	1%	6%	4%	63%
Third	30%	1%	6%	3%	60%

Note: Categories are mutually exclusive.

Reduced Class Size Instructional Strategies and Benefits

Schools using reduced class sizes reported using a variety of instructional strategies (Table 21), including small group instruction (96 percent), one-on-one time with the teacher (80 percent), differentiation of instruction (92 percent), strategic placement of students in groups (81 percent), and, to a lesser extent, strategic placement of students in classrooms (55 percent). Only 3 percent of responding schools reported that they use no additional instructional strategies due to AGR class size reductions.

Table 21: Instructional Strategies Associated with AGR Reduced Class Size

STRATEGIES	%
Small-group instruction	96%
Differentiation of instruction	92%
One-on-one time with the teacher	80%
Strategic placement of students in groups	81%
Strategic placement of students in classrooms	55%
We don't use any specific instructional strategies because of smaller class sizes	3%
Other	4%
Not sure/don't know	0%

Note: 341 respondents.

The survey included an open-ended response item for perceived benefits reduced class size provided. Most schools (N=324) responded to the item, and all responses were positive. There was a wide variety of responses to this item, but the most common theme was that reduced class sizes allow teachers to differentiate instruction via one-on-one time with students or working with students in small groups. These types of answers included:

- Increased one-on-one time for students and the teacher (n=195)
- More opportunities to work with students in small groups (n=132)
- Better opportunity for teachers to provide differentiated and targeted instruction and interventions to meet students' individual needs (n=105)

One representative response reads,

> "It allows us to create an environment where we are able to differentiate our instruction and provide the necessary interventions to help our students succeed at all levels."

Respondents also commonly noted that smaller class sizes allow teachers to build stronger connections and relationships with students, increase student engagement, and build better relationships and strengthen connections with parents and families.

Table 22: Frequency of AGR One-to-One Tutoring

FREQUENCY	%
3 times a week or more	47%
2 times a week	21%
Weekly	13%
Biweekly	0%
Monthly	0%
Other	3%
As needed	17%

Note: 120 respondents to survey item.

Table 23: AGR One-to-One Tutoring Practices

PRACTICE	%
Reviews student data	84%
Communicates regularly with classroom teacher	80%
Provides scaffolding	82%
Adapts to student learning styles	92%
Models appropriate learning behavior	86%
Maintains a focus on equity	60%
Other	2%
Not sure/don't know	0%

Note: 120 respondents to survey item.

One-to-One Tutoring Frequency, Practices, and Benefits

Schools using one-to-one tutoring did so frequently, as seen in Table 22. Eighty-one percent offered tutoring at least weekly, and 47 percent engaged in tutoring 3 times a week or more.

Table 23 shows that schools reported using almost all of the tutoring practices listed on the survey. In 2019-20, schools more frequently cited tutoring as helping to maintain a focus on equity (60 percent), an increase from 2018-19 (50 percent).

The survey incorporated an open-response item for benefits one-to-one tutoring provides. Among the 113 responses, by far the most common were that one-to-one tutoring allows for targeted and individualized instruction (n=75) and allows schools to focus on highest-need students (n=48). For example,

> "Tutoring also allowed participating students to be taught at their current level of proficiency. Lower performing students are often behind on important fundamental skills. These skills need to be developed and practiced before moving ahead. Tutors were able to address areas of weakness. This approach provided students an opportunity to revisit previously covered topics, in order to deepen their understanding of them, which then allowed them to move forward with their learning objectives."

Table 24: Frequency of AGR Instructional Coaches Meeting with Teachers

FREQUENCY	%
Weekly	51%
Monthly	26%
Quarterly	4%
Each semester	0%
Other	7%
As needed	11%
Not sure/don't know	1%

Note: 309 respondents to survey item.

Table 25: Instructional Coaching Practices

PRACTICE	%
Encourages reflective practices	82%
Discusses data with teachers	91%
One-to-one teacher coaching	85%
Observes teacher practices	76%
Team teacher coaching	72%
Advises teachers to set goals	67%
Coaching focuses on teacher goals	69%
Maintains a focus on equity	60%
Keeps a coaching log	49%
Other	5%
Not sure/don't know	1%

Note: 309 respondents to survey item.

Instructional Coaching Frequency, Practices, and Coach Characteristics

Table 24 shows that at 77 percent of schools, instructional coaches met with teachers at least monthly, and 51 percent met weekly.

Schools responded affirmatively to almost all listed coaching practices (Table 25). Similar to practices associated with one-to-one tutoring, 2019-20 respondents were more likely to cite maintaining a focus on equity (60 percent) than they did in 2018-19 (46 percent).

Schools were able to successfully find trained, experienced coaches with content specialization, as seen from Table 26.

The survey also asked an open-ended question regarding the benefits of instructional coaching. The 284 responses included the following categories:

- Coaching provides resources, modeling, and support for teachers in the form of embedded, individualized professional development (n=172)
- Coaching has led to teachers' growth and improved instructional practices (n=99)
- Teachers have increased their use of data, primarily in making data-informed instructional decisions (n=60)

The following quote exemplifies the many schools that cited instructional coaching as providing resources, modeling and support:

> "Our school SST (school support teacher) acts as a coach for all teachers. She is trained in coaching and receives extensive training in the curriculum that the classroom teachers deliver. As a result, she is a thought partner, peer-observer and overall resource for teachers to work with to improve their overall practice."

Table 26: Characteristics of AGR Instructional Coaches

FREQUENCY	%
Coach training	80%
Content specialist in their subject of coaching	69%
Previous instructional coaching experience	74%
Other	7%
Not sure/don't know	3%

Note: 309 respondents to survey item.

Virtual Attendance During Spring 2020 COVID School Closures

As the COVID-19 pandemic began in spring of 2020, educators throughout Wisconsin expressed concerns that students from lower-income backgrounds would be particularly hurt due to fewer educational resources within the home, particularly access to broadband internet. [22]

Because DPI administers the End-of-Year survey in the summer to measure AGR practice in the previous year, the 2020 survey presented an opportunity for DPI to ask schools with concentrations of lower-income families how the COVID-19 pandemic impacted student attendance and how schools attempted to engage all students. As shown in Table 27, spring 2020 attendance at AGR schools suffered substantially. In the 2018-19 evaluation's analysis sample, average attendance was 94 percent. In the spring of 2020, however, just over half of students attended at least 75 percent of the time. Nine percent did not attend at all. Although schools reported similar within-school attendance for all students and students receiving free or reduced-price lunch, the survey results are not able to tell us how attendance may have differed between higher-income and lower-income schools.

Table 27: Percentage of Students Attending 3+ Days per Week of Virtual Schooling, mid-March to June 2020

	NONE	LESS THAN 25%	25-50%	51-75%	MORE
All students	9%	2%	15%	18%	56%
Students who receive FRL	9%	3%	18%	19%	52%

22 Wisconsin Policy Forum. 2020. Wisconsin's Digital Divide and its Impacts on Learning. https://wispolicyforum.org/wp-content/uploads/2020/05/Focus_COVID_Internet_Access.pdf

The survey also included an open-ended item asking about strategies schools used to ensure that all students regularly participate in virtual schooling. Of the 385 responses to this question, the most common focused on providing access to technology and keeping students and families engaged with school staff. For students with remaining deficits regarding technology access, teachers often provided paper options.

- Schools provided students with needed technology equipment and devices such as laptops and iPads (n=240)
- School staff members maintained frequent contact with students and families through various forms of communication (n=240)
- Teachers utilized virtual learning platforms for instructional delivery and communication (Google Classroom, Zoom, Google Meets, etc.) (n=203)
- Schools provided internet access for students, whether through hotspots, working with local internet companies, or providing Wifi access in school parking lots (n=143)
- Teachers provided students with the option of paper materials and packets of work, to supplement or replace online learning (n=126)

It is noteworthy that schools noted the importance of frequent contact with families as frequently as they mentioned technological equipment. Many of the responses cited support networks that schools and communities formed to reach out to students who may be struggling to participate. This relational emphasis also can be found across schools' responses to survey questions about the benefits of instructional strategies. For example, tutoring not only allowed for academic-related outcomes, but it also provided students the opportunity to develop a positive relationship with an additional adult in the building. Coaching helped teachers improve their relationships with students. This relational aspect was especially clear in the class size benefits section. Teachers could build stronger connections and relationships with students and their families, teachers had more time to understand, address, and meet students' social and emotional needs, and there was better interaction among students.

Section 9

Summary/Conclusions

The socioeconomic achievement gap is wide and has remained relatively unchanged over the past fifty years.[23] Researchers and policymakers have hypothesized dozens of causes, including funding deficits for districts located in high-poverty areas.[24] The AGR program seeks to reduce the socioeconomic achievement gap by providing additional funding to districts with large proportions of economically disadvantaged students.

This report provides evidence regarding program impacts on math and reading growth, student attendance, and out-of-school suspensions. These results are presented at the state level and disaggregated by grade and student demographic characteristics. The report also contains data on the AGR strategies schools have implemented, the intensity of strategy use, and preliminary evidence on the relative effectiveness of combinations of strategies.

From 2015-16 to 2018-19, AGR impacts on achievement are limited to strong kindergarten reading growth statewide. Attending an AGR school is associated with a moderate (0.11 standard deviation) increase in PALS growth from fall to spring, relative to a comparison group of students from observably similar schools. Students eligible for free/reduced-price lunch in AGR schools experienced moderate growth (0.13 standard deviations) greater than that of similar students in non-AGR schools. Impacts on PALS growth was large for Hispanic students (0.25 standard deviations), English learner students (0.37 standard deviations), urban students (0.21 standard deviations), and Asian students (0.24 standard deviations). From 2015-16 to 2019-20, math and reading MAP and STAR growth in Grades 1-3 was near zero and insignificant. Although estimates of impacts on math and reading ranged across subgroups, most were not significantly different from zero. Students in special education in AGR schools experienced less growth (0.06 standard deviations) than special education students in non-AGR schools.

The report also estimated impacts for non-testing outcomes. We found some evidence that AGR is associated with fewer out-of-school suspensions. Although suspensions are rare events in Grades K-3, Hispanic students in AGR schools were 0.8 percentage points less likely to receive a suspension relative to similar students in comparable non-AGR schools. Similarly, English learner students at AGR schools were 0.5 percentage points less likely to be suspended. We found very few statistically significant impacts on attendance. Most point estimates of attendance impacts showed decreases in attendance at AGR schools, although these decreases were too small to be significant to state policy.

23 Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessman, L. (2020). Long-run Trends in the U.S. SES-Achievement Gap. NBER Working Paper No. 26764. Retrieved May 14, 2021 from https://www.nber.org/papers/w26764.

24 Jackson, C. K., Wigger, C., & Xiong, H. (2021). Do school spending cuts matter? Evidence from the Great Recession. American Economic Journal: Economic Policy, 13(2), 304-35.

Looking at the AGR strategies that schools chose to implement, we found that most AGR schools took advantage of the program's flexibility and chose to implement instructional coaching and one-to-one tutoring strategies that were not included in the state's previous SAGE policy. Over 60 percent of schools combined two or more strategies.

As in any observational study, this evaluation has several limitations. The PSM methodology matches schools on observable characteristics, but comparison schools may not match AGR schools on unobserved characteristics such as schools' ability to properly implement AGR or instructor quality in the local hiring market. The long history of SAGE, AGR's precursor program that provided funding for reduced class sizes only, also limits the study. Previous school outcomes used for matching likely include SAGE impacts as well, which would bias AGR impacts toward zero. Finally, inconsistent testing patterns in Grades K-3 restricted the sample of AGR and non-AGR schools included in the growth analysis samples, potentially limiting how growth impact estimates can be generalized to schools not in the sample.

Section 10

Technical Appendix

Appendix A
Technical Appendix

Assessment Standardization and Equating

The analysis of reading growth and math growth uses assessment data from three local assessments: PALS, MAP, and STAR. While some readers may be familiar with these assessments' scaling and typical growth, other readers may lack this familiarity. To provide interpretability of results across assessments, we standardized assessment scores to have a mean of zero and standard deviation of one. For each test window, subject, and grade combination (e.g., Fall 2nd grade MAP), we take each individual's score, subtract the mean test score across all test takers, then divide by the standard deviation of the score across all test takers.

For the PALS assessment norms, we calculated the means and standard deviations across the entire state sample provided by DPI for the years 2012-13, 2013-14, and 2014-15 combined. This is the largest sample available for the calculation of means and standard deviations prior to the implementation of AGR. We did not include years after the start of AGR implementation as the impact of the program may also impact the norms. While the predecessor SAGE program may also impact the norms for the 2012-13 through 2014-15 years, the SAGE program originated prior to use of the PALS assessment across the state.

For the MAP assessment, we used 2015 national norms available from the vendor, NWEA.[25] These national norms included means and standard deviations for each subject, grade, and test window combination. NWEA created these 2015 norms from a national sample of data from fall of 2011 through spring of 2014. NWEA recently created 2020 norms, which NWEA calculated from a national sample of data from fall of 2015 through spring of 2018.[26] While the 2020 norms provide a more recent estimate of means and standard deviations throughout the tested MAP population, we continue to use 2015 norms for three main reasons. First, while Wisconsin represents only a fraction of the NWEA testing population, making it unlikely that any impacts of AGR may be present in the national sample of MAP norms, we want to exclude any possibility. Second, we used 2015 norms in prior years of the evaluation. Since 2020 norms are different from 2015 norms, we want to remain consistent in our estimates of AGR impact and not have any estimates of impact vary due to a change in norms. Third, the timeframe of the 2015 norms is relatively similar to that of the PALS norms we calculated for use in this analysis.

25 Thum, Y. M. & Hauser, C.H. (2015). NWEA 2015 MAP norms for student and school achievement status and growth. NWEA Research Report. Portland, OR: NWEA.
26 Thum, Y. M., & Kuhfeld, M. (2020). NWEA 2020 MAP Growth Achievement Status and Growth Norms for Students and Schools. NWEA Research Report. Portland, OR: NWEA.

We did not use STAR norms, because we equated STAR and MAP scores prior to standardization through a process called equipercentile equating. This process first identifies an individual's percentile score for each subject and test window on the STAR assessment. Next, that percentile score is matched to the same percentile score on the MAP assessment in the same subject, test window, and grade. The individual is then assigned the scale score on the MAP assessment that corresponds with the matched percentile score on the MAP assessment. We used the 2015 MAP norms to correspond MAP percentile scores to scale scores. This assigned MAP scale score is then standardized, as described previously.

Predicting Spring 2020 Assessment Scores

Before turning to the primary PSM analysis, we first describe how the 2020-21 evaluation differs from previous AGR evaluations. Due to the COVID-19 pandemic, Wisconsin schools engaged in almost no testing during Spring 2020. While the pandemic lessened the data available for the AGR evaluation, it did not lessen the state's need to understand how AGR impacts the state's academic achievement gap. To address the lack of Spring 2020 assessment scores due to the COVID-19 pandemic, we predicted Spring 2020 test scores using currently available data and past relationships between Fall, Winter, and Spring STAR and MAP assessment scores. The primary advantage of this strategy is that, with Spring 2020 scores, we can use our past evaluation methodology to fit 2019-20 into the overall evaluation framework, for both the 2020-21 evaluation and future evaluations. The primary limitation of the methodology is that, because we are predicting what Spring 2020 test scores would have been had COVID-19 not closed schools, the pandemic's impacts on learning, particularly on disparities in learning by socioeconomic status, are not possible to determine using the prediction methodology. Also, predictions are not possible for PALS, because schools typically do not administer PALS in the winter.

We estimate a predictive model that uses student math and reading scores from Fall 2019 and Winter 2020, student demographics, and school characteristics, to predict what Spring 2020 test scores would have been had the 2020 school year proceeded normally. Specifically, we use data from 2012-13 through 2018-19 in the following specification for each subject, math and reading, to estimate coefficients that we then use to calculate predicted 2019-20 spring test scores for each subject:

Equation (1)

$$Y_{isgy}^{sprin} = \lambda^{fm} Y_{isgy}^{fm} + \lambda^{fr} Y_{isgy}^{fr} + \lambda^{wm} Y_{isgy}^{(w\,m)} + \lambda^{wr} Y_{isgy}^{wr} + \beta X_{iy} + \pi Z_{sy} + \delta_{gy} + \varepsilon_{(isgy)}$$

where Y_{isgy}^{spring} is the standardized STAR or MAP spring test score in either subject for student *i* in school *s*, grade *g*, and academic year *y*. [27] On the right side of Equation (1), the superscripts *f, w, m,* and *r* refer to fall, winter, math, and reading, respectively, so that, for example, *f,m* denotes the fall math test score. X_{iy} represents a vector of student-level covariates, and Z_{sy} represents a vector of school-level covariates (see Table A1 below for a list of variables used in the predictions). We include grade-by-year fixed effects δ_{gy} to control for unobserved, statewide effects that vary by grade and/or academic year. For the final predictions of Spring 2020 assessment scores used in the analysis models, we use both Fall 2019 and Winter 2020 assessments to capture actual test score growth that had occurred before the COVID-19 school closures and use that actual growth trajectory to predict the growth trajectory that is most likely to have occurred between Winter and Spring 2020.

27 STAR and MAP scores are first equated using the methodology described on page A1.

TABLE A1

Characteristics of ACP - Related High School Practices

CONTROL VARIABLE	GROWTH	ATTENDANCE	DISCIPLINE
Student Demographics Gender, Race/Ethnicity*, Free/Reduced-price Lunch, English Learner, Special Education	✓	✓	✓
School Demographic Percentages Gender, Black, Hispanic, White, Other Race/ Ethnicity*, Free/Reduced-price Lunch, English Learner, Special Education	✓	✓	✓
School Population	✓	✓	✓
Local Description (City, Suburb, Town, Rural)*	✓	✓	✓
Student Fall Test Scores**	✓		
School Average Fall Test Scores	✓		

Note: * Due to collinearity, we omitted one Race/Ethnicity category and one Local Description category from the model. ** For math and reading models, both subject pretests are included.

To test the validity of using predicted Spring 2020 test scores in the impact evaluation, we replaced 2018-19 actual test scores with predicted scores and re-estimated analyses from the 2018-19 evaluation. We then compared both 2018-19 actual test scores and resulting impact analyses with the newly-calculated, predicted 2018-19 test scores and impact analyses. Both the test scores and estimated impacts matched well. Although we cannot know how well predicted Spring 2020 test scores match with actual scores, results from this year's evaluation, using actual test scores for 2012-13 through 2018-19, actual test scores for Fall 2019 and Winter 2020, and predicted test scores from Spring 2020, compare favorably to past results.[28]

Evaluation Design

In order to credibly estimate AGR impacts, we must address two primary challenges to identification. First, a plausible comparison or control group must be identified. Schools that receive AGR funding are different from schools statewide (see AGR Demographics above) because those selected for SAGE, and subsequently eligible for AGR, were required to meet certain thresholds of students receiving free and reduced-price lunch. Second, because all AGR schools previously participated in SAGE, total AGR impacts cannot be determined solely through changes over time in AGR schools' outcomes. In most evaluations, schools participating in a program (the treatment) are previously untreated, meaning that, under certain conditions, comparing pre-treatment and post-treatment outcomes results in plausible estimates of the treatment impact. For AGR, however, comparing pre- and post-treatment outcomes only provides estimates of the difference between the AGR and SAGE treatment impacts, not the AGR impact itself.

28 Results available upon request.

To find a plausible control group and identify the AGR impact, we use Propensity Score Matching (PSM). PSM addresses selection bias by choosing a control group with observable characteristics similar to those of the treatment group. As described above (see AGR Demographics), schools that receive AGR funding are observably different from other Wisconsin schools. This is because AGR targets funding to schools with higher percentages of students eligible for free or reduced-price lunch. Coincident with being located in higher poverty environments relative to their non-AGR counterparts, AGR schools have lower pre-program (2013) average test scores and attendance, and higher numbers of suspensions. As a result, naive comparisons of outcomes across non-AGR and AGR schools would find negative program impacts based only on program selection. To address this selection bias, PSM identifies Wisconsin schools that are observably similar to AGR schools in order to create an apples-to-apples comparison when estimating program impacts. Successful matching relies on both the quality of matches and overlap (or common support) of propensity scores between AGR and non-AGR schools.

Comparison schools with high percentages of students eligible for free or reduced-price lunch are not AGR participants for two primary reasons. First, poverty in those schools may have increased since the last SAGE eligibility period. Those schools would be eligible for AGR based on poverty thresholds but are ineligible because they did not participate in SAGE. To test this potential source of bias, we include school-specific time trends in robustness checks below. Impact estimates from these analyses are similar to those from our preferred models. Second, schools may have opted out of SAGE. Opt-out schools would be systematically different from AGR schools due to characteristics of the district or school. Although we cannot test for bias resulting from selection bias associated with opting into or out of SAGE, the final round of SAGE enrollment occurred in 2011-12, and many school and district characteristics, particularly those associated with administration, have since changed.

Despite limitations of PSM regarding unobserved characteristics, it represents the best available methodology given program rollout and available data. Below, we describe the choices of variables to include in the matching model, the overlap in propensity scores between AGR and non-AGR schools, and the covariate balance among the matched sample. In addition, we present multiple robustness checks to provide evidence of whether unobserved school characteristics might bias AGR impact estimates. The primary limitation of PSM is that it rests on the strong assumption that balancing AGR and non-AGR schools on observed characteristics also balances those schools on unobserved characteristics. The most typical method of addressing bias from fixed, unobserved characteristics would be to include school fixed effects in the estimation. For the AGR analysis, however, including school fixed effects would only allow comparisons of AGR to SAGE because all AGR schools previously participated in SAGE. The included robustness checks compare the report's main results to the results of various impact models with partial controls for unobserved school characteristics.

Finally, we present results from the Benjamini-Hochberg correction for multiple comparisons. These corrections adjust the p-values from impact estimates to account for the increased probability of finding statistically significant results due to the large number of models included in the report.

Propensity Score Matching

We estimated the probability of a school receiving AGR with the logit model of treatment shown below. The probability that a school participates in AGR, Pr(EverAGRs), is a function of an intercept term α, a vector of school-level covariates X_s, and a school-specific error term ε_s.

Equation (2)

$$ln\left[\frac{Pr(EverAGR_s)}{1-Pr(EverAGR_s)}\right] = \alpha + \beta X_s + \varepsilon_s$$

In the equation above, matching occurs at the school-level (defined by the grades included in the model, not necessarily all of the grades that a school contains) because AGR is a school-level treatment. We use this matching strategy for both the attendance and discipline models. For the models of test score outcomes, however, we match at the school-grade-year level due to inconsistent testing coverage both across and within schools. As described in Table 5 through Table 7 in the main report, during the 2015-16 through 2019-20 sample period, only a minority of schools used the PALS, MAP, and STAR tests.[29] Underlying Table 5 through Table 7 is even greater variation both across and within schools. Many schools began a new test and/or quit using a test in the middle of the sample period. Other schools tested some of Grades K-3 but not others, and yet others changed which grades they tested during the sample period. Due to this variation, it is not possible to build a sufficiently sized, consistent sample while matching at the school level. To provide DPI with the most complete and generalizable evaluation of AGR impacts, we prioritized the inclusion of as many AGR schools as possible. As a result, we chose to match all models of test outcomes at the school-grade-year level.[30] For these matches, we use school-year averages of demographic and academic characteristics due to instability in school-grade-year level averages, particularly in small schools, but match within school-grade-year to ensure that matches only occur between schools and grades that were tested in the same year. Sample sizes and the percentage of AGR schools and students included in the growth analyses are listed in Table A2 and Table A3 below.

TABLE A2

Number of Growth Analysis AGR Schools and Percentage of All AGR Schools by Grade and Year

	KINDERGARTEN		FIRST		SECOND		THIRD		OVERALL (K-3)	
YEAR	N	%	N	%	N	%	N	%	N	%
2015-16	87	98.9%	12	13.2%	28	30.8%	29	33.0%	33	34.4%
2016-17	203	51.7%	44	11.1%	160	40.2%	207	52.9%	226	55.4%
2017-18	166	42.3%	91	22.9%	155	38.8%	202	51.5%	221	54.0%
2018-19	145	36.8%	107	26.6%	158	39.3%	200	50.8%	212	51.5%
2019-20	N/A	N/A	100	25.0%	144	36.0%	161	40.9%	173	42.0%

29 Schools administered dozens of different types of tests across all grades. PALS, MAP, and STAR were the most common.

30 We tested models that limit the sample to schools that tested throughout 2012-13 to 2019-20, but these models omitted most AGR schools.

TABLE A3

Number of Growth Analysis AGR Students and Percentage of All AGR Students by Grade and Year

	KINDERGARTEN		FIRST		SECOND		THIRD		OVERALL (K-3)	
YEAR	N	%	N	%	N	%	N	%	N	%
2015-16	3,927	94.9	711	15.6%	1,477	31.5%	1,552	34.2%	3,740	27.1%
2016-17	8,989	48.9	1,926	10.0%	7,346	36.6%	10,022	51.4%	19,294	32.8%
2017-18	7,382	40.4	3,543	18.8%	6,443	33.6%	9,378	48.7%	19,364	33.8%
2018-19	6,209	33.9	4,356	23.2%	6,750	35.7%	8,667	47.3%	19,773	35.3%
2019-20	N/A	N/A	4,016	21.7%	5,757	31.0%	6,728	37.2%	16,501	29.9%

Specifying the Propensity Score Model

To determine which variables to include in the propensity score matching model above, we tested the influence of many demographic and academic variables. The final list of covariates appears in Table 1 in the main report. For each of the models, the most important matching variables measure the average outcome in a previous time period (pretests), such as the school's average test scores from the previous time period. The choice of pretest was complicated by both the level of matching (school or school-grade-year) and by the fact that AGR schools previously participated in SAGE. To the greatest extent possible, we aimed to remove previous program impacts from the matching model. Matching schools on post-program data risks biasing the results toward zero, because schools would be matched on previous-period outcomes that already include the treatment impact. However, at the beginning of our sample period, SAGE had been in operation for over 15 years, so it was not possible to include pre-program data. We used two strategies to address matching on post-program outcomes. For the attendance and discipline models, we matched once using school average attendance rate and suspension data from 2012-13, limiting the effect of including a post-program outcome to just one year. For the PALS and MAP/STAR testing models, we focused on growth instead of achievement. Focusing on growth lessens the impact of previous test scores, because, with appropriate pretest controls in the analysis model, the potential for growth is roughly equal regardless of initial pretest score.

In order to find the best PSM model to balance covariates across AGR and comparison schools, retaining as many school observations as possible, and stability of matches, we tested different matching algorithms, including caliper matching with various bandwidths, kernel matching, and Mahalanobis. For the analysis in the report, we used a kernel matching procedure that places higher weights on control observations nearest to a treatment observation and places successively lower weights on control observations as their distance from a treatment observation increases.[31]

31 Specifically, we used Stata's kmatch package with an Epanechnikov kernel and allowed Stata to select the optimal bandwidth.

Prior to matching, we limited the sample using two additional rules. First, we removed any schools that had participated in SAGE but never participated in AGR, including those that declined to participate in AGR. Second, we limited the testing models to schools that tested at least 75 percent of the relevant population in Grades K-3, following previous SAGE evaluations.[32] Table A4 illustrates the matching and subsequent analysis strategies for each outcome.

TABLE A4
Matching and Analysis Strategies

OUTCOME	GRADES	MATCHING LEVEL	MATCHING DATA	ANALYSIS YEARS
PALS Growth	K	School-grade-year	Fall 2012-13 through Fall 2018-19	2012-13 through 2018-19
MAP/STAR Reading Growth	1-3	School-grade-year	Fall 2012-13 through Fall 2019-20	2012-13 through 2019-20
MAP/STAR Math Growth	1-3	School-grade-year	Fall 2012-13 through Fall 2019-20	2012-13 through 2019-20
Absence Rate	K-3	School	2012-13	2013-14 through 2019-20
Suspension Rate	K-3	School	2012-13	2013-14 through 2019-20

When matching is successful, there is sufficient overlap in the propensity scores of treated (AGR) and comparison (non-AGR) schools to ensure that there is a plausible control group for the analysis. Figure A1 through Figure A4 display common support for PALS, Math, Reading, and attendance and discipline (which were matched together), respectively. Each figure shows the number of AGR and non-AGR schools by deciles of the propensity score distributions. For each of the outcomes, there are substantial numbers of non-AGR schools in most propensity score deciles and at least one control school in every decile.

32 The 75 percent threshold helps to ensure that students were tested for benchmarking purposes and not because they had been singled out for testing or had tested at another school before moving. See Meyer, R., Dokumaci, E., Sim, G., Steele, C., Suchor, K., & Vadas, J. (2015). SAGE program evaluation final report. University of Wisconsin-Madison, Value-Added Research Center.

FIGURE A1

Common Support for Matching – PALS Reading (2018-19)

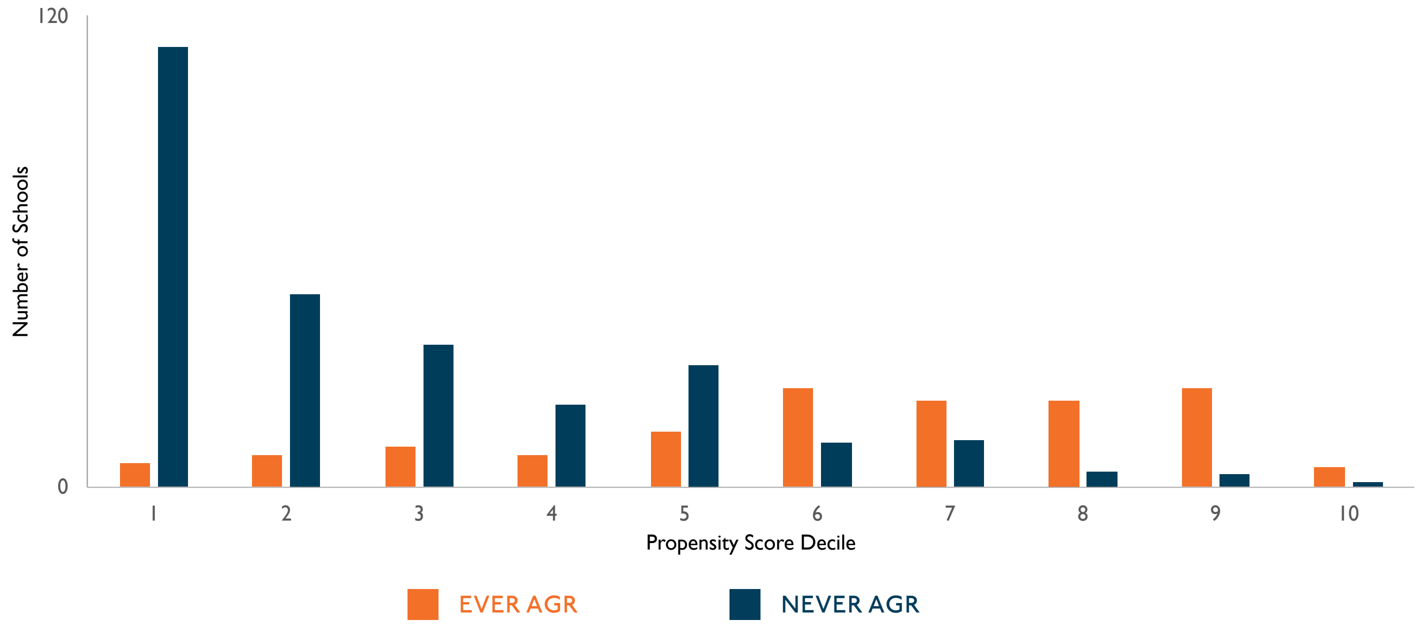


FIGURE A2

Common Support for Matching – MAP/STAR Math (2019-20)

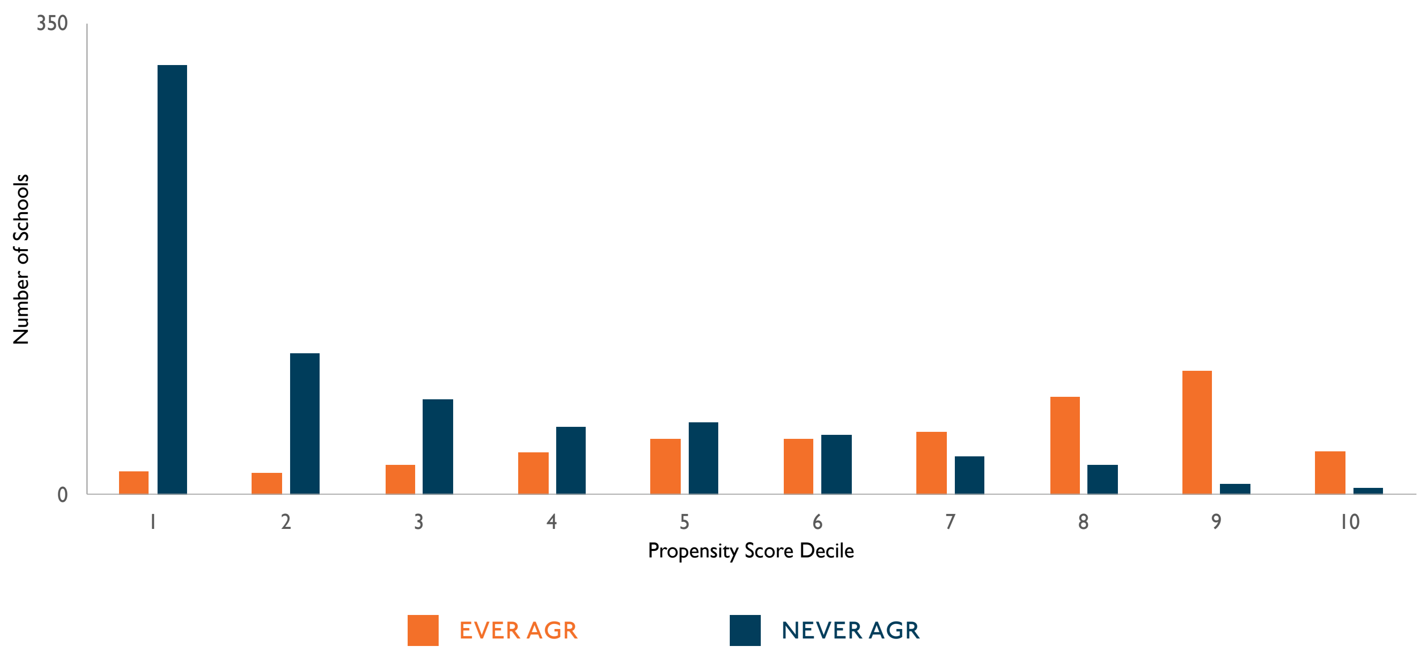


FIGURE A3
Common Support for Matching – MAP/STAR Reading (2019-20)

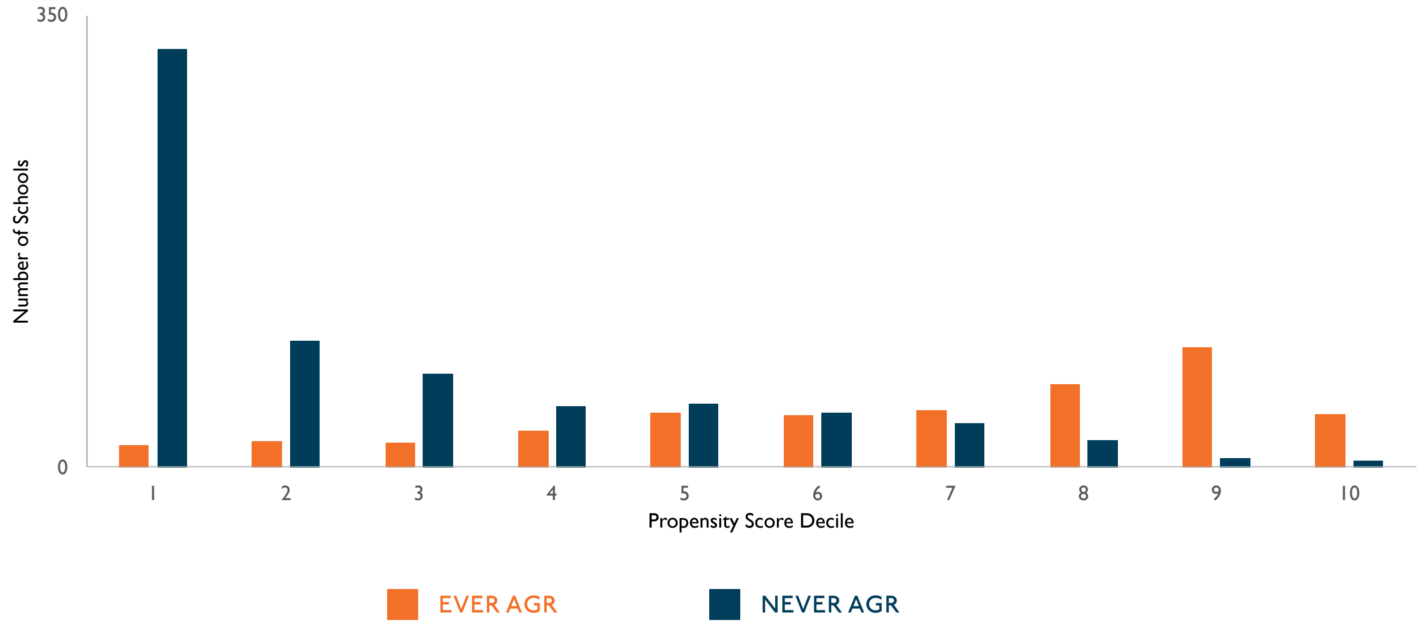
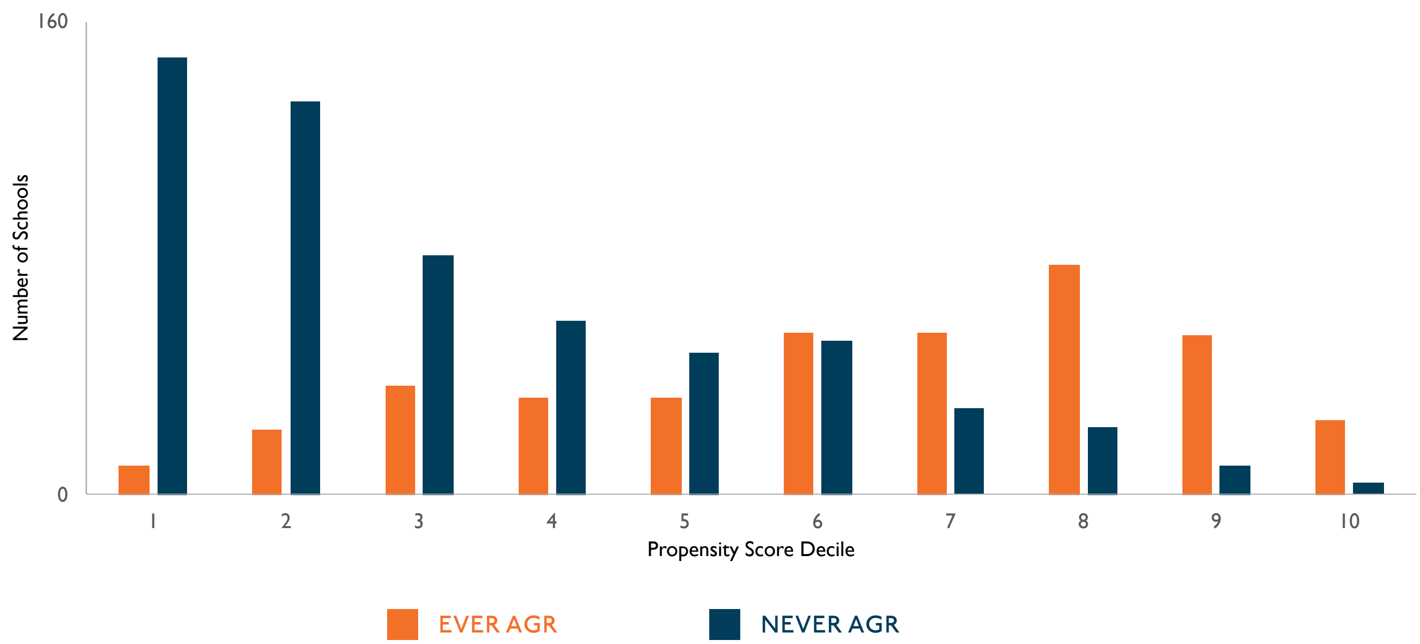


FIGURE A4
Common Support for Matching – Attendance and Discipline (2013-14)



Successful matching should also result in balanced covariates across the treatment and control groups. Table A5 through Table A8 describe student-level balance for each of the matched samples. In keeping with the recommendations of the What Works Clearinghouse (WWC), we assess equivalence using both the p-values from t-tests of differences in means and with standardized differences. The WWC specifies that standardized differences over 0.25 are signals of imbalance, and those between 0.05 and 0.25 require that the covariates be included as covariates in the impact analysis.[33] In Table A5 through Table A8, no standardized differences reach the 0.25 threshold, and we include all covariates in all impact analyses for double robustness.

TABLE A5

Balance of Matched Sample - PALS (from 2018-19)

		TREATMENT	CONTROL	P-VALUE (T/C DIFFERENCE)	EFFECT SIZE
N		98,384	99,454		
Fall PALS Score		-0.17	-0.14	0.00	0.03
	Std. Dev.	1.03	1.04		
School Fall PALS Score		-0.18	-0.15	0.00	0.08
	Std. Dev.	0.39	0.46		
Female		0.49	0.48	0.26	0.01
	Std. Dev.	0.50	0.50		
Black		0.14	0.16	0.00	0.06
	Std. Dev.	0.35	0.37		
Hispanic		0.15	0.14	0.00	0.02
	Std. Dev.	0.35	0.35		
Other Race		0.10	0.11	0.00	0.03
	Std. Dev.	0.30	0.31		
FRL		0.62	0.62	0.59	0.00
	Std. Dev.	0.49	0.49		
SPED		0.14	0.14	0.57	0.00
	Std. Dev.	0.35	0.35		
ELL		0.11	0.11	0.77	0.00
	Std. Dev.	0.31	0.31		
Urban		0.39	0.44	0.00	0.10
	Std. Dev.	0.49	0.50		

33 What Works Clearinghouse. (2020). Standards Handbook (Version 4.1). Retrieved from https://ies.ed.gov/ncee/wwc/handbooks.

TABLE A5, CONTINUED

	AGR	NON-AGR	P-VALUE (T/C DIFFERENCE)	EFFECT SIZE
Suburb	0.09	0.11	0.00	0.05
Std. Dev.	0.29	0.31		
Town	0.18	0.18	0.04	0.01
Std. Dev.	0.39	0.38		
School Population	244.95	244.87	0.89	0.00
Std. Dev.	102.49	101.69		
School Avg Teacher Salary	46,878.27	45,627.94	0.00	0.12
Std. Dev.	8,178.04	13,047.58		
School % Female	0.48	0.48	0.00	0.02
Std. Dev.	0.04	0.04		
School % Black	0.15	0.17	0.00	0.09
Std. Dev.	0.26	0.28		
School % Hispanic	0.15	0.14	0.00	0.04
Std. Dev.	0.20	0.17		
School % Other Race	0.10	0.11	0.00	0.07
Std. Dev.	0.12	0.16		
School % FRL	0.63	0.62	0.21	0.01
Std. Dev.	0.20	0.23		
School % SPED	0.15	0.15	0.00	0.04
Std. Dev.	0.05	0.05		
School % ELL	0.11	0.11	0.00	0.03
Std. Dev.	0.15	0.15		

TABLE A6

Balance of Matched Sample - MAP/STAR Math

	TREATMENT	CONTROL	P-VALUE (T/C DIFFERENCE)	EFFECT SIZE
N	171,052	172,528		
Fall Math Score	-0.04	-0.05	0.01	0.01
Std. Dev.	1.02	1.03		
Fall Reading Score	-0.20	-0.21	0.02	0.01
Std. Dev.	1.06	1.07		
School Fall Reading Score	-0.05	-0.05	0.01	0.01
Std. Dev.	0.39	0.43		
School Fall Math Score	-0.20	-0.21	0.00	0.02
Std. Dev.	0.39	0.42		
Female	0.49	0.48	0.70	0.00
Std. Dev.	0.50	0.50		
Black	0.24	0.22	0.00	0.06
Std. Dev.	0.43	0.41		
Hispanic	0.14	0.14	0.79	0.00
Std. Dev.	0.35	0.35		
Other Race	0.10	0.11	0.00	0.03
Std. Dev.	0.30	0.31		
FRL	0.66	0.64	0.00	0.03
Std. Dev.	0.47	0.48		
SPED	0.15	0.15	0.59	0.00
Std. Dev.	0.36	0.36		
ELL	0.10	0.11	0.00	0.04
Std. Dev.	0.30	0.31		
Urban	0.53	0.54	0.01	0.01
Std. Dev.	0.50	0.50		
Suburb	0.10	0.10	0.00	0.02
Std. Dev.	0.31	0.30		
Town	0.16	0.14	0.00	0.06
Std. Dev.	0.37	0.35		
School Population	232.69	229.62	0.00	0.03
Std. Dev.	93.02	93.29		

TABLE A6, CONTINUED

	AGR	NON-AGR	P-VALUE (T/C DIFFERENCE)	EFFECT SIZE
School Avg Teacher Salary	49,099.79	48,448.12	0.00	0.07
Std. Dev.	8,418.41	11,212.67		
School % Female	0.48	0.48	0.00	0.07
Std. Dev.	0.04	0.04		
School % Black	0.24	0.22	0.00	0.08
Std. Dev.	0.34	0.31		
School % Hispanic	0.14	0.14	0.15	0.01
Std. Dev.	0.18	0.17		
School % Other Race	0.10	0.11	0.00	0.08
Std. Dev.	0.11	0.15		
School % FRL	0.66	0.64	0.00	0.07
Std. Dev.	0.22	0.23		
School % SPED	0.16	0.16	0.00	0.05
Std. Dev.	0.05	0.05		
School % ELL	0.10	0.11	0.00	0.08
Std. Dev.	0.14	0.15		

TABLE A7
Balance of Matched Sample - MAP/STAR Reading

	TREATMENT	CONTROL	P-VALUE (T/C DIFFERENCE)	EFFECT SIZE
N	169,872	172,017		
Fall Math Score	-0.04	-0.05	0.10	0.01
Std. Dev.	1.02	1.03		
Fall Reading Score	-0.20	-0.21	0.11	0.01
Std. Dev.	1.06	1.07		
School Fall Reading Score	-0.05	-0.05	0.16	0.01
Std. Dev.	0.39	0.42		
School Fall Math Score	-0.20	-0.21	0.10	0.01
Std. Dev.	0.39	0.42		
Female	0.49	0.49	0.82	0.00
Std. Dev.	0.50	0.50		
Black	0.24	0.22	0.00	0.06
Std. Dev.	0.43	0.41		
Hispanic	0.14	0.14	0.03	0.01
Std. Dev.	0.35	0.35		
Other Race	0.10	0.11	0.00	0.04
Std. Dev.	0.30	0.31		
FRL	0.66	0.64	0.00	0.04
Std. Dev.	0.47	0.48		
SPED	0.15	0.15	0.85	0.00
Std. Dev.	0.35	0.36		
ELL	0.10	0.11	0.00	0.03
Std. Dev.	0.30	0.31		
Urban	0.53	0.54	0.01	0.01
Std. Dev.	0.50	0.50		
Suburb	0.11	0.10	0.00	0.02
Std. Dev.	0.31	0.30		
Town	0.16	0.14	0.00	0.06
Std. Dev.	0.37	0.35		
School Population	232.76	230.57	0.00	0.02
Std. Dev.	92.96	94.93		

TABLE A7, CONTINUED

	AGR	NON-AGR	P-VALUE (T/C DIFFERENCE)	EFFECT SIZE
School Avg Teacher Salary	49,116.44	48,462.79	0.00	0.07
Std. Dev.	8,412.62	11,344.05		
School % Female	0.48	0.48	0.00	0.06
Std. Dev.	0.04	0.04		
School % Black	0.24	0.22	0.00	0.08
Std. Dev.	0.34	0.31		
School % Hispanic	0.14	0.14	0.00	0.02
Std. Dev.	0.18	0.16		
School % Other Race	0.10	0.11	0.00	0.10
Std. Dev.	0.11	0.15		
School % FRL	0.66	0.64	0.00	0.07
Std. Dev.	0.22	0.23		
School % SPED	0.16	0.15	0.00	0.06
Std. Dev.	0.05	0.05		
School % ELL	0.10	0.11	0.00	0.07
Std. Dev.	0.14	0.14		

TABLE A8

Balance of Matched Sample – Attendance and Discipline

	TREATMENT	CONTROL	P-VALUE (T/C DIFFERENCE)	EFFECT SIZE
N	536,243	501,093		
School Attendance Rate 2012-13	0.95	0.95	0.00	0.03
Std. Dev.	0.02	0.02		
School Discipline Rate 2012-13	0.03	0.03	0.00	0.02
Std. Dev.	0.05	0.05		
Female	0.48	0.48	0.32	0.00
Std. Dev.	0.50	0.50		
Black	0.15	0.16	0.00	0.02
Std. Dev.	0.36	0.36		
Hispanic	0.16	0.14	0.00	0.04
Std. Dev.	0.37	0.35		
Other Race	0.10	0.11	0.00	0.01
Std. Dev.	0.30	0.31		
FRL	0.62	0.60	0.00	0.05
Std. Dev.	0.49	0.49		
SPED	0.16	0.15	0.00	0.01
Std. Dev.	0.36	0.36		
ELL	0.12	0.12	0.81	0.00
Std. Dev.	0.32	0.32		
Urban	0.42	0.42	0.00	0.02
Std. Dev.	0.49	0.49		
Suburb	0.09	0.12	0.00	0.10
Std. Dev.	0.28	0.32		
Town	0.18	0.21	0.00	0.07
Std. Dev.	0.39	0.41		
School Population	247.64	242.41	0.00	0.06
Std. Dev.	98.56	89.93		
Town	0.16	0.14	0.00	0.06
Std. Dev.	0.37	0.35		
School Population	232.76	230.57	0.00	0.02
Std. Dev.	92.96	94.93		

TABLE A8, CONTINUED

	AGR	NON-AGR	P-VALUE (T/C DIFFERENCE)	EFFECT SIZE
School Avg Teacher Salary	47,767.98	47,875.21	0.00	0.01
Std. Dev.	7,803.59	9,559.37		
School % Female	0.49	0.48	0.00	0.01
Std. Dev.	0.04	0.04		
School % Black	0.15	0.16	0.00	0.04
Std. Dev.	0.26	0.27		
School % Hispanic	0.15	0.14	0.00	0.07
Std. Dev.	0.20	0.17		
School % Other Race	0.09	0.08	0.00	0.04
Std. Dev.	0.10	0.08		
School % FRL	0.63	0.62	0.00	0.07
Std. Dev.	0.20	0.22		
School % SPED	0.15	0.15	0.00	0.10
Std. Dev.	0.05	0.05		
School % ELL	0.11	0.11	0.01	0.01
Std. Dev.	0.16	0.14		

Impact Analysis

After matching, we model impact estimates for both SAGE and AGR using the following, student-level specification:

Equation 3

$$Y_{isgy} = \alpha + \gamma_o SAGE_{sy} + \gamma_1 AGR_{sy} + \beta X_{iy} + \pi Z_{sy} + \delta_{gy} + \varepsilon_{isgy}$$

where Y_{isgy} is an outcome for student i in grade g, school s, and year y. $SAGE_{sy}$ and AGR_{sy} are indicators for whether a school received SAGE or AGR funding, respectively, in each year. X_{iy} represents a vector of student-level covariates, including lagged values of the outcome Y, and Z_{sy} represents a vector of school-level covariates. Grade-by-year fixed effects, δ_{gy}, are included to control for any unobserved, statewide effects that vary by grade and/or time.[34] All analysis variables are described in Table 2 in the main report. As described above, the models include all school-level variables from the PSM procedure as well as individual-level controls. For PALS, due to nonlinearity in the pre-post relationship, we include variables for both the fall pretest and a squared measure of the fall pretest.

34 PALS models, which only include kindergarten, and models that estimate differential effects by grade, contain only year fixed effects.

All models include weights generated by the kernel PSM procedure. Standard errors are clustered at the school-level. Models for PALS, MAP/STAR math and reading, and absence rate use Weighted Least Squares, and the suspension rate model, where the outcome is an indicator of whether a student received at least one suspension during the year, uses a logit specification. To account for the non-linearity of absence rate as an outcome, we first converted absence rates onto the standard normal distribution using a probit transformation. To provide meaningful results, we then use an inverse transformation of the raw impact estimates before reporting.

For reference, Table A9 provides information on the average and standard deviation of each of the outcomes of interest in 2019-20 using the appropriate and weighted analysis sample.

TABLE A9
Outcomes Summary Statistics, 2019-20

OUTCOME	MEAN	S.D.
PALS (2018-19)	-0.329	1.346
MAP/STAR Math	0.004	1.037
MAP/STAR Reading	-0.133	1.047
Absence Rate	0.057	0.065
Suspension Rate	0.025	0.157

Table A10 – Table A13 display detailed results for the main impact results that are displayed in Figure 12 – Figure 15 of the main report. Tables A14 – A17 display detailed results from comparisons of AGR and SAGE impacts, as shown above in Figures 16-19 of the main report. In each table, p-values are corrected to account for multiple estimates.

TABLE A10
Overall Impact of AGR on Math Growth

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (APPROX. MAP SCALE)	P-VALUE
MAP/STAR Math	First	0.010	0.14	0.838
	Second	0.013	0.18	0.730
	Third	-0.038	-0.52	0.124
	Overall (1-3)	-0.012	-0.16	0.672

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

TABLE A11
Overall Impact of AGR on Reading Growth

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (APPROX. PALS/MAP SCALE)	P-VALUE
PALS Reading (through 2018-19)	Kindergarten	0.105*	1.44	0.039
MAP/STAR Reading	First	0.007	0.11	0.900
	Second	0.017	0.26	0.498
	Third	-0.013	-0.20	0.508
	Overall (1-3)	0.001	0.02	0.968

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

TABLE A12
Overall Impact of AGR on Absences

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (APPROX. DAYS)	P-VALUE
Absence Rate	Kindergarten	0.41	0.7	0.133
	First	0.38	0.7	0.125
	Second	0.44*	0.8	0.039
	Third	0.41*	0.7	0.042
	Overall (K-3)	0.41	0.7	0.071

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

TABLE A13
Overall Impact of AGR on Discipline

OUTCOME	GRADE	IMPACT (PERCENTAGE POINTS)	P-VALUE
Suspension Rate	Kindergarten	-0.5	0.131
	First	-0.4	0.229
	Second	-0.4	0.238
	Third	-0.4	0.334
	Overall (K-3)	-0.4	0.178

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

TABLE A14
Impact of AGR Compared to SAGE on Math Growth

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (APPROX. MAP SCALE)	P-VALUE
MAP/STAR Math	First	0.045	0.61	0.208
	Second	0.035	0.47	0.239
	Third	-0.042	-0.58	0.104
	Overall (1-3)	0.000	0.00	0.998

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

TABLE A15
Impact of AGR Compared to SAGE on Reading Growth

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (APPROX. PALS/MAP SCALE)	P-VALUE
PALS Reading (through 2018-19)	Kindergarten	0.087	1.20	0.053
MAP/STAR Reading	First	0.042	0.61	0.255
	Second	-0.011	-0.17	0.731
	Third	-0.033	-0.50	0.125
	Overall (1-3)	-0.012	-0.18	0.688

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

TABLE A16
Impact of AGR Compared to SAGE on Absences

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (APPROX. DAYS)	P-VALUE
Absence Rate	Kindergarten	-0.42	-0.7	0.199
	First	-0.31	-0.5	0.247
	Second	-0.27	-0.5	0.254
	Third	-0.29	-0.5	0.224
	Overall (K-3)	-0.32	-0.6	0.222

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

TABLE A17
Impact of AGR Compared to SAGE on Discipline

OUTCOME	GRADE	IMPACT (PERCENTAGE POINTS)	P-VALUE
Suspension Rate	Kindergarten	0.0	0.978
	First	-0.2	0.675
	Second	-0.1	0.768
	Third	0.3	0.505
	Overall (K-3)	0.0	0.991

Note: P-values corrected to account for multiple estimates. * Statistically significant at the 0.05 level.

Subgroup Impact Analysis

Displayed below are results from differential impacts analysis by various demographic subgroups. Table A18 – Table A26 show the impact of AGR on math growth for each subgroup, relative to the same subgroup of students in similar non-AGR schools. In the main report, differential impact analysis is limited to by whether students receive free or reduced-price lunch (Figure 20), consistent with AGR's focus on reducing socioeconomic achievement gaps. Table A16 provides a more detailed description of the results in Figure 20. To provide readers with as much information as possible, Table A18 – Table A26 present differential impacts analysis for the following demographic subgroups: females, Black students, Hispanic students, Asian students, White students, students who are English Learners, students receiving special education services, and students who reside in cities.

For kindergarten PALS, the 2019-20 evaluation found significantly positive impacts of AGR for a variety of subgroups including female students (Table A19), Hispanic students (Table A21), Asian students (Table A22), English learners (Table A24), and students in cities (Table A26). The largest impacts found were for kindergarten English learners who had an average reading growth of 0.37 standard deviations, or 5.1 PALS score points, higher than English learners in similar, non-AGR schools.

Other than for PALS, few differences in impacts across subgroups were significant from both a statistical and policy standpoint. As shown in Table A20, relative to Black students at similar, non-AGR schools, Black students at AGR schools experienced a 1.4 day increase in absences. A similar, though smaller impact was estimated for city students (Table A26). Both Hispanic students (Table A21) and students who are English Learners (Table A24), two subgroups that have substantial overlap, had significantly fewer out-of-school suspensions. AGR students receiving special education services (Table A25) had lower math scores than their non-AGR peers.

TABLE A18

Impact of AGR on FRL Students – All Outcomes

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (SCALED)	P-VALUE
MAP/STAR Math	Overall (1-3)	-0.025	-0.34 points	0.340
PALS (through 2018-19)	Kindergarten	0.134*	1.84 points	0.038
MAP/STAR Reading	Overall (1-3)	-0.001	-0.02 points	0.957
Absence Rate	Overall (K-3)	0.35 percentage points	0.6 days	0.196
Suspension Rate	Overall (K-3)	-0.6 percentage points		0.126

TABLE A19
Impact of AGR on Female Students – All Outcomes

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (SCALED)	P-VALUE
MAP/STAR Math	Overall (1-3)	-0.005	-0.07 points	0.885
PALS (through 2018-19)	Kindergarten	0.103*	1.41 points	0.028
MAP/STAR Reading	Overall (1-3)	0.003	0.04 points	0.915
Absence Rate	Overall (K-3)	0.39 percentage points	0.7 days	0.098
Suspension Rate	Overall (K-3)	-0.2 percentage points		0.222

TABLE A20
Impact of AGR on Black Students – All Outcomes

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (SCALED)	P-VALUE
MAP/STAR Math	Overall (1-3)	-0.058	-0.79 points	0.213
PALS (through 2018-19)	Kindergarten	0.187	2.57 points	0.111
MAP/STAR Reading	Overall (1-3)	-0.039	-0.59 points	0.342
Absence Rate	Overall (K-3)	0.80* percentage points	1.4 days	0.040
Suspension Rate	Overall (K-3)	-0.8 percentage points		0.208

TABLE A21
Impact of AGR on Hispanic Students – All Outcomes

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (SCALED)	P-VALUE
MAP/STAR Math	Overall (1-3)	0.024	0.32 points	0.448
PALS (through 2018-19)	Kindergarten	0.253*	3.48 points	0.005
MAP/STAR Reading	Overall (1-3)	0.012	0.18 points	0.649
Absence Rate	Overall (K-3)	0.34 percentage points	0.6 days	0.219
Suspension Rate	Overall (K-3)	-0.8* percentage points		0.019

TABLE A22

Impact of Impact of AGR on Asian Students – All Outcomes

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (SCALED)	P-VALUE
MAP/STAR Math	Overall (1-3)	-0.021	-0.29 points	0.787
PALS (through 2018-19)	Kindergarten	0.240*	3.31 points	0.007
MAP/STAR Reading	Overall (1-3)	0.004	0.05 points	0.980
Absence Rate	Overall (K-3)	0.57 percentage points	1.0 days	0.317
Suspension Rate	Overall (K-3)	-0.1 percentage points		0.628

TABLE A23

Impact of AGR on White Students – All Outcomes

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (SCALED)	P-VALUE
MAP/STAR Math	Overall (1-3)	0.002	0.02 points	0.986
PALS (through 2018-19)	Kindergarten	0.053	0.73 points	0.189
MAP/STAR Reading	Overall (1-3)	0.015	0.22 points	0.446
Absence Rate	Overall (K-3)	0.33 percentage points	0.6 days	0.230
Suspension Rate	Overall (K-3)	0.0 percentage points		0.994

TABLE A24

Impact of AGR on English Learner Students – All Outcomes

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (SCALED)	P-VALUE
MAP/STAR Math	Overall (1-3)	0.015	0.20 points	0.788
PALS (through 2018-19)	Kindergarten	0.370*	5.08 points	0.000
MAP/STAR Reading	Overall (1-3)	0.014	0.21 points	0.705
Absence Rate	Overall (K-3)	0.33 percentage points	0.6 days	0.234
Suspension Rate	Overall (K-3)	-0.5* percentage points		0.027

TABLE A25

Impact of AGR on Special Education Students – All Outcomes

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (SCALED)	P-VALUE
MAP/STAR Math	Overall (1-3)	-0.060*	-0.82 points	0.038
PALS (through 2018-19)	Kindergarten	0.083	1.14 points	0.313
MAP/STAR Reading	Overall (1-3)	-0.002	-0.04 points	0.984
Absence Rate	Overall (K-3)	0.41 percentage points	0.7 days	0.136
Suspension Rate	Overall (K-3)	-1.1 percentage points		0.098

TABLE A26

Impact of AGR on City Students – All Outcomes

OUTCOME	GRADE	IMPACT (STANDARDIZED)	IMPACT (SCALED)	P-VALUE
MAP/STAR Math	Overall (1-3)	-0.021	-0.28 points	0.591
PALS (through 2018-19)	Kindergarten	0.206*	2.83 points	0.016
MAP/STAR Reading	Overall (1-3)	-0.014	-0.20 points	0.673
Absence Rate	Overall (K-3)	0.64* percentage points	1.1 days	0.016
Suspension Rate	Overall (K-3)	-0.5 percentage points		0.237

Robustness to Alternative Estimation Strategies

To assess the robustness of our findings, we tested several alternative estimation strategies that attempt to address limitations of the matching and estimation strategies described above. These strategies include school fixed effects, school random effects, and school-specific time trends, shown in Table A27 through Table A31 below. In each of the tables, Column (1) displays the results from the preferred specification used in the main analysis above. Columns (2)-(4) of the tables display separate robustness checks.

The matching and estimation strategies described above rely on the assumption that schools matched on observable characteristics (e.g., test scores, demographics) are also matched on unobservable characteristics (e.g., schools' ability to implement AGR, teacher quality available in the local hiring market) that might be related to both outcomes and SAGE/AGR participation, and therefore bias impact estimates. However, there is no way to test this assumption. Including school fixed effects in the estimation would control for differences in unobservable characteristics between schools by comparing outcomes before and after AGR implementation within the same school. For the AGR analysis, however, including school fixed effects only allows comparisons of AGR to SAGE because all AGR schools previously participated in SAGE. With school fixed effects, comparisons to non-SAGE, non-AGR comparison schools would be impossible, because comparison schools, whose program participation does not change over the sample period, would not contribute to the AGR impact estimate. Nevertheless, comparing the AGR-SAGE difference from the preferred specification to a specification with school fixed effects provides useful information about the extent that unobservable school characteristics may bias estimations. To that end, in Table A27 - Table A31, Column (2) contains results of school fixed effects regressions. These results are qualitatively similar to the preferred specification in Column (1), although for PALS the difference between AGR and SAGE is less than half that of the preferred specification.

As an alternative to fixed effects, we also include school-specific random effects, which produce a weighted average of between-school and within-school effects.[35] Random effects, however, do not allow for variation in weights within schools, which occurs when matching testing outcomes within year and grade. As discussed above, our preferred matching strategy enables us to significantly increase the sample and improve generalizability by including schools that did not consistently test throughout the sample period or across Grades 1-3. Conversely, both attendance and discipline outcomes are available statewide in every year, allowing a less restrictive matching strategy that gives control schools the same weights in every year. Column (3) of Table A27 - Table A31 shows results from regressions that include school random effects. For both attendance and discipline, key coefficients are qualitatively similar to those from the preferred specification in Column (1).

Finally, we test for the presence of time trends in outcomes that may differ between AGR/SAGE and control schools and bias results. For example, If AGR/SAGE schools are more likely on positive trajectories unrelated to their participation in the program, estimates of AGR and SAGE impacts would be biased upward. We chose not to include school-specific time trends in our preferred specification because these school trends could be the result of SAGE and AGR, and there is no method to differentiate between unrelated trends and program impacts. However, in Table A27 - Table A31 we include estimates from regressions with school-specific linear time trends to provide readers with as much information as possible. In general, including trends has only small impacts on estimated impacts. For PALS (Table A27), the SAGE impact and the AGR impact decrease. For reading (Table A29), estimated AGR impacts increase substantially, while time trends have little effect on attendance or discipline impacts (Table A30 and Table A31).

35 Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics methods and applications. Cambridge, UK: Cambridge University Press, p. 711.

TABLE A27

Robustness to Alternative Estimation Strategies – PALS (from 2018-19)

	(1)	(2)	(3)	(4)
SAGE vs. non	0.180		N/A	0.053
p-value	0.387		N/A	0.095
AGR vs. non	0.105		N/A	0.086
p-value	0.004		N/A	0.077
AGR vs. SAGE	0.087	0.030	N/A	0.033
p-value	0.011	0.213	N/A	0.217
Individual controls	Yes	Yes	Yes	Yes
School-level controls	Yes	Yes	Yes	Yes
School fixed effects	No	Yes	No	No
School random effects	No	No	Yes	No
School-specific time trends	No	No	No	Yes

TABLE A28

Robustness to Alternative Estimation Strategies – MAP/STAR Math

	(1)	(2)	(3)	(4)
SAGE vs. non	-0.012		N/A	-0.011
p-value	0.430		N/A	0.643
AGR vs. non	-0.012		N/A	-0.004
p-value	0.497		N/A	0.907
AGR vs. SAGE	0.000		N/A	0.007
p-value	0.983		N/A	0.758
Individual controls	Yes	Yes	Yes	Yes
School-level controls	Yes	Yes	Yes	Yes
School fixed effects	No	Yes	No	No
School random effects	No	No	Yes	No
School-specific time trends	No	No	No	Yes

TABLE A29

Robustness to Alternative Estimation Strategies – MAP/STAR Reading

	(1)	(2)	(3)	(4)
SAGE vs. non	0.013		N/A	0.032
p-value	0.351		N/A	0.188
AGR vs. non	0.002		N/A	0.071
p-value	0.875		N/A	0.038
AGR vs. SAGE	-0.010		N/A	0.040
p-value	0.463		N/A	0.026
Individual controls	Yes	Yes	Yes	Yes
School-level controls	Yes	Yes	Yes	Yes
School fixed effects	No	Yes	No	No
School random effects	No	No	Yes	No
School-specific time trends	No	No	No	Yes

TABLE A30

Robustness to Alternative Estimation Strategies – Absences

	(1)	(2)	(3)	(4)
SAGE vs. non	0.105		N/A	0.086
p-value	0.001		N/A	0.031
AGR vs. non	0.062		N/A	0.019
p-value	0.016		N/A	0.748
AGR vs. SAGE	-0.043		N/A	-0.067
p-value	0.061		N/A	0.042
Individual controls	Yes	Yes	Yes	Yes
School-level controls	Yes	Yes	Yes	Yes
School fixed effects	No	Yes	No	No
School random effects	No	No	Yes	No
School-specific time trends	No	No	No	Yes

TABLE A31

Robustness to Alternative Estimation Strategies – Suspensions

	(1)	(2)	(3)	(4)
SAGE vs. non	-0.005		N/A	-0.003
p-value	0.018		N/A	0.322
AGR vs. non	-0.005		N/A	-0.003
p-value	0.039		N/A	0.474
AGR vs. SAGE	-0.001		N/A	0.000
p-value	0.803		N/A	0.926
Individual controls	Yes	Yes	Yes	Yes
School-level controls	Yes	Yes	Yes	Yes
School fixed effects	No	Yes	No	No
School random effects	No	No	Yes	No
School-specific time trends	No	No	No	Yes

Multiple Comparisons Analysis

Estimating multiple impact models, as this report does, increases the likelihood for false positives – results that are statistically significant due to random chance rather than actual program impacts. For example, a 0.05 significance level implies that 5 percent of statistically significant estimates are produced by random chance. To adjust for potential false positives, we apply the Benjamini-Hochberg procedure, a common method of correcting for multiple comparisons by accounting for the total number of statistical tests as well as the strength of the estimates.

According to the Benjamini-Hochberg procedure, impact estimates are ranked in ascending order of p-values. We then calculate a critical value equal to the rank multiplied by a false discovery rate (chosen here to be 5 percent), divided by the total number of comparisons. For each estimate to be statistically significant, its p-value must be less than the critical value. In addition to the critical value, to aid in interpretation for readers accustomed to the 0.05 threshold for statistical significance, we calculate an adjusted p-value from the same formula used to produce the critical value. Full results from the Benjamini-Hochberg procedure are available upon request.

Examination of MAP and STAR for Kindergarten Reading Growth

Due to the lack of Spring 2020 PALS results, the evaluation did not include growth estimation for kindergarten reading through the 2019-20 school year, and instead the evaluation relied on PALS results from the prior year's evaluation (through 2018-19). To supplement the analysis of grades 1-3 reading growth using MAP and STAR assessment results, the evaluation also examined kindergarten reading growth through 2019-20 using MAP and STAR.

In years prior to 2019-20, using PALS for kindergarten growth analysis provided a larger sample than MAP and STAR combined. Coverage on the MAP and STAR reading assessments only reached 20 percent of all Wisconsin schools, and 20 percent of AGR schools, beginning in 2018-19. As a result, we limit the examination of MAP and STAR reading growth for kindergarten to 2018-19 and 2019-20. Similar to the analysis of growth in grades 1-3, the analysis of growth in kindergarten also equated MAP and STAR scores and predicted the Spring 2020 assessment results using Fall and Winter assessment results. Unlike the analysis in grades 1-3, the analysis in kindergarten created predictions based on data from only 2018-19.

After creation of predicted Spring 2020 scores, the analysis of kindergarten reading growth used a similar methodology to the reading growth analysis of grades 1-3: school-level propensity score matching within each year (2018-19 and 2019-20) followed by a linear regression to estimate AGR impacts using the matched sample, weighting by match strength.

Matching for the MAP/STAR kindergarten reading analysis provided a similar sample of students in AGR schools and non-AGR schools, as seen in Figure A5 and Table A32. Figure A5 shows the common support of propensity scores for AGR schools and non-AGR schools and illustrates that there is sufficient overlap across all ten deciles. Table A32 shows a successful balance of covariates used in matching, as none of the variables exceed a standardized difference of 0.25. As with other analyses, we include all covariates in the impact regression for double robustness.

Matching for the MAP/STAR kindergarten reading analysis provided a similar sample of students in AGR schools and non-AGR schools, as seen in Figure A5 and Table A32. Figure A5 shows the common support of propensity scores for AGR schools and non-AGR schools and illustrates that there is sufficient overlap across all ten deciles. Table A32 shows a successful balance of covariates used in matching, as none of the variables exceed a standardized difference of 0.25. As with other analyses, we include all covariates in the impact regression for double robustness.

FIGURE A5

Common Support for Matching – MAP/STAR Reading (2019-20)

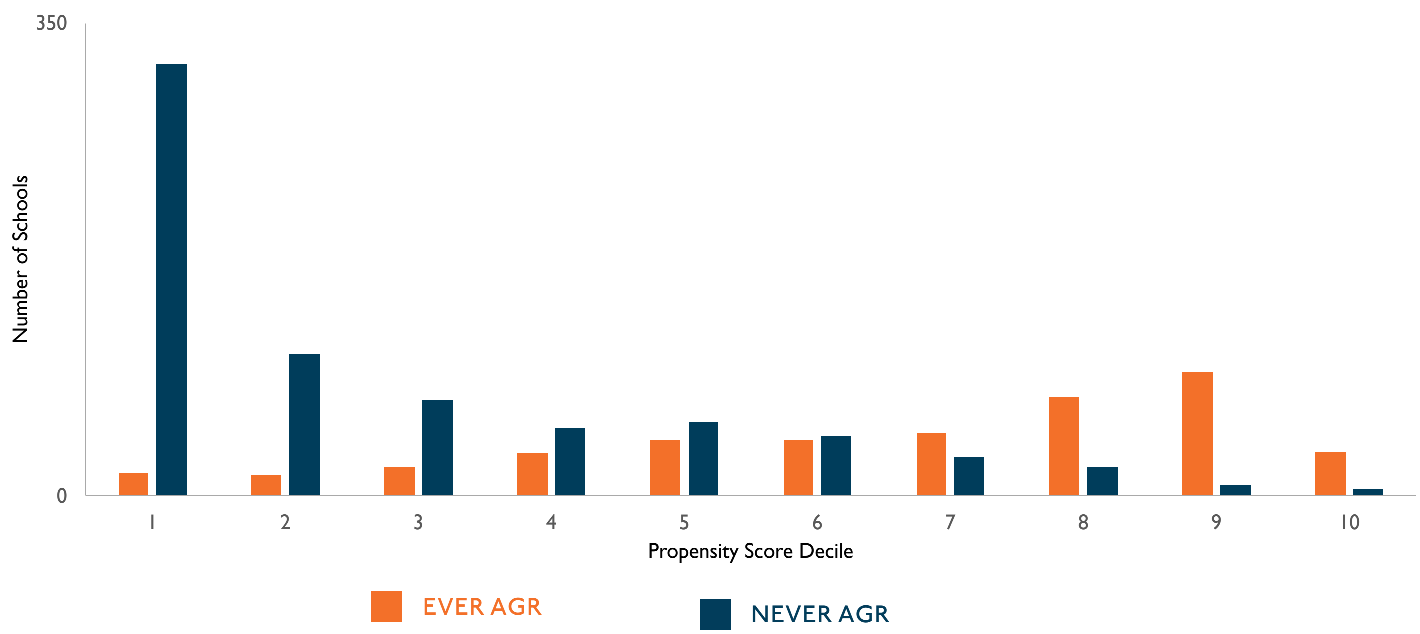


TABLE A32

Balance of Matched Sample – MAP/STAR Reading Kindergarten

	AGR	NON-AGR	P-VALUE (T/C DIFFERENCE)	EFFECT SIZE
N	6,863	7,268		
Fall Reading Score	-0.08	-0.13	0.02	0.06
Std. Dev.	0.97	0.95		
School Fall Reading Score	-0.09	-0.14	0.00	0.12
Std. Dev.	0.38	0.39		
Female	0.48	0.49	0.63	0.01
Std. Dev.	0.50	0.50		
Black	0.42	0.40	0.14	0.04
Std. Dev.	0.49	0.49		
Hispanic	0.13	0.12	0.32	0.02
Std. Dev.	0.33	0.32		
Other Race	0.12	0.13	0.09	0.04
Std. Dev.	0.33	0.34		
FRL	0.74	0.74	0.34	0.02
Std. Dev.	0.44	0.44		
SPED	0.17	0.16	0.47	0.02
Std. Dev.	0.37	0.37		
ELL	0.07	0.07	0.27	0.02
Std. Dev.	0.25	0.26		
City	0.69	0.68	0.55	0.01
Std. Dev.	0.46	0.47		
Suburb	0.13	0.12	0.12	0.03
Std. Dev.	0.33	0.32		
Town	0.07	0.08	0.53	0.01
Std. Dev.	0.26	0.27		
School Population	53.14	54.86	0.00	0.08
Std. Dev.	20.98	21.05		
School Avg Teacher Salary	52,743.75	50,996.74	0.00	0.22
Std. Dev.	7,894.65	7,724.41		

TABLE A32, CONTINUED

	AGR	NON-AGR	P-VALUE (T/C DIFFERENCE)	EFFECT SIZE
School % Black	0.42	0.40	0.08	0.04
Std. Dev.	0.39	0.37		
School % Hispanic	0.13	0.12	0.03	0.04
Std. Dev.	0.14	0.11		
School % Other Race	0.12	0.14	0.00	0.11
Std. Dev.	0.11	0.11		
School % FRL	0.74	0.74	0.20	0.03
Std. Dev.	0.23	0.23		
School % SPED	0.17	0.16	0.00	0.11
Std. Dev.	0.09	0.06		
School % ELL	0.07	0.08	0.01	0.07
Std. Dev.	0.11	0.12		

Table A33 shows the results from the impact regression, using the matched sample created in the previous step. The overall impact of AGR on kindergarten reading is one-tenth of a standard deviation growth, or approximately 1.3 points on the MAP assessment. While this result is not statistically significant, the estimated impact is nearly identical to the estimated impact of AGR in kindergarten for reading using the PALS assessment (0.104 standard deviations versus 0.105 standard deviations). Results by subgroup are also generally similar to previous results using the PALS assessment, though none are statistically significant from zero. This examination, using MAP/STAR results instead of PALS, provides further evidence of a robust impact of AGR on kindergarten reading growth.

TABLE A33

Impacts of AGR Overall and by Subgroup – MAP/STAR Reading Kindergarten

SUBGROUP	IMPACT (STANDARDIZED)	IMPACT (APPROX. MAP SCALE)	P-VALUE
Overall	0.104	1.34	0.070
Female	0.086	1.10	0.122
Asian	0.173	2.22	0.132
Black	0.109	1.40	0.164
Hispanic	0.142	1.83	0.087
White	0.073	0.93	0.142
Other Race/Ethnicity	0.102	1.30	0.151
Free/Reduced Lunch Students	0.104	1.33	0.110
EL Students	0.166	2.14	0.134
Special Ed. Students	0.050	0.64	0.334
City Students	0.102	1.31	0.122

WEC

Supporting PreK-12