WEC
Supporting PreK-12

EVALUATION OF

WISCONSIN'S ACHIEVEMENT GAP REDUCTION PROGRAM

JED RICHARDSON
GRANT SIM
BRIE CHAPA

AUGUST 2019

** The section Differences in Outcomes by Strategy, including Tables 23-26, was updated in June 2020 due to calculation errors in the original report.*

TABLE OF CONTENTS

Executive Summary **4**

Introduction **7**
- This Evaluation 9

Evaluation Data and Methodology **10**
- Data 10
- Identifying Comparison Schools 11
- Limitations 14

AGR Demographics **15**
- Testing Patterns and Growth Analysis Samples 17

AGR Implementation **21**

AGR Impacts **22**
- Overall Impacts 22
- Impacts Compared to SAGE 24
- Impacts by Demographics 26
- Differences in Outcomes by Strategy 30

School Board Report Findings **35**

End-of-Year Report Findings **37**

Summary/Conclusions **45**

Technical Appendix **47**

Survey Appendix **70**

EXECUTIVE SUMMARY

The Achievement Gap Reduction (AGR) program is an initiative of the Wisconsin Department of Public Instruction (DPI) 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 this funding, schools must implement one or more strategies in each participating grade:

- 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 two 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 possesses 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.[1]

This report presents the results of an evaluation completed by the Wisconsin Evaluation Collaborative (WEC) within the Wisconsin Center for Education Research at the University of Wisconsin–Madison. The goal of the 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 these 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 achievement gaps between low-income students and their higher-income peers?
    b. 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?

[1] 2015 Wisconsin Act 53. Wisconsin Senate. Section 118.44.

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 provide biased results. To address this selection bias, WEC used a two-part statistical method in order 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 used propensity score matching to identify non-AGR Wisconsin schools that were similar to those receiving AGR funding. These observationally similar schools then acted as a comparison group for the second part of the analysis, estimating the impact of AGR through multivariate regression techniques.

How are AGR schools implementing the program?

In 2017-18, the most recent year of data, 409 schools implemented the AGR program, serving over 75,000 students in kindergarten through third grades. As previously noted, schools could implement any combination of three different strategies to fulfill AGR obligations: reduced class size, instructional coaching, and/or tutoring.

- Over 60 percent of schools utilized multiple strategies–28.5 percent of schools implemented reduced class size and instructional coaching and 21.5 percent of schools implemented all three.
- Single strategies were employed less frequently, although 18.6 percent of schools implemented instructional coaching alone and 17.3 percent of schools implemented reduced class size alone.
- Comparatively few schools used tutoring as a strategy or in combination with one of the 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. Results from this analysis included:

- An estimated positive and significant impact of the AGR program on statewide reading growth in kindergarten as measured by the PALS assessment. AGR is associated with a 0.12 standard deviation increase in PALS scores relative to similar, non-AGR schools.
- Small and not statistically significant impact of the AGR program on statewide reading and math growth in Grades 1-3 as measured by the MAP and STAR assessments.
- Small and not statistically significant impacts of the AGR program on statewide attendance and behavior as measured by out-of-school suspension rates.

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

- Estimated large, positive, and significant impacts of the AGR program on kindergarten reading growth for English learners, Hispanic students, Asian students, students in urban settings, and low-income students.
- Estimated positive and significant impacts of the AGR program on behavior, as measured through a reduction in suspensions, for English learners and Hispanic students.

AGR program impacts compared to the previously implemented SAGE program were also examined. Results found an estimated positive and significant impact of the AGR program compared to SAGE on kindergarten reading growth but otherwise found few differences.

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

The evaluation provided preliminary evidence of the associations between outcomes and the AGR strategies schools chose. Results included:

- Relative to the reduced class size strategy, tutoring was associated with increased growth in math and reading for Grades 1-3.
- Reduced suspensions associated with reduced class size relative to other strategies.

Future evaluations will further explore relationships between strategies and impacts, calculate AGR's impacts on the statewide achievement gap, and estimate impacts of AGR accumulated throughout Grades K-3 using the third grade, statewide Forward Exam.

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. Starting in 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 AGR 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 plans to completely phase out 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 to be distributed to participating schools. 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 two 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 possesses 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]

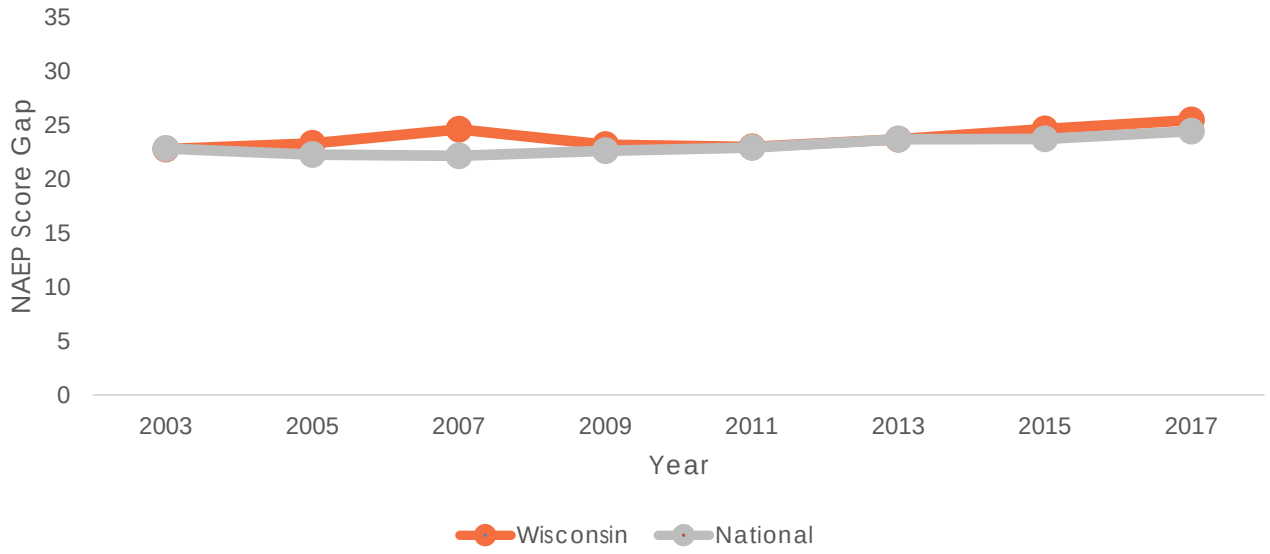
The AGR program prioritizes achievement gap reduction for economically disadvantaged students. Indeed, National Assessment of Educational Progress (NAEP) assessment scores show

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

a persistent gap in performance between economically disadvantaged students and non-economically disadvantaged students in Wisconsin.

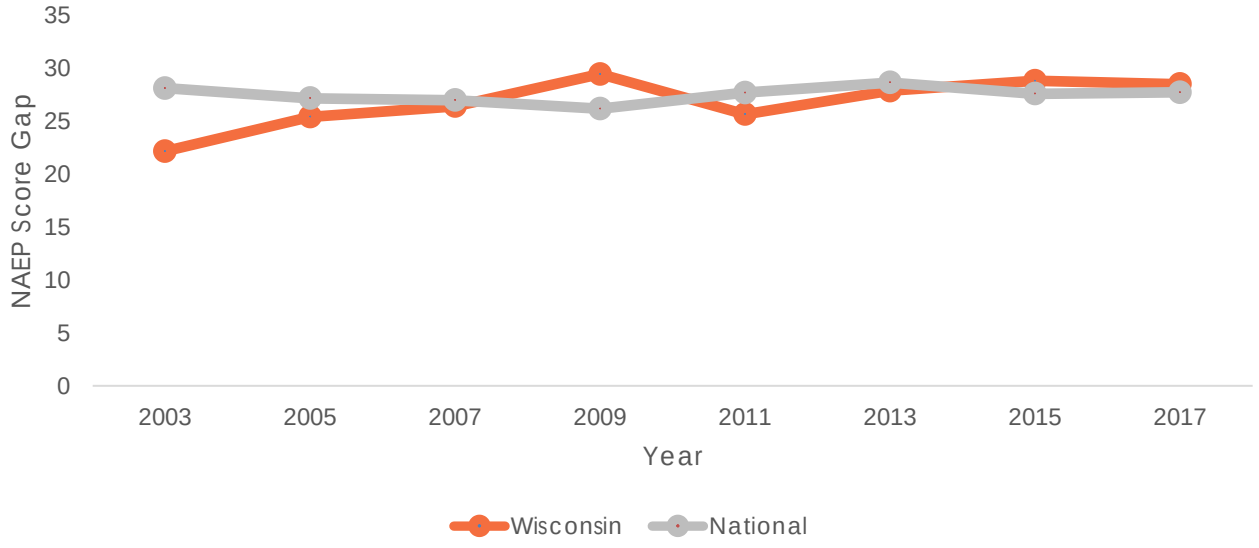
Figure 1 and Figure 2 show the economically disadvantaged achievement gap on the NAEP assessment in math and reading, respectively, for fourth grade. Both figures cover the years 2003 through 2017 and provide gaps for both Wisconsin and the nation as a whole. As displayed, neither Wisconsin nor the nation made substantial progress in reducing this achievement gap over the time period.

Figure 1 | Average Math NAEP Score Gap Between Economically Disadvantaged and Non-Economically Disadvantaged Students in Grade 4, 2003-2017



Source: NAEP Data Explorer. https://www.nationsreportcard.gov/ndecore/

Figure 2 | Average Reading NAEP Score Gap Between Economically Disadvantaged and Non-Economically Disadvantaged Students in Grade 4, 2003-2017



Source: NAEP Data Explorer. https://www.nationsreportcard.gov/ndecore/

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 the results from this initial evaluation of the AGR program from the 2015-16 through 2017-18 school years.

To serve as a foundation for the evaluation, WEC worked in collaboration with DPI to develop the following 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 these 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 achievement gaps between low-income students and their higher-income peers?
    b. 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?

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 compared to the state overall to provide context for later findings. A section on AGR implementation describes strategy usage. The AGR impacts section provides the results of analyses of AGR impacts on math growth, reading growth, attendance, and discipline. 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 2017-18 reports by AGR districts and schools. The End-of-Year Report findings provide the results from the 2017-18 survey of AGR schools. The final section of the report includes a summary of findings and plans for future evaluations. This report also contains two appendices, a technical appendix which provides further details on statistical methodology, and an appendix including the instrument for the 2017-18 End-of-Year Report survey.

EVALUATION DATA & METHODOLOGY

In order to understand how AGR impacts student achievement, attendance, and discipline 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 show 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 2017-18 academic years. In addition, for purposes of comparison, the evaluation includes K-3 students at subsets of non-AGR, non-SAGE schools.

Data

In order 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 2017-18. 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.[3] School AGR strategies are from responses to DPI's End-of-Year Report survey and required school board reports that AGR-funded schools are required to submit.

- **Demographic characteristics** include gender, race/ethnicity, English learner status, special education status, and low-income or economic 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 and spring administrations of the Phonological Awareness Literacy Screening (PALS), MAP, and STAR. For Grades 1-3, MAP and STAR scores were equated and combined into a single test measure in order to include a sufficient student sample.
- **Attendance data** consist of total days absent and total possible attendance days. The associated outcome variable is the absence rate or the total days absent divided by the total possible attendance days.

[3] 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/.

- **Discipline data** consist of the number of out-of-school suspensions. The associated outcome variable, the suspension rate, is an indicator that is one for students with at least one out-of-school suspension during the school year and zero for those who had not been suspended. We use this discipline outcome as a proxy for student behavior throughout the evaluation.
- **Enrollment data** include school attended and grade.
- **Strategies data** include information from AGR school board reports and responses to the End-of-Year Report survey and indicating whether each strategy (class size reduction, instructional coaching, and one-to-one tutoring) were used in each year, grade, subject, and semester. For the strategy analysis, we aggregated strategy data for each school-grade-year. Grades within schools were counted as using a strategy if they indicated use for any subject or semester during the school year.
- **School-level data** include SAGE and AGR funding by year, teacher average salaries, and school location (city, suburb, town, rural).

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 Tables 5-7). Instead, we chose to match at the school-grade-year level using each school's fall data.[4]

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. 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 provided the best balance and sample retention. Covariate balance tables can be found in Tables 44-47 in the Technical Appendix.

[4] This decision is described in detail in the Technical Appendix.

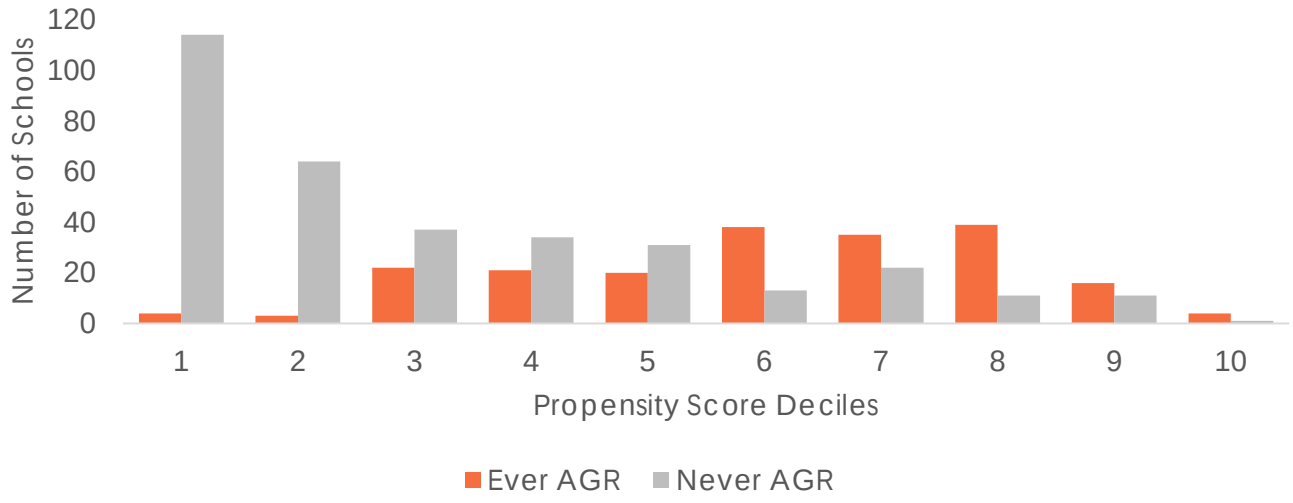
Table 1 | Propensity Score Matching Controls by Analysis Type

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 from the model ** For PALS, only the PALS reading pretest is included, due to low participation in the MAP/STAR math 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 2017-18. In each decile of the propensity score distribution, there is at least one comparison (untreated) school. Most deciles have more than 10 comparison schools, showing sufficient overlap for the analysis. Overlap for all models can be found in the Technical Appendix, Figures 10-13.

Figure 3 | Common Support for Matching – MAP/STAR Math (2017-18)



After matching, 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 Analysis	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	✓	✓	✓
Average Teacher Salary	✓	✓	✓
Local Description (City, Suburb, Town, Rural)	✓	✓	✓
Student Fall Test Scores**	✓		
Student Fall Test Scores Squared***	✓		
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 from the model. ** For math and reading models, both subject pretests are included. For PALS, only the PALS reading pretest is included, due to low participation in the MAP/STAR math exam in kindergarten. *** PALS only.

For further information on methodology and analysis, refer to the Technical Appendix.

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. These limitations and our efforts to address them are described in detail in the Technical Appendix.

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 were 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 was 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.

To the extent that the first two limitations 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). Differences in outcomes by strategy should not be considered causal, because the sample only includes AGR schools that are allowed to select their own strategies. If 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.

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 the tested population of AGR schools used for the evaluation is observationally similar to the untested sample of AGR schools (see Figures 7-9), it is not possible to know whether schools' choices of tests are related to outcomes and participation in AGR.

AGR DEMOGRAPHICS

This section and the sections that follow present the evaluation results aligned to the guiding 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 three 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. A final, small number of schools will join the program in 2018-19.

Table 3 | Number of AGR Schools by Grade and Year

Grade	2015-16	2016-17	2017-18
Kindergarten	88	393	392
First	91	398	398
Second	91	398	399
Third	88	391	392
Any (K-3)	96	408	409

The number of students in AGR schools from 2015-16 to 2017-18, 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.

Table 4 | Number of AGR Students by Grade and Year

Grade	2015-16	2016-17	2017-18
Kindergarten	4,139	18,384	18,797
First	4,571	19,288	18,855
Second	4,682	20,054	19,200
Third	4,544	19,508	19,257
Overall (K-3)	17,936	77,234	75,586

Figures 4 and 5 compare the demographic characteristics of AGR students to all K-3 Wisconsin students in 2017-18. Relative to Wisconsin as a whole, a higher proportion of AGR students were black, Hispanic, English learners, and economically disadvantaged. Students in AGR schools were less likely white.

Figure 4 | Race/Ethnicity of AGR and WI Students, 2017-18

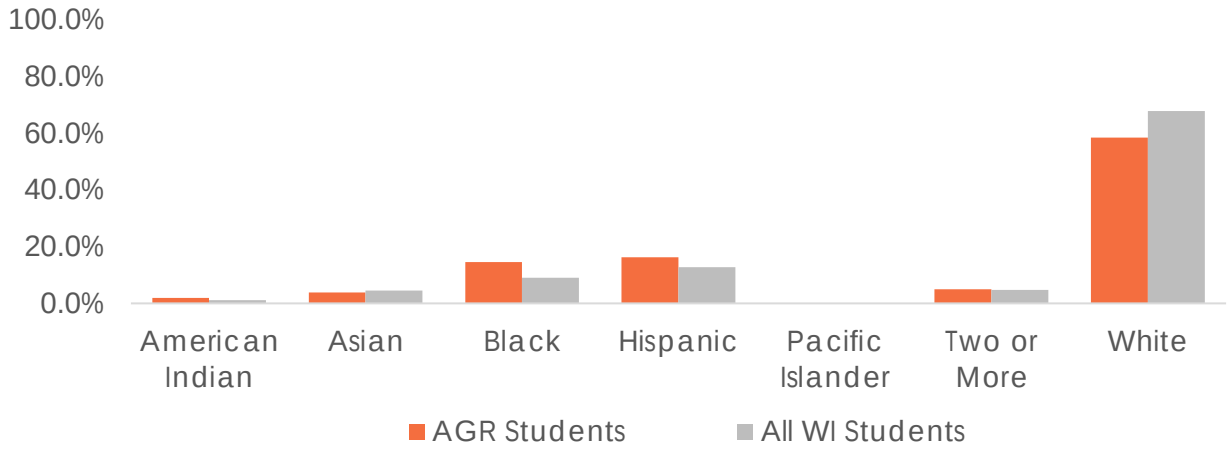
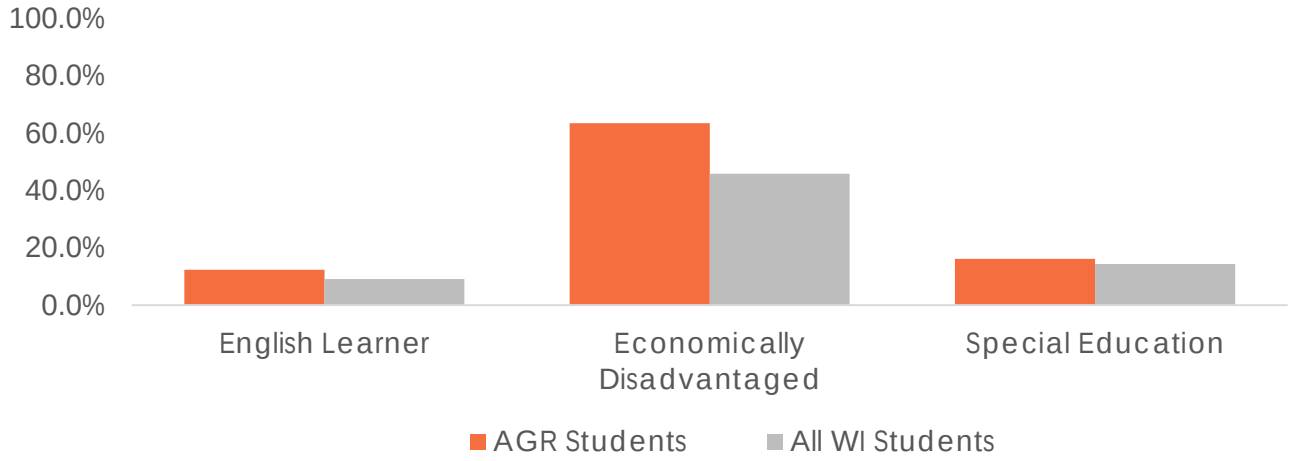


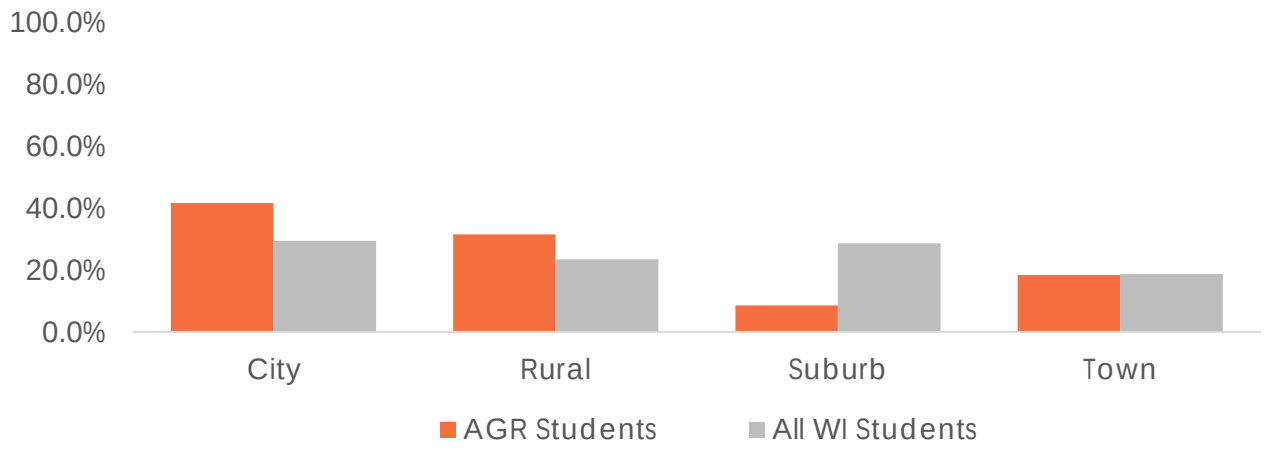
Figure 5 | Percentage of AGR and WI Students that were English Learners, Economically Disadvantaged, and in Special Education, 2017-18



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 economically disadvantaged students, seen previously, as city and rural areas of the state have larger populations with poverty.[5]

[5] https://dpi.wi.gov/news/maps/free-reduced-lunch

Figure 6 | Locale Description of AGR and WI Students, 2017-18



Testing Patterns and Growth Analysis Samples

Shifting testing patterns in Grades K-3 throughout the sample period 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 students are tested throughout Grades K-3, schools 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 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, the sample of AGR schools and students that would be appropriate for use in growth analysis is less than half of the overall population of AGR schools and students.

Given testing patterns, we used two strategies to build sufficient samples. First, we split the growth analysis sample into Grade K and Grades 1-3. Table 5 shows that in 2017-18, nearly half of all Wisconsin kindergarteners took the PALS, which had been a state-mandated reading assessment for the grade through 2015-16. For the purposes of this evaluation, PALS provided sufficient coverage for kindergarten reading, although no combination of assessments resulted in adequate coverage for kindergarten math.

Table 5 | Percentage of Wisconsin Schools Using PALS

Grade	2015-16	2016-17	2017-18
Kindergarten	96%	54%	48%
First	97%	48%	40%
Second	92%	43%	36%

The second strategy we used to build growth analysis samples was to use both the MAP and STAR assessments for Grades 1-3 math and reading. As shown in Table 6 and 7, in 2017-18 between 16-52 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. To combine MAP and STAR into a single measure, we equated assessment scores using national norms.[6]

Table 6 | Percentage of Wisconsin Schools Using MAP or STAR Math Tests

Grade	2015-16	2016-17	2017-18
Kindergarten	13%	11%	10%
First	33%	32%	34%
Second	44%	44%	44%
Third	53%	54%	52%

Table 7 | Percentage of Wisconsin Schools Using MAP or STAR Reading Tests

Grade	2015-16	2016-17	2017-18
Kindergarten	9%	7%	8%
First	18%	16%	16%
Second	42%	42%	41%
Third	53%	53%	51%

Table 8 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 the tested population encompasses. Table 9 shows similar information but for students instead of schools. As seen in these Tables 8 and 9, testing patterns restricted the sample the most in first grade, where the growth analysis was only included 10 to 15 percent of the entire sample of students. This restriction lessened as grade level increased.

[6] 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. Retrieved from https://www.nwea.org/content/uploads/2018/01/2015-MAP-Norms-for-Student-and-School-Achievement-Status-and-Growth.pdf

Table 8 | Number of Growth Analysis AGR Schools and Percentage of All AGR Schools by Grade and Year

Grade	2015-16		2016-17		2017-18	
	Number	%	Number	%	Number	%
Kindergarten	87	98.9	204	51.9	166	42.3
First	12	13.2	44	11.1	49	12.3
Second	28	30.8	160	40.2	158	39.6
Third	28	31.8	211	54.0	202	51.5
Overall (K-3)	93	96.9	310	76.0	296	72.4

Table 9 | Number of Growth Analysis AGR Students and Percentage of All AGR Schools by Grade and Year

Grade	2015-16		2016-17		2017-18	
	Number	%	Number	%	Number	%
Kindergarten	3,927	94.9	9,011	49.0	7,382	40.4
First	712	15.6	1,928	10.0	1,879	10.0
Second	1,477	31.5	7,352	36.7	6,499	33.8
Third	1,458	32.1	10,185	52.2	9,400	48.8
Overall (K-3)	7,574	42.2	28,476	36.9	25,160	33.3

Due to the growth analysis sample of students being smaller than the entire population, as mentioned in the limitations section, the growth analysis results may not apply to all AGR students. To test whether the tested and untested samples differ, Figures 7-9 show comparisons of demographic characteristics between the sample of AGR students used in the growth analysis and the AGR students not used in the growth analysis due to lack of assessment information. Growth analysis AGR students tended to have higher proportions of black and economically disadvantaged students and lower proportions of white students, although the differences are small. The schools included in the growth analysis came more from urban areas and from fewer rural areas.

Figure 7 | Race/Ethnicity of AGR Growth Analysis and Non-Growth Analysis Students, 2017-18

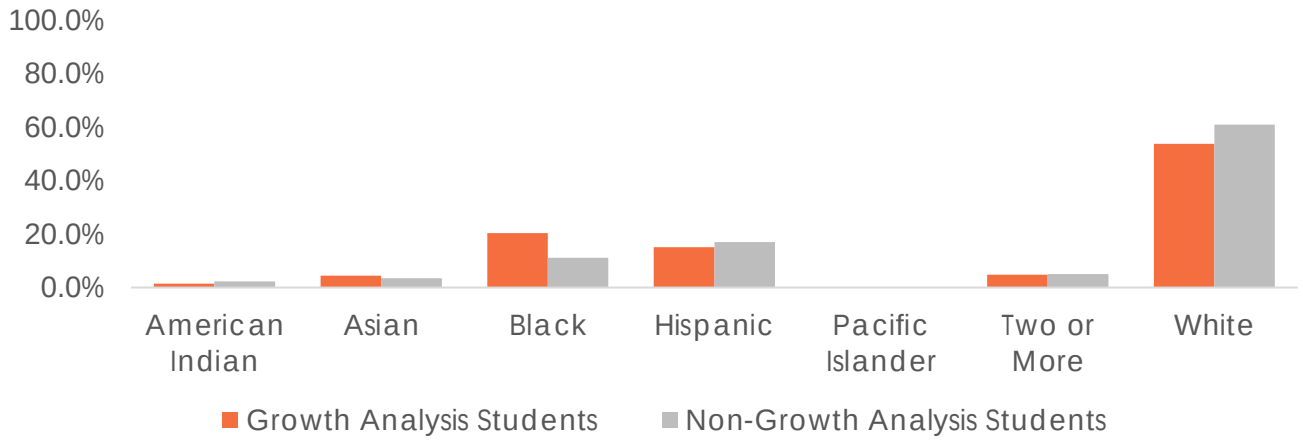


Figure 8 | Percentage of AGR Growth Analysis and Non-Growth Analysis Students that Were English Learners, Economically Disadvantaged, and in Special Education, 2017-18

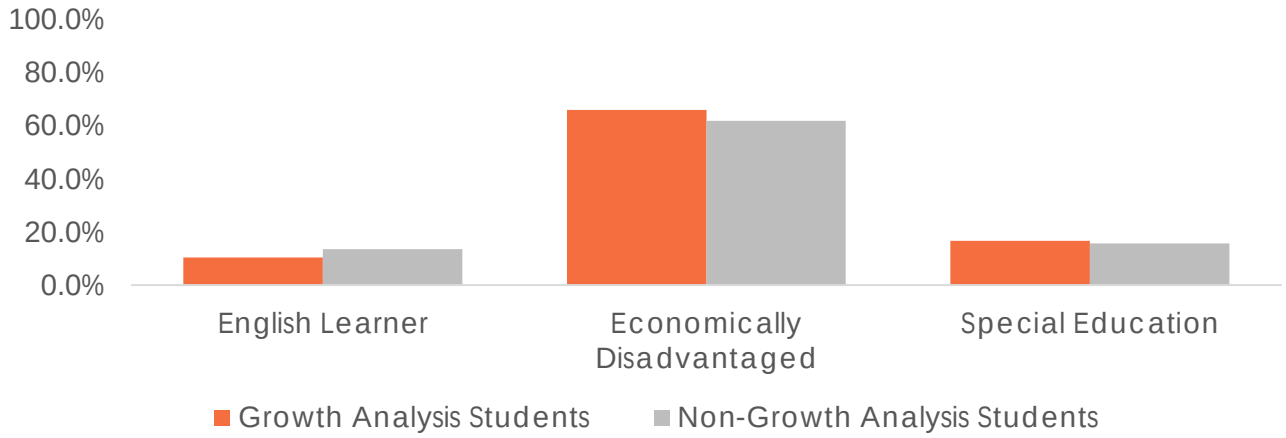
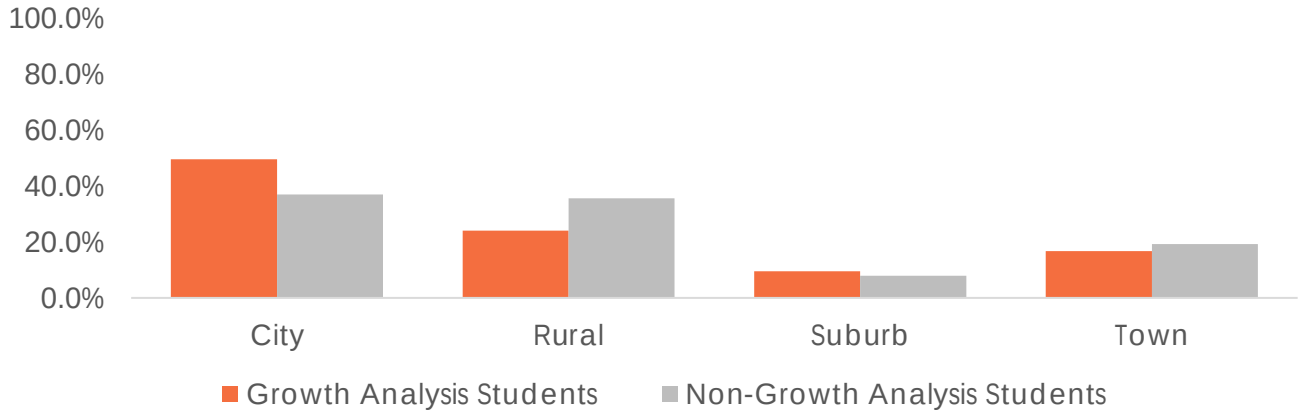


Figure 9 | Locale Description of AGR Growth Analysis and Non-Growth Analysis Students, 2017-18



AGR IMPLEMENTATION

This section of the report examines the usage of the three possible AGR strategies that schools could use in the program. 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 program allowed schools to use more than one strategy within a school, there are seven possible combinations schools could implement AGR: 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. Table 10 provides information on the strategy combinations AGR schools implemented during 2017-18 (the only year with complete data). This table also provides information on the number and percentage of students affected by each strategy combination. The most frequently used strategies included class size reduction and coaching, all three strategies, coaching only, and class size reduction only. Very few schools used only tutoring as a strategy.

Table 10 | Distribution of AGR Strategies, 2017-18

Strategy	# of Schools	% of Schools	# of Students	% of Students
Class size Only	65	17.3%	14,704	21.2%
Coaching Only	70	18.6%	20,836	30.0%
Tutoring Only	4	1.1%	1,673	2.4%
Class Size and Coaching	107	28.5%	15,474	22.3%
Class Size and Tutoring	33	8.8%	2,750	4.0%
Coaching and Tutoring	16	4.3%	4,395	6.3%
All Three	81	21.5%	9,649	13.9%

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? How does AGR impact achievement gaps between low-income students and their higher-income peers? 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 preliminary analyses for different AGR strategy combinations.

All of the impact analyses of AGR examine how students performed on four different outcome measures including math growth, reading growth, absences, and discipline. For each of these outcomes, this report provides a table of results at each applicable grade level and overall (across all applicable grades). These tables all show a measure or measures of the impact of the program and a p-value which indicates the likelihood of observing the reported impact or more extreme assuming that there is no actual impact of the program. A larger p-value indicates weak evidence against an actual result of no impact, while a smaller p-value indicates strong evidence against an actual result of no impact. Throughout the report, the evaluation used a threshold of 0.05 to determine if a result was statistically significant from zero. All p-values presented in this report were corrected to account for multiple estimates (see the Technical Appendix for details).

Overall Impacts

The impact analysis examines how AGR students performed compared to non-AGR students in similar schools, while controlling for student characteristics. Table 11 shows the impacts of AGR on math growth using two different measures. The first measure of impact is on a standardized scale representing the number of standard deviations from zero while the second measure of impact is in approximate MAP scale scores. Both 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.

Table 11 | Overall Impacts of AGR on Math Growth

Outcome	Grade	Impact (Standardized)	Impact (Approximate MAP Scale)	P-Value
MAP/STAR Math	First	0.034	0.46	0.596
	Second	0.023	0.31	0.627
	Third	-0.028	-0.39	0.334
	Overall (1-3)	-0.003	-0.04	0.910

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

Table 12 shows impacts of AGR on reading growth. As with math growth, this table shows average differences in growth on two different scales, a standardized scale and points on the assessment scale (PALS in kindergarten and MAP in Grades 1-3). There were statistically significant impacts of AGR on kindergarten reading growth. On average, AGR students grew 0.12 standard deviations (or 1.6 PALS score points) more than their similar school non-AGR counterparts. The size of this impact is substantive, approximately equal to the impact of a one standard deviation increase in teacher effectiveness.[7] In contrast, results from MAP/STAR reading assessments indicate little difference between AGR and non-AGR students on average growth in Grades 1-3.

Table 12 | Overall Impact of AGR on Reading Growth

Outcome	Grade	Impact (Standardized)	Impact (Approximate PALS/MAP Scale)	P-Value
PALS	Kindergarten	0.121*	1.61	0.024
MAP/STAR Reading	First	-0.010	-0.15	0.859
	Second	0.040	0.61	0.149
	Third	0.010	0.15	0.641
	Overall (1-3)	0.018	0.27	0.386

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

The impacts of AGR on absence rates appear in Table 13. As indicated, while overall there was little impact of AGR on absence rates, there were statistically significantly higher absence rates for 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 0.8 more absence days, a

[7] Hanushek, E. A. & Rivkin, S. G. (2012). *Annual Review of Economics*, 4: 131-157.

relatively small impact that is measured with statistical precision due to the large sample size included in the analysis.

Table 13 | Overall Impact of AGR on Absences

Outcome	Grade	Impact (Percentage Points)	Impact (Approximate Days)	P-Value
Absence Rate	Kindergarten	0.39	0.7	0.139
	First	0.37	0.6	0.125
	Second	0.40	0.7	0.094
	Third	0.44*	0.8	0.043
	Overall (1-3)	0.40	0.7	0.087

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

Table 14 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.

Table 14 | Overall Impact of AGR on Discipline

Outcome	Grade	Impact (Percentage Points)	P-Value
Suspension Rate	Kindergarten	-0.6	0.122
	First	-0.4	0.364
	Second	-0.6	0.273
	Third	-0.3	0.631
	Overall (1-3)	-0.4	0.262

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

Impacts Compared to SAGE

The following results provide information on the impact of the AGR program compared to previous SAGE implementation. As noted in the methodology section, this analysis estimates impacts using AGR schools both before and after their transition from SAGE to AGR, comparing those schools' student-level outcomes to student outcomes from observably similar schools. Table 15 shows the impact of AGR compared to SAGE on math growth with both a standardized measure and a measure on the MAP scale. Overall, differences in math growth between AGR and SAGE were small and non-statistically significant.

Table 15 | Impact of AGR Compared to SAGE on Math Growth

Outcome	Grade	Impact (Standardized)	Impact (Approximate MAP Scale)	P-Value
MAP/STAR Math	First	0.068	0.93	0.173
	Second	0.045	0.61	0.163
	Third	-0.030	-0.41	0.318
	Overall (1-3)	0.010	0.14	0.847

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

Table 16 includes differences between AGR and SAGE from the analysis of reading growth. While the evaluation found few differences in reading growth in first through third grade, the estimated impact on reading growth in kindergarten was that AGR students had higher average growth (by approximately 1.4 PALS score points) than previous SAGE students. Similar to PALS results above, the estimated difference between SAGE and AGR impacts is both statistically significant and of substantive magnitude.

Table 16 | Impact of AGR Compared to SAGE on Reading Growth

Outcome	Grade	Impact (Standardized)	Impact (Approximate MAP Scale)	P-Value
PALS	Kindergarten	0.108*	1.44	0.043
MAP/STAR Reading	First	0.018	0.26	0.816
	Second	0.018	0.27	0.562
	Third	-0.007	-0.11	0.793
	Overall (1-3)	0.007	0.10	0.806

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

Absence rate comparisons between AGR and SAGE reveal few differences, as seen in Table 17. While estimated average AGR absence rates are slightly lower than SAGE, these differences are small and not statistically significant.

Table 17 | Impact of AGR Compared to SAGE on Absences

Outcome	Grade	Impact (Percentage Points)	Impact (Approximate Days)	P-Value
Absence Rate	Kindergarten	-0.49	-0.9	0.092
	First	-0.41	-0.7	0.091
	Second	-0.36	-0.6	0.085
	Third	-0.29	-0.5	0.162
	Overall (K-3)	-0.38	-0.7	0.094

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

Table 18 shows the results of AGR impact compared to SAGE on student discipline. As with math growth, the evaluation found little difference in the suspension rates between AGR students and SAGE students.

Table 18 | Impact of AGR Compared to SAGE on Discipline

Outcome	Grade	Impact (Percentage Points)	P-Value
Suspension Rate	Kindergarten	0.0	0.997
	First	-0.1	0.832
	Second	-0.1	0.846
	Third	0.5	0.235
	Overall (K-3)	0.1	0.817

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

Impacts by Demographics

In addition to examining the overall, statewide impact of the AGR program on student performance, this evaluation also examined whether the AGR program has different impacts for different subgroups of students. Given AGR's focus on closing the achievement gap, examining the impact for specific subgroups, particularly economically disadvantaged students, is important for determining whether the program is meeting its goals. Throughout the following tables, results are pooled across all applicable grade levels to measure the impact of AGR for the following subgroups of students: females, Asian students, black students, Hispanic students, white students, students of other race or ethnicity, economically disadvantaged students (as measured by free or reduced price lunch, or FRL status), English learner (EL) students, and students in schools within cities.

Table 19 shows the impact of the AGR program on math growth for each subgroup, relative to the same subgroup of students in similar non-AGR schools. Across all subgroups, the evaluation found little difference in growth between AGR students and non-AGR students in similar schools.

Table 19 | Impact of AGR on Math Growth by Student Subgroup

Outcome	Grade	Subgroup	Impact (Standardized)	Impact (Approximate MAP Scale)	P-Value
MAP/STAR Math	Overall (1-3)	Female Students	0.000	0.00	0.992
		Asian Students	-0.038	-0.52	0.619
		Black Students	-0.056	-0.76	0.312
		Hispanic Students	0.041	0.56	0.274
		White Students	0.008	0.11	0.790
		Other Race/Ethnicity	-0.016	-0.22	0.666
		FRL Students	-0.015	-0.20	0.628
		EL Students	0.025	0.34	0.636
		City Students	-0.008	-0.11	0.857

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

As shown in Table 20, for kindergarten PALS the evaluation found significantly positive impacts of AGR for a variety of subgroups including females, Asian students, Hispanic students, economically disadvantaged students, English learners, and students in urban areas. The largest impacts found were for kindergarten English learners who had an average reading growth of 0.38 standard deviations, or 5.0 PALS score points, higher than English learners in similar non-AGR schools. The focal subgroup for AGR, economically disadvantaged students, had an average reading growth 0.15 standard deviations higher than economically disadvantaged students in similar non-AGR schools (approximately 2 points higher growth on the PALS assessment). The evaluation found only small, not statistically significant differences in reading growth for first through third grade overall between AGR students and non-AGR students in similar schools, regardless of the type of student.

Table 20 | Impact of AGR on Reading Growth by Student Subgroup

Outcome	Grade	Subgroup	Impact (Standardized)	Impact (Approximate PALS or MAP Scale)	P-Value
PALS	Kindergarten	Female Students	0.122*	1.62	0.026
		Asian Students	0.245*	3.26	0.021
		Black Students	0.230	3.06	0.096
		Hispanic Students	0.258*	3.43	0.000
		White Students	0.066	0.88	0.109
		Other Race/Ethnicity	0.032	0.43	0.647
		FRL Students	0.152*	2.02	0.027
		EL Students	0.376*	5.00	0.000
		City Students	0.232*	3.09	0.016
MAP/STAR Reading	Overall (1-3)	Female Students	0.020	0.30	0.341
		Asian Students	0.004	0.06	0.930
		Black Students	-0.026	-0.39	0.630
		Hispanic Students	0.048	0.72	0.092
		White Students	0.026	0.39	0.256
		Other Race/Ethnicity	0.028	0.42	0.373
		FRL Students	0.014	0.21	0.619
		EL Students	0.043	0.64	0.241
		City Students	0.011	0.16	0.719

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

With the exception of student absences in schools located in cities, AGR impacts on absence rates, as shown in Table 21, show only small differences between AGR students and non-AGR students in similar schools across all subgroups of students. On average, city students at AGR schools had one more absence than their counterparts at non-AGR city schools.

Table 21 | Impact of AGR on Absences by Student Subgroup

Outcome	Grade	Subgroup	Impact (Percentage Points)	Impact (Approximate Days)	P-Value
Absence Rate	Overall (K-3)	Female Students	0.37	0.6	0.105
		Asian Students	0.26	0.5	0.634
		Black Students	0.54	0.9	0.115
		Hispanic Students	0.54	0.9	0.098
		White Students	0.33	0.6	0.267
		Other Race/Ethnicity	0.45	0.8	0.370
		FRL Students	0.30	0.5	0.327
		EL Students	0.42	0.7	0.107
		City Students	0.55*	1.0	0.021

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

Table 22 provides the results of the estimated impact of AGR on student discipline for the same subgroups of students. As seen from this table, although all subgroups of AGR students had lower suspension rates, only estimates for Hispanic and English learner students were statistically significant. Hispanic AGR students had a suspension rate roughly 0.9 percentage points lower than their non-AGR counterparts and English learner AGR students had a suspension rate roughly 0.7 percentage points lower. These are meaningfully large impacts considering the already low rate of suspensions at these grade levels.

Table 22 | Impact of AGR on Discipline by Student Subgroup

Outcome	Grade	Subgroup	Impact (Percentage Points)	P-Value
Suspension Rate	Overall (K-3)	Female Students	-0.2	0.297
		Asian Students	-0.1	0.607
		Black Students	-0.8	0.260
		Hispanic Students	-0.9*	0.038
		White Students	-0.0	0.914
		Other Race/Ethnicity	-0.4	0.625
		FRL Students	-0.6	0.258
		EL Students	-0.7*	0.041
		City Students	-0.6	0.357

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

Differences in Outcomes by Strategy

The final set of student performance results examine the difference in impacts of AGR 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. 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.

The differences in outcomes by strategy shown in Tables 23-26 should be considered only preliminary evidence of how strategy usage might causally impact test score growth, absences, and discipline. The analysis includes only AGR schools with no comparison schools. Because AGR schools are allowed 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 23 shows the impact on math growth for each strategy combination compared to small class size. Strategy combinations resulting in higher average math growth included coaching only in first grade, class size and tutoring in second grade, tutoring only in third grade, and tutoring only across all grades.

Table 23 | Differences in Math Growth by Strategy, Compared to Small Class Size Only

Outcome	Grade	Strategy	Impact (Standardized)	Impact (Approximate MAP Scale)	P-Value
MAP/STAR Math	First	Coaching Only	0.370*	5.04	0.008
		Tutoring Only	N/A	N/A	N/A
		Class Size & Coaching	N/A	N/A	N/A
		Class Size & Tutoring	N/A	N/A	N/A
		Coaching & Tutoring	N/A	N/A	N/A
		All Three	0.129	1.76	0.235
	Second	Coaching Only	0.058	0.78	0.652
		Tutoring Only	N/A	N/A	N/A
		Class Size & Coaching	0.108	1.46	0.430
		Class Size & Tutoring	0.226*	3.06	0.041
		Coaching & Tutoring	0.076	1.02	0.627
		All Three	-0.167	-2.27	0.279
	Third	Coaching Only	-0.038	-0.53	0.632
		Tutoring Only	0.350*	4.84	0.000
		Class Size & Coaching	0.029	0.41	0.691
		Class Size & Tutoring	-0.041	-0.57	0.694
		Coaching & Tutoring	0.032	0.44	0.703
		All Three	-0.032	-0.45	0.764
	Overall (1-3)	Coaching Only	0.053	0.72	0.480
		Tutoring Only	0.400*	5.46	0.000
		Class Size & Coaching	0.097	1.32	0.241
		Class Size & Tutoring	0.064	0,.87	0.377
		Coaching & Tutoring	0.106	1.45	0.240
		All Three	-0.066	-0.90	0.370

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 PALS reading growth for each strategy combination compared to small class size only, the analysis found few differences across strategy combinations in kindergarten, as seen in Table 24. 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 and tutoring only and class size and coaching across first through third grade when compared to small class size only.

Table 24 | Differences in Reading Growth by Strategy, Compared to Small Class Size Only

Outcome	Grade	Strategy	Impact (Standardized)	Impact (Approximate PALS or MAP Scale)	P-Value
PALS	Kindergarten	Coaching Only	-0.112	-1.55	0.374
		Tutoring Only	0.036	0.49	0.878
		Class Size & Coaching	-0.093	-1.29	0.237
		Class Size & Tutoring	-0.020	-0.27	0.849
		Coaching & Tutoring	-0.178	-2.44	0.237
		All Three	-0.129	-1.77	0.232
MAP/STAR Reading	First	Coaching Only	0.231*	3.15	0.014
		Tutoring Only	N/A	N/A	N/A
		Class Size & Coaching	N/A	N/A	N/A
		Class Size & Tutoring	N/A	N/A	N/A
		Coaching & Tutoring	N/A	N/A	N/A
		All Three	0.109	1.49	0.228
	Second	Coaching Only	0.087	1.18	0.297
		Tutoring Only	N/A	N/A	N/A
		Class Size & Coaching	0.124	1.67	0.226
		Class Size & Tutoring	0.130	1.75	0.170
		Coaching & Tutoring	0.020	-0.56	0.834
		All Three	-0.042	-0.56	0.765
	Third	Coaching Only	0.028	0.39	0.628
		Tutoring Only	0.142	1.96	0.155
		Class Size & Coaching	0.093	1.29	0.100
		Class Size & Tutoring	-0.018	-0.25	0.875
		Coaching & Tutoring	0.005	0.07	0.945
		All Three	0.002	0.03	0.973
	Overall (1-3)	Coaching Only	0.077	1.16	0.134
		Tutoring Only	0.197*	2.95	0.020
		Class Size & Coaching	0.126*	1.89	0.008
		Class Size & Tutoring	0.003	0.04	0.974
		Coaching & Tutoring	0.061	0.92	0.367
		All Three	-0.005	-0.07	0.943

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 25 displays differences in absence rates for each of the strategy combinations compared to class size reduction only. As this table illustrates, students in AGR schools using tutoring only, when compared to small class sizes only, had higher absence rates in first grade, second grade, third grade, and in kindergarten through third grade overall.

Table 25 | Differences in Absences by Strategy, Compared to Small Class Size Only

Outcome	Grade	Strategy	Impact (Percentage Points)	Impact (Approximate Days)	P-Value
Absence Rate	Kindergarten	Coaching Only	0.03	0.1	0.950
		Tutoring Only	0.83	1.5	0.241
		Class Size & Coaching	0.25	0.4	0.664
		Class Size & Tutoring	0.89	1.5	0.210
		Coaching & Tutoring	0.21	0.4	0.822
		All Three	0.15	0.3	0.851
	First	Coaching Only	0.42	0.7	0.231
		Tutoring Only	0.86*	1.5	0.005
		Class Size & Coaching	0.60	1.1	0.225
		Class Size & Tutoring	1.04	1.8	0.144
		Coaching & Tutoring	0.14	0.2	0.839
		All Three	0.93	1.6	0.113
	Second	Coaching Only	0.39	0.7	0.278
		Tutoring Only	0.86*	1.5	0.049
		Class Size & Coaching	0.64	1.1	0.232
		Class Size & Tutoring	0.90	1.6	0.237
		Coaching & Tutoring	0.32	0.6	0.660
		All Three	0.46	0.8	0.412
	Third	Coaching Only	0.35	0.6	0.278
		Tutoring Only	0.85*	1.5	0.020
		Class Size & Coaching	0.69	1.2	0.169
		Class Size & Tutoring	0.99	1.7	0.095
		Coaching & Tutoring	0.44	0.8	0.539
		All Three	0.20	0.4	0.701
	Overall (K-3)	Coaching Only	0.33	0.6	0.292
		Tutoring Only	0.85*	1.5	0.005
		Class Size & Coaching	0.55	1.0	0.232
		Class Size & Tutoring	0.97	1.7	0.077
		Coaching & Tutoring	0.29	0.5	0.657
		All Three	0.46	0.8	0.365

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

Table 26 shows differences in student discipline, as measured by the suspension rate, for each combination of AGR strategies as compared to small class size only. Students in schools and grades implementing the tutoring only strategy in third grade and coaching only in all grades had suspension rates significantly higher than students in schools and grades using small class sizes only.

Table 26 | Differences in Discipline by Strategy, Compared to Small Class Size Only

Outcome	Grade	Strategy	Impact (Percentage Points)	P-Value
Suspension Rate	Kindergarten	Coaching Only	2.6	0.231
		Tutoring Only	0.9	0.146
		Class Size & Coaching	0.9	0.466
		Class Size & Tutoring	-0.2	0.629
		Coaching & Tutoring	1.6	0.248
		All Three	0.8	0.213
	First	Coaching Only	1.2	0.482
		Tutoring Only	0.7	0.251
		Class Size & Coaching	2.0	0.370
		Class Size & Tutoring	1.2	0.240
		Coaching & Tutoring	1.1	0.249
		All Three	0.4	0.650
	Second	Coaching Only	1.8	0.293
		Tutoring Only	0.6	0.434
		Class Size & Coaching	0.3	0.864
		Class Size & Tutoring	0.3	0.717
		Coaching & Tutoring	0.7	0.650
		All Three	0.4	0.772
	Third	Coaching Only	1.5	0.627
		Tutoring Only	2.4*	0.001
		Class Size & Coaching	1.7	0.235
		Class Size & Tutoring	0.9	0.281
		Coaching & Tutoring	0.6	0.647
		All Three	0.8	0.468
	Overall (K-3)	Coaching Only	1.2*	0.007
		Tutoring Only	0.4	0.475
		Class Size & Coaching	1.5	0.249
		Class Size & Tutoring	1.2	0.230
		Coaching & Tutoring	0.9	0.245
		All Three	0.6	0.275

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

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.[8] 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 27 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 29). 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 20 percent using all three.

Table 27 | 2017-18 School-level AGR Strategies, School Board Report Data

Strategy	Percentage
Instructional coaching only	11%
Reduced class size only	27%
One-to-one tutoring only	4%
Instructional coaching and reduced class size	23%
Instructional coaching and one-to-one tutoring	5%
Reduced class size and one-to-one tutoring	10%
All 3 strategies	20%

[8] https://dpi.wi.gov/sites/default/files/imce/sage/doc/agr_performance_objectives_and_school_board_report_template.docx

END-OF-YEAR REPORT FINDINGS

This section of the report provides the results from the 2017-18 End-of-Year Report survey of AGR schools. As exemplified earlier in the report, by 2017-18 schools that had transitioned from SAGE to AGR had taken advantage of AGR's increased flexibility. As shown in Table 28, only 21 percent of responding schools employed only the reduced class size strategy. A majority of schools opted for multiple strategies–60 percent used more than one strategy, including 18 percent that used all three strategies. Table 29 shows that reduced class size (76 percent of schools) and instructional coaching (70 percent of schools) were more common than one-to-one tutoring, which was used by only 32 percent of sample schools. As should be expected, the results in Table 28 closely resemble similar results from the school board reports, as shown in Table 27.

Table 28 | Combinations of Strategies Used by AGR Schools, End-of-Year Report

Strategy	Percentage
Instructional coaching only	18%
Reduced class size only	21%
One-to-one tutoring only	1%
Instructional coaching and reduced class size	28%
Instructional coaching and one-to-one tutoring	4%
Reduced class size and one-to-one tutoring	8%
All 3 strategies	18%

Note: 387 respondents to survey item.

Table 29 | Percentage of AGR Schools Using each Strategy, End-of-Year Report

Strategy	Percentage
Instructional coaching	70%
Reduced class size	76%
One-to-one tutoring	32%

Note: 387 respondents to survey item.

In general, schools chose to use the same strategies across classrooms within each grade (Tables 30, 31, and 32). This trend was strongest for grades with reduced size classrooms. Within each grade, over 90 percent of schools chose to use reduced class sizes in at least three-quarters of classrooms or not at all (Table 30). For one-to-one tutoring (Table 31) and instructional coaching (Table 32), approximately 75 to 80 percent of schools used a strategy in at least three-quarters of the classrooms or not at all.

Table 30 | Schools' Distributions of Classrooms with AGR Reduced Class Size, by Grade

Grade	Percentage of Classrooms				
	None	Less than 25%	25-50%	51-75%	More than 75%
Kindergarten	12%	1%	3%	2%	80%
First	23%	1%	3%	3%	69%
Second	27%	1%	4%	3%	64%
Third	29%	3%	3%	3%	61%

Note: 294 respondents to survey item. Results represent the weighted average of semester one and semester two results, which were similar. Categories are mutually exclusive.

Table 31 | Schools' Distribution of Classrooms with AGR One-to-One Tutoring, by Grade

Grade	Percentage of Classrooms				
	None	Less than 25%	25-50%	51-75%	More than 75%
Kindergarten	28%	14%	7%	4%	47%
First	25%	13%	6%	2%	54%
Second	24%	13%	9%	3%	51%
Third	24%	14%	8%	4%	50%

Note: 123 respondents to survey item. Results represent the weighted averages of semester one and semester two results, which were similar. Categories are mutually exclusive.

Table 32 | Schools' Distribution of Classrooms with AGR Instructional Coaching, by Grade

Grade	Percentage of Classrooms				
	None	Less than 25%	25-50%	51-75%	More than 75%
Kindergarten	10%	7%	9%	9%	65%
First	9%	8%	6%	9%	68%
Second	9%	5%	6%	8%	71%
Third	10%	8%	6%	7%	70%

Note: 269 respondents to survey item. Results represent the weighted averages of semester one and semester two results, which were similar. Categories are mutually exclusive.

Schools using reduced class sizes reported using a variety of instructional strategies (Table 33), including small group instruction (93 percent), one-on-one time with the teacher (74 percent), differentiation of instruction (88 percent), strategic placement of students in groups (85 percent), and, to a lesser extent, strategic placement of students in classrooms (62 percent). Only two percent of responding schools reported that they use no additional instructional strategies due to AGR class size reductions.

Table 33 | Instructional Strategies Associated with AGR Reduced Class Size

Strategy	Percentage
Small-group instruction	93%
One-on-one time with the teacher	74%
Differentiation of instruction	88%
Strategic placement of students in groups	85%
Strategic placement of students in classrooms	62%
Other	4%
We don't use any specific instructional strategies because of smaller class sizes	2%
Not sure/don't know	1%

Note: 294 respondents to survey item.

These same schools indicated multiple benefits from class size reduction as seen in Table 34. Eighty percent of schools reported that relationships between students and teachers improved "a lot" due to AGR class size reductions.

Table 34 | Perceived Benefits Derived from AGR Reduced Class Size

Benefit	A lot	A moderate amount	A little	None	Not Sure
Better common planning among teachers	34%	34%	20%	9%	2%
Better interactions among students	65%	31%	2%	1%	0%
Better relationships between teachers and students	80%	17%	1%	0%	0%
Better teacher morale	66%	26%	4%	0%	1%
Increased student ownership of learning	44%	8%	1%	42%	3%
Better teacher performance	46%	41%	8%	2%	2%
Increased use of data among teachers	49%	35%	12%	1%	1%
More participation from students in class	66%	29%	2%	0%	1%
More time for individual interactions	77%	19%	2%	0%	1%
Reduction in student anxiety	41%	36%	11%	1%	10%
Reduction in student behavioral problems	41%	43%	11%	1%	2%
Students engage in student-specific interventions	55%	35%	6%	1%	2%
Teachers value contributions of all students	55%	34%	6%	1%	2%

Note: 294 respondents to survey item.

The survey also included an open-ended response item for other benefits reduced class size provided. Responses to this item included the following categories:

- Description of value, benefits, and impact of class size reduction (n=23)
- Teacher coaching/professional development, capacity building (n=3)
- Improved parent/family communication, interaction, relationships (n=26)
- One-to-one instruction, individualized instruction, differentiated instruction, personalized learning (n=30)
- Building social/emotional relationships with students (n=30)
- Improved opportunities for G/T instruction and/or enrichment activities for students who have met proficiency (n=8)
- Other (n=24)

Schools using one-to-one tutoring did so frequently, as seen in Table 35. Seventy-eight percent offered tutoring at least weekly, and 59 percent engaged in tutoring 3 times a week or more.

Table 35 | Frequency of AGR One-to-One Tutoring

Frequency	Percentage
3 times a week or more	59%
2 times a week	10%
Weekly	9%
Biweekly	3%
Monthly	0%
As needed	11%
Not sure/don't know	3%
3 times a week or more	59%
2 times a week	10%

Note: 123 respondents to survey item.

Table 36 shows that schools reported using almost all of the tutoring practices listed on the survey, although an exception was maintaining a focus on equity (51 percent).

Table 36 | AGR One-to-One Tutoring Practices

Practice	Percentage
Reviews student data	84%
Models appropriate learning behavior	76%
Adapts to student learning styles	82%
Maintains a focus on equity	51%
Provides scaffolding	76%
Communicates regularly with classroom teacher	79%
Not sure/don't know	4%

Note: 123 respondents to survey item.

Similarly, schools answered affirmatively to almost all listed benefits from one-to-one tutoring as seen in Table 37. Relative to reported benefits from reduced class sizes and instructional coaching, responses for one-to-one tutoring were less likely to be "a lot" or "a moderate amount."

Table 37 | Perceived Benefits Derived from AGR One-to-One Tutoring

Benefit	A lot	A moderate amount	A little	None	Not Sure
Better common planning among teachers	20%	31%	24%	20%	4%
Better interactions among students	30%	33%	16%	14%	7%
Better relationships between teachers and students	37%	35%	15%	9%	4%
Better teacher morale	33%	45%	10%	7%	5%
Increased student ownership of learning	42%	37%	11%	7%	2%
Better teacher performance	25%	40%	20%	9%	7%
Increased use of data among teachers	48%	31%	11%	7%	2%
More participation from students in class	33%	41%	12%	10%	5%
More time for individual interactions	63%	22%	7%	6%	2%
Reduction in student anxiety	41%	31%	15%	7%	7%
Reduction in student behavioral problems	32%	41%	15%	8%	3%
Students engage in student-specific interventions	64%	24%	4%	6%	2%
Teachers value contributions of all students	33%	40%	11%	10%	7%

Note: 123 respondents to survey item.

The survey incorporated an open-response item for what other benefits one-to-one tutoring provides. Responses to this item included the following categories:

- Description of value, benefits, and impact of one-to-one tutoring (n=20)
- Teacher professional development, capacity building (n=6)
- Improved communication/planning, collaboration, data use among teachers (n=3)
- One-to-one tutoring, one-to-one instruction, focus on specific interventions (n=10)
- Building/improving student confidence, self-esteem (n=5)
- Building relationships with students (n=9)
- Other (n=7)

Schools were able to successfully find trained, experienced coaches with content specialization, as seen from Table 38.

Table 38 | Characteristics of AGR Instructional Coaches

Characteristic	Percentage
Coach training	78%
Previous instructional coaching experience	67%
Content specialist in their subject of coaching	66%
Not sure/don't know	2%

Note: 269 respondents to survey item.

Table 39 shows that at 73 percent of schools, instructional coaches met with teachers at least monthly, and 47 percent met weekly.

Table 39 | Frequency of AGR Instructional Coaches Meetings with Teachers

Frequency	Percentage
Weekly	47%
Monthly	26%
Quarterly	5%
Each semester	1%
As needed	18%
Not sure/don't know	1%

Note: 269 respondents to survey item.

Table 40 | Instructional Coaching Practices

Practice	Percentage
One-to-one teacher coaching	85%
Team teacher coaching	63%
Keeps a coaching log	42%
Advises teachers to set goals	59%
Coaching focuses on teacher goals	66%
Maintains a focus on equity	46%
Encourages reflective practices	87%
Discusses data with teachers	91%
Observes teacher practices	81%
Not sure/don't know	1%

Note: 269 respondents to survey item.

As with benefits from class size reduction, schools answered affirmatively to most of the benefits listed. As seen in Table 41, increased teacher performance and increased use of data had the strongest affirmative responses, while reductions in student anxiety and behavioral problems were the weakest.

Table 41 | Perceived Benefits Derived from AGR Instructional Coaching

Benefit	A lot	A moderate amount	A little	None	Not Sure
Better common planning among teachers	47%	29%	14%	8%	1%
Better interactions among students	34%	42%	15%	6%	3%
Better relationships between teachers and students	40%	39%	14%	4%	3%
Better teacher morale	48%	35%	11%	3%	2%
Increased student ownership of learning	35%	40%	17%	4%	4%
Better teacher performance	59%	31%	7%	3%	0%
Increased use of data among teachers	68%	26%	3%	3%	0%
More participation from students in class	28%	45%	17%	7%	3%
More time for individual interactions	32%	40%	17%	7%	5%
Reduction in student anxiety	24%	37%	20%	8%	12%
Reduction in student behavioral problems	29%	41%	17%	7%	6%
Students engage in student-specific interventions	48%	32%	14%	4%	2%
Teachers value contributions of all students	40%	40%	11%	4%	5%

Note: 269 respondents to survey item.

The survey also asked an open-ended question regarding what other benefits instructional coaching provides. Responses included the following categories:

- Description of value, benefits, and impact of instructional coaching (n=38)
- Teacher professional development, capacity building (n=64)
- Improved communication/planning, collaboration, data use among teachers (n=32)
- Other (n=13)

SUMMARY/CONCLUSIONS

This report provides evidence regarding program impacts on math and reading growth, student attendance, and out-of-school suspensions. These results are presented on 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 at improving student outcomes.

From 2015-16 to 2017-18, AGR impacts are limited to strong kindergarten reading growth statewide and reductions in out-of-school suspensions for Hispanic students and students classified as English learners. Attending an AGR school is associated with a 0.12 standard deviation increase in PALS growth from fall to spring, relative to a comparison group of students from observably similar schools. Economically disadvantaged students in AGR schools experienced growth 0.15 standard deviations greater than that of similar students in non-AGR schools. Impacts on PALS growth was also higher for Hispanic students (0.26 standard deviations), English learner students (0.38 standard deviations), urban students (0.23 standard deviations), and Asian students (0.25 standard deviations). 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, none were significantly different from zero.

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 0.9 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 0.7 percentage points less likely to be suspended. Statewide, suspensions at AGR schools were 0.4 percentage points lower than at comparable AGR schools, although that result was not statistically significant. 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.

Looking at the AGR strategies that schools chose to implement, we found that most AGR schools took advantage of the 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 2 or more strategies. In a preliminary investigation of how school strategy choices were related to impacts, we found that tutoring was a better strategy for math and reading growth in Grades 1-3, while class size reductions were more likely to be associated with reductions in suspensions. These results should not be interpreted as causal due to selection bias. Schools are free to choose how to implement AGR and the component strategies, and higher capacity schools may choose similar 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.

Future evaluations will further explore relationships between strategies and impacts, taking advantage of End-of-Year Report data on the intensity and quality of tutoring and coaching strategies. In the future, we will also calculate AGR's impacts on the statewide achievement gap. After the program has been in operation for several years, we will estimate impacts on statewide, third grade Forward Exam math and reading scores that Wisconsin uses for school accountability.

TECHNICAL APPENDIX

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 eligible for free or reduce 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.

To find a plausible control group and identify the AGR impact, we use propensity score matching (PSM). PSM address 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 than other Wisconsin schools. This is because AGR targets funding to schools with higher percentages of students receiving free and 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 the 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 reports 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.

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

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.[9] 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 Tables 5 through 7, during the 2015-16 through 2017-18 sample period, only a minority of schools used the PALS, MAP, and STAR tests.[10] Underlying Tables 5 through 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.[11] 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

[9] Stuart, E. (2007). Estimating causal effects using school-level datasets. *Educational Researcher, 36*, 187-198.
[10] Schools administered dozens of different types of tests across all grades. PALS, MAP, and STAR were the most common.
[11] We tested models that limit the sample to schools that tested throughout 2012-13 to 2017-18, but these models omitted most AGR schools.

chose to match all models of test outcomes at the school-grade-year level. 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.

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.

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.[12]

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 and those that will begin AGR in 2018-19. Second, we limited the testing models to schools that tested at least 75 percent of the relevant population in

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

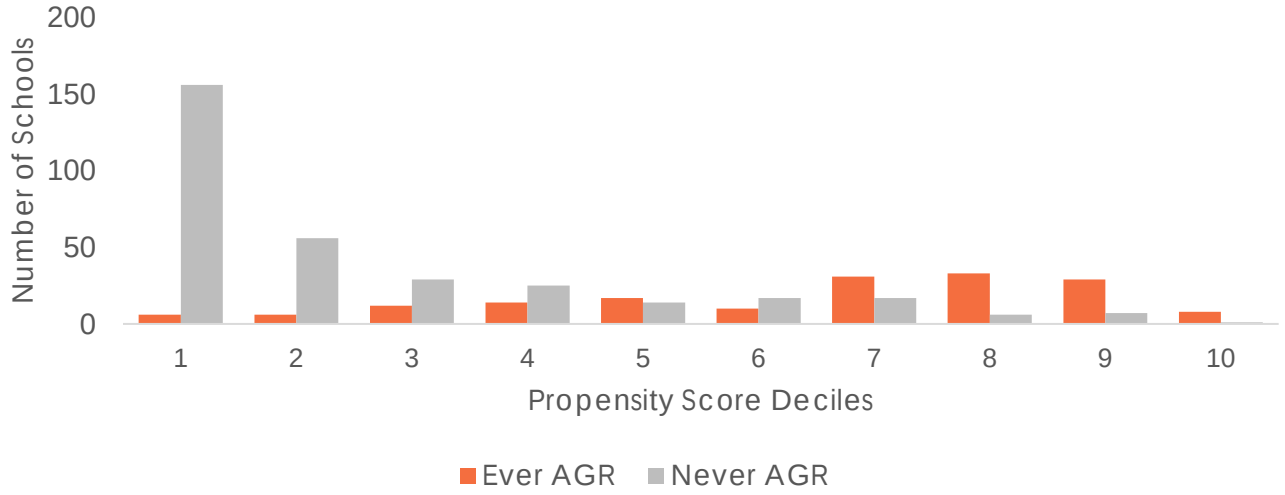
Grades K-3, following previous SAGE evaluations.[13] Table 42 illustrates the matching and subsequent analysis strategies for each outcome.

Table 42 | Matching and Analysis Strategies

Outcome	Grades	Matching Level	Matching Data	Analysis Years
PALS Growth	K	School-grade-year	Fall 2012-13 through Fall 2017-18	2012-13 through 2017-18
MAP/STAR Reading Growth	1-3	School-grade-year	Fall 2012-13 through Fall 2017-18	2012-13 through 2017-18
MAP/STAR Math Growth	1-3	School-grade-year	Fall 2012-13 through Fall 2017-18	2012-13 through 2017-18
Absence Rate	K-3	School	2012-13	2013-14 through 2017-18
Suspension Rate	K-3	School	2012-13	2013-14 through 2017-18

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. Figures 10 through 13 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.

Figure 10 | Common Support for Matching–PALS Reading (2017-18)



[13] 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 11 | Common Support for Matching–MAP/STAR Math (2017-18)

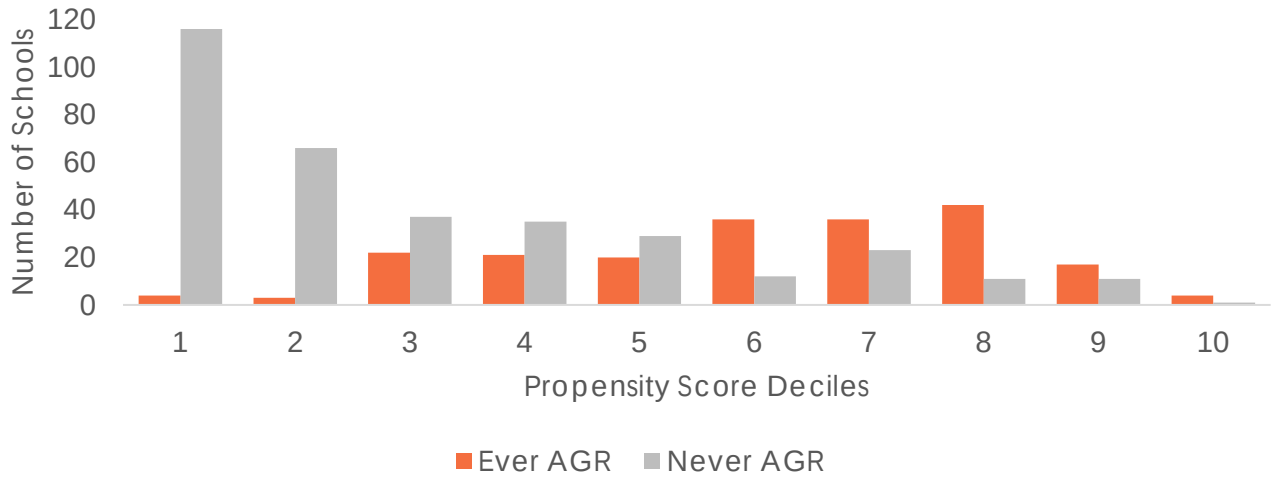


Figure 12 | Common Support for Matching–MAP/STAR Math (2017-18)

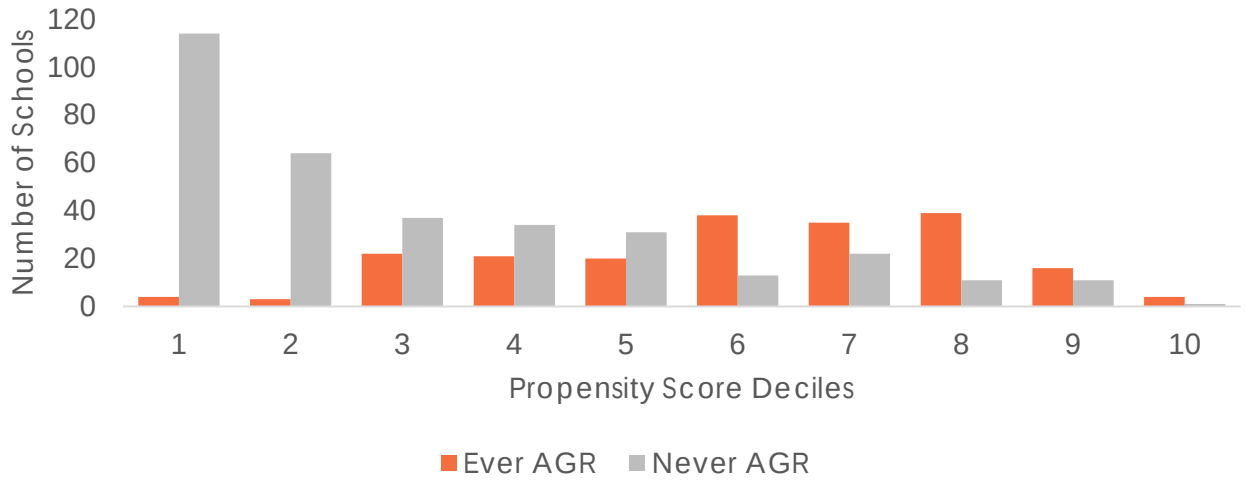
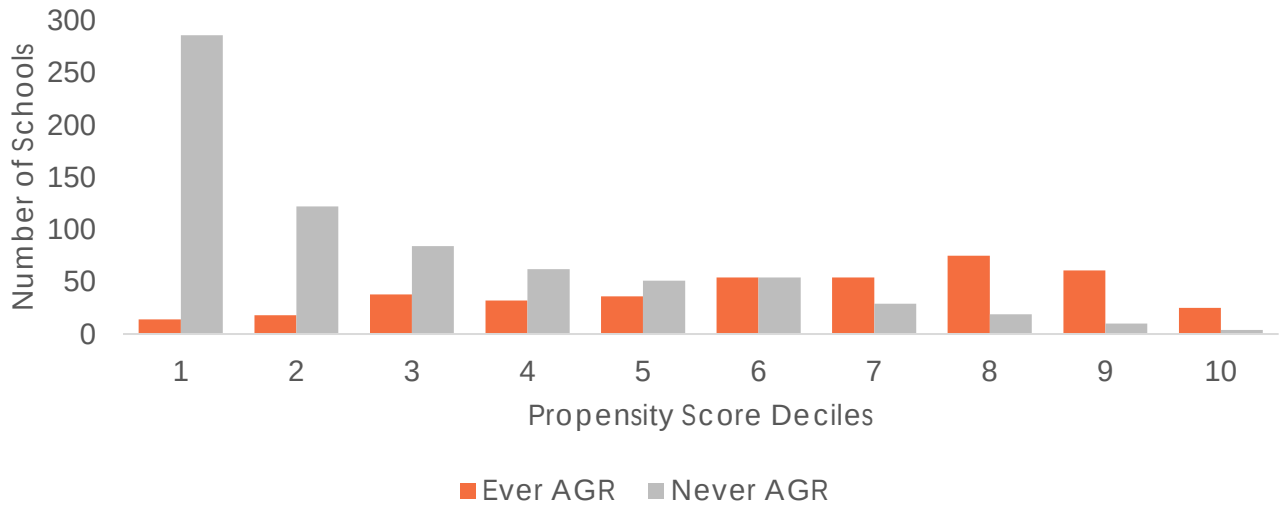


Figure 13 | Common Support for Matching–Attendance and Discipline (2012-13)



Successful matching should also result in balanced covariates across the treatment and control groups. Tables 43 through 46 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.[14] 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. In Tables 43 through 46, no standardized differences reach the 0.25 threshold, and we include all covariates in all impact analyses for double robustness.

Table 43 | Balance of Matched Sample–PALS

	AGR	Non-AGR	P-Value (T/C Difference)	Effect Size
N	90,124	90,663		
Fall PALS Score	-0.17	-0.15	0.01	0.02
Standard Deviation	1.02	1.04		
School Fall PALS Score	-0.18	-0.16	0.00	0.04
Standard Deviation	0.38	0.45		
Female	0.49	0.48	0.34	0.01
Standard Deviation	0.50	0.50		
Black	0.14	0.16	0.00	0.07
Standard Deviation	0.35	0.37		
Hispanic	0.15	0.14	0.00	0.02
Standard Deviation	0.36	0.35		
Other Race	0.10	0.11	0.00	0.04
Standard Deviation	0.30	0.31		
FRL	0.62	0.63	0.00	0.02
Standard Deviation	0.49	0.48		
Special Education	0.14	0.14	0.11	0.01
Standard Deviation	0.34	0.35		
EL	0.11	0.11	0.34	0.01
Standard Deviation	0.31	0.32		
Urban	0.41	0.46	0.00	0.11
Standard Deviation	0.49	0.50		
Suburb	0.09	0.10	0.00	0.05
Standard Deviation	0.28	0.30		
Town	0.18	0.18	0.57	0.00
Standard Deviation	0.38	0.38		

[14] What Works Clearinghouse. (2017). *Procedures and standards handbook* (Version 3.0). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf

Table 43 Continued

	AGR	Non-AGR	P-Value (T/C Difference)	Effect Size
School Population	246.70	245.83	0.15	0.01
Standard Deviation	102.76	99.97		
School Avg Teacher Salary	46,680.76	45,657.18	0.00	0.09
Standard Deviation	8,179.04	12,894.72		
School % Female	0.48	0.48	0.11	0.01
Standard Deviation	0.04	0.04		
School % Black	0.15	0.17	0.00	0.10
Standard Deviation	0.26	0.28		
School % Hispanic	0.15	0.14	0.00	0.04
Standard Deviation	0.19	0.18		
School % Other Race	0.10	0.11	0.00	0.07
Standard Deviation	0.12	0.16		
School % FRL	0.62	0.63	0.00	0.02
Standard Deviation	0.20	0.23		
School % Special Education	0.15	0.15	0.09	0.01
Standard Deviation	0.05	0.05		
School % EL	0.11	0.12	0.00	0.04
Standard Deviation	0.15	0.15		

Table 44 | Balance of Matched Sample–MAP/STAR Math

	AGR	Non-AGR	P-Value (T/C Difference)	Effect Size
N	128,819	130,545		
Fall Math Score	-0.05	-0.06	0.06	0.01
Standard Deviation	1.02	1.03		
Fall Reading Score	-0.19	-0.21	0.00	0.02
Standard Deviation	1.05	1.06		
School Fall Math Score	-0.05	-0.06	0.00	0.02
Standard Deviation	0.40	0.43		
School Fall Read Score	-0.19	-0.21	0.00	0.03
Standard Deviation	0.39	0.43		
Female	0.48	0.49	0.40	0.00
Standard Deviation	0.50	0.50		
Black	0.23	0.21	0.00	0.05
Standard Deviation	0.42	0.41		

Table 44 Continued

	AGR	Non-AGR	P-Value (T/C Difference)	Effect Size
Hispanic	0.15	0.14	0.00	0.02
Standard Deviation	0.35	0.35		
Other Race	0.09	0.10	0.00	0.05
Standard Deviation	0.28	0.30		
FRL	0.65	0.64	0.00	0.03
Standard Deviation	0.48	0.48		
Special Education	0.15	0.15	0.66	0.00
Standard Deviation	0.35	0.35		
EL	0.10	0.10	0.00	0.02
Standard Deviation	0.30	0.31		
Urban	0.54	0.53	0.00	0.02
Standard Deviation	0.50	0.50		
Suburb	0.09	0.09	0.04	0.01
Standard Deviation	0.29	0.29		
Town	0.17	0.15	0.00	0.05
Standard Deviation	0.38	0.36		
School Population	239.74	236.23	0.00	0.04
Standard Deviation	94.55	93.05		
School Avg Teacher Salary	48,408.74	47,746.47	0.00	0.06
Standard Deviation	8,560.55	11,678.57		
School % Female	0.48	0.48	0.00	0.04
Standard Deviation	0.04	0.04		
School % Black	0.23	0.21	0.00	0.06
Standard Deviation	0.33	0.31		
School % Hispanic	0.15	0.14	0.00	0.04
Standard Deviation	0.18	0.17		
School % Other Race	0.09	0.10	0.00	0.10
Standard Deviation	0.10	0.15		
School % FRL	0.66	0.64	0.00	0.07
Standard Deviation	0.21	0.22		
School % Special Education	0.15	0.15	0.00	0.04
Standard Deviation	0.05	0.05		
School % EL	0.10	0.11	0.00	0.05
Standard Deviation	0.14	0.14		

Table 45 | Balance of Matched Sample–MAP/STAR Reading

	AGR	Non-AGR	P-Value (T/C Difference)	Effect Size
N	127,349	129,214		
Fall Math Score	-0.05	-0.06	0.02	0.01
Standard Deviation	1.02	1.03		
Fall Reading Score	-0.19	-0.21	0.00	0.02
Standard Deviation	1.05	1.06		
School Fall Math Score	-0.05	-0.06	0.00	0.02
Standard Deviation	0.40	0.43		
School Fall Read Score	-0.19	-0.20	0.00	0.03
Standard Deviation	0.39	0.43		
Female	0.48	0.49	0.45	0.00
Standard Deviation	0.50	0.50		
Black	0.23	0.21	0.00	0.04
Standard Deviation	0.42	0.41		
Hispanic	0.15	0.14	0.00	0.02
Standard Deviation	0.35	0.35		
Other Race	0.09	0.10	0.00	0.03
Standard Deviation	0.28	0.30		
FRL	0.65	0.64	0.00	0.03
Standard Deviation	0.48	0.48		
Special Education	0.14	0.15	0.75	0.00
Standard Deviation	0.35	0.35		
EL	0.10	0.10	0.00	0.02
Standard Deviation	0.30	0.31		
Urban	0.54	0.53	0.00	0.01
Standard Deviation	0.50	0.50		
Suburb	0.09	0.09	0.00	0.01
Standard Deviation	0.29	0.29		
Town	0.17	0.15	0.00	0.05
Standard Deviation	0.38	0.36		
School Population	239.46	236.15	0.00	0.04
Standard Deviation	94.69	92.79		
School Avg Teacher Salary	48,390.34	47,885.26	0.00	0.05
Standard Deviation	8,584.60	11,419.46		
School % Female	0.48	0.48	0.00	0.03
Standard Deviation	0.04	0.04		

Table 45 Continued

	AGR	Non-AGR	P-Value (T/C Difference)	Effect Size
School % Black	0.23	0.21	0.00	0.05
Standard Deviation	0.33	0.31		
School % Hispanic	0.15	0.14	0.00	0.03
Standard Deviation	0.18	0.17		
School % Other Race	0.09	0.10	0.00	0.08
Standard Deviation	0.10	0.14		
School % FRL	0.66	0.64	0.00	0.07
Standard Deviation	0.21	0.22		
School % Special Education	0.15	0.15	0.00	0.03
Standard Deviation	0.05	0.05		
School % EL	0.10	0.11	0.00	0.06
Standard Deviation	0.14	0.14		

Table 46 | Balance of Matched Sample–Attendance and Discipline

	AGR	Non-AGR	P-Value (T/C Difference)	Effect Size
N	385,013	359,018		
School Attendance Rate 2012-13	0.95	0.95	0.05	0.01
Standard Deviation	0.02	0.02		
School Suspension Rate 2012-13	0.03	0.03	0.00	0.04
Standard Deviation	0.05	0.05		
Female	0.48	0.48	0.27	0.00
Standard Deviation	0.50	0.50		
Black	0.15	0.16	0.00	0.04
Standard Deviation	0.36	0.37		
Hispanic	0.16	0.15	0.00	0.01
Standard Deviation	0.37	0.36		
Other Race	0.10	0.10	0.47	0.00
Standard Deviation	0.29	0.30		
FRL	0.62	0.60	0.00	0.04
Standard Deviation	0.49	0.49		
Special Education	0.15	0.15	0.00	0.01
Standard Deviation	0.36	0.35		

Table 46 Continued

	AGR	Non-AGR	P-Value (T/C Difference)	Effect Size
EL	0.12	0.12	0.00	0.01
Standard Deviation	0.32	0.33		
Urban	0.43	0.44	0.00	0.02
Standard Deviation	0.49	0.50		
Suburb	0.09	0.11	0.00	0.07
Standard Deviation	0.28	0.31		
Town	0.18	0.20	0.00	0.05
Standard Deviation	0.39	0.40		
School Population	248.14	251.69	0.00	0.04
Standard Deviation	98.15	101.29		
School Avg Teacher Salary	47,187.50	47,101.68	0.01	0.01
Standard Deviation	7,801.66	10,372.68		
School % Female	0.48	0.49	0.00	0.05
Standard Deviation	0.04	0.04		
School % Black	0.15	0.17	0.00	0.07
Standard Deviation	0.26	0.27		
School % Hispanic	0.15	0.14	0.00	0.04
Standard Deviation	0.20	0.17		
School % Other Race	0.08	0.08	0.00	0.03
Standard Deviation	0.10	0.08		
School % FRL	0.63	0.62	0.00	0.04
Standard Deviation	0.20	0.22		
School % Special Education	0.15	0.14	0.00	0.12
Standard Deviation	0.05	0.05		
School % EL	0.12	0.12	0.00	0.03
Standard Deviation	0.16	0.14		

Impact Analysis

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

$$Y_{isgy} = \alpha + \gamma_0 SAGE_{sy} + \gamma_1 AGR_{sy} + \beta X_{iy} + \pi Z_{sy} + \partial_{gy} + \varepsilon_{isgy}$$

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.[15] All analysis variables are described in Table 2 above. 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.

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.

Table 47 | Outcome Summary Statistics

Outcome	Mean	S.D.
PALS	-0.232	1.308
MAP/STAR Math	0.066	1.09
MAP/STAR Reading	-0.139	1.078
Absence rate	0.06	0.068
Suspension rate	0.027	0.163

Note: Statistics are weighted and sample-specific.

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 Tables 48 through 52 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

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

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 would only allow 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 Tables 48 through 52, 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 the between school and within school effects.[16] 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 Tables 48 through 52 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 Tables 48 through 52 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 48), the SAGE impact increases and the AGR impact decreases. For both math and reading (Tables 49 and 50, respectively), estimated AGR impacts increase substantially, while time trends have little effect on attendance or discipline impacts (Tables 51 and 52, respectively).

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

Table 48 | Robustness to Alternative Estimation Strategies–PALS

	(1)	(2)	(3)	(4)
SAGE vs. non	0.014		NA	0.047
P-value	0.522		NA	0.153
AGR vs. non	0.121		NA	0.076
P-value	0.003		NA	0.151
AGR vs. SAGE	0.108	0.04	NA	0.029
P-value	0.004	0.123	NA	0.303
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 49| Robustness to Alternative Estimation Strategies–Math

	(1)	(2)	(3)	(4)
SAGE vs. non	-0.013		NA	0.031
P-value	0.414		NA	0.275
AGR vs. non	-0.003		NA	0.051
P-value	0.867		NA	0.257
AGR vs. SAGE	0.01	-0.021	NA	0.02
P-value	0.609	0.26	NA	0.414
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 50 | Robustness to Alternative Estimation Strategies–Reading

	(1)	(2)	(3)	(4)
SAGE vs. non	0.011		NA	0.069
P-value	0.014		NA	0.018
AGR vs. non	0.018		NA	0.117
P-value	0.253		NA	0.005
AGR vs. SAGE	0.007	-0.013	NA	0.048
P-value	0.655	0.409	NA	0.013
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 51 | Robustness to Alternative Estimation Strategies–Absences

	(1)	(2)	(3)	(4)
SAGE vs. non	0.115		0.1	0.11
P-value	0.001		0.01	0.006
AGR vs. non	0.062		0.065	0.078
P-value	0.019		0.071	0.12
AGR vs. SAGE	-0.053	-0.034	-0.034	-0.032
P-value	0.019	0.101	0.099	0.226
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 52 | Robustness to Alternative Estimation Strategies–Suspensions

	(1)	(2)	(3)	(4)
SAGE vs. non	-0.005		-0.003	-0.003
P-value	0.007		0.106	0.435
AGR vs. non	-0.004		-0.003	0
P-value	0.127		0.188	0.915
AGR vs. SAGE	0.001	-0.001	0	0.002
P-value	0.689	0.718	0.983	0.337
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, as measured by p-values.[17]

Table 53 below shows results of the Benjamini-Hochberg procedure for AGR's main and subgroup impacts. According to the 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 (in this case, 64). 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 significant, we calculate an adjusted p-value from the same formula used to produce the critical value.[18]

After correcting for multiple comparisons, half of the AGR impact estimates with unadjusted p-values below 0.05 are no longer statistically significant. Only the strongest estimates, those with unadjusted p-values less than 0.01, remain significant. These results include mostly PALS estimates along with some subgroup impacts for reading, attendance, and discipline.

[17] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300.
[18] Specifically, the number of comparisons (64), multiplied by the false discover rate (0.05), divided by the rank.

Table 53 | Results of the Benjamini-Hochberg Procedure for Multiple Comparisons–AGR Statewide and Subgroup Impacts

Outcome	Model	Coeff.	P-Value	Rank	Critical Value	Adj. P-Value	Statistically Significant after Multiple Comparisons Correction
PALS	Subgroup–EL	0.376	0.000	1	0.001	0.000	Yes
PALS	Subgroup–Hispanic	0.258	0.000	2	0.002	0.000	Yes
PALS	Subgroup–Asian	0.245	0.001	3	0.002	0.021	Yes
PALS	Subgroup–City	0.232	0.001	4	0.003	0.016	Yes
PALS	Subgroup–Female	0.122	0.002	5	0.004	0.026	Yes
Absence Rate	Subgroup–City	0.078	0.002	6	0.005	0.021	Yes
PALS	Subgroup–FRL	0.152	0.003	7	0.005	0.027	Yes
PALS	Statewide	0.121	0.003	8	0.006	0.024	Yes
Absence Rate	Subgroup–Grade 3	0.074	0.006	9	0.007	0.043	Yes
Suspensions	Subgroup–Hispanic	-0.009	0.006	10	0.008	0.038	Yes
Suspensions	Subgroup–EL	-0.007	0.007	11	0.009	0.041	Yes
PALS	Subgroup–Black	0.23	0.018	12	0.009	0.096	No
Absence Rate	Subgroup–Grade 2	0.066	0.019	13	0.010	0.094	No
Absence Rate	Statewide	0.062	0.019	14	0.011	0.087	No
Absence Rate	Subgroup–Hispanic	0.073	0.023	15	0.012	0.098	No
Reading	Subgroup–Hispanic	0.048	0.023	16	0.013	0.092	No
Absence Rate	Subgroup–Female	0.057	0.028	17	0.013	0.105	No
Absence Rate	Subgroup–EL	0.091	0.03	18	0.014	0.107	No
Absence Rate	Subgroup–Black	0.07	0.034	19	0.015	0.115	No
PALS	Subgroup–White	0.066	0.034	20	0.016	0.109	No
Suspensions	Subgroup–Grade K	-0.006	0.04	21	0.016	0.122	No
Absence Rate	Subgroup–Grade 1	0.055	0.043	22	0.017	0.125	No

Table 53 Continued

Outcome	Model	Coeff.	P-Value	Rank	Critical Value	Adj. P-Value	Statistically Significant after Multiple Comparisons Correction
Absence Rate	Subgroup–Grade K	0.05	0.05	23	0.018	0.139	No
Reading	Subgroup–Grade 2	0.04	0.056	24	0.019	0.149	No
Reading	Subgroup–EL	0.043	0.094	25	0.020	0.241	No
Reading	Subgroup–White	0.026	0.104	26	0.020	0.256	No
Suspensions	Subgroup–Grade 2	-0.006	0.115	27	0.021	0.273	No
Absence Rate	Subgroup–White	0.055	0.117	28	0.022	0.267	No
Suspensions	Subgroup–Black	-0.008	0.118	29	0.023	0.260	No
Suspensions	Subgroup–FRL	-0.006	0.121	30	0.023	0.258	No
Suspensions	Statewide	-0.004	0.127	31	0.024	0.262	No
Math	Subgroup–Hispanic	0.041	0.137	32	0.025	0.274	No
Suspensions	Subgroup–Female	-0.002	0.153	33	0.026	0.297	No
Math	Subgroup–Black	-0.056	0.166	34	0.027	0.312	No
Absence Rate	Subgroup–FRL	0.039	0.179	35	0.027	0.327	No
Reading	Subgroup–Female	0.02	0.192	36	0.028	0.341	No
Math	Subgroup–Grade 3	-0.028	0.193	37	0.029	0.334	No
Suspensions	Subgroup–City	-0.006	0.212	38	0.030	0.357	No
Suspensions	Subgroup–Grade 1	-0.004	0.222	39	0.030	0.364	No
Reading	Subgroup–Other Race/Ethnicity	0.028	0.233	40	0.031	0.373	No
Absence Rate	Subgroup–Other Race/Ethnicity	0.061	0.237	41	0.032	0.370	No
Reading	Statewide	0.018	0.253	42	0.033	0.386	No
Absence Rate	Subgroup–Asian	0.051	0.426	43	0.034	0.634	No
Math	Subgroup–Grade 2	0.023	0.431	44	0.034	0.627	No

Table 53 Continued

Outcome	Model	Coeff.	P-Value	Rank	Critical Value	Adj. P-Value	Statistically Significant after Multiple Comparisons Correction
Suspensions	Subgroup–Grade 3	-0.003	0.444	45	0.035	0.631	No
Suspensions	Subgroup–Other Race/Ethnicity	-0.004	0.449	46	0.036	0.625	No
Reading	Subgroup–Black	-0.026	0.463	47	0.037	0.630	No
Reading	Subgroup–FRL	0.014	0.464	48	0.038	0.619	No
Suspensions	Subgroup–Asian	-0.001	0.465	49	0.038	0.607	No
Math	Subgroup–Grade 1	0.034	0.466	50	0.039	0.596	No
Math	Subgroup–EL	0.025	0.507	51	0.040	0.636	No
Math	Subgroup–FRL	-0.015	0.51	52	0.041	0.628	No
Math	Subgroup–Asian	-0.038	0.513	53	0.041	0.619	No
Reading	Subgroup–Grade 3	0.01	0.541	54	0.042	0.641	No
PALS	Subgroup–Other Race/Ethnicity	0.032	0.556	55	0.043	0.647	No
Math	Subgroup–Other Race/Ethnicity	-0.016	0.583	56	0.044	0.666	No
Reading	Subgroup–City	0.011	0.64	57	0.045	0.719	No
Math	Subgroup–White	0.008	0.716	58	0.045	0.790	No
Math	Subgroup–City	-0.008	0.79	59	0.046	0.857	No
Reading	Subgroup–Grade 1	-0.01	0.805	60	0.047	0.859	No
Math	Statewide	-0.003	0.867	61	0.048	0.910	No
Suspensions	Subgroup–White	0	0.885	62	0.048	0.914	No
Reading	Subgroup–Asian	0.004	0.915	63	0.049	0.930	No
Math	Subgroup–Female	0	0.992	64	0.050	0.992	No

In addition, we apply the Benjamini-Hochberg procedure to estimates of the differences in impacts between AGR and SAGE. Table 54 displays these results. Similar to the AGR impacts in Table 53, few of the AGR-SAGE comparisons remain statistically significant post-procedure. PALS statewide comparisons, as well as several PALS subgroups, the results with large, positive estimates of AGR impacts, remain significant.

Table 54 | Results of the Benjamini-Hochberg Procedure for Multiple Comparisons–AGR-SAGE Comparisons

Outcome	Model	Coeff.	P-Value	Rank	Critical Value	Adj. P-Value	Statistically Significant after Multiple Comparisons Correction
PALS	Subgroup–EL	0.415	0.000	1	0.001	0.000	Yes
PALS	Subgroup–Hispanic	0.319	0.000	2	0.002	0.000	Yes
PALS	Subgroup–City	0.196	0.000	3	0.002	0.000	Yes
PALS	Subgroup–FRL	0.137	0.001	4	0.003	0.016	Yes
PALS	Subgroup–Female	0.11	0.003	5	0.004	0.038	Yes
PALS	Statewide	0.108	0.004	6	0.005	0.043	Yes
Absence Rate	Subgroup–White	-0.064	0.006	7	0.005	0.055	No
Absence Rate	Subgroup–Other Race/Ethnicity	-0.072	0.013	8	0.006	0.104	No
Absence Rate	Subgroup–Grade K	-0.06	0.013	9	0.007	0.092	No
Absence Rate	Subgroup–Female	-0.055	0.015	10	0.008	0.096	No
PALS	Subgroup–Asian	0.163	0.016	11	0.009	0.093	No
Math	Subgroup–Black	-0.067	0.017	12	0.009	0.091	No
Absence Rate	Statewide	-0.053	0.019	13	0.010	0.094	No
Absence Rate	Subgroup–Grade 1	-0.055	0.02	14	0.011	0.091	No
Absence Rate	Subgroup–Grade 2	-0.052	0.02	15	0.012	0.085	No
Math	Subgroup–EL	0.058	0.03	16	0.013	0.120	No
Reading	Subgroup–Black	-0.049	0.04	17	0.013	0.151	No
Absence Rate	Subgroup–FRL	-0.045	0.047	18	0.014	0.167	No
Absence Rate	Subgroup–Grade 3	-0.045	0.048	19	0.015	0.162	No
Math	Subgroup–Grade 2	0.045	0.051	20	0.016	0.163	No

Table 54 | Continued

Outcome	Model	Coeff.	P-Value	Rank	Critical Value	Adj. P-Value	Statistically Significant after Multiple Comparisons Correction
PALS	Subgroup–Other Race/Ethnicity	0.091	0.052	21	0.016	0.158	No
PALS	Subgroup–Black	0.104	0.059	22	0.017	0.172	No
Reading	Subgroup–White	0.03	0.061	23	0.018	0.170	No
Math	Subgroup–Grade 1	0.068	0.065	24	0.019	0.173	No
PALS	Subgroup–White	0.059	0.073	25	0.020	0.187	No
Absence Rate	Subgroup–Black	-0.049	0.073	26	0.020	0.180	No
Absence Rate	Subgroup–Asian	-0.051	0.102	27	0.021	0.242	No
Suspensions	Subgroup–Grade 3	0.005	0.103	28	0.022	0.235	No
Absence Rate	Subgroup–City	-0.038	0.109	29	0.023	0.241	No
Math	Subgroup–Grade 3	-0.03	0.149	30	0.023	0.318	No
Math	Subgroup–White	0.029	0.151	31	0.024	0.312	No
Math	Subgroup–Asian	0.047	0.222	32	0.025	0.444	No
Math	Subgroup–Hispanic	0.03	0.222	33	0.026	0.431	No
Reading	Subgroup–EL	0.024	0.277	34	0.027	0.521	No
Reading	Subgroup–Other Race/Ethnicity	0.024	0.302	35	0.027	0.552	No
Reading	Subgroup–Grade 2	0.018	0.316	36	0.028	0.562	No
Reading	Subgroup–City	-0.018	0.334	37	0.029	0.578	No
Math	Subgroup–Other Race/Ethnicity	0.027	0.339	38	0.030	0.571	No
Absence Rate	Subgroup–EL	-0.025	0.405	39	0.030	0.665	No

Table 54 | Continued

Outcome	Model	Coeff.	P-Value	Rank	Critical Value	Adj. P-Value	Statistically Significant after Multiple Comparisons Correction
Suspensions	Subgroup–Hispanic	-0.001	0.437	40	0.031	0.699	No
Math	Subgroup–Female	0.012	0.535	41	0.032	0.835	No
Suspensions	Subgroup–EL	-0.001	0.549	42	0.033	0.837	No
Suspensions	Subgroup–FRL	0.001	0.565	43	0.034	0.841	No
Suspensions	Subgroup–City	0.002	0.576	44	0.034	0.838	No
Reading	Subgroup–Female	0.009	0.583	45	0.035	0.829	No
Math	Statewide	0.01	0.609	46	0.036	0.847	No
Suspensions	Subgroup–Grade 1	-0.001	0.611	47	0.037	0.832	No
Reading	Subgroup–Grade 1	0.018	0.612	48	0.038	0.816	No
Suspensions	Subgroup–Black	0.002	0.626	49	0.038	0.818	No
Absence Rate	Subgroup–Hispanic	-0.012	0.639	50	0.039	0.818	No
Suspensions	Subgroup–White	0.001	0.642	51	0.040	0.806	No
Reading	Statewide	0.007	0.655	52	0.041	0.806	No
Reading	Subgroup–Grade 3	-0.007	0.657	53	0.041	0.793	No
Suspensions	Statewide	0.001	0.689	54	0.042	0.817	No
Suspensions	Subgroup–Grade 2	-0.001	0.727	55	0.043	0.846	No
Reading	Subgroup–Hispanic	-0.006	0.778	56	0.044	0.889	No
Suspensions	Subgroup–Female	0	0.787	57	0.045	0.884	No
Reading	Subgroup–FRL	-0.004	0.818	58	0.045	0.903	No
Math	Subgroup–FRL	-0.003	0.865	59	0.046	0.938	No
Reading	Subgroup–Asian	0.005	0.869	60	0.047	0.927	No
Suspensions	Subgroup–Grade K	0	0.95	61	0.048	0.997	No
Suspensions	Subgroup–Other Race/Ethnicity	0	0.954	62	0.048	0.985	No

Table 54 | Continued

Outcome	Model	Coeff.	P-Value	Rank	Critical Value	Adj. P-Value	Statistically Significant after Multiple Comparisons Correction
Math	Subgroup–City	-0.001	0.955	63	0.049	0.970	No
Suspensions	Subgroup–Asian	0	0.957	64	0.050	0.957	No

Finally, we also conducted the Benjamini-Hochberg procedure to estimates of AGR by strategy. Due to size, a table containing these results is available upon request.

SURVEY APPENDIX

1. Which of the following strategies were used in K-3 classrooms during the school year? (Please select all that apply.) [*forced answer*]
   - Reduced class size (either 18:1 or 30:2) Yes No
   - One-to-one tutoring Yes No
   - Instructional coaching Yes No

2. [*If indicated reduced class size*] What percentage of classrooms in each grade have a reduced class size? If your school does not use semesters, use the first half of the year and the second half of the year as your indicators.

Grade	Semester	Percent of Classrooms				
		None	Less than 25%	25-50%	51-75%	More than 75%
Kindergarten	1					
	2					
First	1					
	2					
Second	1					
	2					
Third	1					
	2					

3. [*If indicated reduced class size*] Because of the AGR program in your school, what instructional strategies are reduced class size teachers using with students? (Please select all that apply.)
   - We don't use any specific instructional strategies because of smaller class sizes
   - Small-group instruction
   - One-on-one time with the teacher
   - Differentiation of instruction
   - Strategic placement of students in groups
   - Strategic placement of students in classrooms
   - Other _____
   - Not sure/don't know

4. [*If indicated reduced class size*] To what extent does the reduced class size provide the following benefits for your school?

Benefit	A lot	A moderate amount	A little	None	Not sure
Better common planning among teachers					
Better interactions among students					
Better relationships between teachers and students					
Better teacher morale					
Increased student ownership of learning					
Better teacher performance					
Increased use of data among teachers					
More participation from students in class					
More time for individual interactions					
Reduction in student anxiety					
Reduction in student behavioral problems					
Students engage in student-specific interventions					
Teachers value contributions of *all* students					

5. [*If indicated reduced class size*] What other benefits does the reduced class size provide for your school? _____

6. [*If indicated one-to-one tutoring*] What percentage of classrooms in each grade have one-to-one tutoring? If your school does not use semesters, use the first half of the year and the second half of the year as your indicators.

Grade	Semester	Percent of Classrooms				
		None	Less than 25%	25-50%	51-75%	More than 75%
Kindergarten	1					
	2					
First	1					
	2					
Second	1					
	2					
Third	1					
	2					

7. [*If indicated one-to-one tutoring*] On average, how often are AGR one-to-one tutors meeting with students?
   - 3/week or more
   - 2/week
   - Weekly
   - Biweekly
   - Monthly
   - As needed
   - Not sure/don't know

8. [*If indicated one-to-one tutoring*] What practices does your one-to-one tutor use? (Please select all that apply.)
   - Reviews student data
   - Models appropriate learning behavior
   - Adapts to student learning styles
   - Maintains a focus on equity
   - Provides scaffolding
   - Communicates regularly with classroom teacher
   - Not sure/don't know

9. [*If indicated one-to-one tutoring*] To what extent does the one-to-one tutoring from the AGR program provide the following benefits for your school?

Benefit	A lot	A moderate amount	A little	None	Not sure
Better common planning among teachers					
Better interactions among students					
Better relationships between teachers and students					
Better teacher morale					
Increased student ownership of learning					
Better teacher performance					
Increased use of data among teachers					
More participation from students in class					
More time for individual interactions					
Reduction in student anxiety					
Reduction in student behavioral problems					
Students engage in student-specific interventions					
Teachers value contributions of *all* students					

10. [*If indicated one-to-one tutoring*] What other benefits does the one-to-one tutoring provide for your school? _____

11. [*If indicated instructional coaching*] What percentage of teachers in each grade have received instructional coaching? If your school does not use semesters, use the first half of the year and the second half of the year as your indicators.

Grade	Semester	Percent of Classrooms				
		None	Less than 25%	25-50%	51-75%	More than 75%
Kindergarten	1					
	2					
First	1					
	2					
Second	1					
	2					
Third	1					
	2					

12. [*If indicated instructional coaching*] Which of the following characteristics do your AGR instructional coaches have? (Please select all that apply.)
    - Coach training
    - Previous instructional coaching experience
    - Content specialist in their subject of coaching
    - Not sure/don't know

13. [*If indicated instructional coaching*] On average, how often are AGR instructional coaches meeting with the teachers they coach?
    - Weekly
    - Monthly
    - Quarterly
    - Each semester
    - As needed
    - Not sure/don't know

14. [*If indicated instructional coaching*] What practices do your instructional coaches use? (Please select all that apply.)
    - One-to-one teacher coaching
    - Team teacher coaching
    - Keeps a coaching log
    - Advises teachers to set goals
    - Coaching focuses on teacher goals
    - Maintains a focus on equity
    - Encourages reflective practices
    - Discusses data with teachers
    - Observes teacher practices
    - Not sure/don't know

15. [*If indicated instructional coaching*] To what extent does the instructional coaching provide the following benefits for your school?

Benefit	A lot	A moderate amount	A little	None	Not sure
Better common planning among teachers					
Better interactions among students					
Better relationships between teachers and students					
Better teacher morale					
Increased student ownership of learning					
Better teacher performance					
Increased use of data among teachers					
More participation from students in class					
More time for individual interactions					
Reduction in student anxiety					
Reduction in student behavioral problems					
Students engage in student-specific interventions					
Teachers value contributions of *all* students					

16. [*If indicated instructional coaching*] What other benefits does the instructional coaching program provide for your school? _____

17. Which assessments are you using to measure student progress in each grade during the school year (ex. MAP, PALS, STAR)? (Please list all that apply.)
    - Kindergarten _____
    - First _____
    - Second _____
    - Third _____

Acknowledgments

The authors would like to acknowledge Brielle Glatzel, Sharon Suchla, and Jonas Zuckerman at the Department of Public Instruction. Their collaboration and insight was essential to this report. The authors would also acknowledge the contributions of Alison Bowman, who designed and formatted this report and accompanying documents.

About the Authors

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

Brie Chapa is a Knowledge Manager with the Wisconsin Evaluation Collaborative. She holds a B.S. in Secondary Mathematics Education 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.