Wisconsin Evaluation Collaborative

August 2022

Evaluation of the Achievement Gap Reduction Program

2015-16 through 2020-21

for the Wisconsin Department of Public Instruction

WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

Wisconsin Center for Education Research
UNIVERSITY OF WISCONSIN-MADISON

About the Authors

Jed Richardson

Jed Richardson is a Scientist with the Wisconsin Evaluation Collaborative. He holds a Ph.D. in Economics from the University of California, Davis.

Grant Sim

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

About the Wisconsin Evaluation Collaborative

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

Contact

Jed Richardson

jed.richardson@wisc.edu

Acknowledgements

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

Copyright Information

Contents

Section 1

Executive Summary

Executive Summary

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

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

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

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

1 Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessman, L. (2020). Long-run Trends in the U.S. SES-Achievement Gap [Working paper 26764]. National Bureau of Economic Research. https://www.nber.org/papers/w26764

2 2015 Wisconsin Act 53. Wisconsin Senate. Section 118.44. https://docs.legis.wisconsin.gov/2015/related/acts/53.pdf

b. How does AGR's impact on outcomes compare to impacts associated with the SAGE program?

c. Are there differences between the three AGR strategies' relationships to intended outcomes?

3. Are there differences between the three AGR strategies' impacts on intended outcomes?

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

This year's evaluation methodology improves on previous evaluations by taking advantage of the growing number of cohorts that have received AGR funds. By 2020-21, two cohorts of Wisconsin students had received AGR throughout grades K-3, enabling estimation of the impacts of four years of AGR on statewide, third grade Forward exam scores. Relative to previous estimations of AGR impacts on MAP/STAR test scores, the revised methodology removes a potential source of statistical bias, provides results for the more policy-relevant statewide Forward exam, and includes a greater cross-section of AGR students.

The evaluation methodology also makes several adjustments to address complexities arising from the COVID-19 pandemic that have the potential to bias estimates of AGR impacts. These complexities include missing test score data, attendance and suspension outcomes that changed due to the pandemic, and differences in virtual and in-person learning models across AGR and non-AGR schools.

How are AGR schools implementing the program?

In 2020-21, the most recent year of data, 408 schools implemented the AGR program, serving over 71,000 students in kindergarten through third grades. As previously noted, to fulfill AGR obligations schools could implement any combination of three strategies: reduced class size, instructional coaching, and/or tutoring.

- From 2017-18 to 2019-20, the use of multiple strategies steadily increased from 63 percent of schools to 73 percent. In 2020-21, 64 percent of schools utilized multiple strategies, perhaps in response to the COVID-19 pandemic.
- Over 80 percent of schools use reduced class sizes in at least one grade. At the school level, the three most common strategy choices are class size reduction and instructional coaching combined, class size reduction alone, and using all three strategies.
- Despite strong evidence in the education literature indicating the effectiveness of tutoring for improving achievement, comparatively few AGR schools chose tutoring, either on its own or in combination with other strategies.

To what extent is AGR meeting intended outcomes?

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

- There is no estimated impact of the AGR program on third grade Forward reading or math for both the statewide sample and the sample of students who receive free or reduced-price lunch. Seen in conjunction with previous evaluations showing positive and significant impacts of AGR on kindergarten reading, the lack of estimated impacts on third grade Forward implies that AGR reading impacts in kindergarten fade out by third grade or that there is misalignment between kindergarten and Forward tests.
- Preliminary evidence is consistent with an "implementation dip" after AGR began, then improving impacts thereafter, particularly for third grade Forward math.
- There is no estimated impact of the AGR program on statewide attendance or out-of-school suspension rates.
- For all outcomes, AGR has similar impacts to SAGE, its predecessor program.

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

The evaluation provides evidence of impacts depending on the AGR strategies that schools choose. We find preliminary evidence that choosing reduced class sizes results in fewer suspensions, but no evidence that strategy choice impacts other outcomes.

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

Section 2

Introduction

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 with the enactment of Wisconsin Acts 53 and 71. Wisconsin began a gradual phase-in of AGR in 2015-16 by transitioning schools from SAGE to AGR, with the final phase out of previous SAGE programs by the end of the 2017-18 school year. Like SAGE, AGR targets funding to schools with economically disadvantaged students through contracts to implement the program in kindergarten through third grade. Each year, the state provides approximately $110,000,000 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:

1. 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.
2. Provide data-driven instructional coaching for one or more teachers of one or more participating grades. The instruction shall be provided by licensed teachers who possess appropriate content knowledge to assist classroom teachers in improving instruction in math or reading and possess expertise in reducing the achievement gap.
3. Provide data-informed, one-to-one tutoring to pupils in the class who are struggling with reading or mathematics or both subjects. Tutoring shall be provided during regular school hours by a licensed teacher using an instructional program to be found effective by the What Works Clearinghouse of the Institute of Education Sciences.[4]

4 2015 Wisconsin Act 53. Wisconsin Senate. Section 118.44. https://docs.legis.wisconsin.gov/2015/related/acts/53.pdf

Context

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

5 Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). Equality of Educational Opportunity. National Center for Educational Statistics.

6 Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessman, L. (2020). Long-run Trends in the U.S. SES-Achievement Gap [Working paper 26764]. National Bureau of Economic Research. https://www.nber.org/papers/w26764

7 Burdick-Will, J., Ludwig, J., Raudenbush, S. W., Sampson, R. J., Sanbonmatsu, L., & Sharkey, P. (2011). Converging Evidence for Neighborhood Effects on Children's Test Scores: An Experimental, Quasi-Experimental, and Observational Comparison. In G. J. Duncan & R. J. Murnane (Eds.), Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances (pp. 255-276). Russell Sage Foundation.

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

9 Guryan, J., Hurst E., & Kearney, M. (2008). Parental Education and Parental Time with Children. Journal of Economic Perspectives 22(3), 23-46. https://doi.org/10.1257/jep.22.3.23

10 Nelson, C. A., & Sheridan, M. A. (2011). Lessons from Neuroscience Research for Understanding Causal Links Between Family and Neighborhood Characteristics and Educational Outcomes. In G. J. Duncan & R. J. Murnane (Eds.), Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances (pp. 27-46). Russell Sage Foundation.

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

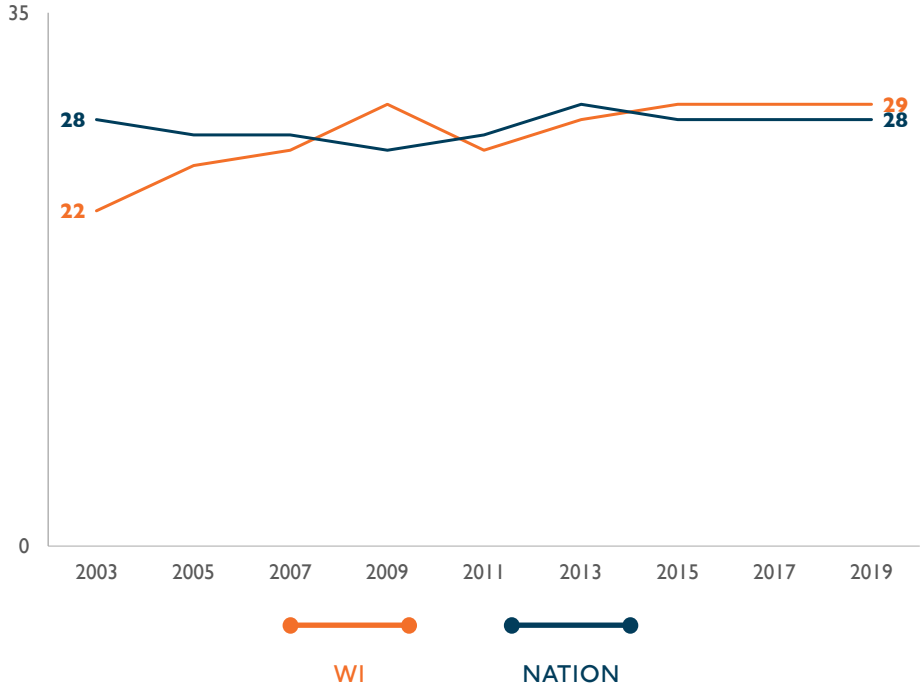
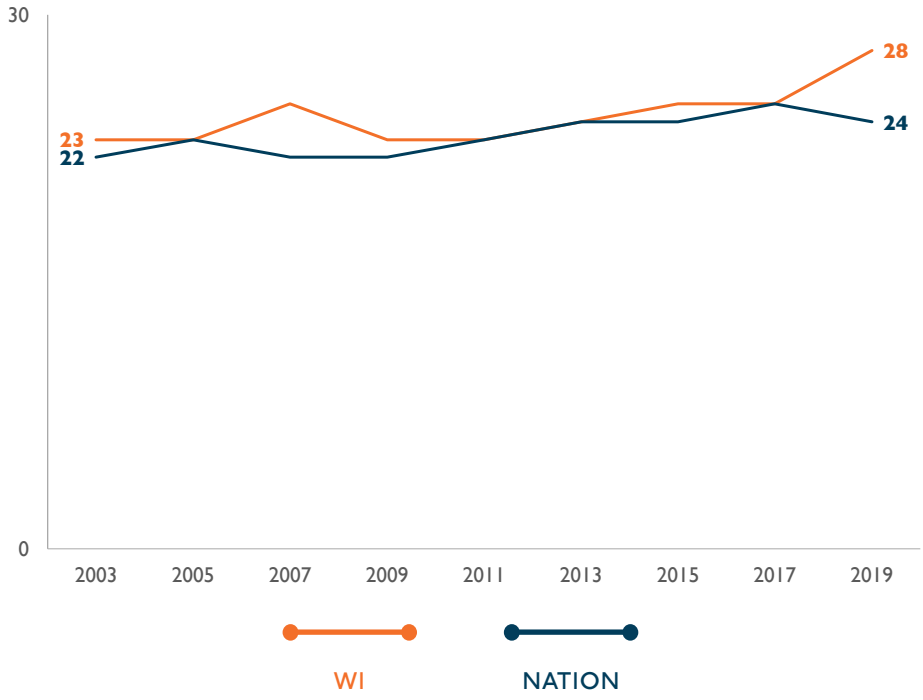


Figure 2: Grade 4 Math Socioeconomic Achievement Gaps, 2003-2019



AGR's strategy for closing the socioeconomic achievement gap is to provide additional funding to districts with large proportions of economically disadvantaged students. This strategy is consistent with the academic literature concerning how school funding impacts student outcomes. In recent research that is especially pertinent to AGR, Jackson, et al. (2021) use plausibly exogenous variation in school funding associated with the Great Recession to show that decreases in school resources widened the socioeconomic achievement gap.[11]

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

11 Jackson, C. K., Wigger, C., & Xiong, H. (2021). Do School Spending Cuts Matter? Evidence from the Great Recession. American Economic Journal: Economic Policy, 13(2), 304-35. https://doi.org/10.1257/pol.20180674

12 Jackson, C. K., & Mackevicius, C. (2021). The Distribution of School Spending Impacts [Working paper 28517]. National Bureau of Economic Research. https://www.nber.org/papers/w28517

13 Jackson, C. K., Johnson, R. C., & Persico, C. (2016). The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms. The Quarterly Journal of Economics, 131(1), 157-218. https://doi.org/10.1093/qje/qjv036

This Evaluation

2015 Wisconsin Acts 53 and 71 include a provision for an annual evaluation of the AGR program starting in the 2018-19 school year. DPI contracted with the Wisconsin Evaluation Collaborative (WEC) within the Wisconsin Center for Education Research at the University of Wisconsin–Madison for these evaluation services. This report provides results from the evaluation of the AGR program from 2015-16 through 2020-21.

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

1. How are AGR schools implementing the AGR program as specified by 2015 Wisconsin Acts 53 and 71?
    a. What is the breakdown of implementation with regard to the three strategies?
    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 impact vary by student characteristics?
    c. 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?

Although the evaluation questions remain the same as in other years, this year's evaluation methodology and report differ. In previous years, we evaluated the program's year-to-year impact on math and reading learning by using fall and spring MAP and STAR exams. A limitation of this strategy is that statistical bias may occur because prior year's AGR impacts are "baked in" to previous outcomes used for matching AGR schools to similar, non-AGR schools.[14] By 2020-21, however, two cohorts of Wisconsin students had received AGR throughout grades K-3, enabling estimation of the impacts of 4 years of AGR on statewide, third grade Forward exam scores. In addition to removing the potential statistical bias that can occur with year-to-year estimations of AGR impacts, the new evaluation methodology using the statewide Forward exam includes a greater cross-section of AGR students than MAP and STAR, which are used by a minority of districts.

This evaluation, like last year's, is more complex due to the COVID-19 pandemic. School shutdowns from the pandemic prevented testing for all districts in spring 2020 and for many in fall 2020 and spring 2021, complicating measurement of test score growth. At the same time, districts chose different learning models (in-person, virtual, and hybrid) depending on public health and learning considerations in their local contexts. AGR schools, which are more likely than non-AGR schools to be located in urban areas, were more likely to choose hybrid and virtual instruction. In addition, there is evidence that in-person learning was more effective for academic achievement growth during the pandemic.[15] The correlation between schools' learning models and participation in AGR can create a negative statistical bias that would result in underestimation of AGR's effects on achievement. Therefore, to best estimate AGR's impacts, the evaluation must address differences in learning models. This is not to second-guess districts' decisions during the pandemic. Districts chose learning models within their unique local contexts, taking into account the prevalence of COVID-19, transportation infrastructure, internet availability, and many other factors, not just to maximize learning, but also to protect student and public

14 Note that this limitation does not apply to previous years' PALS results, which use a pre-treatment baseline measure of reading performance (fall PALS).

15 Goldhaber, D., Kane, T. J., McEachin, A., Morton, E., Patterson, T., & Staiger, D. O. (2022). The Consequences of Remote and Hybrid Instruction During the Pandemic [Working Paper 30010]. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w30010; Jack, R., Halloran, C., Okun, J., & Oster, E. (2021). Pandemic Schooling Mode and Student Test Scores: Evidence from US States [Working Paper 29497]. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w29497/w29497.pdf

health. As a result, addressing differences in learning models as part of this evaluation does not imply that some districts' choices were better than others, only that learning may have systematically differed across AGR and non-AGR schools for reasons unrelated to AGR itself.

This report has eight main sections including the Introduction. Evaluation Data and Methodology includes details on data, analysis designs, and statistical models used to evaluate program impacts, as well as the limitations of this evaluation. AGR Demographics describes characteristics of AGR students and schools and the resulting analysis samples. AGR Implementation describes the strategies schools choose. AGR Impacts provides the results of analyses of AGR impacts on Forward reading and math growth, as well as impacts on absences and out-of-school suspensions (OSS). This section provides overall impacts, impacts of AGR compared to SAGE, and impacts for low-income students. Longitudinal Analyses of End-of-Year (EOY) and School Board Reports looks at these data from 2017-18 through 2020-21. The last two sections are comprised of the Summary and a Technical Appendix.

Section 3

Evaluation Data and Methodology

Evaluation Data and Methodology

In order to understand how AGR impacts student achievement outcomes, and to compare AGR's impacts to those of its predecessor, SAGE, we must identify a plausible comparison group of schools and students. Because AGR targets higher-poverty schools where outcomes are lower on average, naïve comparisons of AGR schools' outcomes to those of other Wisconsin schools would 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 2020-21 academic years. In addition, for purposes of comparison, the evaluation includes K-3 students at subsets of non-AGR, non-SAGE schools.

Data

To identify plausibly equivalent, non-AGR schools for a comparison group, and to estimate impacts, the evaluation combines several sources of student- and school-level data for the academic years 2012-13 through 2020-21. Student-level achievement test data, student demographics, and enrollment records come from DPI administrative data.

DPI also provided school-level data on AGR and SAGE funding by year. School-level teacher average salaries are sourced from DPI Public Staff Reports, and school location information comes from the National Center for Education Statistics (NCES).[16] Finally, data on district-level median income and poverty levels are also sourced from NCES, which uses data from the U.S. Census Bureau's American Community Survey.[17]

- 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 (FRL) eligibility. School- and grade-level measures of demographic characteristics are calculated from student-level data.
- Achievement test data include third grade spring Forward test scores and fall and spring kindergarten reading scores from the Phonological Awareness Literacy Screening (PALS). Achievement test data also include fall and spring administrations of the MAP and STAR. For Grades 1-3, MAP and STAR scores were equated and combined into a single test measure in order to attain a sufficient student sample.[18]
- Attendance data consist of total days absent and total attendance days. The associated outcome variable is the absence rate, the total days absent divided by total possible attendance days.

16 Public Staff Reports are available at https://publicstaffreports.dpi.wi.gov/PubStaffReport/Public/PublicReport. School location information available at https://nces.ed.gov/ccd/elsi/.

17 District-level American Community Survey results available at https://nces.ed.gov/programs/edge/demographic/acs.

18 In the Spring of 2020, nearly all Wisconsin schools opted to forgo assessments due to the COVID-19 pandemic. To address the lack of these scores, we estimated a model that uses scores from Fall 2019 and Winter 2020 along with student and school characteristics to predict Spring 2020 scores as if the school year proceeded normally. Further information may be found in the Technical Appendix.

- Discipline data consist of the number of OSS. The associated outcome variable is an indicator that is one for students with at least one OSS during the school year and zero for those who were not suspended. We use this outcome as a proxy for student behavior.
- Enrollment data include school attended and grade.
- School-level data include SAGE and AGR funding by year, average teacher compensation, and school location (city, suburb, town, rural).
- District-level data include median income and poverty levels.

Identifying Comparison Schools

Using the data described above, we aggregate each school's K-3 data to find a comparison group of non-AGR schools. For each of the outcomes, we explored multiple variations of PSM in order to, (1) achieve the best match between AGR and comparison schools, (2) retain as many AGR observations as possible, and (3) ensure that there are sufficient control schools matched to each AGR school. To do so, we tested combinations of demographic and academic variables and several matching algorithms. As a result of this testing process, we selected 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.

Matching followed two strategies, depending on the outcomes. For third grade Forward math and reading, we matched schools within cohorts based on the year students started kindergarten, using school characteristics measured during the fall of each cohort's kindergarten year. These characteristics, shown in Table 1, include the school-level mean of fall kindergarten PALS, which performs equally well as a pretest for both Forward reading and math.

During matching, we selected the sample to address testing gaps and differences in learning models due to the COVID-19 pandemic. Due to a lack of Forward testing during the 2019-20 school year, the analysis omits the cohort of students in third grade during that year (the 2017 kindergarten cohort). For the cohort that attended third grade in 2020-21 (the 2018 kindergarten cohort), we omit any schools that were not in-person at least 75 percent of the days between September and April, when Forward testing occurs. We made this decision for two reasons. First, the primary policy question that this evaluation seeks to inform is whether AGR is effective under conditions that we expect to see in the future. Statewide, virtual schooling was temporary and is unlikely to return, thereby diminishing the value of knowing whether AGR is effective in virtual school environments. Second, AGR schools were more likely to choose virtual learning. Tests of several samples and analysis specifications showed that, relative to previous years, schools that were primarily virtual in 2020-21 experienced larger decreases in Forward scores relative to schools that spent more time in-person. This finding is consistent with national evidence of a negative relationship between achievement and hybrid and virtual learning.[19] These test score differences do not imply that districts opting for virtual schooling made the "wrong" decision. Districts chose learning models to balance public health and learning concerns in their unique, local contexts. The goal of this evaluation, however, is to evaluate the impact of AGR, and removing schools that were heavily virtual in 2020-21 avoids conflating AGR impacts with the impacts of differing learning models.

19 Goldhaber, D., Kane, T. J., McEachin, A., Morton, E., Patterson, T., & Staiger, D. O. (2022). The Consequences of Remote and Hybrid Instruction During the Pandemic [Working Paper 30010]. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w30010;
Jack, R., Halloran, C., Okun, J., & Oster, E. (2021). Pandemic Schooling Mode and Student Test Scores: Evidence from US States [Working Paper 29497]. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w29497/w29497.pdf

Table 1: Propensity Score Matching Controls, Third Grade Forward Reading and Math

CONTROL VARIABLE (MEASURED FALL OF KINDERGARTEN)
School Average Fall Kindergarten PALS
School % Black, Hispanic, White, Other Race/Ethnicity*
School % Free/Reduced-price Lunch
School % Mobile from Prior Year
Locale Description (City, Suburb, Town, Rural)*
District Median Income
School K-3 Student Population
School Number of Cohort Students in Analysis Sample

* Due to collinearity, we omit one Race/Ethnicity category and one Locale Description category from the model.

Table 2: Propensity Score Matching Controls, Attendance and Discipline

CONTROL VARIABLE (MEASURED 2012-13)	ATTENDANCE	DISCIPLINE
School % Black, Hispanic, White, Other Race/ Ethnicity*	✓	✓
School % Free/Reduced-price Lunch	✓	✓
Locale Description (City, Suburb, Town, Rural)*	✓	✓
District Median Income	✓	✓
School K-3 Student Population	✓	✓
2012-13 School Attendance Rate (Model 2 only)	✓	
2012-13 School Suspension Rate (Model 2 only)		✓

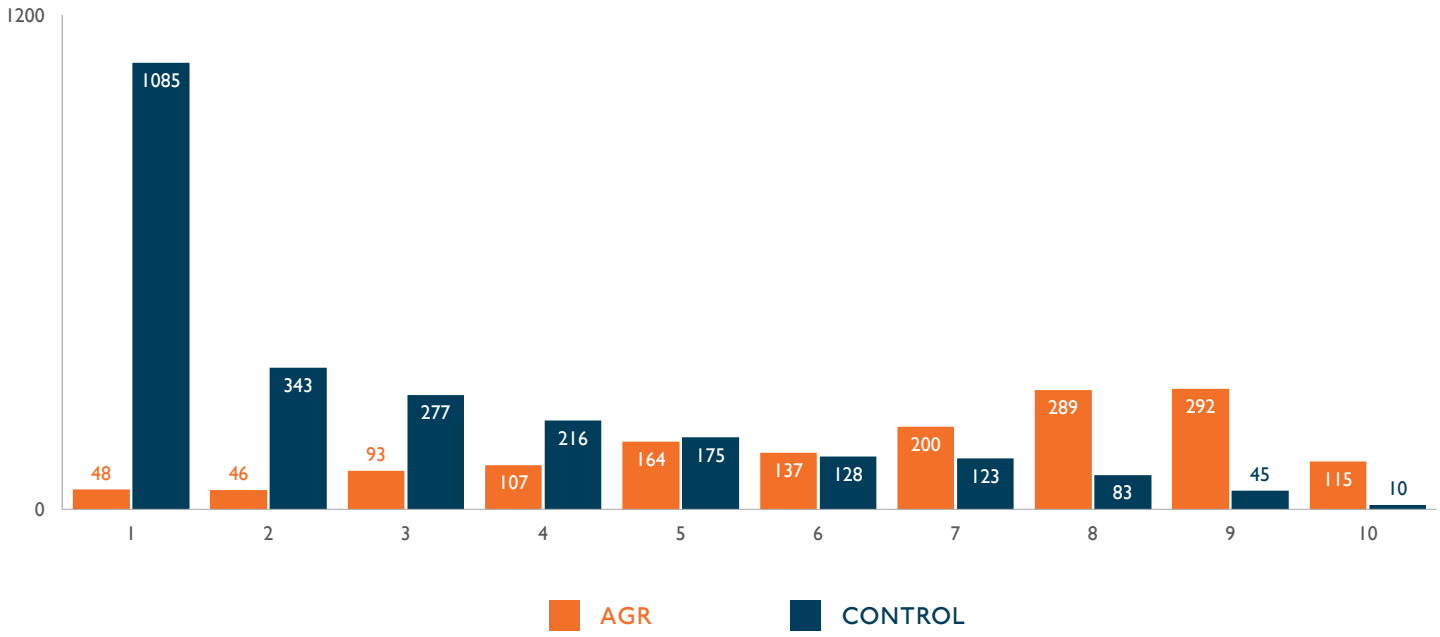
* Due to collinearity, we omit one Race/Ethnicity category and one Locale Description category from the model.

Matching for attendance and discipline outcomes follows a different strategy, matching at the school level based on K-3 demographic characteristics in 2012-13, the first year in our data. The matching controls appear in Table 2. Previous AGR evaluations included 2012-13 school-level means of attendance rates and OSS rates in both matching and analysis to ensure that AGR schools were matched to control schools with similar attendance and OSS rates. This practice may cause bias, however, depending on whether SAGE, which was in place in 2012-13, impacted those 2012-13 outcomes. All AGR schools previously participated in SAGE, which had been in operation for over 15 years in 2012-13. As a consequence, matching AGR schools to non-AGR schools based on post-SAGE outcomes risks biasing the results toward zero (toward estimating smaller impacts), because schools would be matched on previous-period outcomes that already include the treatment impact (in this case, the SAGE program is similar enough to AGR to raise similar concerns). If SAGE had no impact on 2012-13 outcomes, however, omitting the outcomes from matching risks bias from poor matches. With available data, it is not possible to know whether SAGE impacted 2012-13 absence and OSS outcomes and, subsequently, which matching model is correct. This year, we choose to present results from models with and without 2012-13 outcomes in order to provide upper and lower bounds of the AGR impact. To improve the quality of school matches without the outcomes, we include district-level median income.

Absence and discipline models omit 2020-21 data. Due to the COVID-19 pandemic, absence and discipline fluctuated significantly relative to previous years.

When matching is successful, there should be sufficient overlap in the propensity scores of treated (AGR) and untreated (non-AGR) schools to ensure that there is a plausible comparison group for analysis. Figure 3 shows the overlap between AGR and non-AGR schools for the third grade reading matching analysis. In each decile of the propensity score distribution, there are at least 10 comparison (untreated) schools. Most deciles have more than 40 comparison schools, showing sufficient overlap for the analysis. This overlap is similar across all models.

Figure 3: Common Support for Matching – Third Grade Forward Reading



Analysis

After matching, we estimate AGR impacts via multivariate regression models. These models include all school-level matching covariates listed in Tables 1 and 2 above, as well as other student- and school-level variables that are not necessary for successful matching but improve analysis model fit. A full listing of analysis variables can be found in Table 3 below. All models include weights generated by the kernel PSM procedure. Standard errors are clustered at the school level.

Table 3: Analysis Model Controls

CONTROL VARIABLE	GROWTH	ATTENDANCE	DISCIPLINE
Student Demographics Race/Ethnicity,* School % Free/Reduced-price Lunch, English Learner, Special Education	✓	✓	✓
School Demographic Percentages Race/Ethnicity,* English Learner, Special Education	✓	✓	✓
Student Year-to-Year Mobility	✓		
School % Mobile from Prior Year	✓		
Locale Description (City, Suburb, Town, Rural)*	✓	✓	✓
Student Fall Kindergarten PALS	✓		
School Average Fall Kindergarten PALS	✓		
Average Teacher Compensation		✓	✓
District Median Income	✓	✓	✓
District % Under the Poverty Line		✓	✓
School K-3 Student Population	✓	✓	✓
School Number of Cohort Students in Analysis Sample	✓		
2012-13 School Attendance Rate (Model 2 only)		✓	
2012-13 School Suspension Rate (Model 2 only)			✓
Cohort Indicators**	✓		
Grade-by-year indicators ***		✓	✓

Notes: * Due to collinearity, we omit one Race/Ethnicity category and one Locale Description category from the model.
** Indicators equal one if the student is in that cohort, zero otherwise.
*** Indicators for each grade-year combination equal one if a student's grade and year in school match the indicator variable, zero otherwise.

Limitations

The methodology outlined above provides the most rigorous possible evaluation given the rollout of AGR, available data, and the COVID-19 pandemic's impact on school attendance, learning models, and testing. There are several limitations, however, that could impact this report's results and conclusions.

The primary limitation stems from PSM's 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. In particular, if AGR schools are systematically more (less) effective than schools in the matched comparison group, impact estimates will be biased upward (downward).

The second limitation occurs due to the phase out of PALS testing. Through 2015-16, Wisconsin mandated PALS for kindergarten students. When that mandate ended prior to the 2016-17 school year, PALS usage dropped significantly, although approximately 40 percent of schools continue to use the assessment. Although overall the tested population of AGR schools used for the evaluation is observationally similar to the untested sample of AGR schools, available data cannot support analysis of whether schools' choice of PALS testing is related to outcomes and participation in AGR.

Finally, readers should note that the COVID-19 pandemic's impact on student attendance, learning models, and testing may impact results from 2019-20 and 2020-21. For example, the 2020-21 sample of third graders at AGR schools included in Forward analyses is significantly different from the general AGR population. Although the evaluation methodology attempts to address potential biases from the pandemic, as described above, results from the post-pandemic time period should be interpreted with care.

Section 4

AGR Demographics

AGR Demographics

This section and the sections that follow present evaluation results aligned with the evaluation questions above. We begin with information on the characteristics of AGR students and schools. Table 4 shows the number of AGR schools for each of the six 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. After a small increase in the subsequent years, the overall number of AGR schools dropped slightly back to 408 in 2020-21.

The numbers of students in AGR schools from 2015-16 to 2020-21, overall and by grade, are presented in Table 5. The first cohort of AGR schools included approximately 18,000 students, while the addition of the second cohort in 2016-17 brought the total to over 77,000 students. Since, student participation has declined, decreasing to 71,102 in 2020-21.

Figure 4 and Figure 5 compare the demographic characteristics of AGR students to all K-3 Wisconsin students (including AGR students) in 2020-21. Relative to Wisconsin as a whole, a higher proportion of AGR students are Black, Hispanic, English learners, and eligible for FRL. A lower proportion of students in AGR schools were White. AGR schools are more likely to be located in urban or rural settings and less likely to be in suburban areas compared to the state overall, as shown in Figure 6.

Table 4: Number of AGR Schools by Grade and Year

GRADE	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21
Kindergarten	88	393	392	394	396	391
First	91	398	398	402	400	395
Second	91	398	399	402	400	395
Third	88	391	392	394	394	389
Any (K-3)	96	408	409	412	412	408

Table 5: Number of AGR Students by Grade and Year

GRADE	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21
Kindergarten	4,139	18,384	18,274	18,340	18,429	17,658
First	4,571	19,288	18,855	18,741	18,533	17,950
Second	4,682	20,054	19,200	18,911	18,592	18,003
Third	4,544	19,508	19,257	18,328	18,092	17,491
Any (K-3)	17,936	77,234	75,586	74,320	73,646	71,102

Figure 4: Race/Ethnicity of AGR and WI Students, 2020-21

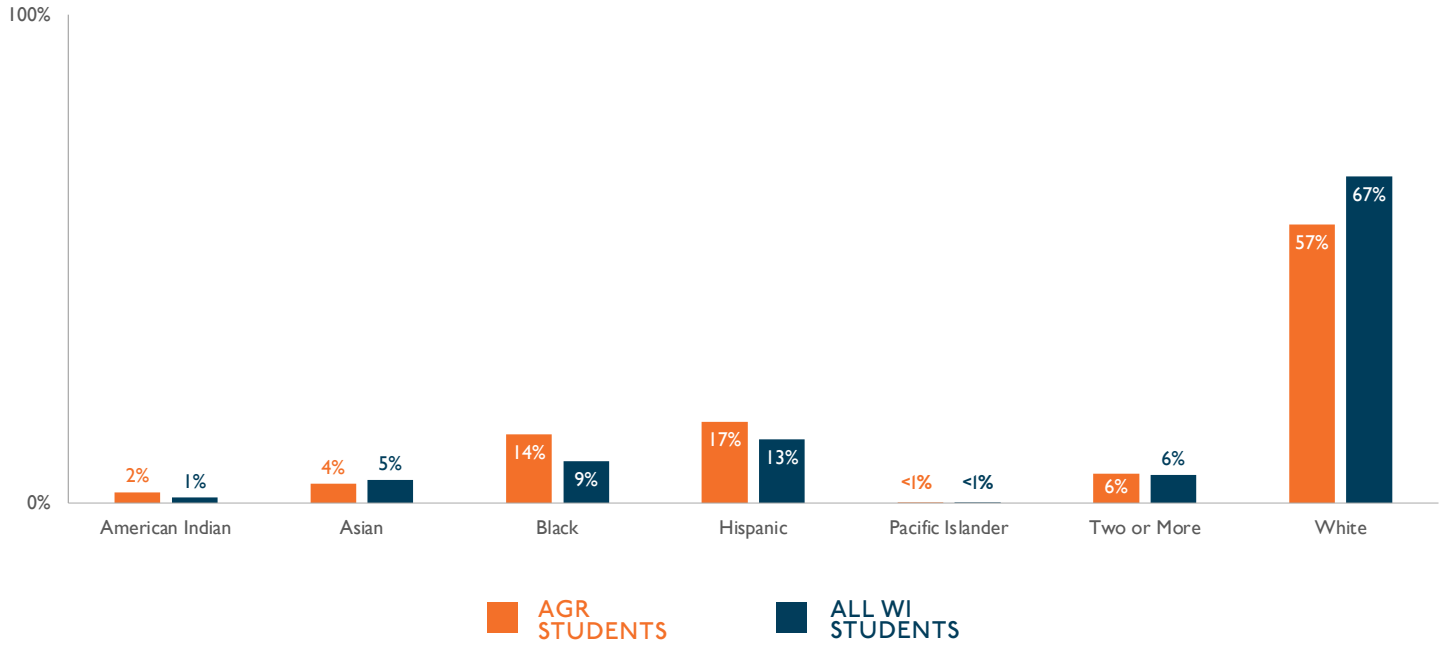


Figure 5: Percentage of AGR and WI Students who were English Learners, Eligible for FRL, and in Special Education, 2020-21

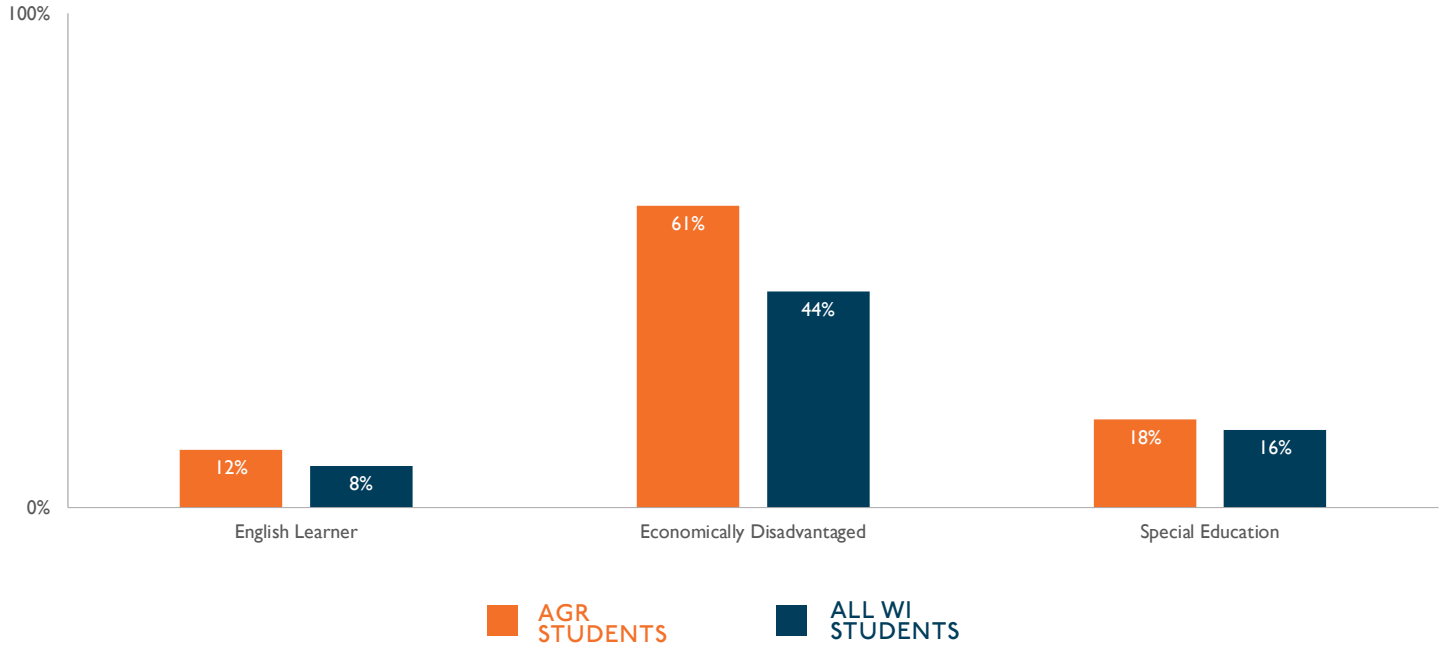
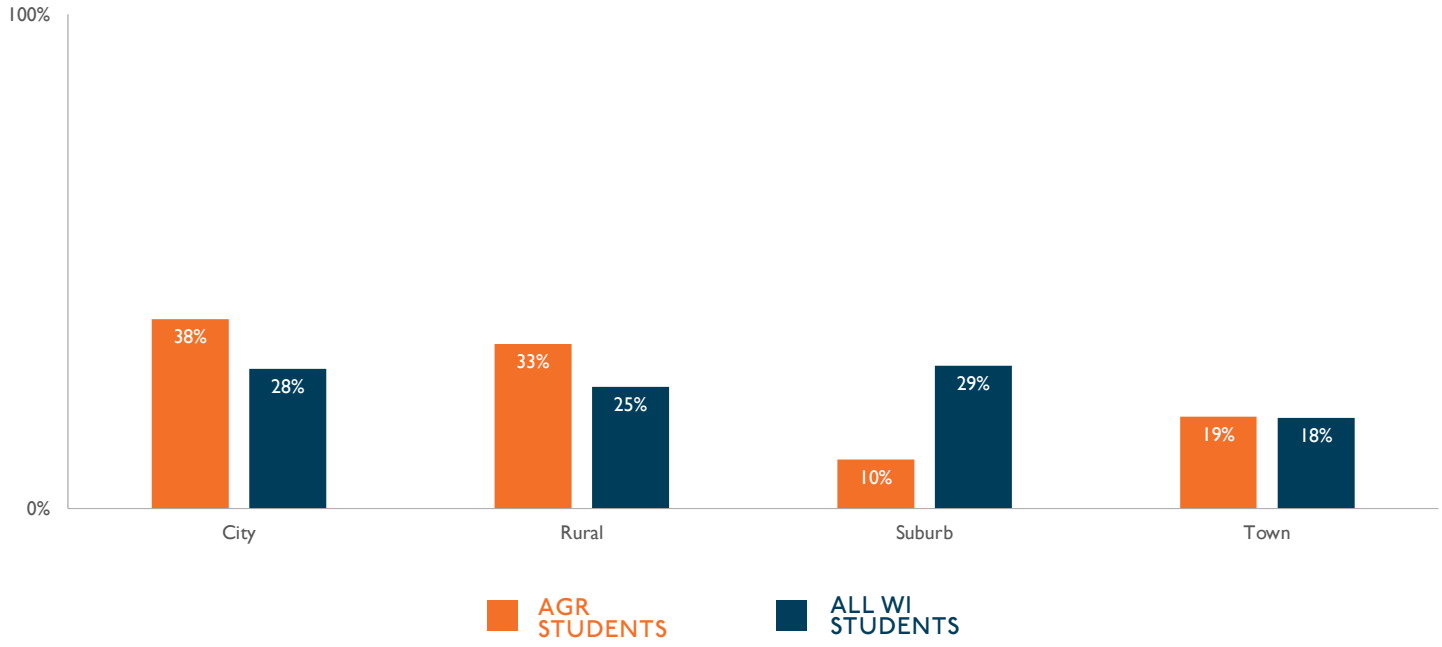


Figure 6: Locale Description of AGR and WI Students, 2020-21



Third Grade Forward Analysis Sample

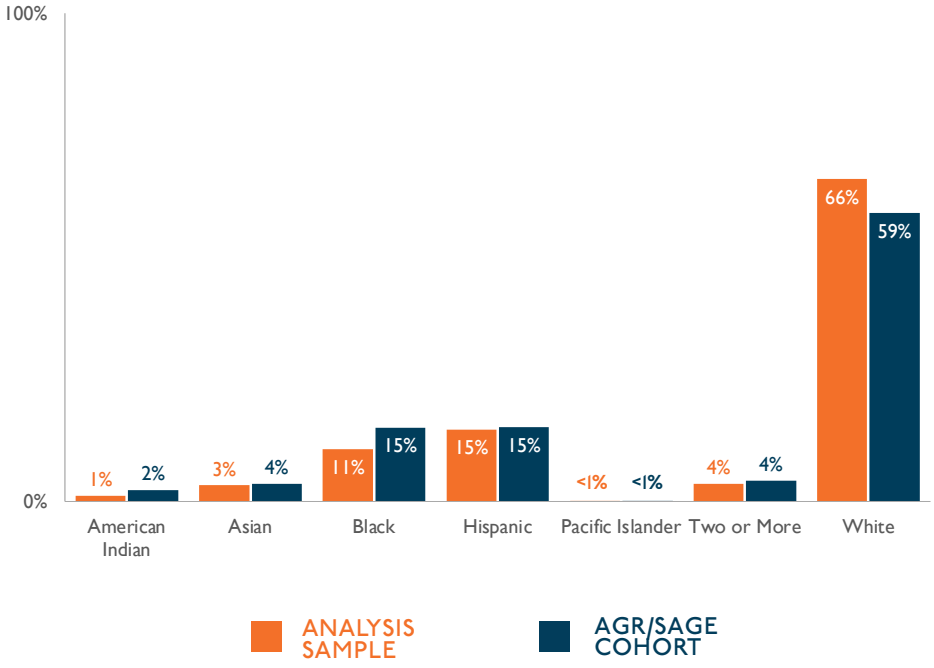
The examination of third grade Forward requires several population restrictions to create a sample of students for analysis. First, we include only students with third grade Forward results as well as students with information on each of the controls used in the analysis (Table 3). As noted above, this results in dropping the cohort of students starting kindergarten in 2016-17 who would have taken the Forward exam in 2019-20 had it not been cancelled due to the COVID-19 pandemic. Since PALS was the state-mandated reading assessment for kindergarten from 2012-13 through 2015-16, nearly all students in the first four cohorts have Fall kindergarten PALS information, but after 2015-16, participation in PALS dropped to less than 50 percent of kindergarteners. The resulting sample drops accordingly. Other restrictions on the number of students used in the analysis sample include a student needing to appear in the data all four years (kindergarten through third grade), a student following the typical grade progression from kindergarten to third grade, and a student not switching between an AGR and a non-AGR school from kindergarten through third grade. Finally, for the cohort of students starting kindergarten in 2017-18 and attending third grade in 2020-21, we omit any schools that were not in person at least 75 percent of the days between September and April, as noted above. At the school level, we omit schools that participated in SAGE but never participated in AGR, schools that do not include all grades K-3, and schools with less than five students who otherwise would have been included in the analysis sample. The resulting sample of analysis students compared to the total number of AGR students is presented in Table 6.

Table 6: Number of Forward Analysis AGR Students and Percentage of All AGR Students by Kindergarten Cohort

	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	ALL COHORTS
All AGR Students	21,915	21,025	20,518	19,414	18,762	18,623	120,257
Analysis AGR Students	12,910	12,929	12,522	11,387	0	2,507	52,255
Percentage in Analysis	58.9%	61.5%	61.0%	58.7%	0.0%	13.5%	43.5%

Because the Forward analysis sample of students is smaller than the entire population, the Forward analysis results may not apply to all AGR students. To observe whether the analysis sample differs from the overall population, the following figures compare demographic characteristics between AGR students used in the Forward analysis and AGR students overall. While across all cohorts the demographic characteristics of AGR analysis students and AGR students overall is similar (Figure 7 - Figure 9), for the 2017-18 kindergarten cohort the differences are larger. Analysis students in the 2017-18 cohort are less likely to be Black, Hispanic, English learners, and eligible for FRL, and more likely to be White (Figure 10 and Figure 11). The schools included in the Forward analysis in the 2017-18 cohort are more likely rural and far less likely urban (Figure 12).[20]

Figure 7: Race/Ethnicity of AGR Forward Analysis Students and AGR Students Overall, All Cohorts



20 The lack of students in the city locale in the 2017-18 analysis cohort is primarily due to the prevalence of virtual schooling in and around Madison and Milwaukee during 2020-21.

Figure 8: AGR Forward Analysis Students and AGR Students Overall who were English Learners, Eligible for FRL, and in Special Education, All Cohorts

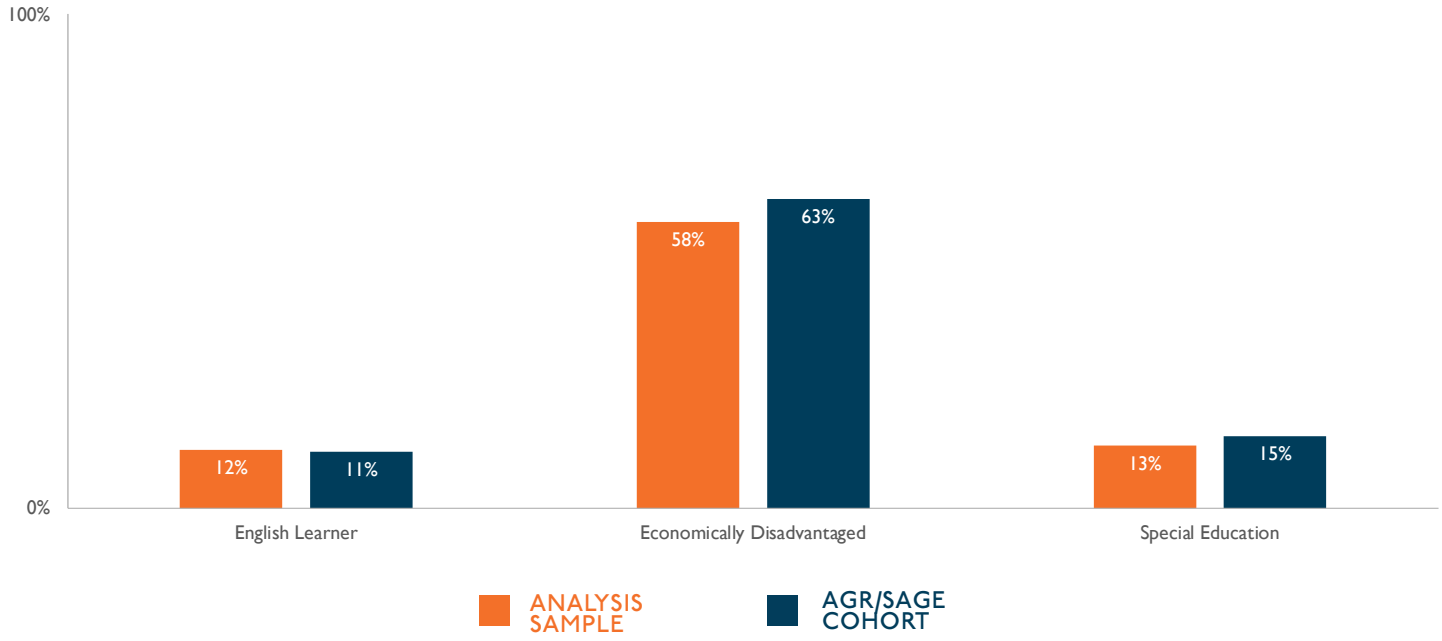


Figure 9: Locale Description of AGR Forward Analysis Students and AGR Students Overall, All Cohorts

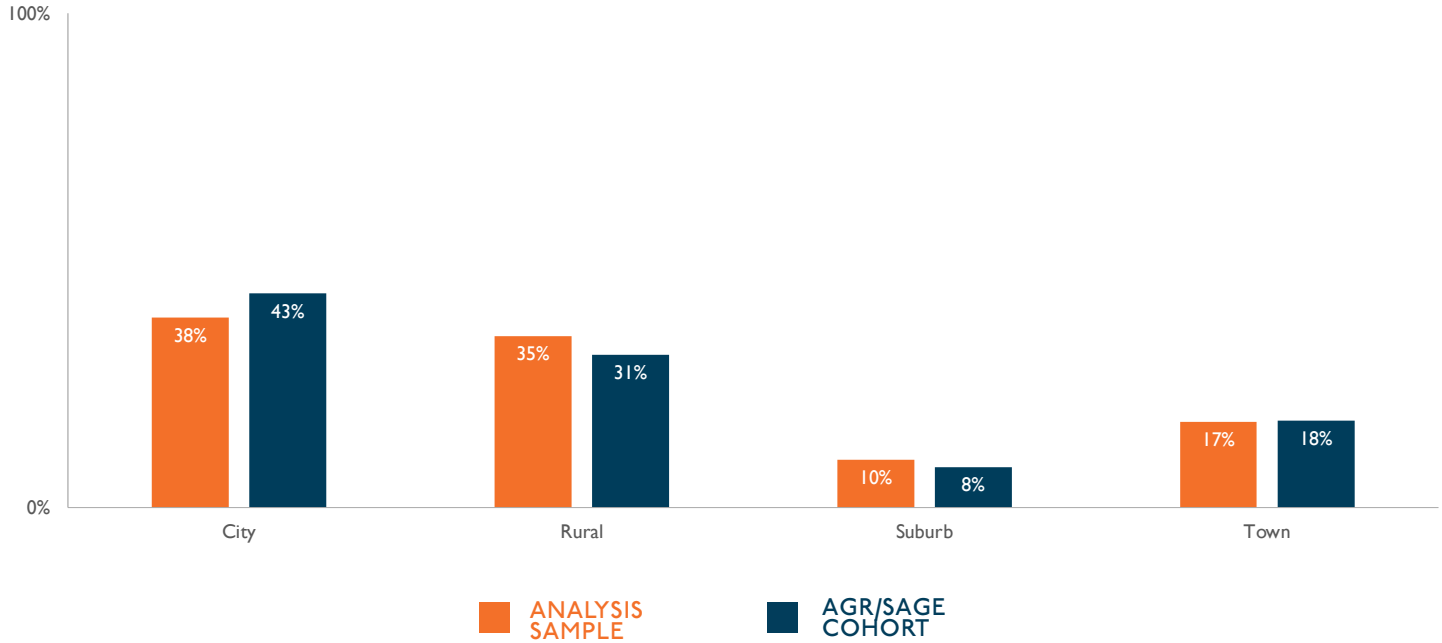


Figure 10: Race/Ethnicity of AGR Forward Analysis Students and AGR Students Overall, 2017-18 Cohort

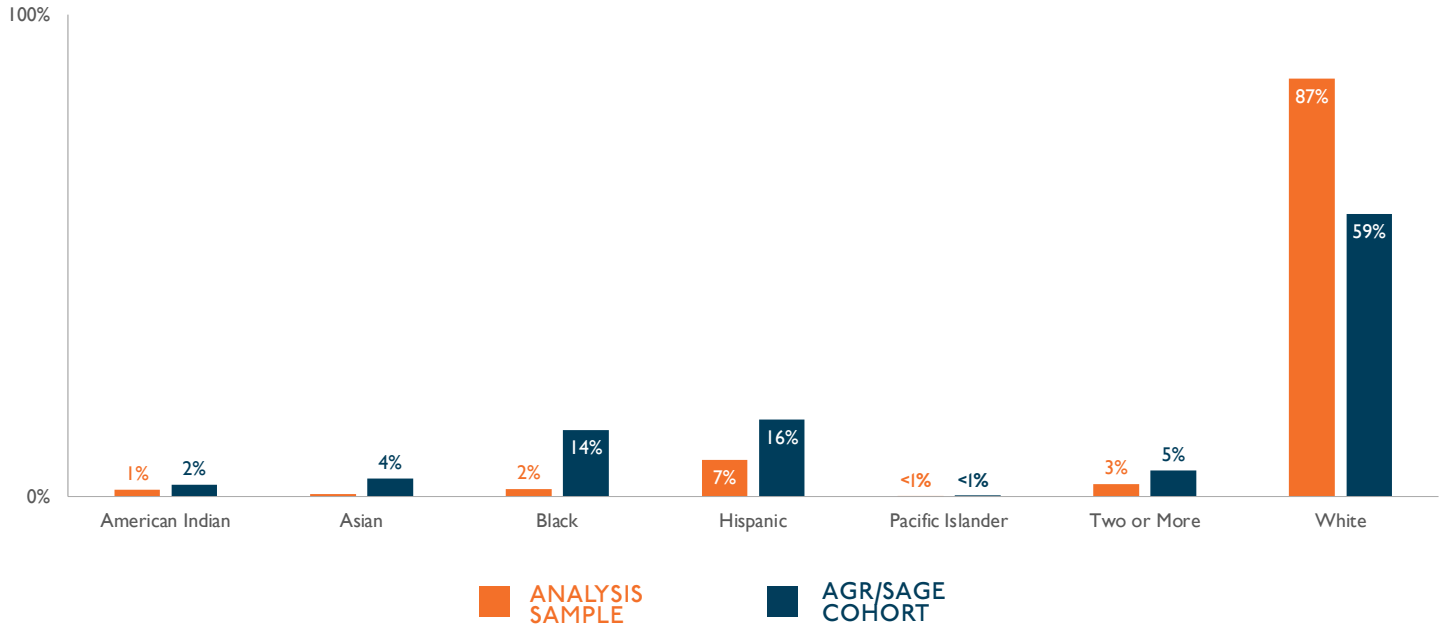


Figure 11: AGR Forward Analysis Students and AGR Students Overall who were English Learners, Eligible for FRL, and in Special Education, 2017-18 Cohort

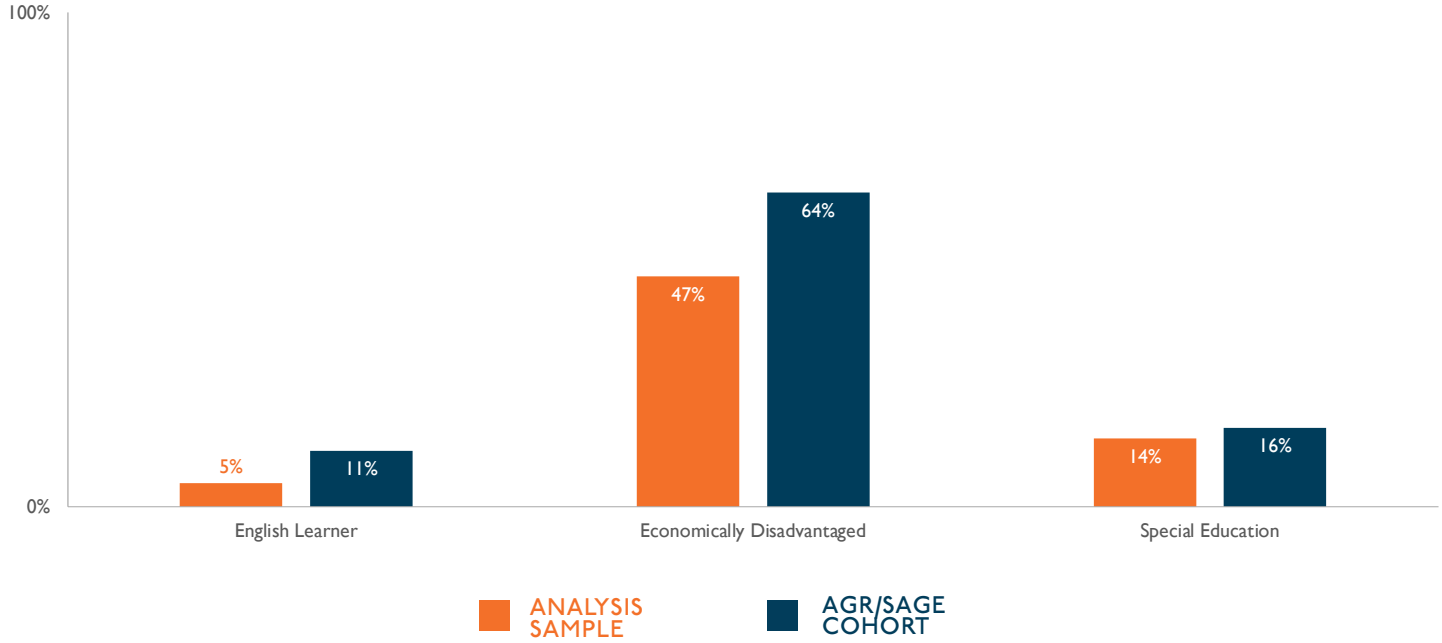
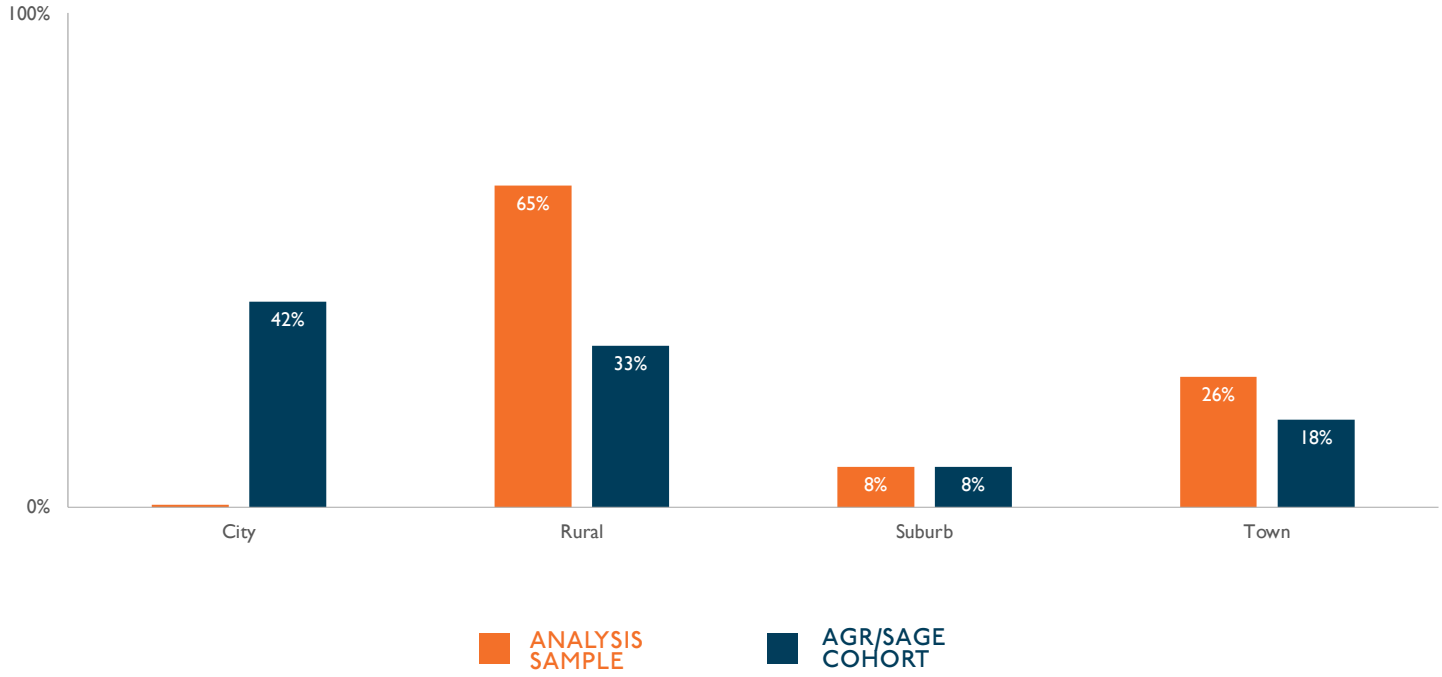


Figure 12: Locale Description of AGR Forward Analysis Students and AGR Students Overall, 2017-18 Cohort



Section 5

AGR Implementation

AGR Implementation

This section of the report examines the usage of the three possible AGR strategies. As noted previously, the three strategies include:

1. 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.
2. Provide data-driven instructional coaching for the class teachers.
3. Provide data-informed, one-to-one tutoring to pupils in the class who are struggling with reading or mathematics or both subjects.

As the AGR program allows schools to use more than one strategy within a school, there are seven possible combinations schools could implement: class size reduction only, coaching only, tutoring only, class size reduction and coaching, class size reduction and tutoring, coaching and tutoring, and all three strategies. Table 7 and Table 8 provide information on the strategy combinations AGR schools implemented during 2020-21. These tables also provide information on the number and percentage of students affected by each strategy combination. The most frequently used strategies were class size reduction and coaching combined, all three strategies, class size reduction only, and coaching only. Only one school used only tutoring as a strategy.

Figure 13: Implementation of AGR Strategies

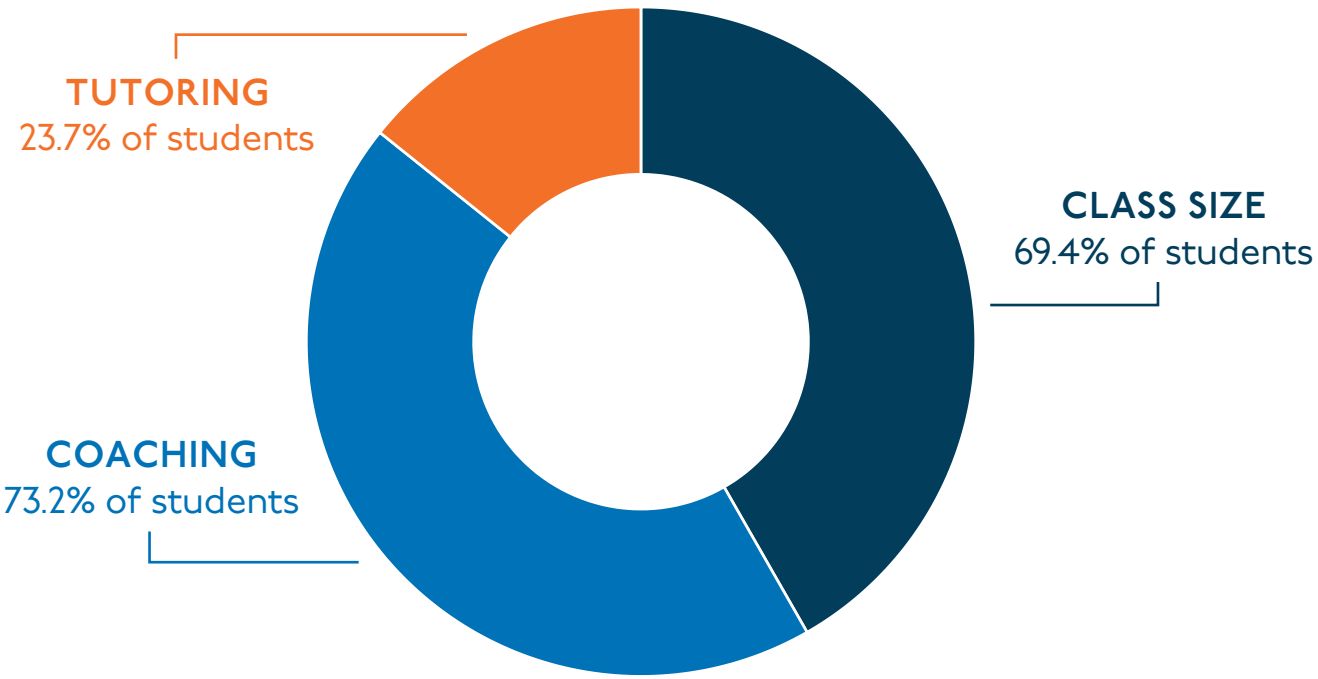
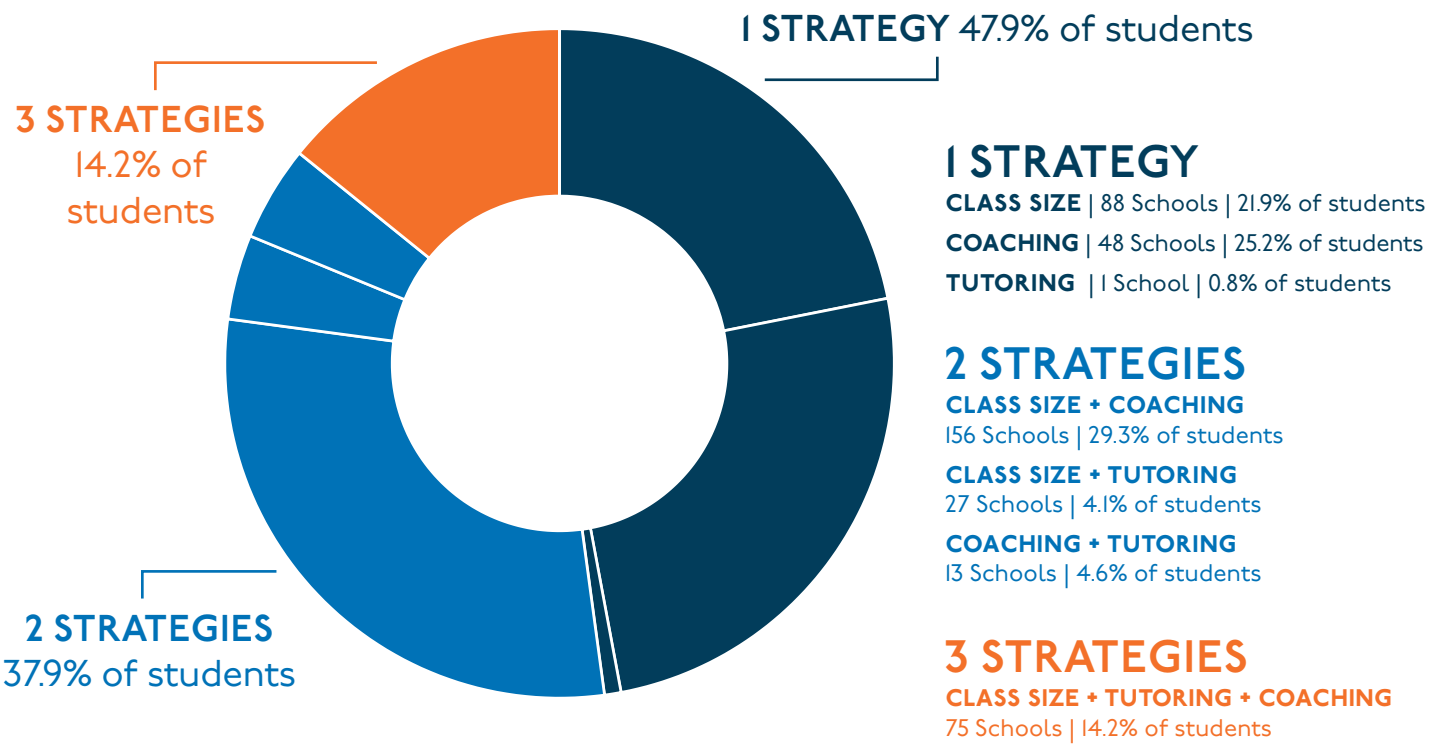


Figure 14: Schools' AGR Strategy Choices



Section 6

AGR Impacts

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?
    a. How does AGR impact achievement gaps between low-income students and their higher-income peers?
    b. How does AGR impact vary by student characteristics?
    c. 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 of AGR's impacts by comparing outcomes of AGR students relative to students in similar, non-AGR schools. We provide AGR impact results statewide and for the subpopulation of low-income students. We also present AGR impacts relative to impacts from previous SAGE implementation. All impact analyses examine how students performed on four outcome measures: third grade Forward reading, third grade Forward math, absences, and OSS. For each of these outcomes, this report provides a table of results for AGR impacts. These tables show a measure or measures of the program impact and a p-value that indicates the likelihood of observing the reported impact or a more extreme impact assuming that there is no actual impact of the program. Larger p-values indicate weaker evidence of an impact, while smaller p-values indicate stronger evidence of an impact. Throughout the report, the evaluation uses a threshold of 0.05 to determine if a result was statistically significant from zero. All p-values presented in this report are corrected to account for multiple estimates (see the Technical Appendix for details).

Third Grade Forward Impacts

Table 9 and Table 10 present the impacts of AGR on third grade Forward reading and math, respectively. Impacts show the difference between average AGR student growth, non-AGR student growth for students in similar schools, and SAGE student growth. Tables 9 and 10 display results for all AGR students and for the subset of AGR students who receive FRL (for whom impacts are measured relative to FRL students in non-AGR schools). For AGR comparisons to non-AGR/non-SAGE schools, results are presented for one year of AGR and for four years of AGR, which reflects students who attended AGR schools throughout grades K-3.

As shown in Table 7, AGR impacts on third grade Forward reading are positive but small and not statistically different from zero, both for all students and for the subset of students who receive FRL. For context, the one-year impact (0.106 scale score points) represents 0.0022 standard deviations.[21] Table 7 also shows that AGR impacts are statistically indistinguishable from SAGE impacts. These results are interesting in light of previous evaluations' results which show consistently substantive, statistically significant impacts of AGR on Kindergarten PALS Reading test score growth (see the Technical Appendix) but zero impacts on MAP/STAR reading growth in grades 1-3. Taken together, the Forward, PALS, and MAP/STAR results indicate that either (a) kindergarten AGR impacts fade out by the time students reach third grade, a common phenomenon among education interventions, and/or (b) improvements in PALS skills do not translate to improvements in the skills tested on Forward.

AGR impacts on third grade Forward math, as shown in Table 8, are similar to those of reading. The one-year impact (-0.019 scale score points) is equal to 0.0003 standard deviations.[22] There are no statistically significant impacts on Forward math for all students or the subset of students who receive FRL. AGR impacts are not statistically distinguishable from SAGE impacts.

21 Data Recognition Corporation. (2021). Wisconsin Forward Exam Spring 2021 Technical Report. https://dpi.wi.gov/sites/default/files/imce/assessment/pdf/WI_Forward_Exam_Spring_2021_Technical_Report.pdf. See Table 10-1.

22 Ibid.

Table 7: Impact of AGR on Third Grade Forward Reading Growth

	AGR VS. NON-AGR/NON-SAGE				AGR VS. SAGE	
SAMPLE	FORWARD SCORE IMPACT (1 YEAR OF AGR)	P-VALUE	FORWARD SCORE IMPACT (4 YEARS OF AGR)	P-VALUE	FORWARD SCORE IMPACT (1 YEAR OF AGR)	P-VALUE
All Students	0.106	0.887	0.425	0.887	-0.016	0.972
FRL Students	0.037	0.999	0.150	0.999	0.064	0.999

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

Table 8: Impact of AGR on Third Grade Forward Math Growth

	AGR VS. NON-AGR/NON-SAGE				AGR VS. SAGE	
SAMPLE	FORWARD SCORE IMPACT (1 YEAR OF AGR)	P-VALUE	FORWARD SCORE IMPACT (4 YEARS OF AGR)	P-VALUE	FORWARD SCORE IMPACT (1 YEAR OF AGR)	P-VALUE
All Students	-0.019	0.999	-0.078	0.999	-0.102	0.992
FRL Students	-0.592	0.403	-2.370	0.403	-0.920	0.326

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

Table 9: Impact of AGR on Third Grade Forward Reading Growth, by Cohort

	AGR VS. NON-AGR/NON-SAGE		
KINDERGARTEN COHORT	FORWARD SCORE IMPACT (1 YEAR OF AGR)	DIFFERENCE FROM THE 2013 COHORT	P-VALUE (DIFFERENCE FROM 2013 COHORT
2013 Cohort	-2.208	N/A	N/A
2014 Cohort	0.280	2.488	0.315
2015 Cohort	-0.703	1.505	0.452
2016 Cohort	0.057	2.265	0.395
2018 Cohort	1.208	3.416	0.229

* Statistically significant at the 0.05 level. P-values are corrected to account for multiple estimates. The 2017 cohort is not included because they did not take third grade Forward due to the COVID-19 pandemic.

Table 10: Impact of AGR on Third Grade Forward Math Growth, by Cohort

	AGR VS. NON-AGR/NON-SAGE		
KINDERGARTEN COHORT	FORWARD SCORE IMPACT (1 YEAR OF AGR)	DIFFERENCE FROM THE 2013 COHORT	P-VALUE (DIFFERENCE FROM 2013 COHORT
2013 Cohort	-3.160	N/A	N/A
2014 Cohort	-1.434	1.726	0.375
2015 Cohort	-1.384	1.776	0.359
2016 Cohort	-0.121	3.039	0.185
2018 Cohort	1.748*	4.908*	0.050

* Statistically significant at the 0.05 level. P-values are corrected to account for multiple estimates. The 2017 cohort is not included because they did not take third grade Forward due to the COVID-19 pandemic.

During educational change, such as the transition from SAGE to AGR, outcomes can suffer an "implementation dip" as teachers and administrators create new systems, learn new skills, and integrate new programs into existing school structures.[23] When schools transitioned from SAGE, which only allowed for class size reduction, to AGR, most schools chose to change their implementation to include instructional coaching and/or tutoring (see Figure19). If an "implementation dip" occurred, we would expect to observe increasing test score growth with successive AGR cohorts. To investigate this, we present AGR impacts by cohort for third grade Forward reading and math scores (Tables 9 and 10, respectively). Tables 11 and 12 show some evidence of an "implementation dip" and improvement thereafter. Table 11 shows that AGR impacts on Forward reading were lowest for the 2013 kindergarten cohort, approximately 2.2 scale score points less than the comparison group. For the 2018 cohort, however, one-year impacts had were 1.2 scale score points higher than the comparison group. While by-cohort AGR impacts on Forward reading could be consistent with improving implementation, impacts for the 2014 through 2016 cohorts show no consistent pattern, and the 2018 cohort is limited to schools that attended in-person during 2020-21.

The pattern of AGR impacts on third grade Forward math are more consistent with an "implementation dip" and later improvement. Impacts increase steadily over time, and AGR impacts for the 2018 cohort are nearly 5 scale score points higher and statistically different than impacts for the 2013 cohort. These results are preliminary, however, due to the limited sample for the 2018 cohort.

23 Fullan, M. (2001). Leading in a Culture of Change. Jossey-Bass.

AGR Impacts on Absences and OSS

Table 11 and Table 12 present the impacts of AGR on absences and OSS, respectively. Similar to the reading and math growth impacts above, results in Table 13 and Table 14 show the average difference in outcomes between AGR students and non-AGR students (or SAGE students) in similar schools. Results include the sample of all students and the sample of students who receive FRL. In each of these tables, negative impacts are desirable, reflecting that AGR reduces absences or OSS. As discussed above, we present findings from two impact models in order to provide upper and lower bounds for our results. Model 1 does not use 2012-13 attendance or OSS outcomes in matching or as analysis controls. Model 2 includes these outcomes in matching and analysis.

Table 11 shows AGR impacts on absences, which are presented both in terms of percentage points and as absence days in a typical school year. For all AGR students and AGR students who receive FRL, the AGR impact is small and not statistically significant. The results from Models 1 and 2 are very similar when comparing AGR schools to non-AGR schools. In Model 2, relative to SAGE AGR is associated with fewer absences. This difference in impacts is statistically significant, but the policy impact, only 0.8 days per year, is not large enough to be meaningful.

For OSS, there are no statistically significant differences between AGR schools and non-AGR schools, or AGR schools and SAGE schools. In addition, OSS models are statistically noisy – available student and school-level controls are unable to explain much of the variance in the outcome.

Table 11: Impact of AGR on Absences

	AGR VS. NON-AGR/NON-SAGE			AGR VS. SAGE		
SAMPLE	IMPACT (PERCENTAGE POINTS)	IMPACT (DAYS)	P-VALUE	IMPACT (PERCENTAGE POINTS)	IMPACT (DAYS)	P-VALUE
All Students - Model 1 (no controls for prior outcomes)	0.33	0.6	0.306	-0.23	-0.4	0.315
All Students - Model 2 (controls for 2012-13 attendance)	0.41	0.7	0.189	-0.45*	-0.8*	0.050
FRL Students – Model 1 (no controls for prior outcomes)	0.27	0.5	0.358	-0.12	-0.2	0.627
FRL Students – Model 2 (controls for 2012-13 attendance)	0.33	0.6	0.380	-0.34	-0.6	0.241

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

Table 12: Impact of AGR on OSS

	AGR VS. NON-AGR/NON-SAGE		AGR VS. SAGE	
SAMPLE	IMPACT (PERCENTAGE POINTS)	P-VALUE	IMPACTS (PERCENTAGE POINTS)	P-VALUE
All Students - Model 1 (no controls for prior outcomes)	-0.3	0.415	0.4	0.387
All Students - Model 2 (controls for 2012-13 attendance)	-0.2	0.456	0.1	0.651
FRL Students – Model 1 (no controls for prior outcomes)	-0.4	0.368	0.4	0.361
FRL Students – Model 2 (controls for 2012-13 attendance)	-0.3	0.352	0.1	0.711

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

Differences in Outcomes by AGR Strategies

This set of student performance results examines how AGR strategies impact outcomes. For schools that changed their AGR strategies over time, we estimate how those strategy changes impact outcomes. For the four following tables, the impact is the difference in the outcome between AGR students in schools before and after a strategy is employed. For instance, if a school switched from not using class size reduction to using class size reduction, the resulting estimated change in outcome is the impact presented below. Due to the number of combinations of possible strategies used, the estimates provided only examine the extent to which each individual type of strategy (class size reduction, coaching, and tutoring) changed, regardless of the other strategies used within a school. The analysis uses annual MAP and STAR reading and math data to capture year-to-year changes in test score growth over time in grades 1-3, which would not be possible using Forward, which is first administered in third grade.

The analysis includes only AGR schools with no comparison schools for 2017-18 through 2019-20. Prior to 2017-18, high quality information on strategy use was not available and after 2019-20, the pandemic resulted in lower MAP and STAR participation as well as fluctuations in attendance and discipline rates, as noted earlier in the report. For more information on the strategies analysis methodology, refer to the Technical Appendix.

Table 13 shows the impact of a strategy change for class size, coaching, or tutoring on reading growth, and Table 14 shows the impact on math growth. For each subject, none of the strategies are associated with a statistically significant impact on academic growth. Similarly, Table 15 shows no associated impact on absence rates for changing to any particular AGR strategy. For discipline, however, Table 16 highlights that changing to using class size reduction as an AGR strategy is associated with reducing OSS, a statistically significant and positive impact.

Table 13: Impact of AGR Strategies on MAP/STAR Reading Growth

STRATEGY	IMPACT (STANDARDIZED)	IMPACT (APPROX. MAP SCALE)	P-VALUE
Class Size	0.018	0.27	0.461
Coaching	0.008	0.11	0.867
Tutoring	-0.034	-0.50	0.452

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

Table 14: Impact of AGR Strategies on MAP/STAR Math Growth

STRATEGY	IMPACT (STANDARDIZED)	IMPACT (APPROX. MAP SCALE)	P-VALUE
Class Size	-0.012	-0.17	0.776
Coaching	-0.043	-0.59	0.377
Tutoring	0.016	0.22	0.781

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

Table 15: Impact of AGR Strategies on Absences

STRATEGY	IMPACT (PERCENTAGE POINTS)	IMPACT (APPROX. DAYS)	P-VALUE
Class Size	-0.01	0.0	0.998
Coaching	-0.01	0.0	0.989
Tutoring	0.16	0.3	0.392

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

Table 16: Impact of AGR Strategies on OSS

STRATEGY	IMPACT (PERCENTAGE POINTS)	P-VALUE
Class Size	-0.6*	0.009
Coaching	0.0	0.982
Tutoring	0.2	0.738

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

Section 7

Longitudinal Analyses of End-of-Year and School Board Reports

Longitudinal Analyses of End-of-Year and School Board Reports

At the conclusion of each school year, DPI surveys AGR schools to produce an EOY report describing schools' AGR strategy choices and implementation. DPI also requires that schools submit twice-yearly reports of AGR implementation to their district school boards and to DPI.

Previous years' evaluation reports have focused on the most recent EOY and school board reports. In this year's report, we present longitudinal analyses of both reports to better understand whether there are any observable trends in AGR implementation. The longitudinal analyses below include four years of EOY data collected from 2017-18 through 2020-21, and three years of school board report data, collected from 2017-18 through 2019-20 (due to the pandemic and the pressures it put on schools, DPI did not require schools to submit their 2020-21 school board reports to DPI).

Table 17 shows AGR strategy combinations schools chose during each of the four sample years. Schools are listed as employing a strategy if they so indicate on either their EOY or school board reports. As can be seen in Table 17, strategy choices have been stable over time. One exception is that schools combining class size and instructional coaching have increased by about 10 percentage points since 2017-18. This increase occurred in 2018-19, after which the number stabilized around 40 percent.

Table 15: School-level AGR Strategies from End-of-Year and School Board Reports

STRATEGY	2017-18	2018-19	2019-20	2020-21
Coaching Only	19%	12%	11%	12%
Class Size Only	17%	19%	12%	22%
Tutoring Only	1%	0%	1%	0%
Coaching and Class Size	29%	38%	40%	38%
Coaching and Tutoring	4%	3%	3%	3%
Class Size and Tutoring	9%	7%	7%	7%
All Three	22%	22%	23%	18%

Figure 15: Schools Using Each Strategy in Any Grade, from End-of-Year and School Board Reports

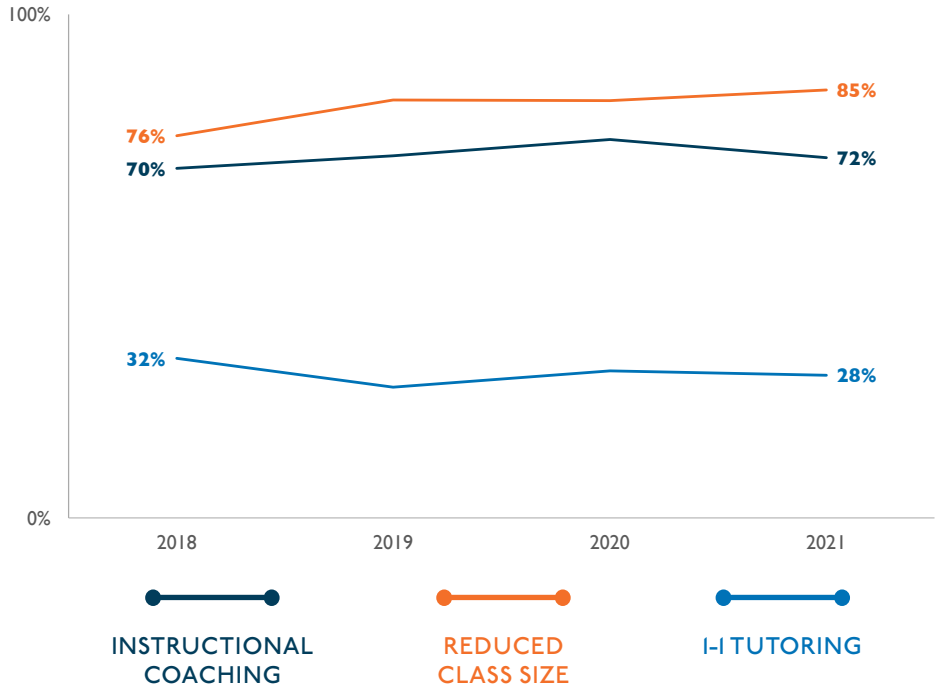
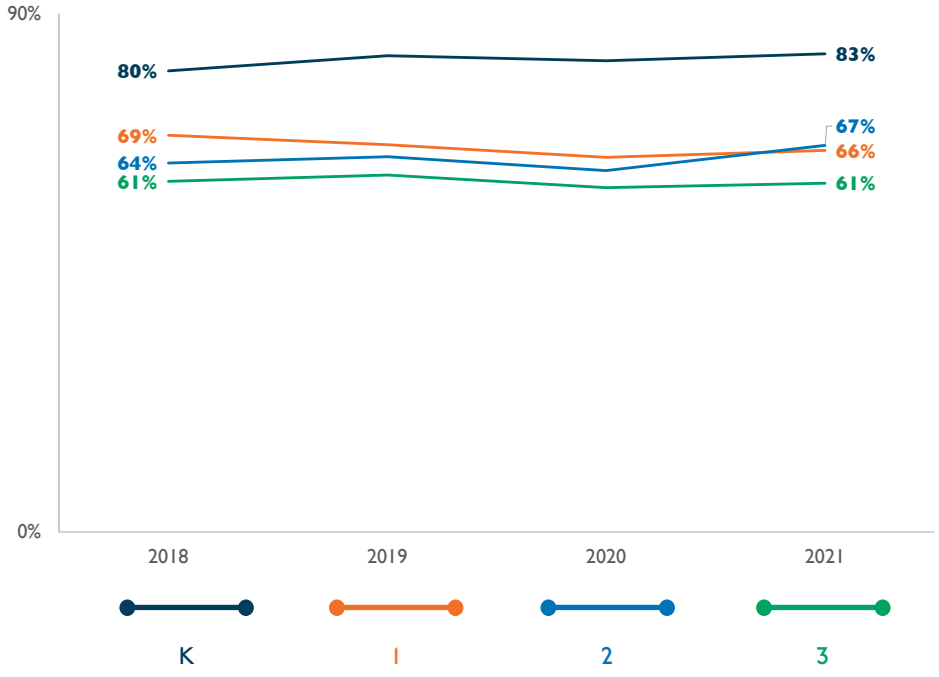


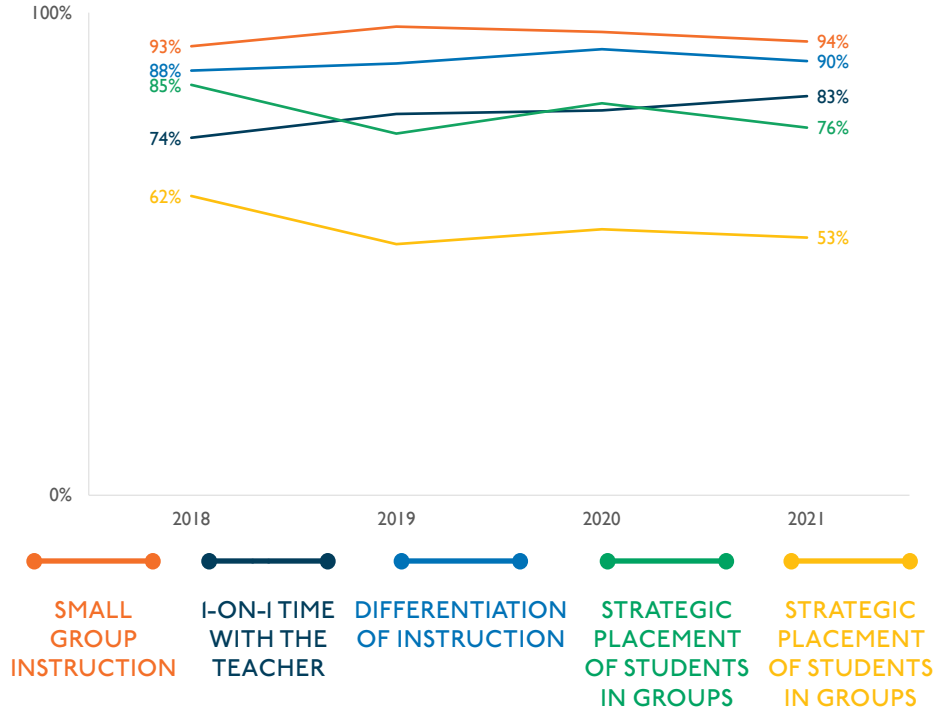
Figure 16: Schools Using Class Size Reduction Doing So in at Least 75% of Classrooms, by Grade



Reduced class size has remained AGR's most used strategy since the program's beginning, when its predecessor program SAGE required that all AGR schools reduce class sizes in grades K-3. As seen in Figure 15, about 80 percent of schools use class size reduction in at least one grade, slightly higher than the percentage that use instructional coaching. Throughout the sample, less than 40 percent of schools have used tutoring in any grade, despite the consensus in the research literature demonstrating the benefits of tutoring.

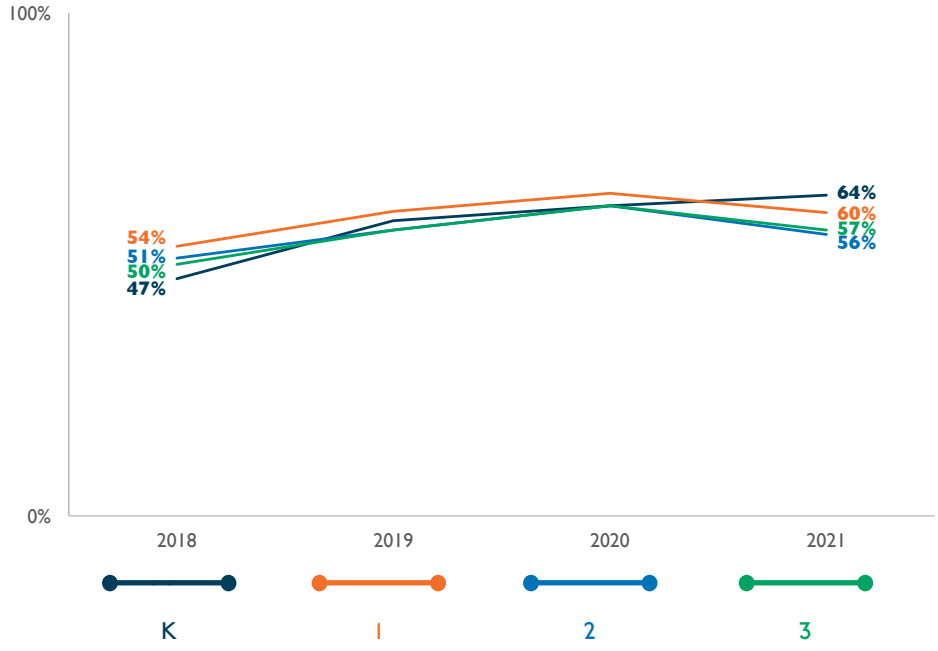
AGR statewide policy allows schools to choose how to distribute each strategy across grades and across classrooms within grades. As a result, some schools choose to implement particular strategies in just some grades. Figure 16 shows that schools using class size reduction do not always use that strategy uniformly across grades K-3. Class size reduction is more common in kindergarten than in grades 1-3. In kindergarten, over 80 percent of schools using class size reduction do so in at least three quarters of classrooms.These percentages have been stable over time.

Figure 17: Instructional Strategies in Schools with Class Size Reduction



*Not Pictured: No specific instructional strategies, Other, Not Sure/Don't Know

Figure 18: Schools Using Tutoring Doing So in at Least 75% of Classrooms, by Grade



Schools using reduced class sizes reported using a variety of instructional strategies (Figure 17). Reported use of each strategy was similar in each of the four sample years. The most common instructional strategies associated with small class sizes were small group instruction, one-on-one time with the teacher, differentiation of instruction, and strategic placement of students in groups. Strategic placement of students in classrooms was less common, but was reported by nearly 60 percent of schools. Only 2-3 percent of schools reported that they use no additional instructional strategies due to reduced class sizes (not shown).

Figure 18 shows that, unlike schools' distribution of class size reduction, tutoring is equally common in all grades. Within grades, since 2017-18 schools have been more likely to use tutoring in at least 75 percent of classrooms. Figure 16 also reflects that schools implementing AGR tutoring in a particular grade do not necessarily do so for all classrooms, with approximately 60 percent of schools implementing in at least 75 percent of classrooms.

Table 16: School-Reported Tutoring Frequency

	2018	2019	2020	2021
3 times a week or more	59%	62%	47%	56%
2 times a week	10%	11%	21%	13%
Weekly	9%	11%	13%	14%
Biweekly	3%	2%	0%	1%
As needed	0%	14%	17%	14%
None	11%	0%	0%	0%
Other	3%	0%	3%	3%

Table 17: School-Reported Tutoring Practices

	2018	2019	2020	2021
Reviews student data	84%	84%	84%	88%
Models appropriate learning behavior	76%	78%	86%	79%
Adapts to student learning styles	82%	79%	92%	81%
Maintains a focus on equity	51%	50%	60%	63%
Provides scaffolding	76%	82%	82%	83%
Communicates regularly with classroom teacher	79%	83%	80%	78%
Other	0%	6%	2%	9%
Not sure/don't know	4%	0%	0%	1%

Studies show that more frequent tutoring results in larger impacts on standardized exam scores.[24] Most schools implementing AGR tutoring do so frequently, as seen in Figure 19. A little less than 60 percent of schools provide tutoring at least three times per week, and a little less than 20 percent provide tutoring two times per week. In 2019-20, perhaps due to the COVID-19 pandemic, fewer schools tutored students three times per week, but there was a parallel increase in schools tutoring two times per week.

Table 18 shows that schools reported frequent use of all of the practices listed on the EOY survey. While the fractions of schools reporting each practice remained mostly stable over time, during the sample period, schools became increasingly likely to use AGR tutoring to maintain a focus on equity. In 2020-21, 63 percent of schools used tutoring for equity purposes, up from 50 percent in 2018-19.

24 Nickow, A. J., Oreopoulos, P., & Quan, V. (2020). The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence [EdWorkingPaper 20-67]. Annenberg Institute at Brown University. https://doi.org/10.26300/eh0c-pc52

Figure 19: Schools Using Instructional Coaching Doing So in at Least 75% of Classrooms, by Grade

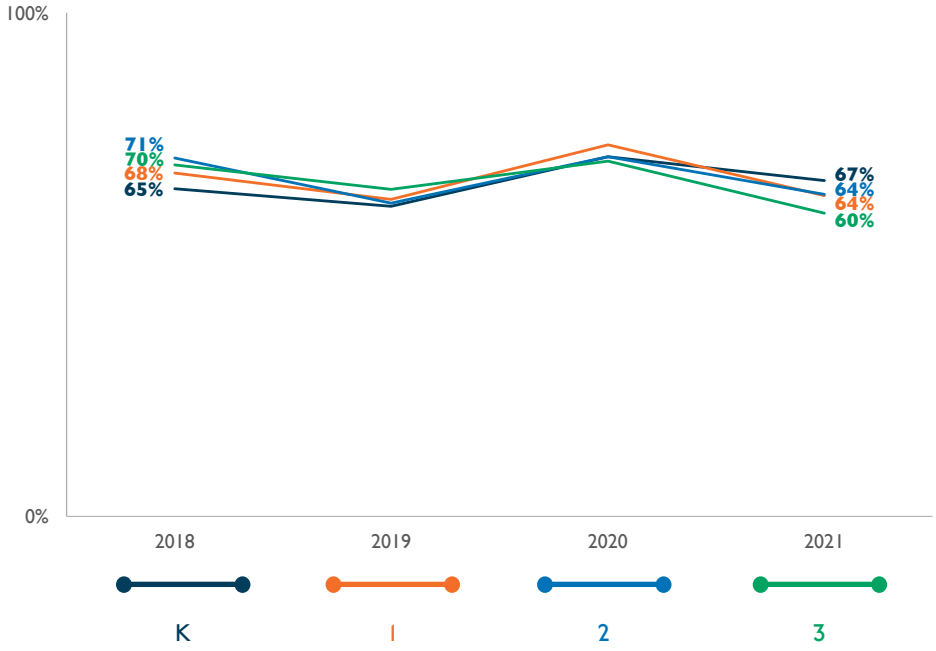


Figure 20: School-Reported Characteristics of Instructional Coaches

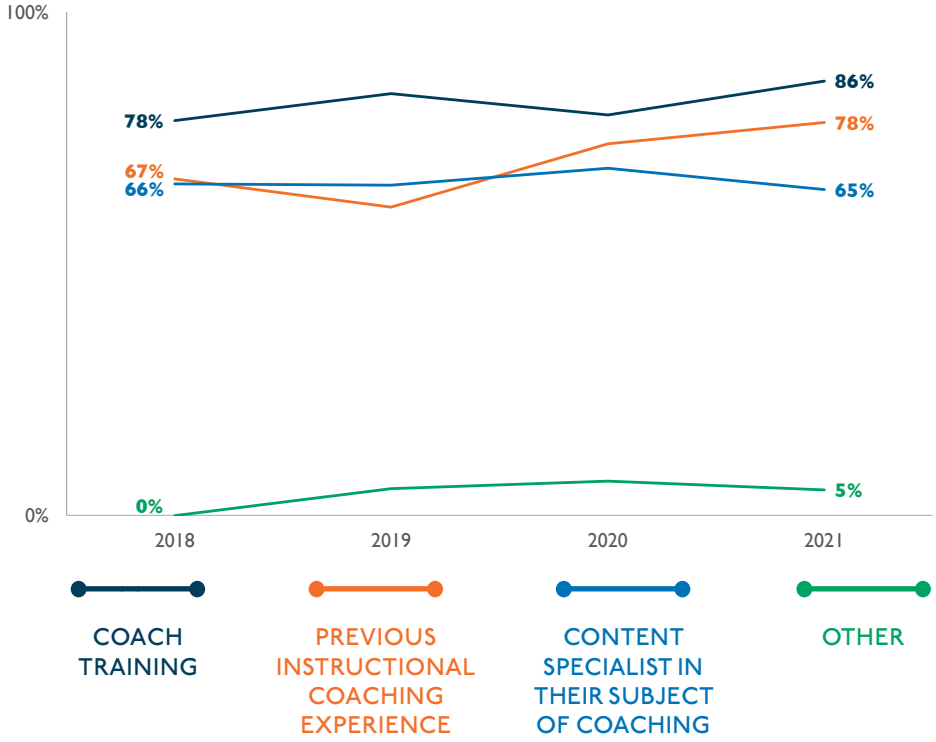


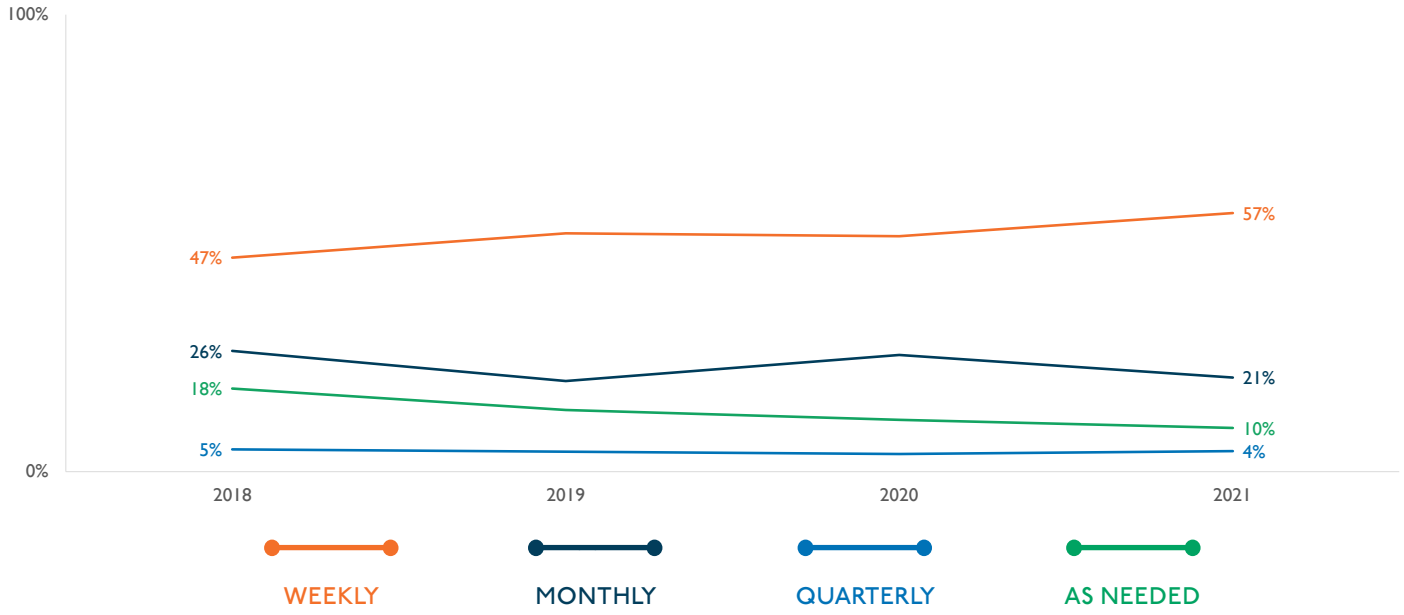
Figure 20 shows that instructional coaching is equally common across grades K-3. Of the schools implementing coaching, about 70 percent do so in at least three quarters of classrooms in any particular grade. These patterns have been stable since 2017-18.

Schools responded affirmatively to almost all listed instructional coaching characteristics (Figure 21). As expected, over the sample period an increasing percentage of schools reported that their coaches had previous coaching experience, from 61 percent of schools in 2018-19 to 78 percent in 2020-21. The percentage of schools whose coaches had previous training and were content specialists in their subjects of coaching remained static over the sample period.

From 2017-18 to 2020-21, schools reported increasing frequency of instructional coaching (Figure 22). Schools providing weekly coaching increased from 47 percent in 2017-18 to 57 percent in 2020-21, while the fraction of schools providing coaching on an as-needed basis steadily declined.

Table 19 shows that, during the sample period, schools reported increasing use of most of the instructional coaching practices listed on the EOY survey. Similar to practices associated with one-to-one tutoring, the most substantial growth occurred for maintaining a focus on equity. In 2020-21, 68 percent of respondents cited maintaining a focus on equity, 22 percentage points more than in 2017-18 (46 percent). Keeping a coaching log and advising teachers to set goals also saw substantial growth.

Figure 21: School-Reported Frequency of Instructional Coaching



*Not shown: Each Semester, Other, Not Sure/Don't Know

Table 18: School-Reported Instructional Coaching Practices

	2018	2019	2020	2021
One-to-one teacher coaching	85%	81%	85%	88%
Team teacher coaching	63%	65%	72%	71%
Keeps a coaching log	42%	45%	49%	55%
Advises teachers to set goals	59%	61%	67%	72%
Coaching focuses on teacher goals	66%	61%	69%	70%
Maintains a focus on equity	46%	46%	60%	68%
Encourages reflective practices	87%	85%	82%	90%
Discusses data with teachers	91%	85%	91%	94%
Observes teacher practices	81%	71%	76%	79%
Other	0%	3%	5%	5%
Not sure/don't know	1%	1%	1%	0%

Section 8

Summary and Conclusions

Summary and Conclusions

The socioeconomic achievement gap is wide and has remained relatively unchanged over the past fifty years.[25] Researchers and policymakers have hypothesized dozens of causes, including funding deficits for districts located in high-poverty areas.[26] The AGR program seeks to reduce the socioeconomic achievement gap by providing additional funding to districts with large proportions of economically disadvantaged students.

This report provides evidence regarding program impacts on math and reading growth, student attendance, and OSS. These results are presented at the state level and disaggregated for students who receive FRL. The report also contains data on the AGR strategies schools have implemented, the intensity of strategy use, and preliminary evidence on the relative effectiveness the three AGR strategies.

This year's evaluation methodology improves on previous evaluations by taking advantage of the growing number of cohorts that have received AGR funds to investigate AGR's K-3 impacts on third grade Forward reading and math. The revised methodology removes a potential source of statistical bias, provides results for the more policy-relevant statewide Forward exam, and includes a greater cross-section of AGR students. The evaluation methodology makes several adjustments to address complexities arising from the COVID-19 pandemic that have the potential to bias estimates of AGR impacts.

For the cohorts entering kindergarten from 2013 through 2016 and 2018, AGR impacts on third grade Forward reading and math are small and not statistically different from zero, for both the statewide sample and for students receiving FRL. These results, particularly for reading, stand in contrast to previous evaluations' findings that AGR has strong impacts on PALS reading growth in kindergarten. These results suggest that either AGR impacts fade out by third grade and/or that PALS and Forward reading are not well aligned. Fade out of test score impacts in early grades is a common phenomenon in early education programs, including programs that have been shown to have meaningful impacts on later life outcomes.

Looking at AGR impacts by cohort shows some evidence that the introduction of AGR was accompanied by an "implementation dip" in outcomes, a well-known feature of some education policies. AGR impacts on third grade Forward reading and math are negative for the first cohort to experience AGR but grow steadily thereafter, particularly for math. The impact on Forward math for the 2018 cohort is statistically different than the impact for the 2013 cohort.

The report also estimates impacts for non-testing outcomes through the 2019-20 academic year, omitting 2020-21 from the sample due to changes in absence and suspension rates due to the COVID-19 pandemic. We find no consistent evidence of AGR impacts on absences or OSS.

25 Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessman, L. (2020). Long-run Trends in the U.S. SES-Achievement Gap [Working paper 26764]. National Bureau of Economic Research. https://www.nber.org/papers/w26764

26 Jackson, C. K., Wigger, C., & Xiong, H. (2021). Do School Spending Cuts Matter? Evidence from the Great Recession. American Economic Journal: Economic Policy, 13(2), 304-35. https://doi.org/10.1257/pol.20180674

Looking at the AGR strategies that schools chose to implement, we find that most AGR schools took advantage of the program's flexibility and chose to implement multiple strategies, including coaching and one-to-one tutoring strategies that were not available under the state's previous SAGE policy. Over 80 percent of schools use reduced class sizes in at least one grade. Despite strong evidence in the research literature that tutoring can positively impact achievement score growth, relatively few AGR schools implemented tutoring, either alone or in combination with other strategies. Examinations of schools that change strategies over time find preliminary evidence that class size reduction reduces OSS.

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 labor market. The long history of SAGE, AGR's precursor program that provided funding for reduced class sizes only, also limits the study. For absences and OSS, previous school outcomes used for matching likely include SAGE impacts as well, which would bias AGR impacts toward zero. Inconsistent testing patterns in grades K-3, including a lack of testing in spring of 2020, restricted the sample of AGR and non-AGR schools included in the growth analysis samples, potentially limiting how growth impact estimates can be generalized to schools not in the sample.

Section 9

Technical Appendix

Technical Appendix Evaluation Design

In order to credibly estimate AGR impacts, we must address two primary challenges to identification. First, we must identify a plausible comparison or control group. Schools that receive AGR funding are different from schools statewide (see AGR Demographics above) because those selected for SAGE, and subsequently eligible for AGR, were required to meet certain thresholds of students receiving FRL. Second, because all AGR schools previously participated in SAGE, total AGR impacts cannot be determined solely through changes over in AGR schools' outcomes over time. 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 AGR and SAGE treatment impacts, not the AGR impact itself.

To find a plausible control group and identify AGR impacts, we use Propensity Score Matching (PSM). PSM addresses selection bias by choosing a control group with observable characteristics similar to those of the treatment group. As described above (see AGR Demographics), schools that receive AGR funding are observably different from other Wisconsin schools. This is because AGR targets funding to schools with higher percentages of students eligible for FRL. 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 OSS. As a result, naïve comparisons of outcomes across non-AGR and AGR schools would find negative program impacts based only on program selection effects. 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 FRL are not AGR participants for two primary reasons. First, poverty in those schools may have increased since the last SAGE eligibility period. Those schools currently 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 mostly similar to those from our preferred models, although impact estimates for Forward math (lower than the main impact estimate) and attendance (fewer absences than the impact estimate) differ slightly. Second, schools may have opted out of SAGE. Opt-out schools would be systematically different from AGR schools due to characteristics of the district or school. Although we cannot test for bias resulting from selection bias associated with opting into or out of SAGE, the final round of SAGE enrollment occurred in 2011-12, and many school and district characteristics, particularly those associated with administration, have since changed.

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

In the remainder of this appendix, we describe the propensity score matching and analysis methodology, the methodology for estimating impacts of individual AGR strategies, multiple comparisons corrections of p-values, and robustness checks. We also present previous evaluations' impact estimates for kindergarten PALS, which, due to their use of a pre-program pretest in fall of kindergarten, are well-identified and substantial. Finally, we present preliminary evidence from an analysis of students leaving Wisconsin public schools during the COVID-19 pandemic and whether AGR mitigated students' leaving choices.

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, $\Pr(\text{EverAGR}_s)$, is a function of an intercept term α, a vector of school-level covariates X_s, and a school-specific error term ε_s.

Equation (1)

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

In the equation above, matching occurs at the school-level (defined by the grades included in the model, not necessarily all of the grades that a school contains) because AGR is a school-level treatment. We use this matching strategy for both the attendance and discipline models. For the models of Forward test score outcomes, however, we match at the school-cohort level, taking advantage of fall kindergarten PALS as a pre-program measure of student achievement. For these matches, we use school-year averages of demographic and school-cohort characteristics of fall kindergarten PALS. Sample sizes and the percentages of AGR students included in the growth analyses are listed in Table 6 above.

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 and Table 2 of the main report. For each of the models, the most important matching variables measure the average outcome in a previous time period (pretests), such as the school's average test scores from the previous time period. The choice of pretest was complicated by both the level of matching (school or school-grade-year) and by the fact that AGR schools previously participated in SAGE. To the greatest extent possible, we aimed to remove previous program impacts from the matching model.

However, at the beginning of our sample period, SAGE had been in operation for over 15 years. As a result, it was not possible to include pre-program data. We use two strategies to address matching on post-program outcomes. For Forward, we match separately for each school-cohort, using pre-program kindergarten fall PALS measures along with other, school-level controls (see Table 2). For the attendance and discipline models, we produced two models (see Table 3 for included covariates). Both match once using school-level controls. Model 1 does not include school-level controls for outcomes. Model 2 includes average attendance rate (or suspension rate), limiting the effect of including a post-program outcome to just one year. The inclusion of school-level outcome controls may cause bias, however, depending on whether SAGE, which was in place in 2012-13, impacted those 2012-13 outcomes. All AGR schools previously participated in SAGE, which had been in operation for over 15 years in 2012-13. As a consequence, matching AGR schools to non-AGR schools based on post-SAGE outcomes risks biasing the results toward zero (toward estimating smaller impacts), because schools would be matched on previous-period outcomes that already include the treatment impact (in this case, the SAGE program is similar enough to AGR to raise similar concerns). If SAGE had no impact on 2012-13 outcomes, however, omitting the outcomes from matching risks bias from poor matches. With available data it is not possible to know whether SAGE impacted 2012-13 attendance and OSS outcomes and, subsequently, which matching model is correct. This year, we choose to present results from models with and without 2012-13 outcomes in order to provide upper and lower bounds of the AGR impact.

In order to find the best PSM model to balance covariates across AGR and comparison schools, retain as many observations as possible, and find the best stability of matches, we tested different matching algorithms, including caliper matching with various bandwidths, kernel matching, and Mahalanobis. For the analysis appearing in the report, we opted for a kernel matching procedure that places higher weights on control observations whose propensity scores are closest to that of a treatment observation and successively lower weights on control observations as their propensity scores increase in distance from a treatment observation.[27]

27 We use Stata's kmatch package with an Epanechnikov Kernel and allow Stata to select the optimal bandwidth based on the outcome.

Prior to matching, we limited the sample using several 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 schools that did not include all grades K-3. We further limited the Forward sample to omit the 2017 kindergarten cohort, who did not take third grade Forward due to COVID-19 shutdowns, students from the 2018 kindergarten cohort whose schools did not host at least 75 percent of class days in person through April, students who did not follow the typical grade progression from K-3, students who did not appear in the data during all four years K-3, students who transferred between AGR and non-AGR schools during K-3, and schools with fewer than five students with pre- and post-tests in their cohort. Table AI illustrates the matching and subsequent analysis strategies for each outcome.

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. For each of the outcomes, there are substantial numbers of non-AGR schools in most propensity score deciles and at least ten control schools in every decile.

Successful matching should also result in balanced covariates across the treatment and control groups. In keeping with the recommendations of the What Works Clearinghouse (WWC), we assess equivalence using both the p-values from t-tests of differences in means and with standardized differences. The WWC specifies that standardized differences over 0.25 are signals of imbalance, and those between 0.05 and 0.25 require that the covariates be included as covariates in the impact analysis.[28] In examining the matching balance of covariates, no standardized differences reach the 0.25 threshold, and we include all covariates in all impact analyses for double robustness.

Full results from common support and matching balance checks are available upon request.

Table AI: Matching and Analysis Strategies and Samples

OUTCOME	GRADE(S)	MATCHING LEVEL	MATCHING DATA	ANALYSIS COHORTS	ANALYSIS YEARS
Forward Reading Growth	3	School-cohort	Fall 2013 through Fall 2016, 2018	Kindergarten in 2013 through 2016, 2018	N/A
Forward Math Growth	3	School-cohort	Fall 2013 through Fall 2016, 2018	Kindergarten in 2013 through 2016, 2018	N/A
Absence Rate	K-3	School	2012-13	N/A	2012-13 through 2019-20
Suspended 1 or More Times	K-3	School	2012-13	N/A	2012-13 through 2019-20

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

Impact Analysis

After matching, we model impact estimates separately for each outcome. The analysis model for Forward differs from the model for absences and OSS, although both cases model impact estimates for SAGE and AGR in the same regression. All models include weights generated by the kernel PSM procedure. Standard errors are clustered at the school-level. Models for Forward 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 convert 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. Outcome statistics are available upon request.

Forward Reading & Math

Equation (2)

$$Forward_{i,g3,s,c} = \propto + \lambda PALS_{i,gK} + \beta_1 AGRdose_i + \beta_2 SAGEdose_i + \gamma X_i + \theta Z_s + \delta_c + \varepsilon_{i,g3,s,c}$$

where $Forward_{i,g3,s,c}$ is an outcome for student *i* in grade 3, school *s*, and year *y*. $PALS_{i,gK}$ is student i's fall kindergarten PALS score, which acts as a pretest for both reading and math. $SAGEdose_i$ and $AGRdose_s$ are counts of the number of years during kindergarten, first, second, and third grades that student *i* spent in SAGE or AGR schools, respectively. X_i represents a vector of student-level covariates, and Z_s represents a vector of school-level co-variates measured during the student's kindergarten year. Cohort fixed effects, δc, are included to control for any unobserved, statewide effects that vary by time. All analysis variables are described in Table 3 in the main report. As described above, the models include all school-level variables from the PSM procedure.

Absences & OSS

Equation (3)

$$Y_{i,s,g,y} = \alpha + \beta_1 SAGE_{s,y} + \beta_2 AGR_{s,y} + \gamma X_{i,y} + \theta Z_{s,y} + \delta_{g,y} + \varepsilon_{i,s,g,y}$$

where $Y_{i,s,g,y}$ is an outcome for student *i* in grade *g*, school *s*, and year *y*. $SAGE_{s,y}$ and $AGR_{s,y}$ are indicators for whether a school received SAGE or AGR funding, respectively, in each year. $X_{i,y}$ represents a vector of student-level covariates, including lagged values of the outcome *Y*, and $Z_{s,y}$ represents a vector of school-level co-variates. Grade-by-year fixed effects, $\delta_{g,y,}$ are included to control for any unobserved, statewide effects that vary by grade and/or time. All analysis variables are described in Table 3 in the main report. As described above, the models include all school-level variables from the PSM procedure as well as individual-level controls.

Strategy Analysis

With multiple years of school strategy data (from the EOY survey and school board reports) now available, this year's evaluation improves on the previous cross-sectional methodology by exploring how within-school strategy changes impact outcomes. This analysis only includes AGR schools with no comparison schools. Each school serves as its own control observation – the analysis compares outcomes before the strategy change to outcomes after the change. The analysis includes the years 2017-18 through 2019-20. In the years prior to 2017-18, high quality information on strategy use was not available. For the 2020-21 school year (the most recent year of data available), the pandemic resulted in lower MAP and STAR participation as well as fluctuations in attendance and discipline rates. As a result, this analysis does not include the 2020-21 school year.

Because the analysis identifies strategies' impacts on year-to-year changes in outcomes, reading and math growth outcomes use assessment data from two local assessments, MAP and STAR, which many schools administer in fall and spring each year. While some readers may be familiar with these assessments' scaling and typical growth, other readers may lack this familiarity. To provide interpretability of results across assessments, we standardize assessment scores to have a mean of zero and standard deviation of one. For each test window, subject, and grade combination (e.g., Fall 2nd grade MAP), we take each individual's score, subtract the mean test score across all test takers, then divide by the standard deviation of the score across all test takers.

For the MAP assessment, we use 2015 national norms available from the vendor, NWEA.[29] These national norms include means and standard deviations for each subject, grade, and test window combination. NWEA created these 2015 norms from a national sample of data from fall of 2011 through spring of 2014. NWEA recently created 2020 norms, calculated from a national sample of data from fall of 2015 through spring of 2018.[30] While the 2020 norms provide a more recent estimate of means and standard deviations throughout the tested MAP population, we continue to use 2015 norms for two main reasons. First, while Wisconsin represents only a fraction of the NWEA testing population, making it unlikely that any impacts of AGR may be present in the national sample of MAP norms, we want to exclude any possibility. Second, we used 2015 norms in prior years of the evaluation. Since 2020 norms are different from 2015 norms, we select the 2015 norms to retain consistency in our estimates of AGR impacts.

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

29 Thum, Y. M., & Hauser, C.H. (2015). NWEA 2015 MAP Norms for Student and School Achievement Status and Growth. Northwest Evaluation Association.

30 Thum, Y. M., & Kuhfeld, M. (2020). NWEA 2020 MAP Growth: Achievement Status and Growth Norms for Students and Schools. NWEA. https://teach.mapnwea.org/impl/normsResearchStudy.pdf

Due to the COVID-19 pandemic, Wisconsin schools engaged in almost no testing during Spring 2020. While the pandemic lessened the data available for the AGR evaluation, it did not lessen the state's need to understand how AGR impacts the state's academic achievement gap. To address the lack of Spring 2020 assessment scores due to the COVID-19 pandemic, we predict Spring 2020 test scores using available data and past relationships between Fall, Winter, and Spring STAR and MAP assessment scores. The primary limitation of the methodology is that, because we are predicting what Spring 2020 test scores would have been had COVID-19 not closed schools, the pandemic's impacts on learning, particularly on disparities in learning by socioeconomic status, are not possible to determine using the prediction methodology. We estimate a predictive model that uses student math and reading scores from Fall 2019 and Winter 2020, student demographics, and school characteristics, to predict what Spring 2020 test scores would have been had the 2019-20 school year proceeded normally. Specifically, we use data from 2012-13 through 2018-19 in the following specification for each subject, math and reading, to estimate coefficients that we then use to calculate predicted 2019-20 spring test scores for each subject:

Equation (4)

$$Y_{i,s,g,y}^{spring} = \lambda^{f,m} Y_{i,s,g,y}^{f,m} + \lambda^{f,r} Y_{i,s,g,y}^{f,r} + \lambda^{w,m} Y_{i,s,g,y}^{w,m} + \lambda^{w,r} Y_{i,s,g,y}^{w,r} + \gamma X_{i,y} + \theta Z_{s,y} + \delta_{g,y} + \varepsilon_{i,s,g,y}$$

where $Y_{i,s,g,y}^{spring}$ is the standardized STAR or MAP spring test score in either subject for student *i* in school *s*, grade *g*, and academic year *y*.[31] On the right side of Equation (4), the superscripts *f, w, m*, and *r* refer to fall, winter, math, and reading, respectively, so that, for example, *f,m* denotes the fall math test score. $X_{i,y}$ represents a vector of student-level covariates, and $Z_{s,y}$ represents a vector of school-level covariates. Student-level covariates included gender, race/ethnicity, FRL status, English learner status, special education status. School-level covariates included percentages of the student-level covariates, school population, and average fall test scores. We include grade-by-year fixed effects $\delta_{g,y}$ to control for unobserved, statewide effects that vary by grade and/or academic year. For the final predictions of Spring 2020 assessment scores used in the analysis models, we use both Fall 2019 and Winter 2020 assessments to capture actual test score growth that occurred before the COVID-19 school closures and use that actual growth trajectory to predict the growth trajectory that is most likely to have occurred between Winter and Spring 2020.

31 STAR and MAP scores are first equated using the methodology described above.

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

After the overall standardization and equating as well as the prediction of the Spring 2020 test scores, this analysis estimates strategy impact on reading and math growth using the following student-level specification:

Equation (5)

$$Y_{i,s,g,y} = \propto + \beta_1 \text{class size}_{s,g,y} + \beta_2 \text{ coaching}_{s,g,y} + \beta_3 \text{tutoring}_{s,g,y} + \gamma X_{iy} + \pi_s + \delta_{g,y} + \varepsilon_{i,s,g,y}$$

where $Y_{i,s,g,y}$ is an outcome for student *i* in grade *g*, school *s*, and year *y*. *class size*$_{s,g,y}$, *coaching*$_{s,g,y}$, and *tutoring*$_{s,g,y}$ are indicators for whether a school employed that particular strategy, in each grade and year. $X_{i,y}$ represents a vector of student-level covariates, including lagged values of the outcome *Y*. Grade-by-year fixed effects, $\delta_{g,y}$, are included to control for any unobserved, statewide effects that vary by grade and/or time. School fixed effects, π_s, are included to control for any unobserved effects that vary by school. Standard errors are clustered at the school level. Models for MAP/STAR math and reading and absence rate use a linear regression, 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, then after estimation used an inverse transformation for interpretable results.

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

Robustness to Alternative Estimation Strategies

To assess the robustness of our findings, we tested several alternative estimation strategies that attempt to address limitations of the matching and estimation strategies described above. These strategies include school fixed effects, school random effects, and school-specific time trends, shown in Table A2 through Table A7. In each of the tables, Column (1) displays the results from the preferred specification used in the main analysis above. Columns (2)-(4) display separate robustness checks.

The matching and estimation strategies described above rely on the standard propensity score 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, the quality of teachers available in the local labor market) that might be related to both outcomes and AGR/SAGE participation, and therefore bias impact estimates. However, there is no way to test this assumption for AGR using available data. Including school fixed effects in the estimation controls for differences in unobservable characteristics between schools by comparing outcomes before and after AGR implementation within the same school. For the AGR analysis, however, including school fixed effects only allows comparisons of AGR to SAGE because all AGR schools previously participated in SAGE. With school fixed effects, comparisons to non-SAGE, non-AGR comparison schools are impossible, because comparison schools' program participation does not change and therefore 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, Columns (2) in Table A2 - Table A7 contain the results of school fixed effects regressions. These results are qualitatively similar to the preferred specification in Column (1).

As an alternative to fixed effects, we also consider school-specific random effects, which produce a weighted average of between-school and within-school effects.[33] Random effects, however, do not allow for variation in weights within schools, which occurs when matching testing outcomes. As discussed above, our preferred matching strategy matches (and assigns weights) within cohorts to maximize comparability of AGR and non-AGR schools. Conversely, attendance and discipline outcomes are best matched using outcome data from as early as possible (2012-13), resulting in schools retaining their matching weights across years. Column (3) of Table A2 - Table A7 shows results from regressions that include random effects. For both attendance and discipline, key coefficients are qualitatively similar to those from the preferred specification in Column (1).

32 Benjamini, Y., & Hochberg, Y. (1995). 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. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

33 Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics Methods and Applications (p. 711). Cambridge University Press.

Finally, we test for the presence of time trends that may differ between AGR/SAGE and control schools and therefore bias results, shown in Column (4). For example, if AGR/SAGE schools are more likely on positive trajectories unrelated to their participation in the program, estimates of AGR/SAGE impacts would be biased upward. We choose not to include school-specific time trends in our preferred specification because these trends could be the result of SAGE and AGR, and there is no way to differentiate between unrelated trends and program impacts. However, in Table A2 - A7 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 a small influence on estimated AGR/SAGE impacts. The exceptions are Forward math, which becomes statistically significant and negative when including time trends, and absences in Model 2, where the positive and significant coefficient (before correcting for multiple estimates) disappears with the inclusion of trends. Given the concerns with including time trends and negating previous AGR and SAGE impacts, however, these differences in coefficients do not cause major concern with the preferred specification.

Table A2: Robustness to Alternative Estimation Strategies – Forward Reading

SAMPLE	(1)	(2)	(3)	(4)
AGR vs. non	0.106	N/A	N/A	-0.281
p-value	0.727	N/A	N/A	0.833
AGR vs. SAGE	-0.016	1.420	N/A	-0.367
p-value	0.447	0.359	N/A	0.734
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

* Statistically significant at the 0.05 level. Note that p-values have not been corrected for multiple estimates.

Table A3: Robustness to Alternative Estimation Strategies – Forward Math

SAMPLE	(1)	(2)	(3)	(4)
AGR vs. non	-0.019	N/A	N/A	-3.274*
p-value	0.960	N/A	N/A	0.016
AGR vs. SAGE	-0.102	-0.587	N/A	-2.717*
p-value	0.853	0.632	N/A	0.016
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

* Statistically significant at the 0.05 level. Note that p-values have not been corrected for multiple estimates.

Table A4: Robustness to Alternative Estimation Strategies – Absences, Model I

SAMPLE	(1)	(2)	(3)	(4)
AGR vs. non	0.33	N/A	0.08	0.17
p-value	0.098	N/A	0.814	0.645
AGR vs. SAGE	-0.23	-0.15	-0.16	-0.31
p-value	0.082	0.145	0.167	0.127
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

* Statistically significant at the 0.05 level. Note that p-values have not been corrected for multiple estimates.

Table A5: Robustness to Alternative Estimation Strategies – Absences, Model 2

SAMPLE	(1)	(2)	(3)	(4)
AGR vs. non	0.41*	N/A	0.18	-0.01
p-value	0.034	N/A	0.582	0.980
AGR vs. SAGE	-0.45*	-0.24*	-0.01*	-0.61*
p-value	0.005	0.037	0.041	0.012
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

* Statistically significant at the 0.05 level. Note that p-values have not been corrected for multiple estimates.

Table A6: Robustness to Alternative Estimation Strategies – OSS, Model 1

SAMPLE	(1)	(2)	(3)	(4)
AGR vs. non	-0.4	N/A	-0.3	-0.4
p-value	0.235	N/A	0.293	0.279
AGR vs. SAGE	0.3	0.2	0.2	0.2
p-value	0.322	0.596	0.439	0.482
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

* Statistically significant at the 0.05 level. Note that p-values have not been corrected for multiple estimates.

Table A7: Robustness to Alternative Estimation Strategies – OSS, Model 2

SAMPLE	(1)	(2)	(3)	(4)
AGR vs. non	-0.3	N/A	-0.3	0.1
p-value	0.256	N/A	0.124	0.951
AGR vs. SAGE	0.1	-0.1	-0.0	-0.0
p-value	0.5145	0.554	0.824	0.951
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

* Statistically significant at the 0.05 level. Note that p-values have not been corrected for multiple estimates.

Students Leaving Wisconsin Public Schools Due to COVID-19

This section explores how COVID-19 impacted students leaving WI Public Schools by comparing leavers in 2020-21 to leavers in previous years for both students that were in AGR schools and students that were in non-AGR schools. There is anecdotal evidence to suggest that in 2020-21, due to COVID-19, districts were unable to contact some students and that many students left for private schools. To investigate, this section analyzes (a) whether higher fractions of students left the Wisconsin public schools database in 2020-21, and (b) whether students left AGR schools at higher rates than their peers at comparable schools.

We begin with Table A8, which shows the percentages of students that left the data by year and AGR status. Students who continue in Wisconsin public schools should reappear in the administrative data the next year. Therefore, we define leavers as students who were in a Wisconsin public school in the previous year but did not appear in the administrative data in the current year. As seen in Table A8 both students in AGR and non-AGR schools left public schools at higher rates in 2020-21, lending credence to the anecdotal evidence.

Table A8: Percentage of K-3 Students Leaving WI Public Schools by Year and AGR/SAGE Status

GROUP	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21
AGR/SAGE	4.5%	4.5%	4.5%	4.8%	4.4%	5.2%	4.5%	5.2%
Non-AGR/SAGE	3.5%	3.6%	3.5%	3.5%	3.2%	3.3%	3.2%	4.7%

To more thoroughly examine whether AGR mitigated students leaving Wisconsin public schools during 2020-21, we match AGR schools to similar, comparison schools. In creating the matching process, we sought to control for locale and COVID learning model, factors that are correlated with AGR and may be correlated with students' decisions to leave public schools during 2020-21. Students attending schools in urban areas, who are more likely to attend an AGR school (see Figure 6), likely have more nearby private school options, which might increase the probability that those students leave their schools. Similarly, anecdotal evidence indicates that some students and their families were dissatisfied with virtual learning, which was more common among AGR schools, making them more likely to switch to private schools.

To control for locale and learning models in our analysis, we match AGR schools to non-AGR schools using exact matching on indicators for locale (urban, suburb, town, rural) and September learning model (virtual, hybrid, in-person), under the assumption that students who left during 2020-21 made their decisions early in the school year or before. Otherwise, matching for this analysis uses the same matching controls as described for attendance and discipline (see Table 2), with the exception that for Model 2 we replace school-level averages of attendance or discipline with each schools' 2012-13 leaver rate.

For the leaver analysis, the matching process was unable to find quality matches while exact matching on locale and learning model. The highest quality model results in unacceptably high effect sizes for the difference in school percent FRL between AGR and comparison schools. Despite this, we present results below, with the caution that the analysis should be considered suggestive evidence. Balance and common support tables from the leaver analysis are available upon request.

The analysis uses difference-in-differences with a pre-trend common to the AGR and non-AGR groups to determine whether there are meaningful differences in leaving between AGR and "similar" non-AGR schools. Results of the analysis are displayed in Table A9, showing no differences in leaving between the AGR and comparison groups.

Table A9: Preliminary Impact of AGR on 2020-21 Leaving Rates

SAMPLE	IMPACT (PERCENTAGE POINTS)	P-VALUE
All Students - Model 1 (no controls for prior outcomes)	0.01	0.952
All Students - Model 2 (controls for 2012-13 leaver rate)	0.07	0.776

Wisconsin Evaluation Collaborative