Wisconsin Evaluation Collaborative

November 2024

Evaluation of the Achievement Gap Reduction Program

2015-16 through 2022-23

for the Wisconsin Department of Public Instruction

WISCONSIN UNIVERSITY OF WISCONSIN-MADISON

Wisconsin Center for Education Research
UNIVERSITY OF WISCONSIN-MADISON

About the Authors

Jed Richardson

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

Grant Sim

Grant Sim is an Evaluation and Research Scientist at the Wisconsin Evaluation Collaborative. He holds a Master's degree in Public Affairs from the University of Wisconsin–Madison's La Follette School of Public Affairs.

McKenna Goetz

McKenna Goetz is a Research Intern with WEC. She is currently pursuing a Master of Public Affairs degree at the La Follette School of Public Affairs.

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 https://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 report 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 two regular classroom teachers
- Provide data-driven instructional coaching for one or more teachers of one or more participating grades. The instruction shall be provided by licensed teachers who possess appropriate content knowledge to assist classroom teachers in improving instruction in math or reading and 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? How does AGR impact vary by student characteristics?
3. Are there differences between the three AGR strategies' relationships to intended outcomes?
4. What best practices can districts use to maximize AGR impacts on achievement gaps?

1 Hanushek, E.A., Light, J. D., Peterson, P. E., Talpey, L. M., & Woessman, L. (2022). Long-run Trends in the U.S. SES-Achievement Gap. *Education Finance & Policy, 17*(4), 608-640. https://doi.org/10.1162/edfp_a_00383

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

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. 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.

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. 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 2022-23, the most recent year of data, 400 schools implemented the AGR program, serving over 70,000 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. There was a slight decrease in 2021-22, when 64 percent of schools utilized multiple strategies, but this returned to 75 percent of schools in 2022-23.
- Over 70 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 positive impacts of K-3 SAGE on eventual high school persistence and completion.[3]

3 Meyer, R., Dokumaci, E., Sim, G., Steele, C., Suchor, K., & Vadas, J. (2015). *SAGE Program Evaluation Final Report*. Value-Added Research Center. https://www.google.com/url?client=internal-element-cse&cx=012012553761441775853:itqvwm_yb2u&q=https://dpi.wi.gov/sites/default/files/imce/sage/pdf/sage_2015_evaluation.pdf&sa=U&ved=2ahUKEwiUoNfmtfyGAxUbjYkEHU8kAAUQFnoECAYQAQ&usg=AOv-VawIX8zU4-doBGLjVknR5IgIz&fexp=72519171,72519168

Results from the current analysis show that there is no estimated impact of the AGR program on third, fourth, or fifth 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 Forward implies that AGR reading impacts in kindergarten fade out by third grade or that there is misalignment between kindergarten and Forward tests.

Impact estimates should be interpreted with caution. Inconsistent testing patterns in grades K-3, including diminishing use of kindergarten PALS, a lack of testing in spring 2020, and reduced testing participation in spring 2021, restricted the sample of AGR and non-AGR schools included in the growth analysis samples. The evaluation also limits cohorts to schools that were mostly in-person during 2020-21. All of these sample restrictions potentially limit how growth impact estimates can be generalized to schools not in the sample.

How do schools with high academic growth implement AGR?

This year's evaluation identified AGR schools with high math and reading growth and interviewed principals at those schools in an attempt to identify best practices. Several themes emerged from these interviews:

- Principals are grateful for the funding and think that it has substantial impacts on their schools. As one stated, "We are so successful … we score so well on …. any measure, standardized and non-standardized of reading and math … there's no doubt that AGR is a big part of it."
- Principals see AGR funding as helping the students who need it most.
- According to these principals, AGR funds benefit the neediest students through several mechanisms. Reduced class sizes allow both classroom teachers and interventionists to work more closely with individual students. Schools using one-to-one tutoring focus their efforts on the neediest students.
- AGR funding, though helpful, is not sufficient to meet current needs in some schools. One principal noted that AGR funds have remained stable over time while staff health care costs have increased, meaning that fewer staff can be hired with the same dollars.

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. Beginning 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 two 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, Section 118.44 (2015). https://docs.legis.wisconsin.gov/2015/related/acts/53.pdf

Figure 1: Socioeconomic Achievement Gaps: NAEP Reading

Grade 4, 2003 – 2022

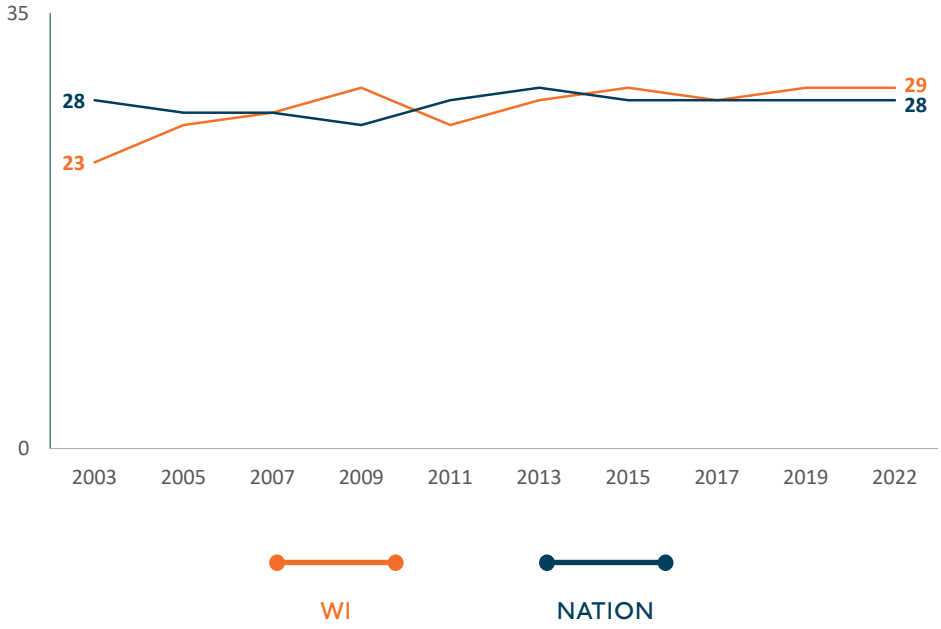
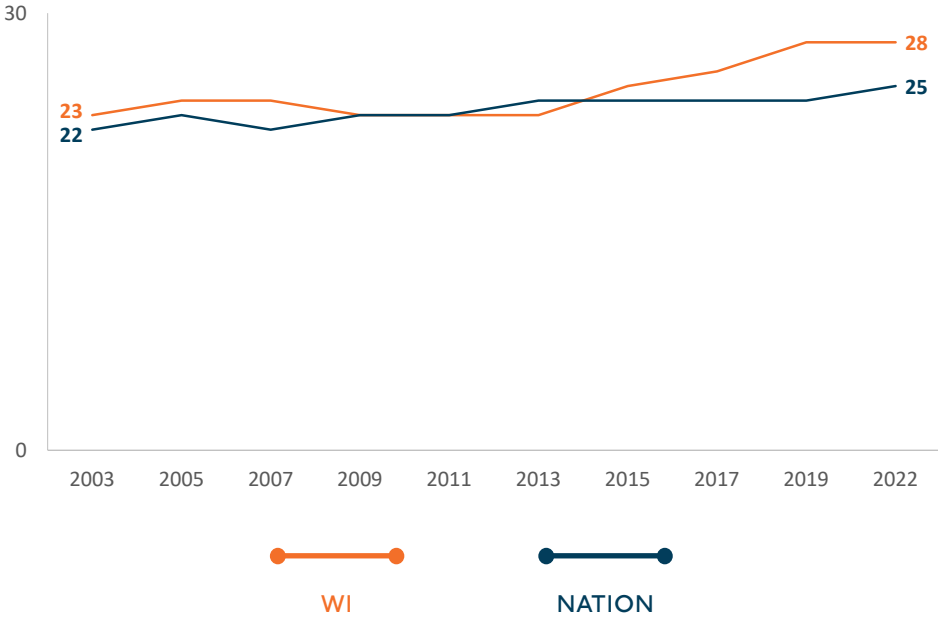


Figure 2: Socioeconomic Achievement Gaps: NAEP Math

Grade 4, 2003 – 2022



Context

The AGR program seeks to reduce the achievement gap for economically disadvantaged students. Over the past fifty years, however, nationwide achievement gaps by socioeconomic status have been stagnant.[5] Wisconsin is no exception. As shown in Figures 1 and 2, during the 2000s neither Wisconsin nor the nation made progress reducing gaps for economically disadvantaged 4th graders on NAEP math and reading, respectively.

5 Hanushek, E.A., Light, J. D., Peterson, P. E., Talpey, L. M., & Woessman, L. (2022). Long-run Trends in the U.S. SES-Achievement Gap. *Education Finance & Policy, 17*(4), 608-640. https://doi.org/10.1162/edfp_a_00383

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,[6] differences in summer opportunities,[7] differences in the amount of time parents are able to spend with their children,[8] and differences in neural development owing to exposure to high-poverty, and potentially, traumatic environments.[9]

AGR's strategy for closing socioeconomic achievement gaps is to provide additional funding to districts with large proportions of economically disadvantaged students. This strategy is consistent with research concerning how school funding impacts student outcomes. In findings that are especially pertinent to AGR, Jackson et al. (2021) use a high-quality research design to show that decreases in school resources widened the socioeconomic achievement gap.[10] 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.[11] Furthermore, an individual study shows that for low-income students, increased spending increases educational attainment, increases adult wages, and lowers the incidence of poverty.[12]

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 2022-23.

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

1. How are AGR schools implementing the AGR program as specified by 2015 Wisconsin Acts 53 and 71?
    a. What is the breakdown of strategy usage across the state?
    b. How does implementation of the three strategies differ across schools?
2. To what extent is AGR meeting intended outcomes, including impacts on standardized test scores? How does AGR impact vary by student characteristics?
3. Are there differences between the three AGR strategies' relationships to intended outcomes?
4. What best practices can districts use to maximize AGR impacts on achievement gaps?

6 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.

7 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), 549-584. https://digitalcommons.law.byu.edu/cgi/viewcontent.cgi?article=1374&context=elj

8 Guryan, J., Hurst, E., & Kearney, M. (2008). Parental Education and Parental Time with Children. *Journal of Economic Perspectives, 22*(3), 23-46. https://www.aeaweb.org/articles?id=10.1257/jep.22.3.23

9 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.

10 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), 304-335. https://www.aeaweb.org/articles?id=10.1257/pol.20180674

11 Jackson, C. K., & Mackevicius, C. (2021). *The Distribution of School Spending Impacts* (NBER Working Paper No. 28517). National Bureau of Economic Research. https://www.nber.org/papers/w28517

12 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

After 2020, the AGR evaluation became 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. Taking into account evidence that in-person learning was more effective for academic achievement growth during the pandemic, 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.[13] 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, educator, and public 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 ten 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. Effective Implementation in AGR Schools describes the analytical framework used to identify high-growth AGR schools and the interview process that evaluators used to collect data from identified schools. Longitudinal Analyses of End-of-Year (EOY) and School Board Reports looks at these data from 2017-18 through 2022-23. The last three sections provide Summary and Conclusions, Appendix A describing technical aspects of the methodology, and Appendix B, which includes the principal interview protocol.

13 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* (NBER Working Paper No. 30010). National Bureau of Economic Research. https://www.nber.org/papers/w30010; Jack, R., Halloran, C., Okun, J.C., & Oster, E. (2021). *Pandemic Schooling Mode and Student Test Scores: Evidence from US States* (NBER Working Paper No. 29497). National Bureau of Economic Research. https://www.nber.org/papers/w29497

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 2022-23 academic years.[14] 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 2022-23. 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).[15] Finally, data on district-level median income and school poverty levels are also sourced from NCES, which uses data from the U.S. Census Bureau's American Community Survey.[16]

- Demographic characteristics include gender, race/ethnicity, English learner (EL) 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, fourth, and fifth grade spring Forward reading and math test scores and fall and spring kindergarten reading scores from the Phonological Awareness Literacy Screening (PALS).
- Enrollment data include school attended and grade.
- School-level data include SAGE and AGR funding by year, average teacher compensation, school location (city, suburb, town, rural), and school learning model (in-person, virtual, mixed) during 2020-21.
- District-level data include median income and poverty levels.

14 Note that although SAGE impacts are accounted for in the evaluation methodology, the report no longer includes SAGE impact estimates.

15 Public Staff Reports are available at https://publicstaffreports.dpi.wi.gov/PubStaffReport/Public/PublicReport. NCES school location information is available at https://nces.ed.gov/ccd/elsi/.

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

Identifying Comparison Schools

Using the data described above, we aggregated each school's K-3 data to find a comparison group of non-AGR schools. 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.

For Forward math and reading outcomes, 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. The matching is the same regardless of whether the outcome is measured for third, fourth, or fifth grade. As a result, the matching and analysis omit the 2016-17 cohort because those students were in third grade during 2019-20 when there was no Forward testing. 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. For the primary specification, we omit any affected students who attended schools that were not in-person at least 75 percent of the days between September 2020 and April 2021, 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. 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.

During matching, we selected the sample to address testing gaps and differences in learning models due to the COVID-19 pandemic. For the primary specification, we omit any affected students who attended schools that were not in-person at least 75 percent of the days between September 2020 and April 2021, 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.[17] 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.[18]

17 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* (NBER Working Paper No. 30010). National Bureau of Economic Research. https://www.nber.org/papers/w30010; Halloran, C., Jack, R., Okun, J.C., & Oster, E. (2021). *Pandemic Schooling Mode and Student Test Scores: Evidence from US States* (NBER Working Paper No. 29497). National Bureau of Economic Research. https://www.nber.org/papers/w29497

18 For the previous year's evaluation, we tested and presented models of third and fourth grade Forward scores that included all students, regardless of their school's 2020-21 learning model. These models match on learning model but likely do not fully control for the impacts of COVID shutdowns.

Table 1: Propensity Score Matching Controls

Third, Fourth, and Fifth Grade Forward Reading and Math

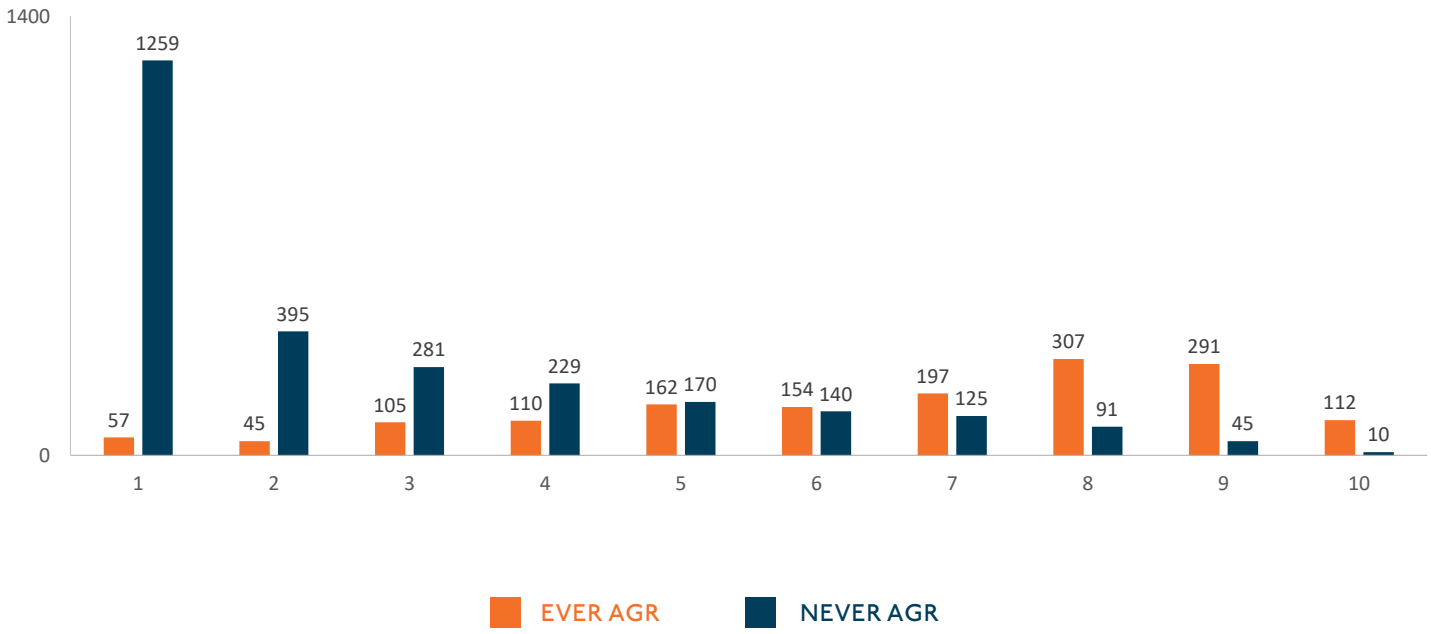
MATCHING VARIABLE (MEASURED FALL OF KINDERGARTEN)
School Average Fall Kindergarten PALS
School Percent Black, Hispanic, White, Other Race/Ethnicity*
School Percent Free/Reduced-Price Lunch
School Percent Mobile from Prior Year
Locale Description (City, Suburb, Town, Rural)*
District Median Income
School 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 .

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 untreated schools. Most deciles have more than 40 comparison schools, showing sufficient overlap for the analysis. This overlap is similar across in the math model.

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 Table 1 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 2. All models include weights generated by the kernel PSM procedure. Standard errors are clustered at the school level.

Table 2: Analysis Model Controls

CONTROL VARIABLE
Student Demographics Race/Ethnicity*, Free/Reduced-Price Lunch, English Learner, Special Education
School Demographic Percentages Race/Ethnicity*, English Learner, Special Education
Student Year-to-Year Mobility
School Percent Mobile from Prior Year
Locale Description (City, Suburb, Town, Rural)*
Student Fall Kindergarten PALS
School Average Fall Kindergarten PALS
District Median Income
School K-3 Student Population
School Number of Cohort Students in Analysis Sample
Cohort Indicators**

* 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.

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 substantially and continues to decline. As districts, particularly the state's largest, have adopted alternative kindergarten assessments, the overall tested population of AGR schools is less and less representative of the untested sample of AGR schools. It should also be noted that available data cannot support analysis of whether schools' choice of PALS testing is related to outcomes and participation in AGR.

Readers should also 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 through 2022-23 samples of AGR students and schools included in Forward analyses are markedly different from the general AGR population. In particular, less than one-quarter of AGR schools, those that began AGR the earliest, had a cohort experience four years of AGR without also experiencing the pandemic in grades K-3. Due to this timing, it is not possible to analyze the impacts of the full, four-year AGR exposure without including the impacts of the pandemic. 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.

Finally, the phase out of PALS and the need to omit schools that were mainly virtual during 2020-21 results in a limited sample that does not represent the full sample of AGR schools and produces results that are not consistently robust to changes in model specification. Viewed together, the limitations described above imply that the results of this evaluation should be viewed as suggestive of AGR impacts but insufficient for determining the true impacts of the program.

Section 4

AGR Demographics

AGR Demographics

This section of the report provides details on the characteristics of AGR students and schools. Table 3 shows the number of AGR schools for each of the eight 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 subsequent years, the overall number of AGR schools dropped slightly. In 2022-23, there were a total of 400 AGR schools.

The numbers of students in AGR schools from 2015-16 to 2022-23, overall and by grade, are presented in Table 4. The first cohort of AGR schools included approximately 18,000 students, while the addition of the second cohort in 2016-17 brought the total to over 77,000 students. Since, student participation has declined, decreasing to 70,714 in 2022-23.

Table 3: Number of AGR Schools

By Grade and Year

GRADE	2015-16	2016-17	2017-18	2017-18	2019-20	2020-21	2021-22	2022-23
Kindergarten	88	393	391	395	396	391	386	386
First	91	398	397	403	400	395	391	386
Second	91	398	398	403	400	395	391	386
Third	88	391	391	395	394	389	384	378
Overall (K-3)	96	408	408	413	412	408	404	400

Table 4: Number of AGR Students

By Grade and Year

GRADE	2015-16	2016-17	2017-18	2017-18	2019-20	2020-21	2021-22	2022-23
Kindergarten	4,138	18,382	18,262	18,406	18,430	17,659	17,739	17,591
First	4,570	19,294	18,813	18,786	18,539	17,950	17,749	17,921
Second	4,682	20,053	19,158	18,959	18,594	18,002	17,826	17,866
Third	4,544	19,506	19,227	18,532	18,093	17,493	17,440	17,366
Overall (K-3)	17,934	77,235	75,460	74,683	73,656	71,104	70,754	70,714

Figure 4 and Figure 5 compare the demographic characteristics of AGR students to all K-3 Wisconsin students (including AGR students) in 2022-23. 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 are 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.

Figure 4: Race/Ethnicity of AGR and WI Students

2022-23

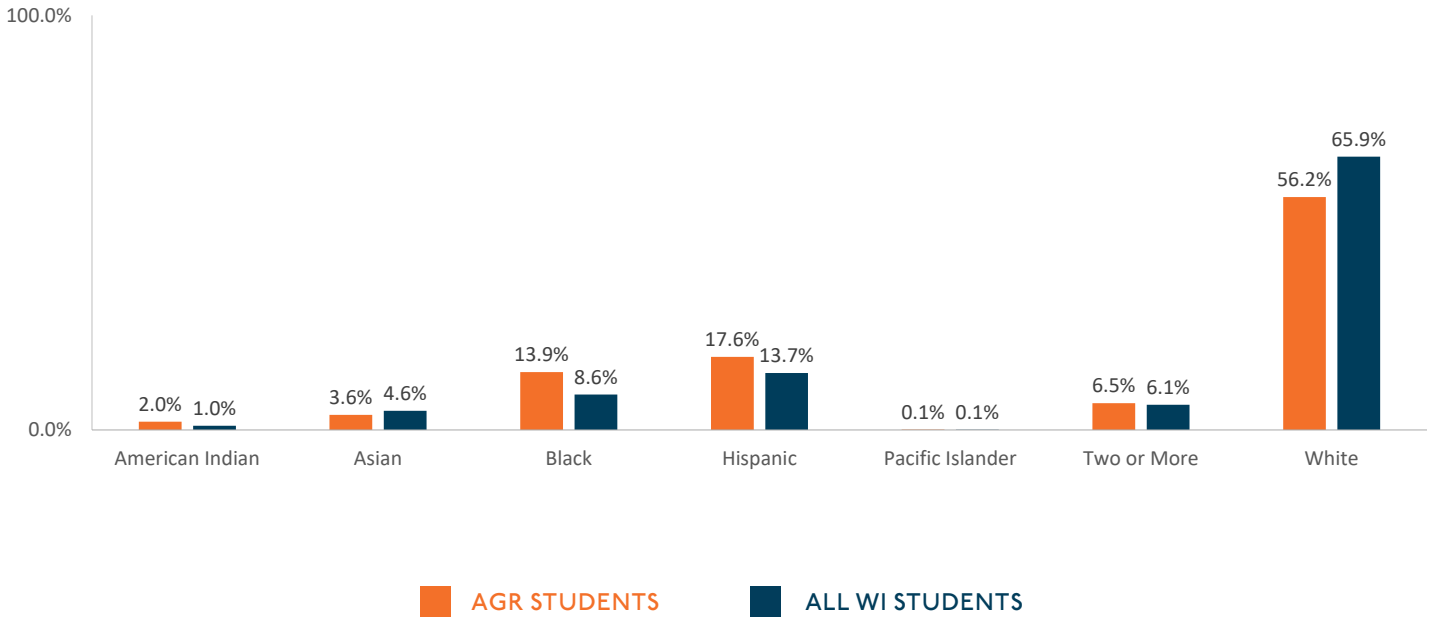


Figure 5: Percentage of AGR and WI Students who were English Learners, Eligible for FRL, and in Special Education

2022-23

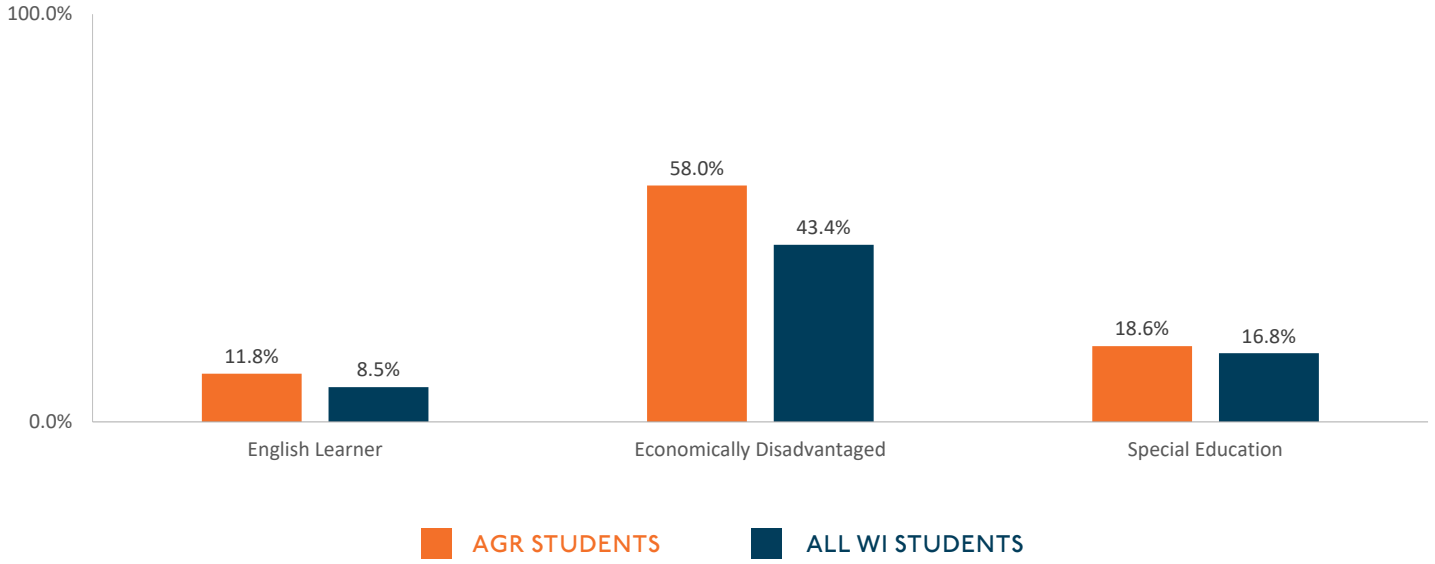
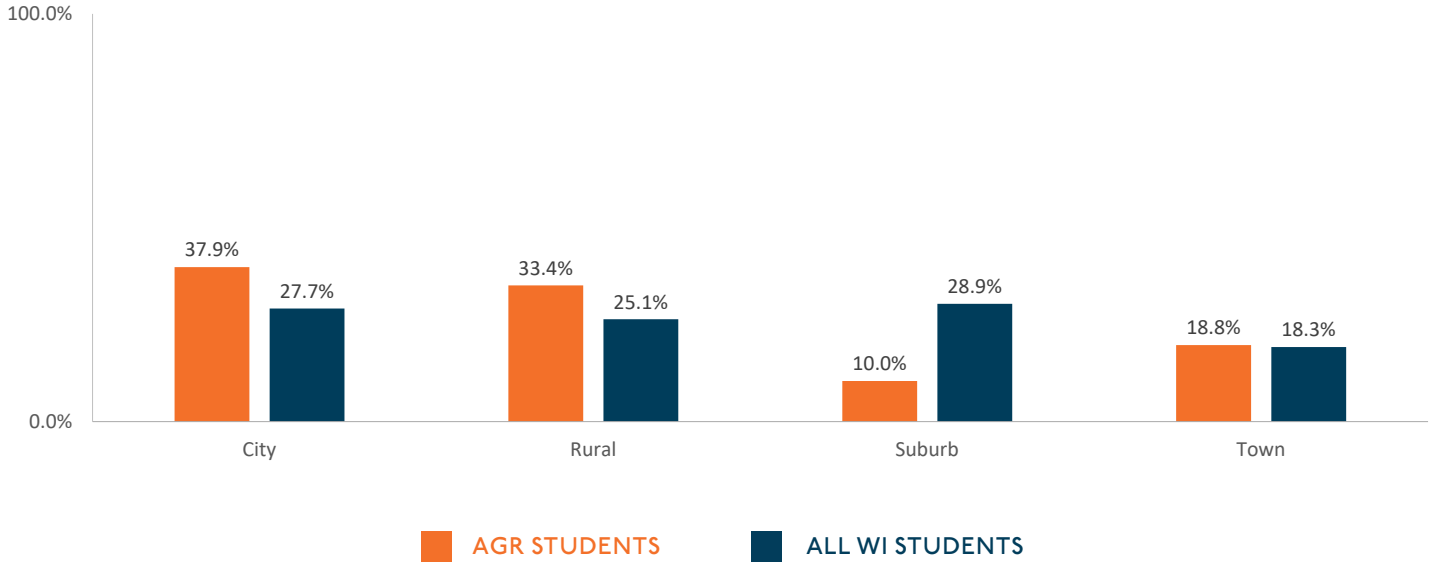


Figure 6: Locale Description of AGR and WI Students

2022-23



Third Grade Forward Analysis Sample

Examining third grade Forward as an outcome of AGR 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 2). 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 and has continued to decrease since. The resulting sample drops accordingly. Other restrictions on the number of students used in the analysis sample include: (a) students need to appear in the data all four years (kindergarten through third grade); (b) students must follow the typical grade progression from kindergarten to third grade; and, (c) students must not switch between an AGR and a non-AGR school during kindergarten through third grade. 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 fewer than five students who otherwise would have been included in the analysis sample. Finally, for the cohorts of students starting kindergarten in 2017-18 through 2019-20 and attending third grade in 2020-21 through 2022-23 respectively, we omit any schools that were not in person at least 75 percent of the days between September and April in 2020-21, as noted above. The resulting sample of analysis students compared to the total number of AGR students is presented in Table 5.

Table 5: 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	2018-19	2019-20	ALL YEARS
All AGR Students	22,380	21,464	20,884	19,782	19,121	18,930	18,667	18,674	159,902
Analysis AGR Students	12,622	12,310	12,166	11,240	0	2,397	1,625	1,602	53,962
Percentage in Analysis	56.4%	57.4%	58.3%	56.8%	0.0%	12.7%	8.7%	8.6%	33.7%

Because not all AGR students are included in the Forward analysis sample, the results likely do not generalize to all AGR students, particularly for students in the 2017-18 through 2019-20 kindergarten cohorts. To better understand 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 are similar (Figure 7 - Figure 9), for the 2019-20 kindergarten cohort the differences are larger. Analysis students in the 2019-20 cohort are less likely to be Asian, 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 2019-20 cohort are more likely rural and none are urban (Figure 12).[19] The 2017-18 and 2018-19 cohorts are similar to the 2019-20 cohort – for more information see the 2020-21 AGR evaluation report and 2021-22 AGR evaluation report. Based on the data in Figures 10-12 and in prior reports, results from the 2017-18 through 2019-20 cohorts are likely not representative of the overall AGR impact. This is particularly important because, other than approximately 3,000 students from the 2016-17 cohort, the 2017-18 through 2019-20 cohorts were the first to experience AGR throughout kindergarten, first, second, and third grades. As a result, impact estimates of the full, four-year AGR treatment rely on a smaller number of students who mostly attended school in-person during COVID and do not adequately represent AGR students as a whole.

19 The lack of students in the city locale in the 2019-20 analysis cohort is primarily due to the prevalence of virtual schooling in and around Madison and Milwaukee during 2020-21.

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

All Cohorts

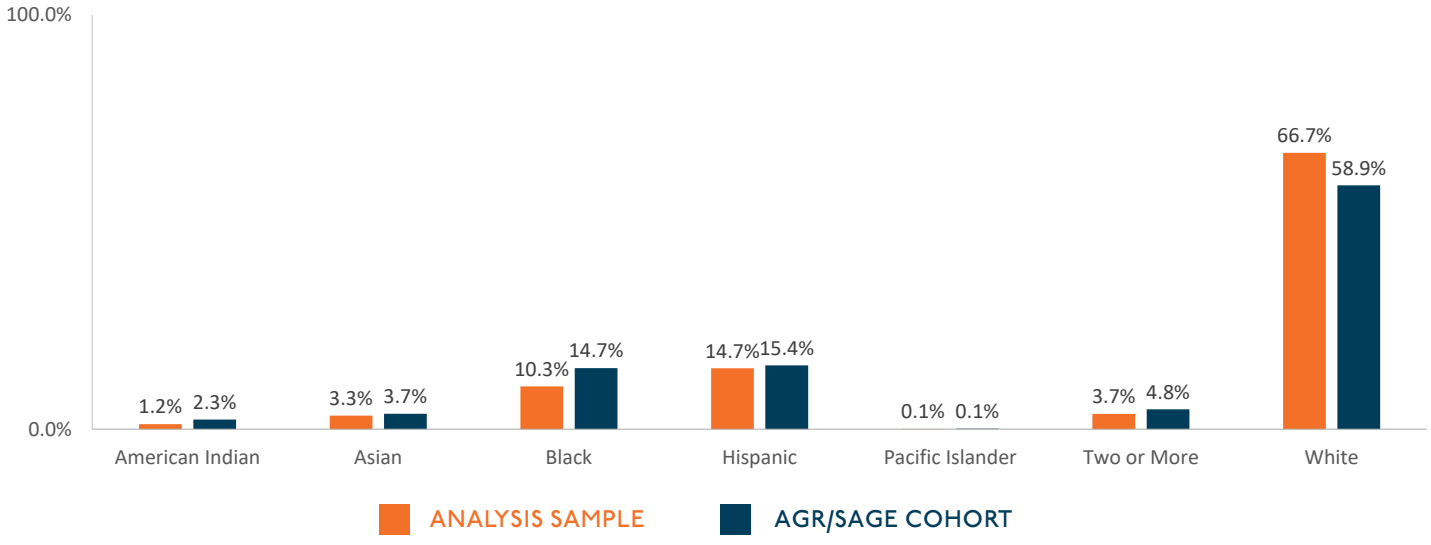


Figure 8: AGR Forward Analysis Students and AGR Students Overall who were English Learners, Eligible for FRL, and in Special Education

All Cohorts

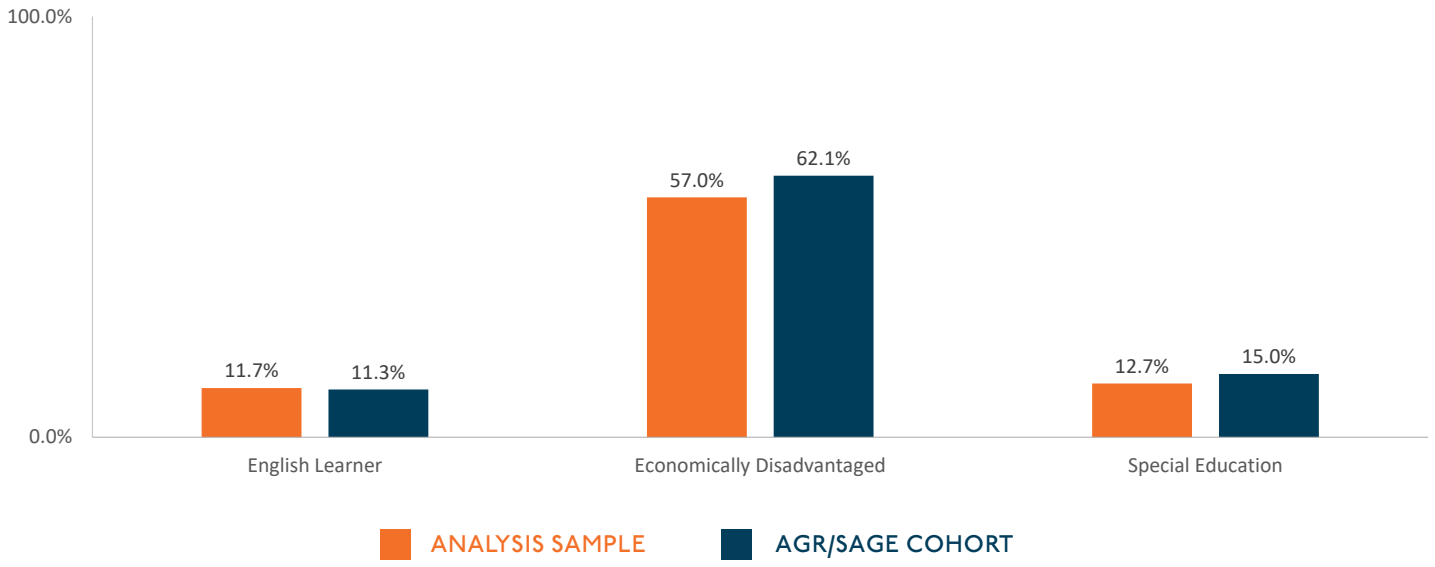


Figure 9: Locale Description of AGR Forward Analysis Students and AGR Students Overall

All Cohorts

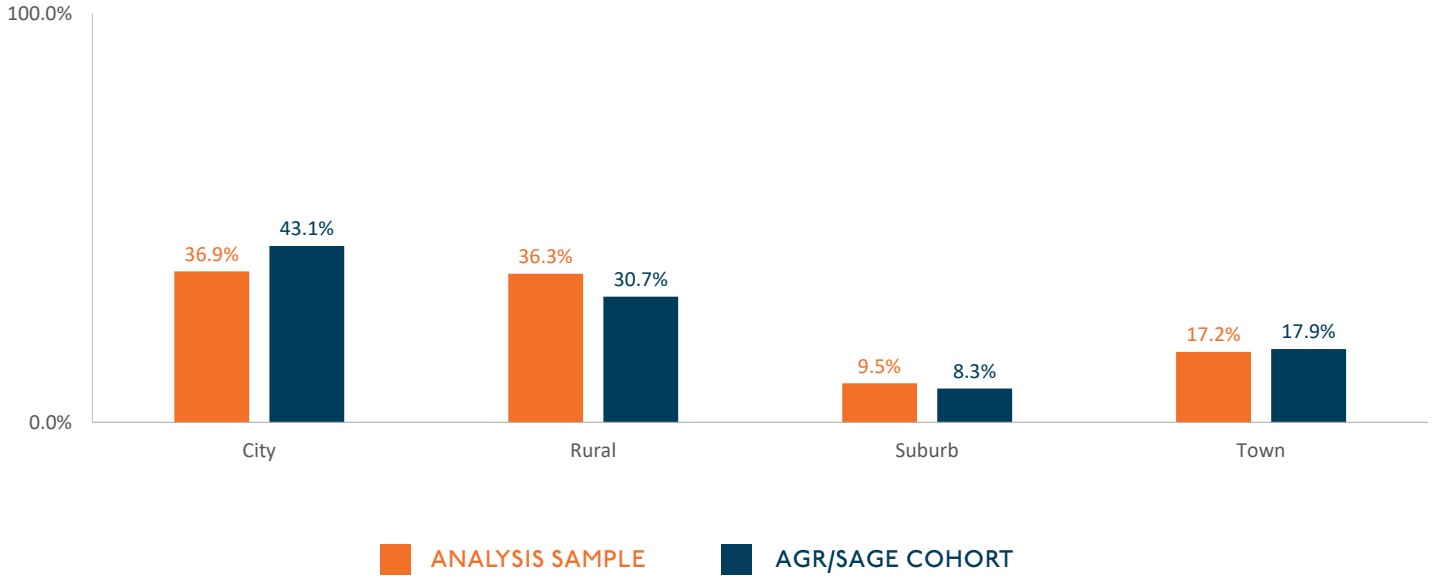


Figure 10: Race/Ethnicity of AGR Forward Analysis Students and AGR Students Overall

2019-20 Kindergarten Cohort

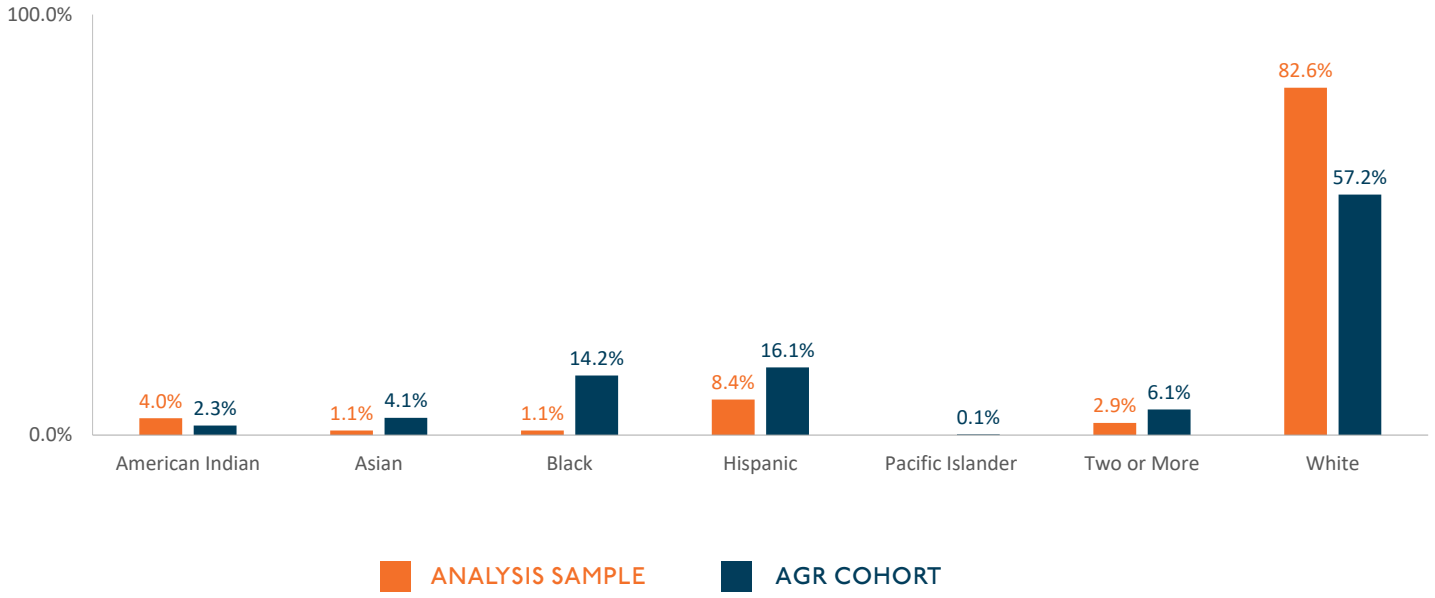


Figure 11: AGR Forward Analysis Students and AGR Students Overall who were English Learners, Eligible for FRL, and in Special Education

2019-20 Kindergarten Cohort

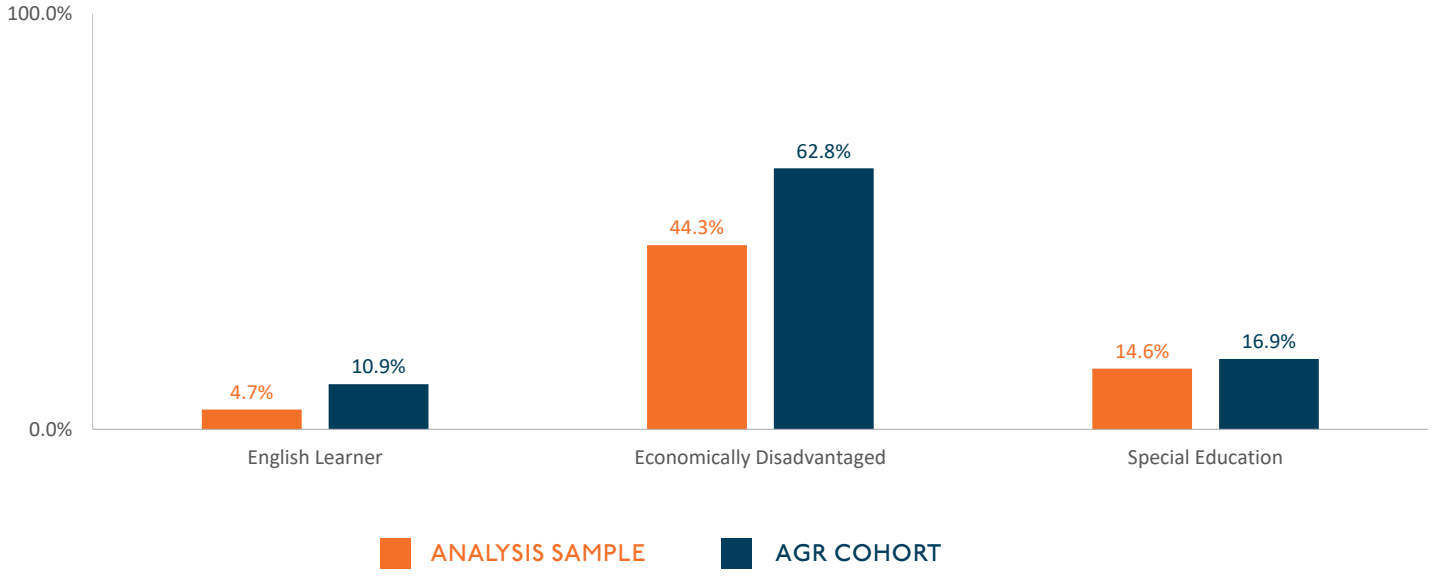
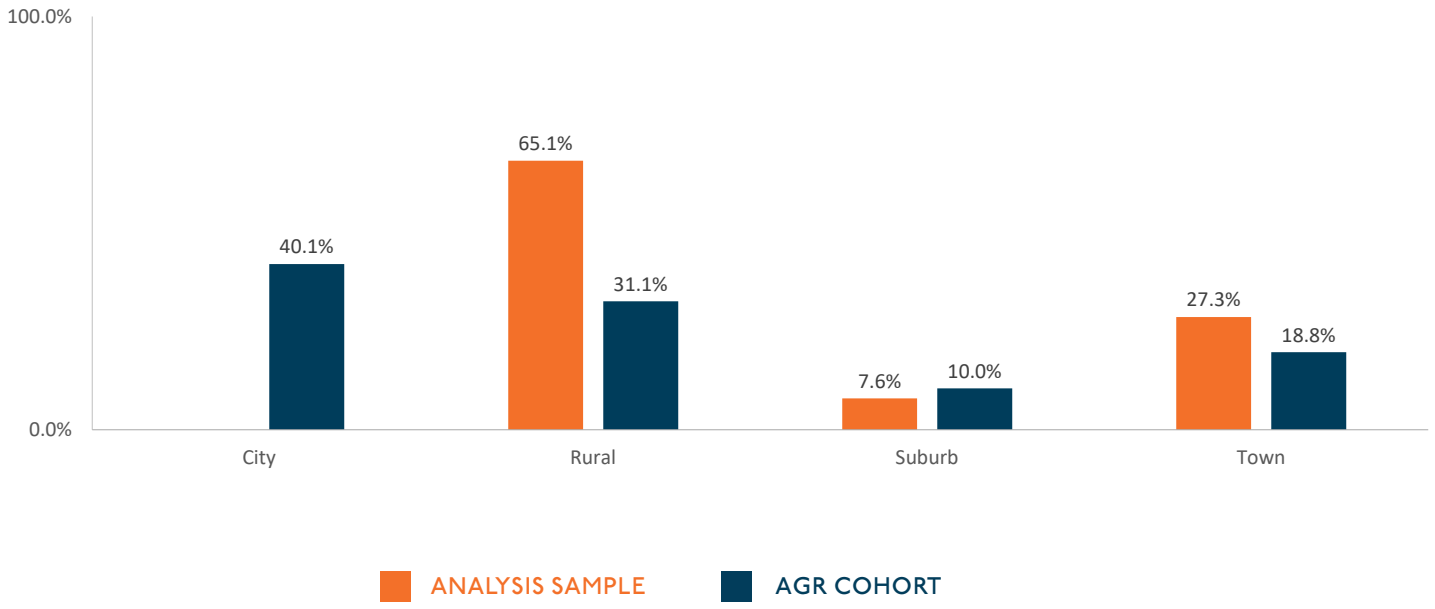


Figure 12: Locale Description of AGR Forward Analysis Students and AGR Students Overall

2019-20 Kindergarten Cohort



Section 5

AGR Implementation

AGR Implementation

This section and the sections that follow present evaluation results aligned with the evaluation questions above. This section examines implementation of the three possible AGR strategies to answer the first evaluation question: How are AGR schools implementing the AGR program? As noted previously, the three strategies include:

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

As the AGR program allows schools to use more than one strategy within a school, there are seven possible combinations schools could implement: class size reduction only, coaching only, tutoring only, class size reduction and coaching, class size reduction and tutoring, coaching and tutoring, and all three strategies. Figure 13 and Figure 14 show data on the strategy combinations AGR schools implemented during 2022-23. These figures also provide information on the number and percentage of students affected by each strategy combination. Schools most frequently used coaching only, class size reduction and coaching combined, all three strategies, and class size reduction only. Only three schools used only tutoring as a strategy.

Figure 13: Implementation of AGR Strategies

Percentage of Students, 2022-23

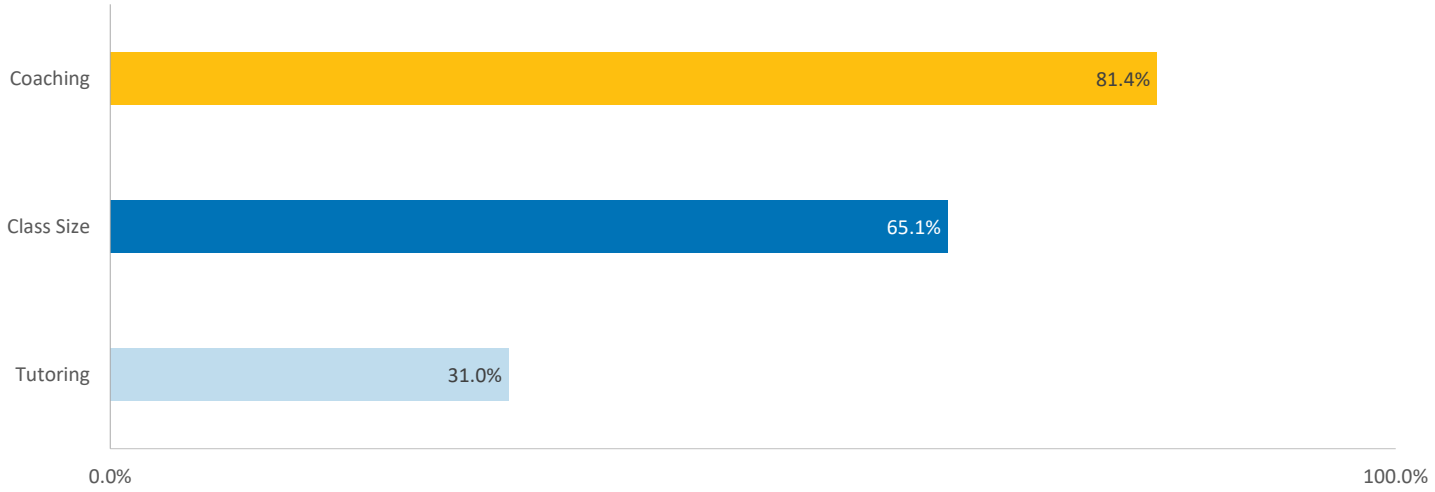
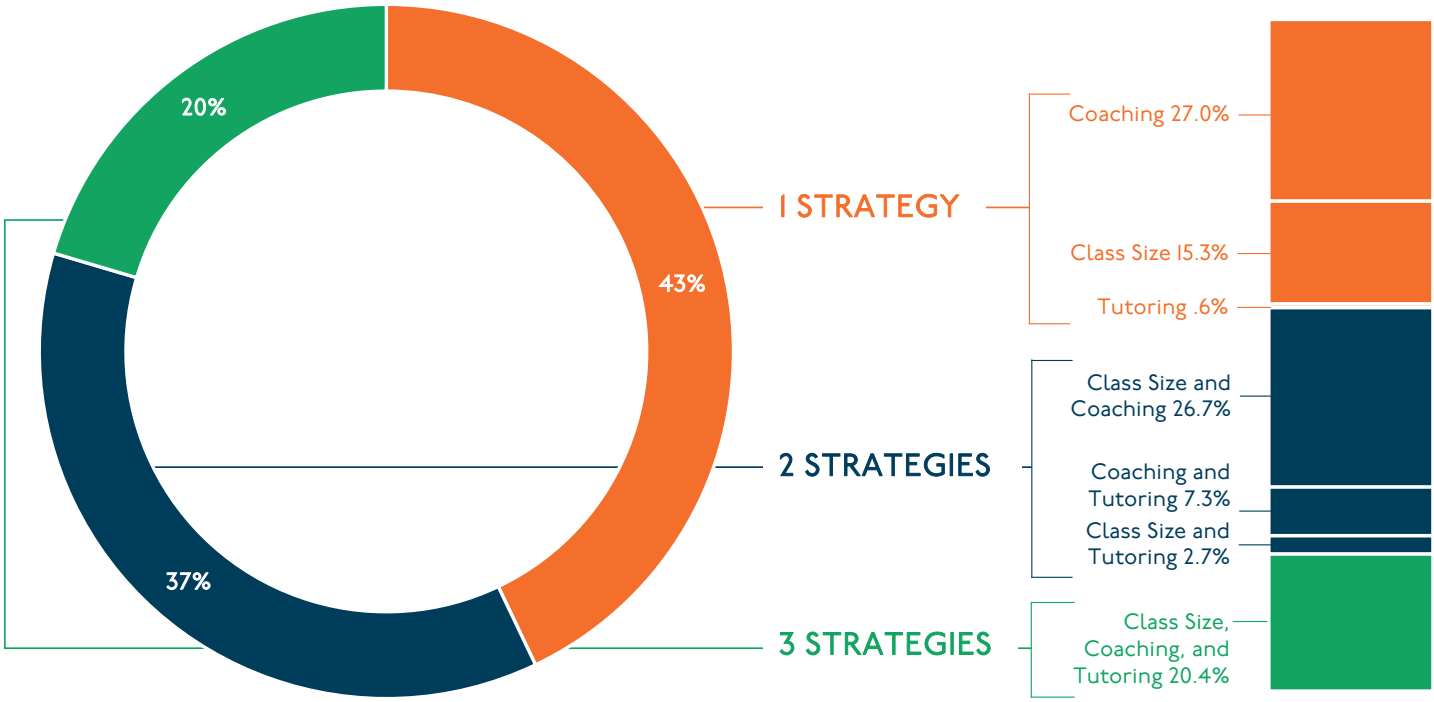


Figure 14: Implementation of AGR Strategies

Percentage of Students by Strategies, 2022-23



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 question listed above: To what extent is AGR meeting intended outcomes?

To answer this question, 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 various subpopulations of students, including low-income students. All impact analyses examine how students performed on six outcome measures:

- Third grade Forward reading,
- Third grade Forward math,
- Fourth grade Forward reading,
- Fourth grade Forward math,
- Fifth grade Forward reading, and
- Fifth grade Forward math.

To avoid over-interpretation of results based on small fractions of the overall AGR population due to COVID and diminishing PALS usage, we opt only to report results for subgroups with at least 15 percent of their AGR population included in the analysis sample.

For each outcome, this report provides a table of impact results. 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.

Third Grade Forward Impacts

Table 6 and Table 7 present the impacts of AGR on third grade Forward reading and math, respectively. Impacts show the differences between students who received four years of AGR, or some combination of AGR and SAGE, and students who received zero years of AGR or SAGE. Results are broken down by student demographics.

As shown in Table 6, AGR impacts on third grade Forward reading are generally positive but small and not statistically different from zero, both for all students and for all subsets of students. For context, the impact from four years of AGR (1.62 scale score points) represents 0.03 standard deviations of Forward growth.[20] Table 6 also shows that impacts from having a mix of both AGR and SAGE are statistically indistinguishable from zero.

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 but zero impacts on MAP/STAR reading growth in grades 1-3.[21] 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, (b) improvements in PALS skills do not translate to improvements in the skills tested on Forward, and/or (c) there were too few students who received the full four years of AGR with assessment information for the evaluation to provide the full picture on estimates of impact.

20 Data Recognition Corporation (2023). *Wisconsin Forward Exam Spring 2023 Technical Report*. https://dpi.wi.gov/sites/default/files/imce/assessment/pdf/EWI270_WIFW_SPRG_23_TECH.pdf. See Table 10-1.

21 See Richardson, J., & Sim, G. (2020). *Evaluation of the Achievement Gap Reduction Program: 2015-16 through 2018-19*. Wisconsin Evaluation Collaborative. https://dpi.wi.gov/sites/default/files/imce/assessment/pdf/EWI270_WIFW_SPRG_23_TECH.pdf

Table 6: Impact of AGR on Third Grade Forward Reading Growth

SAMPLE	FORWARD SCORE IMPACT (4 YEARS OF AGR)	P-VALUE	FORWARD SCORE IMPACT (MIX OF AGR/SAGE)	P-VALUE
All	1.62	0.69	0.02	>0.99
FRL	2.89	0.44	-0.67	0.78
Non-Urban	2.31	0.58	0.95	0.76
FRL & Non-Urban	2.78	0.63	0.93	0.79
Non-Urban & Black	0.08	>0.99	2.67	0.73
Non-Urban & Hispanic	7.34	0.13	-1.20	0.87
Non-Urban & White	2.06	0.67	0.99	0.76
Non-Urban & Other Race	-0.01	>0.99	3.49	0.60

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

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

SAMPLE	FORWARD SCORE IMPACT (4 YEARS OF AGR)	P-VALUE	FORWARD SCORE IMPACT (MIX OF AGR/SAGE)	P-VALUE
All	2.19	0.67	-0.53	0.87
FRL	0.94	0.87	-1.21	0.65
Non-Urban	2.29	0.59	-0.12	>0.99
FRL & Non-Urban	-0.68	0.97	-0.57	0.93
Non-Urban & Black	2.47	0.95	0.75	1.02
Non-Urban & Hispanic	7.06	0.29	-2.10	0.70
Non-Urban & White	2.40	0.62	0.17	>0.99
Non-Urban & Other Race	-5.93	0.62	-1.53	0.87

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

AGR impacts on third grade Forward math, as shown in Table 7, are similar to those of reading. The impact of four years of AGR (2.19 scale score points) is equal to 0.04 standard deviations.[22] Again, there are no statistically significant impacts on Forward math for all students or any subset of students. Similarly, there were no statistically significant impacts on Forward math for students with a mix of AGR and SAGE.

Fourth Grade Forward Impacts

Table 8 and Table 9 present the impacts of AGR on fourth grade Forward reading and math, respectively. Impacts show the difference between students who received four years of AGR, or some combination of AGR and SAGE, and students who received zero years of AGR or SAGE. Results are broken down by student demographics.

As shown in Table 8, AGR impacts on fourth grade Forward reading are negative but small and not statistically different from zero, both for all students and for all reported subgroups. For context, the impact of four years of AGR (-1.66 scale score points) represents 0.03 standard deviations.[23] Table 8 shows several negative impacts of a mix of AGR and SAGE, although none of these results are statistically significant.

22 Ibid.

23 Ibid.

Table 8: Impact of AGR on Fourth Grade Forward Reading Growth

SAMPLE	FORWARD SCORE IMPACT (4 YEARS OF AGR)	P-VALUE	FORWARD SCORE IMPACT (MIX OF AGR/SAGE)	P-VALUE
All	-1.66	0.71	-2.47	0.12
FRL	-0.44	>0.99	-3.68	0.06
Non-Urban	-1.53	0.76	-2.05	0.43
FRL & Non-Urban	-1.25	0.85	-2.95	0.31
Non-Urban & Black	6.13	0.70	3.87	0.65
Non-Urban & Hispanic	4.75	0.59	-3.47	0.65
Non-Urban & White	-2.73	0.66	-2.45	0.27
Non-Urban & Other Race	5.68	0.72	5.79	0.57

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

Table 9: Impact of AGR on Fourth Grade Forward Math Growth

SAMPLE	FORWARD SCORE IMPACT (4 YEARS OF AGR)	P-VALUE	FORWARD SCORE IMPACT (MIX OF AGR/SAGE)	P-VALUE
All	4.97	0.47	-2.77	0.14
FRL	3.72	0.58	-3.44	0.15
Non-Urban	5.02	0.46	-3.73	0.05
FRL & Non-Urban	2.41	0.76	-4.68	0.07
Non-Urban & Black	9.25	0.56	-0.53	>0.99
Non-Urban & Hispanic	7.26	0.47	-4.63	0.38
Non-Urban & White	4.35	0.68	-3.90	0.08
Non-Urban & Other Race	7.88	0.64	-0.43	>0.99

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

AGR impacts on fourth grade Forward math, as shown in Table 9, are slightly different from reading impacts. The overall impact of AGR and the impacts of AGR on various subgroups of students are small, positive, and not statistically significant. The impact of four years of AGR (4.97 scale score points) is equal to 0.09 standard deviations.[24] As with fourth grade reading, there are several negative impacts of a mix of AGR and SAGE, although none of these results are statistically significant.

Fifth Grade Forward Impacts

Table 10 and Table 11 present the impacts of AGR on fifth grade Forward reading and math, respectively. Impacts show the difference between students who received four years of AGR, or some combination of AGR and SAGE, and students who received zero years of AGR or SAGE. Results are broken down by student demographics.

AGR impacts on fifth grade Forward reading are positive but small and not statistically different from zero, both for all students and for all reported subgroups (except for Non-Urban and Black students for which AGR had larger impacts, but which are still not statistically significant). For context, the impact of four years of AGR (0.62 scale score points) represents 0.01 standard deviations.[25] Table 10 shows some positive and some negative impacts of a mix of AGR and SAGE, depending on the subgroup, although none of these results are statistically significant.

24 Ibid.

25 Ibid.

Table 10: Impact of AGR on Fifth Grade Forward Reading Growth

SAMPLE	FORWARD SCORE IMPACT (4 YEARS OF AGR)	P-VALUE	FORWARD SCORE IMPACT (MIX OF AGR/SAGE)	P-VALUE
All	0.62	>0.99	-1.50	0.45
FRL	0.13	>0.99	-2.29	0.28
Non-Urban	1.07	0.91	-0.18	>0.99
FRL & Non-Urban	0.19	>0.99	-0.18	>0.99
Non-Urban & Black	10.01	0.36	3.11	0.77
Non-Urban & Hispanic	3.35	0.66	-2.43	0.54
Non-Urban & White	0.78	0.97	-0.15	>0.99
Non-Urban & Other Race	0.74	>0.99	1.58	0.87

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

Table II shows the AGR impacts on fifth grade Forward math, which are mostly similar to reading impacts. The overall impact of AGR is very close to zero and the impacts of AGR on various subgroups of students are small, generally positive, and not statistically significant. There are several negative impacts of a mix of AGR and SAGE, although again, none of these results are statistically significant.

Table II: Impact of AGR on Fifth Grade Forward Math Growth

SAMPLE	FORWARD SCORE IMPACT (4 YEARS OF AGR)	P-VALUE	FORWARD SCORE IMPACT (MIX OF AGR/SAGE)	P-VALUE
All	0.01	>0.99	-2.49	0.37
FRL	0.08	>0.99	-3.34	0.25
Non-Urban	0.21	>0.99	-1.95	0.60
FRL & Non-Urban	-0.32	>0.99	-2.61	0.63
Non-Urban & Black	3.93	0.92	6.78	0.72
Non-Urban & Hispanic	5.75	0.61	-3.95	0.57
Non-Urban & White	-0.68	>0.99	-2.17	0.62
Non-Urban & Other Race	4.58	0.76	-0.01	>0.99

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

Section 7

Effective Implementation in AGR Schools

Effective Implementation in AGR Schools

New to this year's evaluation is an examination of which AGR schools have most effectively implemented AGR funding to improve test score outcomes and the implementation practices those schools utilize to achieve those outcomes. This report details the process used to identify AGR schools with high growth on assessment outcomes and the process for interviewing principals at these schools.

Identification of High-Growth Schools

The first step in learning more about effective implementation was to identify AGR schools that demonstrated high growth on district-used assessments in math and reading. The evaluation used a value-added methodology[26] to estimate the relative growth of each AGR school. To identify the estimated average Fall to Spring student growth attributable to a school, a value-added methodology controls for prior student academic achievement and other student- and school-level characteristics associated with academic achievement. Table 12 shows a full list of the student- and school-level characteristics controlled for in the model.

26 Specifically, a two-step value-added methodology suggested in Koedel, C., Mihaly, K., and Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180-195. https://www.sciencedirect.com/science/article/pii/S0272775715000072

Table 12: Value-Added Model Controls

CONTROL VARIABLE
Fall Math Score*
Fall Reading Score*
Student Demographics Race/Ethnicity**, Gender, Free/Reduced-Price Lunch, English Learner, Special Education
School Demographic Percentages Race/Ethnicity**, Free/Reduced-Price Lunch, English Learner, Special Education
School Average Fall Math Score
School Average Fall Reading Score
Grade-by-year indicators ***

* The model also includes quadratic and cubic fall assessment variables.
** Due to collinearity, we omit one Race/Ethnicity category from the model.
*** Indicators for each grade-year combination equal one if a student's grade and year in school match the indicator variable, zero otherwise.

The growth estimation focused on within-year (Fall to Spring) math and reading growth on the MAP and STAR assessments in Grades 1-3 for the school years 2017-18 through 2019-20 and 2021-22. The 2017-18 school year was the first school year when all schools transitioned to AGR away from SAGE. The 2020-21 school year was omitted due to low assessment rates during COVID. Since local assessment did not occur in the Spring of 2019-20, the evaluation estimated Spring scores for that school year with a predictive model using scores from the Fall and Winter assessments.[27] To attain sufficient sample across both assessments, MAP and STAR scores were equated and combined into a single score.

After producing growth estimates for each subject and school, the evaluation averaged the estimates across math and reading to produce a single combined growth estimate for each school. Schools were then ranked from highest to lowest according to their combined growth estimates, and the highest 45 schools were identified for potential further qualitative examination via principal interviews.

Interview Process

After the identification of high-growth AGR schools using the value-added methodology, the evaluation narrowed the list of potential schools to those that would be valuable for interviews. Schools without all grades K-3 were removed from the list of possible interview candidates since the evaluation wanted to get information on effective implementation that would be generalizable to most other AGR schools. Since the estimation of growth used data from 2017-18 through 2019-20 and 2021-22, schools that were no longer in operation or no longer AGR schools were removed from the list as well. To help verify the growth status of these schools, the evaluation also examined their School Report Card Growth Scores.[28] Schools with lower Report Card Growth and without more recent (after 2019-20) data for value-added growth were also removed. This resulted in removing 11 of the 45 schools.

The evaluation then reached out to principals of these high-growth AGR schools by order of priority, prioritizing those schools with higher value-added growth that also represented a diversity in both region (CESA) and locale (City, Suburb, Town, or Rural) to schedule interviews. Of the principals contacted, the evaluation conducted a total of five interviews.

Interview Findings

The interviews examined how schools implement AGR and the impact AGR has had on K-3 students in their schools. Findings are presented for AGR impacts, challenges, class size reduction, instructional coaching, and one-to-one tutoring. The interview protocol is included in Appendix B.

"We are so successful ... we score so well on .... any measure, standardized and non-standardized of reading and math ... there's no doubt that AGR is a big part of it."

Principal

27 For more information, refer to Richardson, J., & Sim, G. (2020). *Evaluation of the Achievement Gap Reduction Program: 2015-16 through 2019-20*. Wisconsin Evaluation Collaborative. https://dpi.wi.gov/sites/default/files/imce/sage/pdf/WEC_Evaulation_of_AGR_2015-16_-_2019-20_Report_2021.pdf

28 For more information, refer to https://dpi.wi.gov/accountability/report-cards

Overall Impacts

The five participating principals had overwhelmingly positive views of AGR's impacts at their schools, particularly for math and reading growth and for students who need the most help. For example, one principal said, "In the last three years … we have seen great growth in our reading and math. And I think because of this, the AGR supports, we have been able to focus more at the early stages in K, one, two, and hope to help fill in that gap and build those foundational skills for our kids." A second principal echoed that idea – "We are so successful … we score so well on …. any measure, standardized and non-standardized of reading and math … there's no doubt that AGR is a big part of it." A third principal was more specific – "It's [AGR has] been really helpful with those kids that have significant [challenges] … it's been huge."

"We're really able to focus on the whole child, because our classes are smaller in size. … And I speak from having been in a building for eight years that was not AGR."

Principal

Given the principals' positive outlook on AGR, it is not surprising that they foresee significant problems if the funding were ever removed. Schools reported using AGR funding in conjunction with other funds and programs (e.g., funds from districts and CESAs, Title I, ESSER, Reading Corps), but a lack of AGR funding would result in immediate reductions in staff and decreased use of teaching strategies that have proven effective. Class sizes would increase and there would be less differentiation in the classroom, with schools reverting to Tier I universal instruction: "[With] ESSER going away, we're like many [schools] with a big funding cliff on the horizon. [No AGR] would just mean less people. So we'd have to scale back tutoring. We would lose a person. … So we lose." As another principal describes their school's current situation – "[Without the AGR funding] then I just retire. That'd be it. … The AGR money is huge. It allows us to have small class sizes."

The most common criticism of AGR is that the funding is not enough – "In a building like mine, there is so much more that is needed. … I think it helps, because if we didn't have that, we'd be, you know, maybe even further behind." Another principal suggested that part of the financial issue is that AGR funds are fixed at a certain level while costs are increasing – "The problem … with all education funding in the State of Wisconsin is, it doesn't increase … And yet, healthcare costs have gone so high up hiring a teacher is, I mean, it's twice what it looks like on paper. So is it enough money? I would hazard a guess probably not. But surely we're thankful for it."

Two principals connected the effectiveness of AGR funding to high-quality staff and/or minimal turnover. Speaking of AGR's impact, one principal thought that the program was effective and followed up by describing their school's instructional coach as "wonderful." Another principal went into detail connecting staffing and AGR impacts, wondering whether the program would work as well elsewhere: "I think AGR is the best thing ever, because of the small class sizes. When I think about that, I think, well, yeah, of course, it's the best thing ever here, because I have the best teachers with no turnover. It's [AGR funding] almost like a smoke screen to say that something like that's going to help in a place like Milwaukee where there are so many social problems with the education system, with the idea of throwing a little money at small class sizes."

Class-Size Reduction

All five principals who granted interviews use AGR to reduce class sizes in grades K-3. Principals cited many benefits and impacts from smaller classes. Class sizes help address both academic and social-emotional needs, particularly for the students who need the most help: "We're really able to focus on the whole child, because our classes are smaller in size. … And I speak from having been in a building for eight years that was not AGR. I taught first and second grade, and I had up to 24 kids in my classroom. So hard to connect one-on-one with kids and hard to meet the varying needs of all of them." This same principal sees more behavioral issues in grades four and five, where class sizes cannot be reduced via AGR. A second principal said, "The school directly across the border that has the same population of kids, and they do horribly because their class sizes are 23, 24, 25, and our class sizes are 18 or less."

Across the interviews, class-size reduction was most frequently cited as being used in combination with other AGR strategies – "I also think that the less kids you have, the more you get to know everybody. And truly, the more intentional you can be about whatever extra programming you're having." Three of the five schools utilized class-size reduction in combination with one-to-one tutoring and small group instruction. For example, one principal thought that class size reduction allowed for more one-to-one tutoring and small group interactions after the introduction of a new building schedule that sets aside intervention times for reading and math. By staggering grade-level intervention times, the school can push in many support staff in addition to the classroom teachers. Three schools also used instructional coaching and other professional development to maximize the benefits of reduced class sizes. Schools coached teachers as much as possible to implement the small class sizes, sometimes using professional learning communities or one-day workshops.

Although all five principals use reduced class sizes, three of the interviewees said that some of their classes would meet AGR thresholds regardless of the program. For these principals, the ability to choose instructional coaching or one-to-one tutoring was beneficial when they could not easily reduce class sizes below 18 students. One challenge that emerged from the interviews was the difficulty ensuring that kindergarten class sizes met AGR thresholds because principals do not have enough information on how many families plan to attend. By the time this information is known, it is often too late to hire additional teachers.

Instructional Coaching

Two of the five principals interviewed indicated that they used instructional coaching as an AGR strategy at their schools. Both schools using coaching were complimentary of their coaches' contributions ("The coach has been phenomenal"). Also, both schools implement coaching similarly. They fund their coaches through AGR and other sources, including their CESAs, Title I, and, previously, ESSER (one school that does not fund coaching via AGR indicated that they use instructional coaches but fund them exclusively through other sources). The schools differentiate coaching based on teachers' needs for support: "We identify teachers who are in need of extra support … Make her [the coach's] schedule up based on that to provide supports in those rooms daily." Schools identify teachers who can benefit from coaching support via literacy coaches, the principal, and assessment data. Both schools' instructional coaches work with teachers on setting and achieving goals, and one principal singled out classroom management as a common need. One school looks to the teachers themselves for feedback on what types of professional development would help most. The most significant difference between the schools' implementation was frequency – one school's coach meets with teachers six to eight times per year, while the other school's coach meets with teachers weekly.

One-to-One Tutoring

Three principals use AGR to fund one-to-one tutoring. Like coaching, schools fund tutoring through a variety of sources, often combining AGR with other funding to hire tutors. All three principals indicated that their schools use assessment data to identify students who are in most need of academic support – "We do a variety of measures. ... We'll do a STAR screening. We'll look at their unit assessments. We'll look at any other data we have on the kid. And then the kids that have the greatest demand. ... It's really those kids that have the greatest need [who receive] intensive intervention. And usually there's not a question as to whether or not ... we have the need. It's just which students express the greatest need." Schools also focus tutoring on students who need help with a certain skill for a short period. Identification results in remarkably similar numbers of students receiving tutoring, with principals reporting four to five per grade, six per grade, and six to eight per grade. Frequency of tutoring sessions varied more, ranging between two to four days per week, four days per week, and five days per week. Tutoring sessions occur during specific times throughout the school day. One principal indicated that AGR class size reduction and one-to-one tutoring go hand in hand (see Class-Size Reduction above).

Finally, one principal suggested that AGR may be more effective if it allowed tutors to expand their interactions beyond one-to-one meetings. This principal feels that the state might get more "bang for their buck" by enabling AGR tutors to work with two or three students at a time.

Section 8

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 survey 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.

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 six years of EOY data collected from 2017-18 through 2022-23, and five years of school board report data, collected from 2017-18 through 2019-20, and 2021-22 through 2022-23. 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 13 shows AGR strategy combinations schools chose during each of the five sample years on the school board reports, while Table 14 shows strategy combinations as reported in the EOY. As can be seen in both Table 13 and Table 14, strategy choices have been mostly stable over time. There are some differences in strategies reported across the school board reports and EOY reports, although for the most part the reporting follows similar patterns.

Table 13: School-level AGR Strategies (School Board Reports)

STRATEGY	2017-18	2018-19	2019-20	2020-21	2021-22	2022-23
Coaching Only	11%	16%	15%	N/A	8%	12%
Class Size Only	27%	26%	23%	N/A	24%	21%
Tutoring Only	4%	2%	2%	N/A	2%	1%
Coaching and Class Size	23%	33%	36%	N/A	38%	37%
Coaching and Tutoring	5%	3%	4%	N/A	6%	5%
Class Size and Tutoring	10%	6%	6%	N/A	3%	5%
All Three	20%	14%	14%	N/A	20%	18%

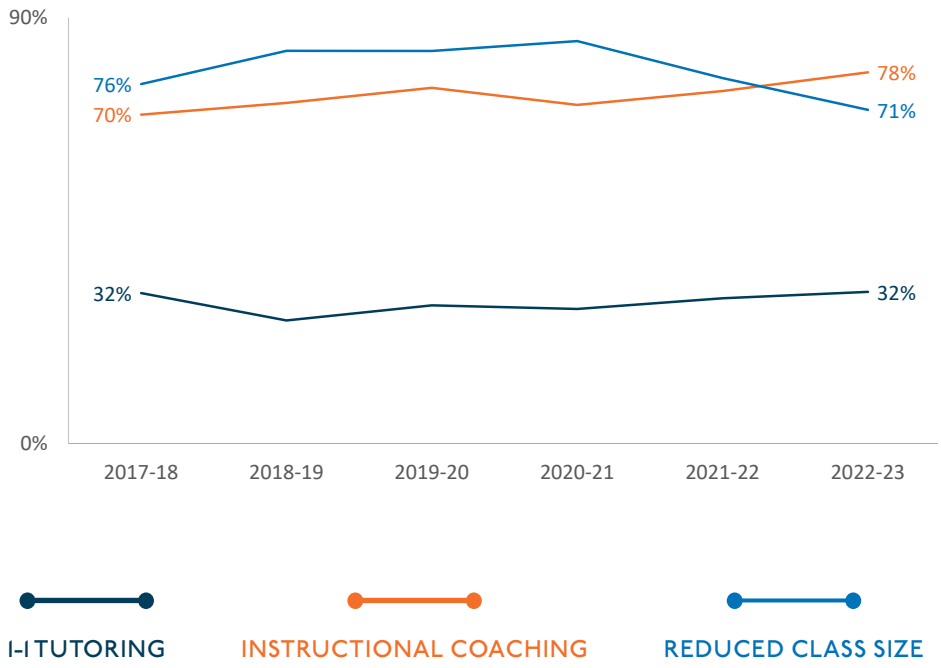
Note: In 2020-21, DPI did not require schools to submit school board reports.

Table 14: School-level AGR Strategies (End-of-Year Reports)

STRATEGY	2017-18	2018-19	2019-20	2020-21	2021-22	2022-23
Coaching Only	18%	13%	12%	12%	17%	23%
Class Size Only	21%	22%	18%	22%	19%	16%
Tutoring Only	1%	0%	1%	0%	0%	1%
Coaching and Class Size	28%	39%	41%	38%	34%	29%
Coaching and Tutoring	4%	4%	4%	3%	5%	6%
Class Size and Tutoring	8%	6%	6%	7%	6%	4%
All Three	18%	16%	18%	18%	18%	21%

From AGR's inception (its transition from SAGE, which required that all AGR schools reduce class sizes in grades K-3) through 2021-22, reduced class size was AGR's most used strategy. That changed in 2022-23, when schools reporting class size reductions decreased. As seen in Figure 15, about 71 percent of schools used class size reduction in at least one grade, lower than the approximately 78 percent that used 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 intensive tutoring.

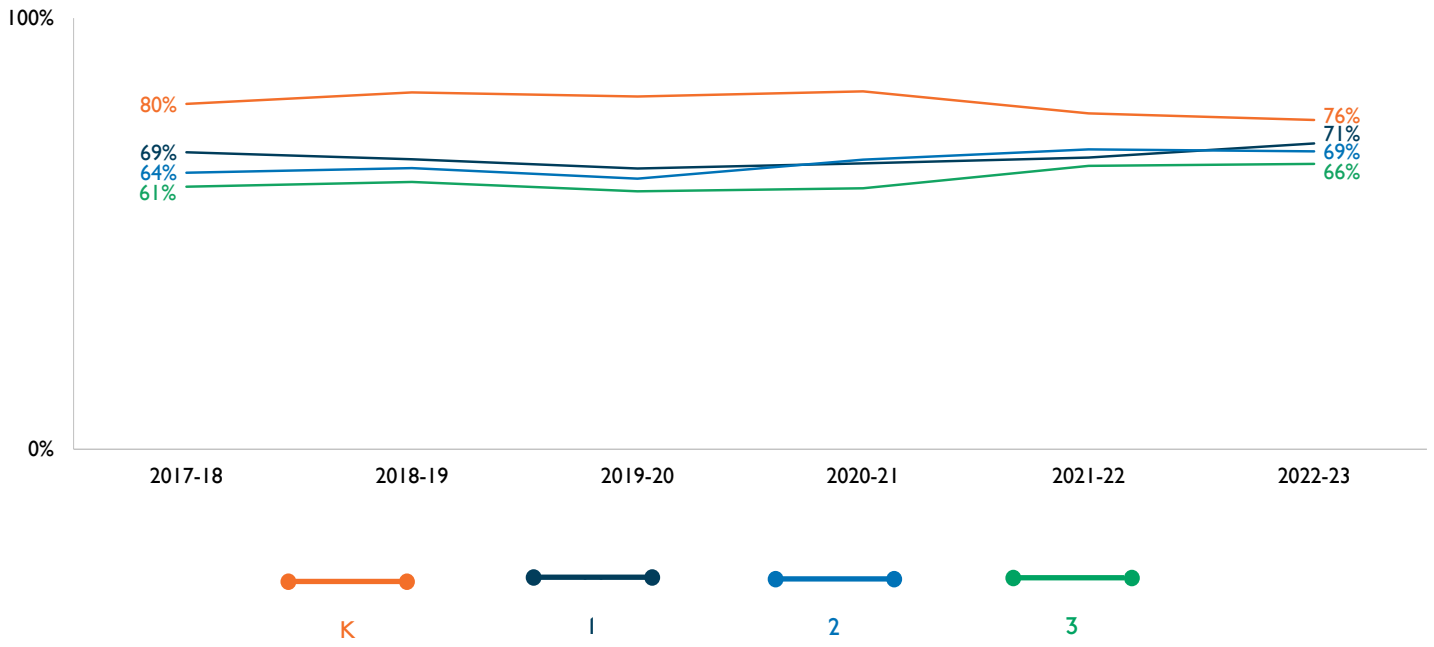
Figure 15: Schools Using Each Strategy in Any Grade



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. For example, Figure 16 shows that schools using class size reduction did not always use that strategy uniformly across grades K-3. Class size reduction was more common in kindergarten than grades 1-3; in kindergarten, over 75 percent of schools using class size reduction did so in at least three quarters of their classrooms. These percentages have been stable over time.

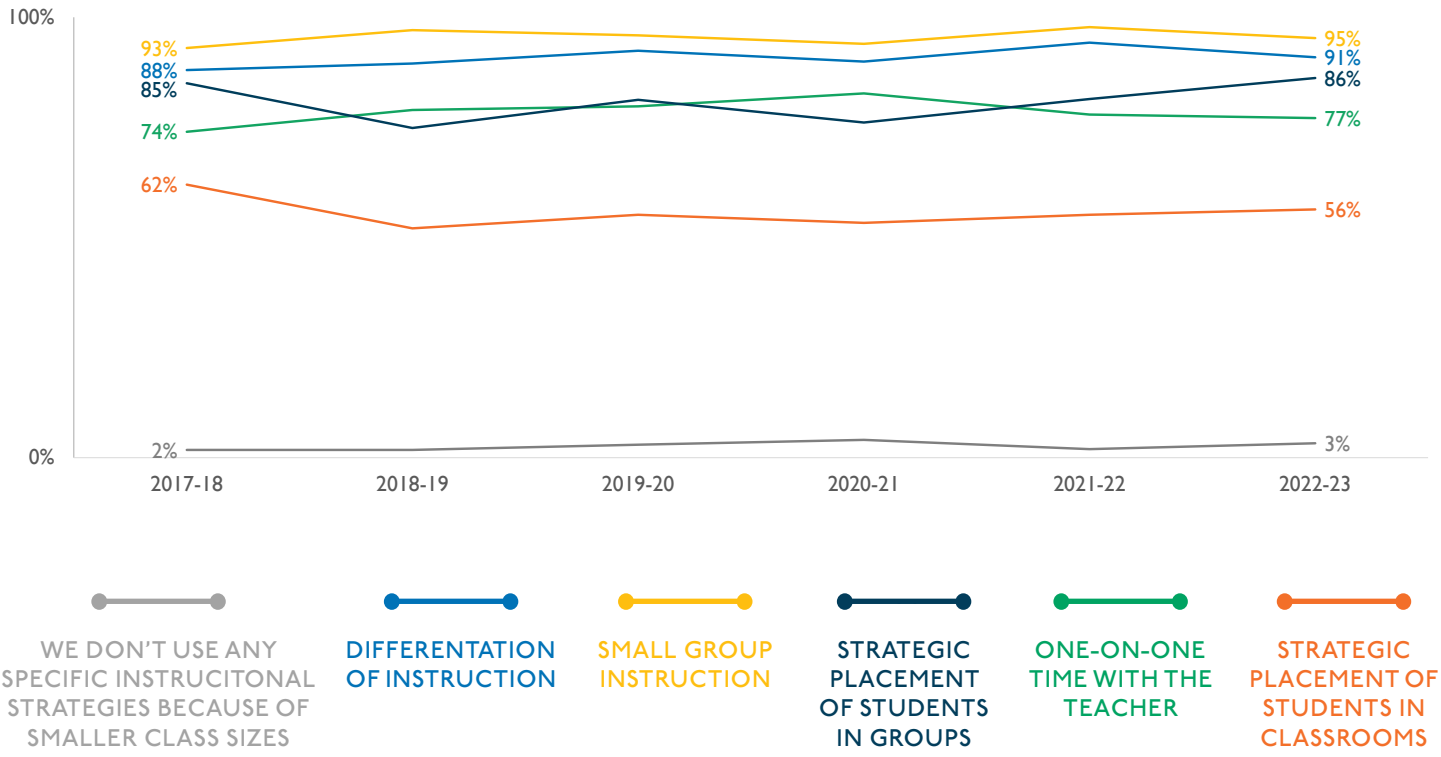
Figure 16: Schools Using Class Size Reduction Doing So in at Least 75% of Classrooms

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 six sample years. The most common instructional strategies associated with small class sizes were small group instruction, differentiation of instruction, strategic placement of students in groups, and one-on-one time with the teacher. Strategic placement of students in classrooms was less common but was reported by about 56 percent of schools.

Figure 17: Instructional Strategies in Schools with Class Size Reduction



* Note: 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

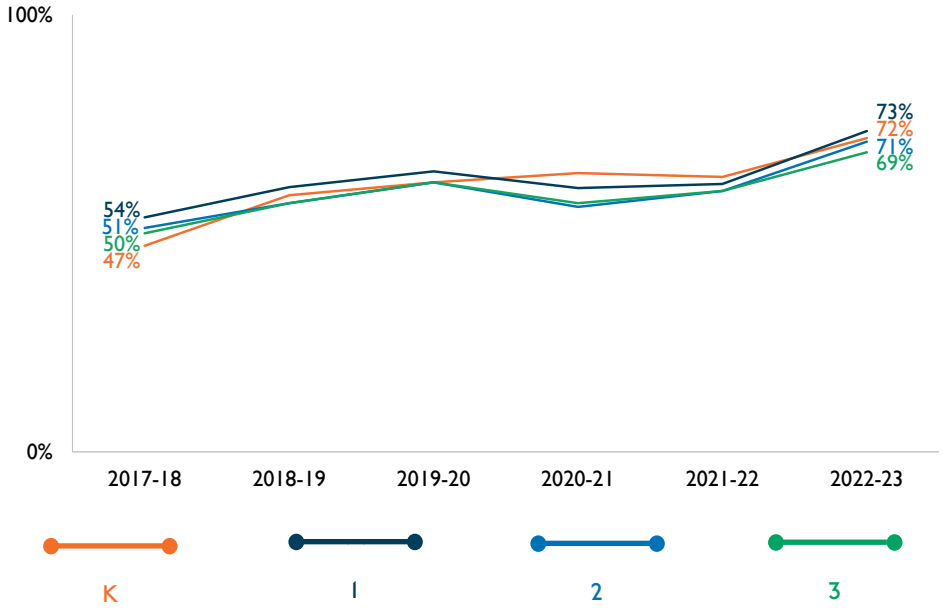
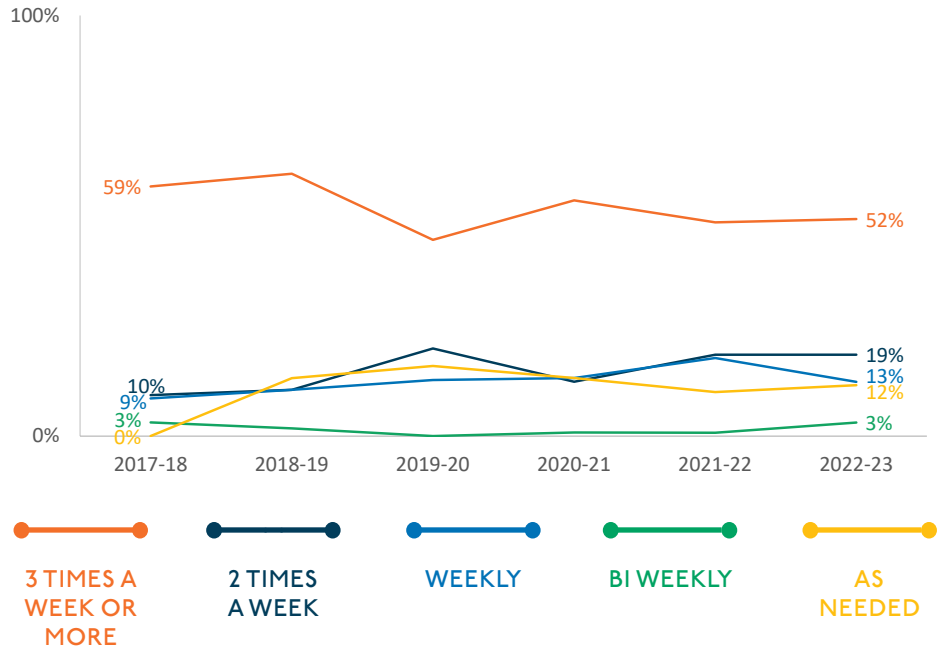


Figure 19: School-Reported Tutoring Frequency



* Note: Not Pictured – None, Other

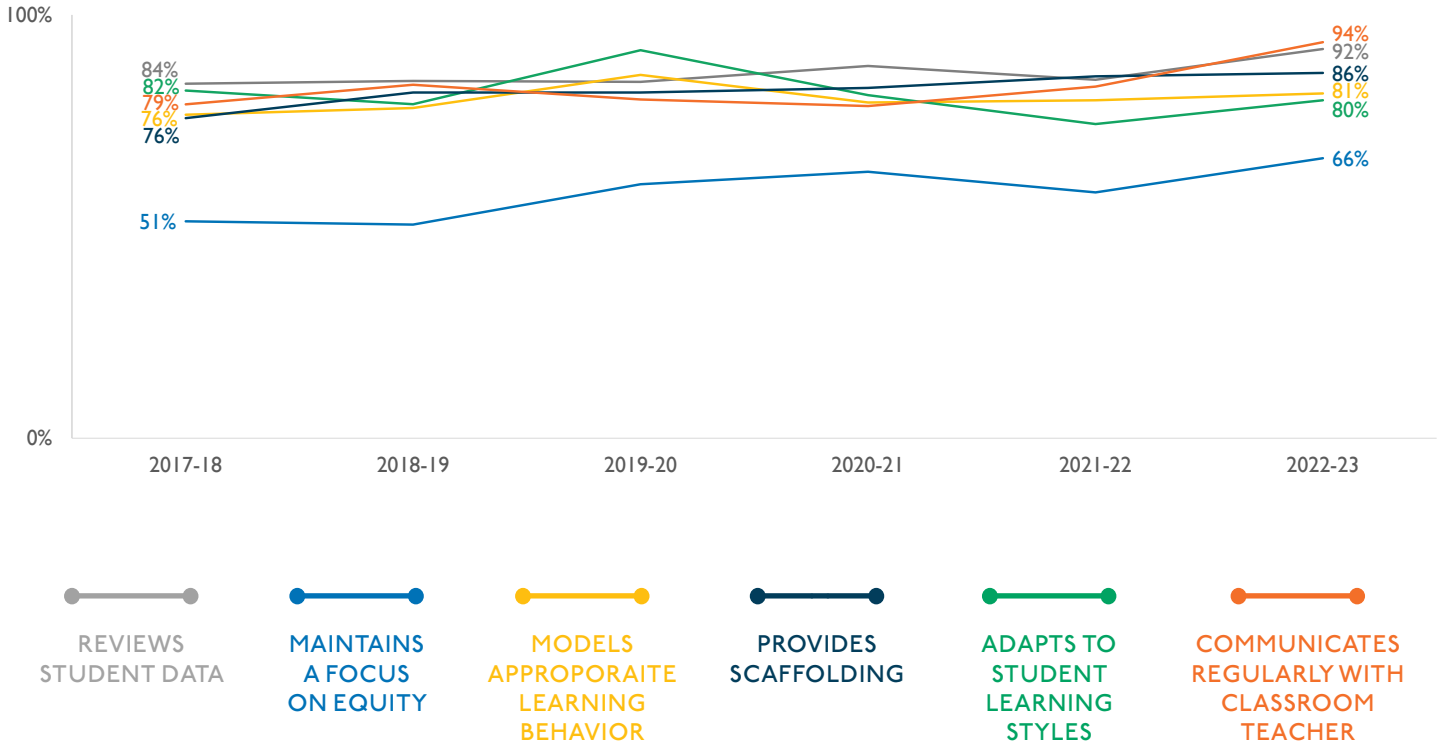
Figure 18 shows that, unlike schools' distribution of class size reduction, tutoring is equally common in all grades. Within grades, since 2017-18, schools that use tutoring have been increasingly likely to use tutoring in at least 75 percent of classrooms. In 2022-23, the percentage of schools using tutoring in at least 75 percent of their classrooms was around 70 percent.

Studies show that more frequent tutoring results in larger impacts on standardized exam scores.[29] Most schools implementing AGR tutoring do so frequently, as seen in Figure 19. Approximately 71 percent of schools provide tutoring at least two times per week, with over 50 percent providing tutoring three 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.

29 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 No. 20-267). Annenberg Institute at Brown University. https://doi.org/10.26300/eh0c-pc52

Figure 20 shows that schools reported frequent use of all of the tutoring practices listed on the EOY survey. While the proportions 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 2022-23, 66 percent of schools used tutoring for equity purposes, up from 51 percent in 2017-18.

Figure 20: School-Reported Tutoring Practices

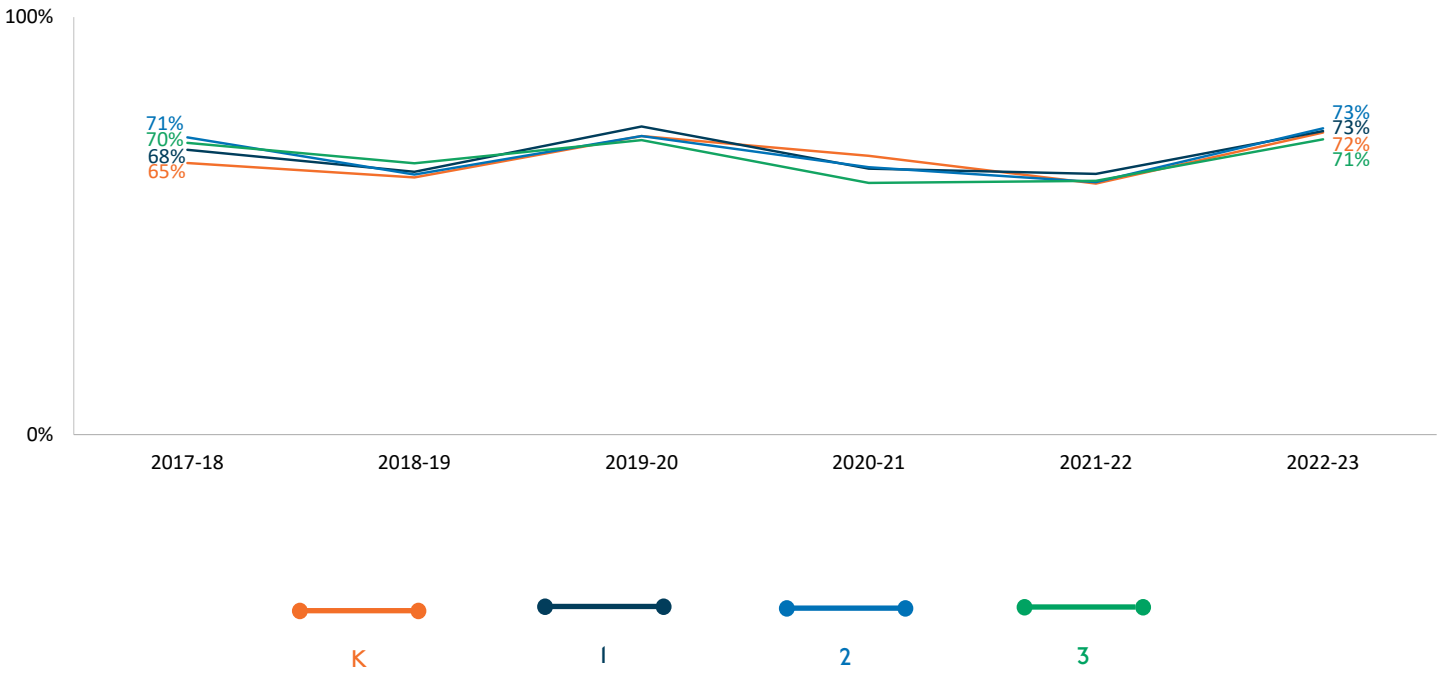


* Note: Not Pictured – Other, Not Sure/Don't Know

Figure 21 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. This is an increase from about 60 percent of classrooms in 2021-22, more in line with pre-pandemic levels in 2019-20.

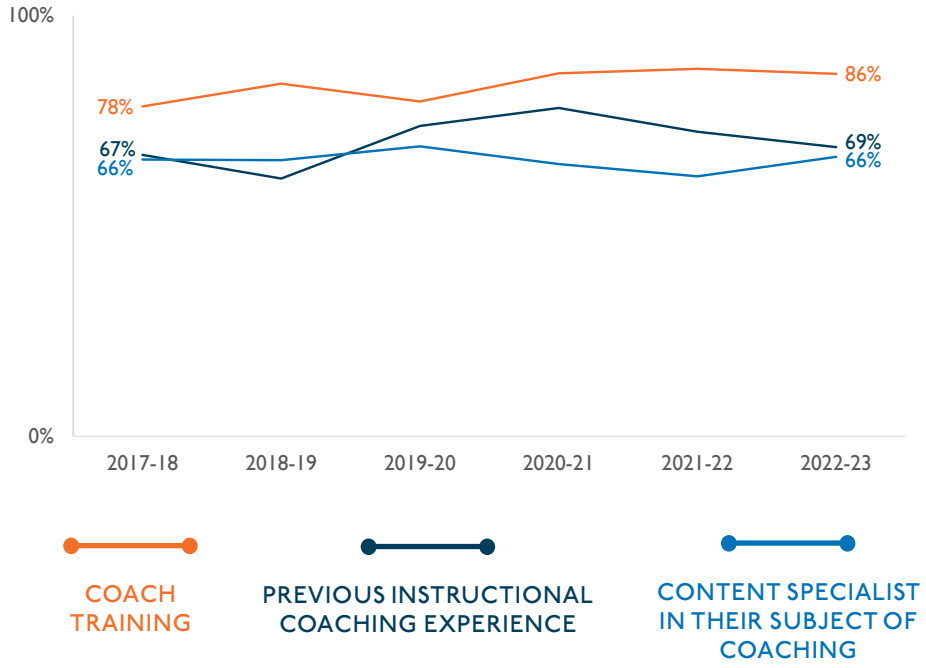
Figure 21: Schools Using Instructional Coaching Doing So in at Least 75% of Classroom

By Grade



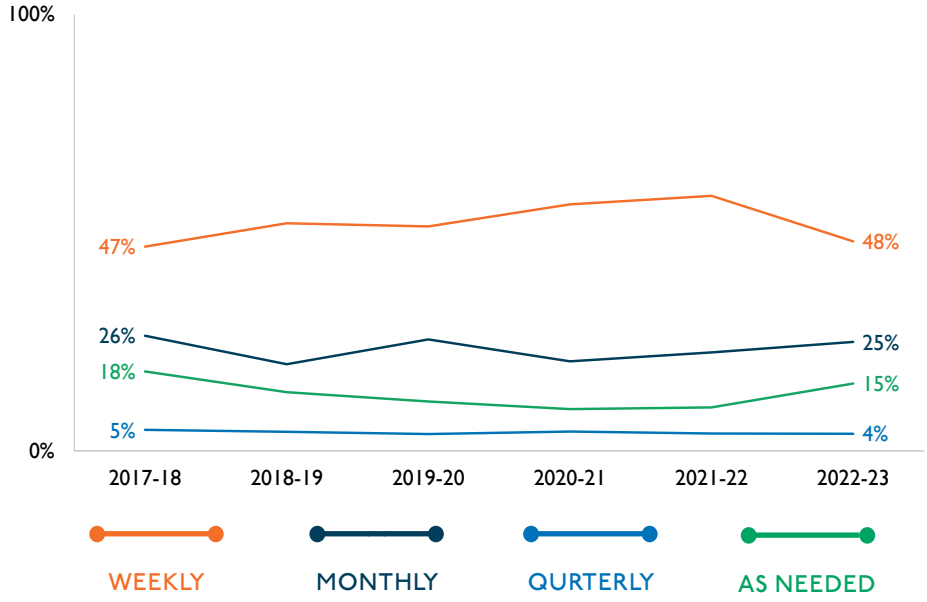
* Note: Not Pictured – Other, Not Sure/Don't Know

Figure 22: School-Reported Characteristics of Instructional Coaches



* Note: Not Pictured – Other, Not Sure/Don't Know

Figure 23: School-Reported Instructional Coaching Frequency



* Note: Not Pictured – Other, Not Sure/Don't Know, None

Schools responded affirmatively to almost all listed instructional coaching characteristics (Figure 22). Over the sample period, reported characteristics have been stable. From 2017-18 to 2021-22, schools reported increasing frequency of instructional coaching (Figure 23), although in 2022-23 that percentage dropped to 2017-18 levels. As-needed coaching increased by 5 percentage points in 2022-23.

Section 9

Summary and Conclusions

Summary and Conclusions

The socioeconomic achievement gap is wide and has remained relatively unchanged over the past fifty years.[30] Researchers and policymakers have hypothesized dozens of causes, including funding deficits for districts located in high-poverty areas.[31] 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. These results are presented at the state level and disaggregated for certain subgroups of students, including those who receive FRL. The report also contains data on the AGR strategies schools have implemented, the intensity of strategy use, and an examination of the implementation practices of AGR schools that have most effectively implemented AGR funding to improve test score outcomes.

This year's evaluation methodology takes advantage of the growing number of cohorts that have received AGR funds to investigate AGR's K-3 impacts on third grade, fourth grade, and fifth grade Forward reading and math. 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 through 2020, AGR impacts on third grade Forward reading and math are small and not statistically different from zero, for both the statewide sample and for all student subgroups, including students receiving FRL. Impacts for fourth and fifth grade Forward reading and math are similar. 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. The 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 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 85 percent of schools used reduced class sizes in at least one grade. Despite strong evidence in the research literature that tutoring can positively impact achievement score growth, only close to a third of AGR schools implemented tutoring, either alone or in combination with other strategies.

30 Hanushek, E.A., Light, J. D., Peterson, P. E., Talpey, L. M., & Woessman, L. (2022). Long-run Trends in the U.S. SES-Achievement Gap. *Education Finance & Policy, 17*(4), 608-640. https://doi.org/10.1162/edfp_a_00383

31 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), 304-335. https://www.aeaweb.org/articles?id=10.1257/pol.20180674

New to this year's evaluation was an examination of high-growth AGR schools and their AGR implementation practices. Principals interviewed overwhelmingly had positive views of AGR's impacts on math and reading growth and were thankful for the program's funding support. These schools implemented reduced class sizes often in combination with other AGR strategies and found that smaller classes allowed them to connect more one-on-one with students. The schools using instructional coaching had similar approaches, with coaches working frequently with teachers on setting and achieving goals. The schools using one-to-one tutoring all used assessment data to identify students who were in need of academic support and then provided tutoring between two and five days per week.

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. Inconsistent testing patterns in grades K-3, including diminishing use of PALS, a lack of testing in spring 2020, and reduced testing participation in spring 2021, restricted the sample of AGR and non-AGR schools included in the growth analysis samples. The evaluation also limits cohorts to schools that were mostly in-person during 2020-21. All of these sample restrictions potentially limit the extent to which growth impact estimates can be generalized to schools not in the sample.

Section 10

Appendix A: Technical Appendix

Appendix A: Technical Appendix

Evaluation Design

In order to credibly estimate AGR impacts, we must address two primary challenges to identification. First, because all AGR schools previously participated in SAGE, total AGR impacts cannot be determined solely through changes 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. Second, 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 the percentage of their students receiving FRL.

To find a plausible control group and identify AGR impacts separate from SAGE, we use a cohort model of treatment impacts and Propensity Score Matching (PSM). The cohort model differentiates between cohorts that received the full, 4-year AGR impact and those that received a mix of AGR and SAGE or 4 full years of SAGE.

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. 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. 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 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.
The primary limitation of PSM is that it relies 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. Past evaluations have included robustness checks to address these concerns. Because robustness checks of this sample are similar to those in previous evaluations, we omit them from this report.

In the remainder of this appendix, we describe the propensity score matching and analysis methodology and the multiple comparisons corrections of p-values.

Propensity Score Matching

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

Equation (I)

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

In the equation above, matching occurs at the school-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 5 above.

Specifying the Propensity Score Model

To determine which variables to include in the propensity score matching model, 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 the model, 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 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.

At the beginning of our sample period, however, SAGE had been in operation for over 15 years. As a result, it was not possible to include pre-program data. To address matching on post-program outcomes, we matched separately for each school-cohort, using pre-program kindergarten fall PALS measures along with other, school-level controls (see Table 1).

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.[32]

32 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 2018-2020 kindergarten cohorts whose schools did not host at least 75 percent of class days in person from September through April of the 2020-21 school year, 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 A-1 illustrates the matching and subsequent analysis strategies for each grade and 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.[33] 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 A-1: Matching and Analysis Strategies and Samples

OUTCOME	GRADE(S)	MATCHING LEVEL	MATCHING DATA	ANALYSIS COHORTS
Forward Reading & Math Growth	3	School-cohort	Fall 2013-2016, 2018-2020	Kindergarten in 2013 through 2016, 2018-2020
Forward Reading & Math Growth	4	School-cohort	Fall 2013-2015, 2018-2019	Kindergarten in 2013 through 2015, 2018-2019
Forward Reading & Math Growth	5	School-cohort	Fall 2013, 2014, 2016, 2018	Kindergarten in 2013, 2014, 2016, 2018

33 What Works Clearinghouse. (2022). *Procedures and Standards Handbook, Version 5.0.* https://files.eric.ed.gov/fulltext/ED621928.pdf

Impact Analysis

After matching, we model impact estimates separately for each outcome. All models include weights generated by the kernel PSM procedure. Standard errors are clustered at the school level. Models for Forward use Weighted Least Squares.

Forward Reading & Math

Equation (2)

$$Forward_{i,g3,s,c} = \propto +\lambda PALS_{i,gK} + \beta_1 FullAGR_i + \beta_2 FullSAGE_i + \beta_3 AGR/SAGEMix_i + \gamma X_i + \theta Z_s + \delta_c + \varepsilon_{i,g3,s,c}$$

where Forwardi,g3,s,c is an outcome for student i in grade 3 (or 4), school s, and cohort c. PALSi,gK is student i's fall kindergarten PALS score, which acts as a pretest for both reading and math. FullAGRi, FullSAGEi, and AGR/SAGEMixi are indicators of treatment type. Xi represents a vector of student-level covariates, and Zs represents a vector of school-level covariates measured during the student's kindergarten year. Cohort fixed effects, c, are included to control for any unobserved, statewide effects that vary by time. All analysis variables are described in Table 2 in the main report. As described above, the models include all school-level variables from the PSM procedure.

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.[34] 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.

34 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

Section II

Appendix B: Implementation Interview Protocol

Appendix B: Implementation Interview Protocol

Thank you for agreeing to talk to me today. I am part of a group working with DPI to study the AGR program in schools in Wisconsin. We would like to understand how schools implement AGR and what impact AGR has had on K-3 students.

Your participation in this interview is voluntary. Nothing you say will be connected to your name or any identifiable information in reports, including the name of your school. With your consent, I would like to record the interview to help us accurately collect what you say. The recording will be destroyed after we write up the summary report. We expect this interview to last about 45 minutes to 1 hour. Are there any questions?

Introduction

To get some context, I have a few questions about you and your school.

1. Tell me a little about yourself. What is your background? a. How long have you been a principal at this school?
2. How would you describe your school community?

AGR Program Overall Effectiveness

I'd like to hear what you think about AGR in your school.

3. In your opinion, is the AGR program impactful for your school, particularly regarding impacts on math and reading growth?
    a. If yes, why do you think AGR is impactful for your school?
    b. Do you think that the AGR program is reducing achievement gaps within your school? Why or why not?

AGR Implementation and Funding

(For these questions, the interviewer will have data on the school's previous replies to the End-of-Year and School Board Reports.)

4. To implement the AGR program, how does your school use the money provided by AGR funds?
    a. Which AGR strategies do you use in each grade (class-size reduction/instructional coaching/1-1 tutoring)?
    b. Within each grade, do these strategies differ by classroom or subject (math/reading)?
        i. If yes, how?

5. Of the AGR strategies your school uses, which of these would you use even if your school received no AGR funding?
    a. (If there are strategies the school would use regardless of AGR funding) Does AGR funding free up funding for other initiatives and projects?
        i. If yes, please describe these other initiatives and projects that AGR funding enables.
6. Does the extra funding you receive through the AGR program cover the expenses of participating in AGR?
    a. If not, how else are you covering the expenses?

AGR Strategies

Next, I'm going to ask you about the implementation of each of your school's AGR strategies.

If the school uses class-size reduction:

7. What is the typical class-size in grades you are implementing class-size reduction? What is the typical class-size in the other grades?
8. Does your school use any specific instructional strategies to take advantage of reduced-size classrooms? (Can list examples – small group instruction, increased one-on-one time with the teacher, differentiation of instruction, strategic placement of students in groups or classrooms, something else)
    a. For each instructional strategy named:
        i. Please tell me more about how your school implements [instructional strategy] in reduced-size classrooms?
        ii. Does your school also use [instructional strategy] in larger-sized classrooms?
9. Does your school offer any professional development to help teachers take advantage of reduced-sized classrooms?

If the school uses instructional coaching:

10. How frequently do AGR instructional coaches meet with teachers?
    a. Do any teachers receive more/less instructional coaching (e.g. less experienced teachers, or struggling teachers)? If so [and not already answered], how are those teachers identified?

11. Which practices do instructional coaches use? (Can list examples – one-to-one coaching, team teacher coaching, coaching logs, goal setting, coaching focuses on teacher goals, maintaining a focus on equity, encouraging reflective practices, discussing data, observing teacher practice, something else)
    a. For each coaching practice named:
        i. Please tell me more about how your school implements [coaching practice]?
        ii. Does your school also use [coaching practice] for non K-3 teachers?
    b. [If not discussed] How do your instructional coaches make use of data in their practices?
12. Does your school evaluate the effectiveness of your instructional coaching?
    a. If so, how?

If the school uses I-I tutoring:

13. Are you using a specific instructional program for your tutoring?
    a. If so, which one(s) and how were they selected?
    b. If not, please describe what a typical tutoring session looks like in reading and math?
14. How are students selected for tutoring?
15. Approximately how many students in each grade receive AGR tutoring (grades asked will depend on the answer to Question 5)?
    a. Kindergarten?
    b. First grade?
    c. Second grade?
    d. Third grade?
16. How often do AGR tutors meet with students?
    a. Does the frequency of AGR coaching differ by student need?

If the school uses multiple strategies:

17. Does your school try to integrate your strategies in any way? For example, a school using both class-size reduction and instructional coaching might coach teachers on how to take advantage of smaller class sizes.

K-3 Additional Programs and Practices

18. Are there other programs or practices that your school is using in grades K-3 to make your students successful, particularly in regard to math and reading growth or closing the achievement gap?

Those are all of the questions I have for you today. Is there anything else you would like to add about your school's implementation of AGR program that you think would be useful to know? Thank you for your time today.

Wisconsin Evaluation Collaborative