



Wisconsin
Evaluation
Collaborative

September 2023

Evaluation of the Achievement Gap Reduction Program

2015-16 through 2021-22

for the Wisconsin Department of Public Instruction



About the Authors



Jed Richardson

Jed Richardson is an Assistant Scientist with the Wisconsin Evaluation Collaborative, where he is the Principal Investigator on multiple evaluations of state policies and district academic programs. He holds a Ph.D. in Economics from the University of California, Davis.

Grant Sim

Grant Sim is a Scientist at the Wisconsin Evaluation Collaborative, specializing in mixed methods program evaluation of educational initiatives at both the state and district levels. He holds a master's degree in public affairs from the University of Wisconsin–Madison.

McKenna Goetz

McKenna Goetz is a Research Intern with WEC. She is currently pursuing a Master of Public Affairs degree at the LaFollette School of Public Affairs, and her interests focus on education policy.



About the Wisconsin Evaluation Collaborative

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



Contact

Jed Richardson

jed.richardson@wisc.edu

Acknowledgements

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

Copyright Information

©2023 the Wisconsin Evaluation Collaborative

Contents

- 5 Executive Summary**
- 9 Introduction**
- 15 Evaluation Data and Methodology**
- 22 AGR Demographics**
- 31 AGR Implementation**
- 33 AGR Impacts**
- 46 Longitudinal Analyses of End-of-Year and School Board Reports**
- 55 Future AGR Evaluation**
- 57 Summary and Conclusions**
- 60 Technical Appendix**

Section I

Executive Summary

Executive Summary

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

- Provide professional development related to small group instruction and reduce class size to one of the following:
 - No more than 18
 - No more than 30 in a combined classroom having at least 2 regular classroom teachers
- Provide data-driven instructional coaching for one or more teachers of one or more participating grades. The instruction shall be provided by licensed teachers who possess appropriate content knowledge to assist classroom teachers in improving instruction in math or reading and possess expertise in reducing the achievement gap.
- Provide data-informed, one-to-one tutoring to pupils in the class who are struggling with reading or mathematics or both subjects.

Tutoring shall be provided during regular school hours by a licensed teacher using an instructional program to be found effective by the What Works Clearinghouse of the Institute of Education Sciences.²

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

1. How are AGR schools implementing the AGR program as specified by 2015 Wisconsin Acts 53 and 71?
 - a. What is the breakdown of strategy usage across the state?
 - b. How does implementation of the three strategies differ across schools?
 - c. Did schools change their use of the three strategies during or after the COVID-19 pandemic?
2. To what extent is AGR meeting intended outcomes, including impacts on standardized test scores, attendance, and disciplinary events?
 - a. How does AGR impact vary by student characteristics?
 - b. How does AGR's impact on outcomes compare to impacts associated with the SAGE program?
3. Are there differences between the three AGR strategies' relationships to intended outcomes?

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, attendance, and discipline outcomes, and to compare AGR's impacts to those of its predecessor, SAGE. The first part of the analysis uses propensity score matching to identify non-AGR Wisconsin schools that are similar to those receiving AGR funding. These observationally similar schools function as a comparison group for the second step of the analysis, estimating the impact of AGR through multivariate regression techniques.

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 2021-22, the most recent year of data, 404 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. In 2021-22, 64 percent of schools utilized multiple strategies, perhaps in response to the COVID-19 pandemic.
- 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 large impacts of K-3 SAGE on eventual high school persistence and completion.³ Results from the current analysis included:

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

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

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

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

Section 2

Introduction

Introduction

The Achievement Gap Reduction (AGR) program is an initiative of the Wisconsin Department of Public Instruction (DPI) that provides funding to improve the academic performance of students in schools with high concentrations of low-income or economically disadvantaged students. AGR functions as a revision and continuation of the Student Achievement Guarantee in Education (SAGE) program, which the Wisconsin legislature and DPI initiated in 1995 to address the need for additional resources for economically disadvantaged students, particularly in urban areas. 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 2 regular classroom teachers.
2. Provide data-driven instructional coaching for one or more teachers of one or more participating grades. The instruction shall be provided by licensed teachers who possess appropriate content knowledge to assist classroom teachers in improving instruction in math or reading and possess expertise in reducing the achievement gap.
3. Provide data-informed, one-to-one tutoring to pupils in the class who are struggling with reading or mathematics or both subjects. Tutoring shall be provided during regular school hours by a licensed teacher using an instructional program to be found effective by the What Works Clearinghouse of the Institute of Education Sciences.⁴

4 2015 Wisconsin Act 53, Section 118.44 (2015). <https://docs.legis.wisconsin.gov/2015/related/acts/53.pdf>

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 any progress reducing gaps for economically disadvantaged 4th graders on NAEP math and reading, respectively.

Researchers and policymakers have hypothesized dozens of causes for the socioeconomic achievement gap. These causes include neighborhood factors such as exposure to entrenched poverty and violent crime,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

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

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.

Figure 1: Socioeconomic Achievement Gaps: NAEP Reading

Grade 4, 2003 - 2022

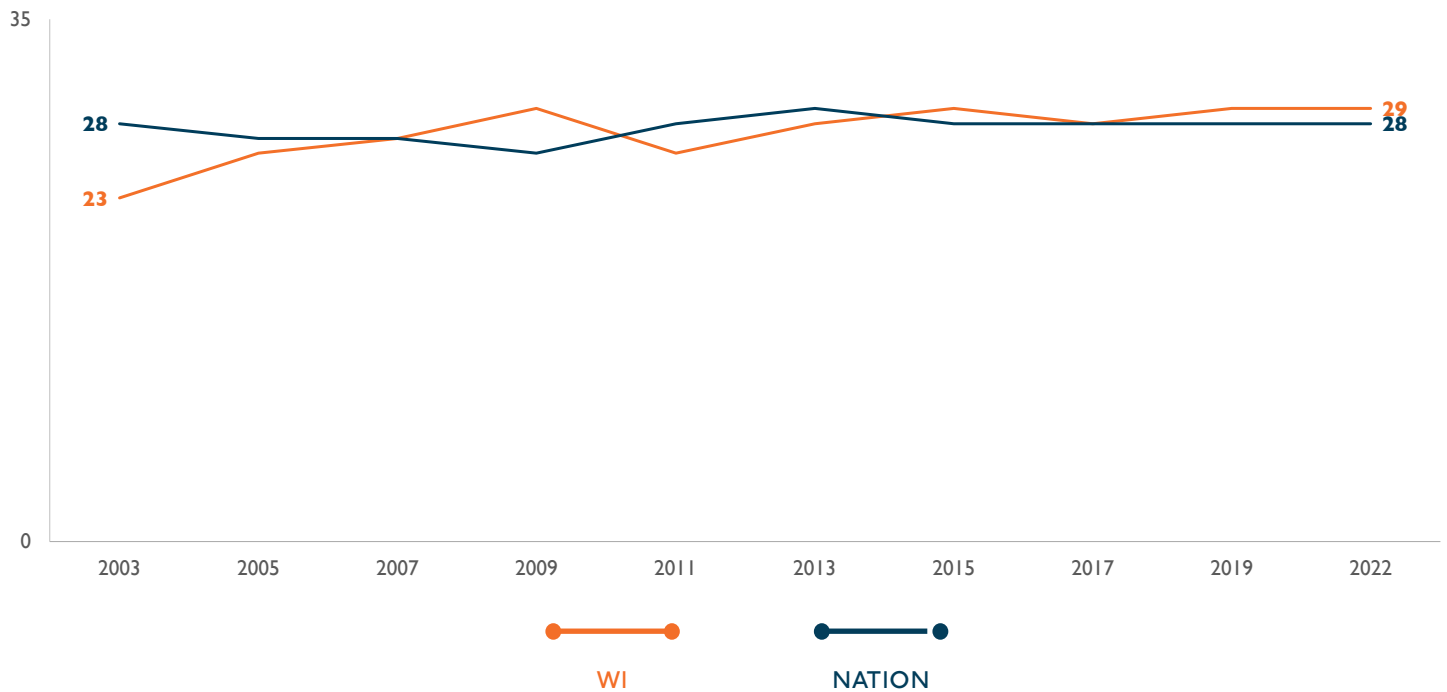
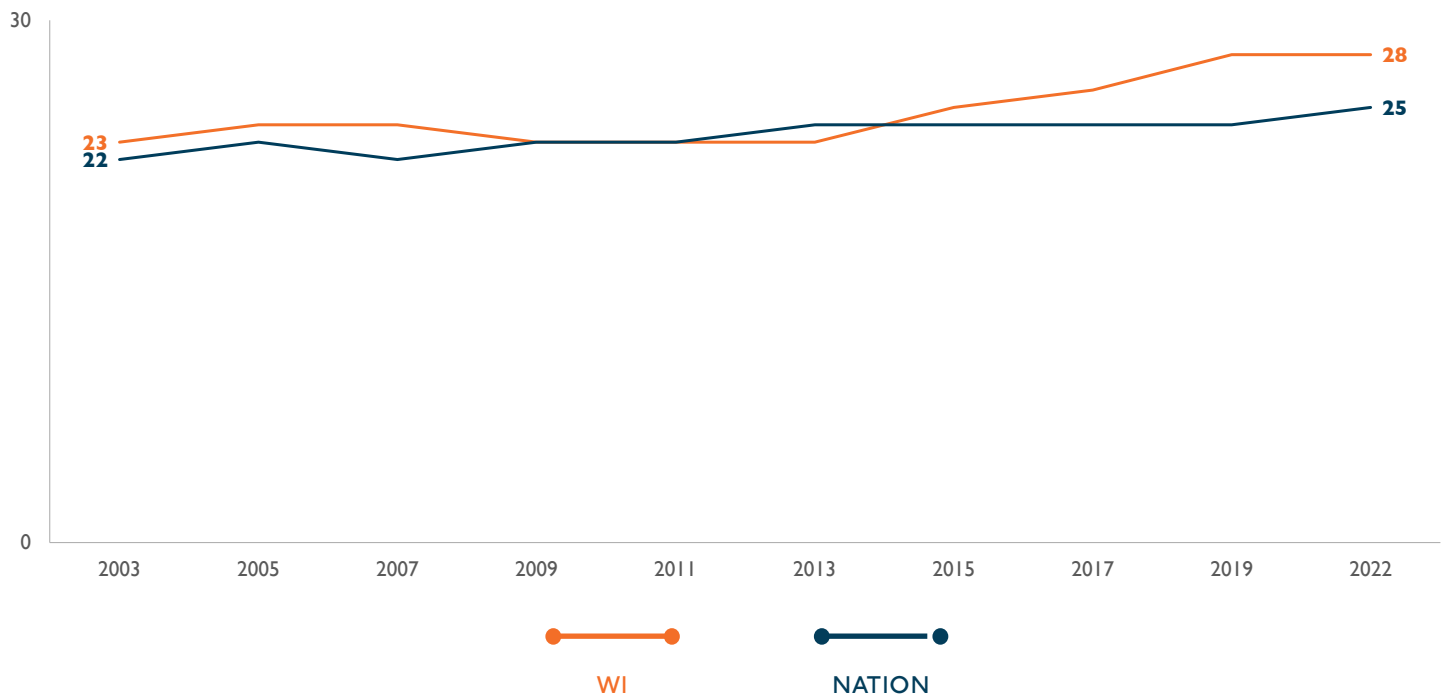


Figure 2: Socioeconomic Achievement Gaps: NAEP Math

Grade 4, 2003-2022



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 recent 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 2021-22.

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?
 - c. Did schools change their use of the three strategies during or after the COVID-19 pandemic?
2. To what extent is AGR meeting intended outcomes, including impacts on standardized test scores, attendance, and disciplinary events?
 - a. How does AGR impact vary by student characteristics?
 - b. How does AGR's impact on outcomes compare to impacts associated with the SAGE program?
3. Are there differences between the three AGR strategies' relationships to intended outcomes?

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 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. In addition, there is evidence that in-person learning was more effective for academic achievement growth during the pandemic.¹³ The correlation between schools' learning models and participation in AGR can create a negative statistical bias that would result in underestimation of AGR's effects on achievement. Therefore, to best estimate AGR's impacts, the evaluation must address differences in learning models. This is not to second-guess districts' decisions during the pandemic. Districts chose learning models within their unique local contexts, taking into account the prevalence of COVID-19, transportation infrastructure, internet availability, and many other factors, not just to maximize learning, but also to protect student and public health. As a result, addressing differences in learning models as part of this evaluation does not imply that some districts' choices were better than others, only that learning may have systematically differed across AGR and non-AGR schools for reasons unrelated to AGR itself.

This report has eight main sections including the Introduction. Evaluation Data and Methodology includes details on data, analysis designs, and statistical models used to evaluate program impacts, as well as the limitations of this evaluation. AGR Demographics describes characteristics of AGR students and schools and the resulting analysis samples. AGR Implementation describes the strategies schools choose. AGR Impacts provides the results of analyses of AGR impacts on Forward reading and math growth, as well as impacts on absences and out-of-school suspensions (OSS). This section also provides overall impacts, impacts of AGR compared to SAGE, and impacts for various subgroups of students, including low-income students. Longitudinal Analyses of End-of-Year (EOY) and School Board Reports looks at these data from 2017-18 through 2021-22. The last two sections provide Summary and Conclusions and a Technical Appendix.

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/system/files/working_papers/w29497/w29497.pdf

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 2021-22 academic years. In addition, for purposes of comparison, the evaluation includes K-3 students at subsets of non-AGR, non-SAGE schools.

Data

To identify plausibly equivalent, non-AGR schools for a comparison group, and to estimate impacts, the evaluation combines several sources of student- and school-level data for the academic years 2012-13 through 2021-22. 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).¹⁴ Finally, data on district-level median income and poverty levels are also sourced from NCES, which uses data from the U.S. Census Bureau's American Community Survey.¹⁵

- 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 and fourth grade spring Forward test scores and fall and spring kindergarten reading scores from the Phonological Awareness Literacy Screening (PALS). Achievement test data also includes fall and spring administrations of the MAP and STAR. For Grades 1-3, MAP and STAR scores were equated and combined into a single test measure in order to attain a sufficient student sample.¹⁶
- Attendance data consist of total days absent and total attendance days. The associated outcome variable is the absence rate, the total days absent divided by total possible attendance days.

¹⁴ Public Staff Reports are available at <https://publicstaffreports.dpi.wi.gov/PubStaffReport/Public/PublicReport>. School location information available at <https://nces.ed.gov/ccd/elsi/>.

¹⁵ District-level American Community Survey results available at <https://nces.ed.gov/programs/edge/demographic/acs>

¹⁶ In the Spring of 2020, nearly all Wisconsin schools opted to forgo assessments due to the COVID-19 pandemic. To address the lack of these scores, we estimated a model that uses scores from Fall 2019 and Winter 2020 along with student and school characteristics to predict Spring 2020 scores as if the school year proceeded normally. Further information may be found in the Technical Appendix.

- Discipline data consists of the number of OSS. The associated outcome variable is an indicator that is one for students with at least one OSS during the school year and zero for those who were not suspended. We use this outcome as a proxy for student behavior.
- Enrollment data include school attended and grade.
- School-level data include SAGE and AGR funding by year, average teacher compensation, school location (city, suburb, town, rural), charter school indicators, and school learning model (in-person, virtual, mixed) during 2020-21.
- District-level data include median income and poverty levels.

Identifying Comparison Schools

Using the data described above, we aggregate each school's K-3 data to find a comparison group of non-AGR schools. For each of the outcomes, we explored multiple variations of PSM in order to, (1) achieve the best match between AGR and comparison schools, (2) retain as many AGR observations as possible, and (3) ensure that there are sufficient control schools matched to each AGR school. To do so, we tested combinations of demographic and academic variables and several matching algorithms. As a result of this testing process, we selected a kernel matching procedure. Kernels place higher weights on untreated observations nearest to a treatment observation and assign successively lower weights to untreated observations as their distance from a treatment observation increases.

Matching followed two strategies, depending on the outcomes. For third and fourth grade Forward math and

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

During matching, we selected the sample to address testing gaps and differences in learning models due to the COVID-19 pandemic. Due to a lack of Forward testing during the 2019-20 school year, the analysis omits the cohort of students in third grade and fourth grade during that year (the 2017 and 2016 kindergarten cohorts, respectively). 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 and April, when Forward testing occurs. We made this decision for two reasons. First, the primary policy question that this evaluation seeks to inform is whether AGR is effective under conditions that we expect to see in the future. Statewide, virtual schooling was temporary and is unlikely to return, thereby diminishing the value of knowing whether AGR is effective in virtual school environments. Second, AGR schools were more likely to choose virtual learning. Tests of several samples and analysis specifications showed that, relative to previous years, schools that were primarily virtual in 2020-21 experienced larger decreases in Forward scores relative to schools that spent more time in-person. This finding is consistent with national evidence of a negative relationship between achievement and hybrid and virtual learning.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>; 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/system/files/working_papers/w29497/w29497.pdf

¹⁸ For this year's evaluation, we tested models of third and fourth grade Forward scores that include 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. We present the results from these models in Appendix Tables A-2 through A-5.

Table 1: Propensity Score Matching Controls

Third and Fourth Grade Forward Reading and Math

MATCHING VARIABLE (MEASURED FALL OF KINDERGARTEN)

School Average Fall Kindergarten PALS

School % Black, Hispanic, White, Other Race/Ethnicity*

School % Free/Reduced-price Lunch

School % Mobile from Prior Year

Locale Description (City, Suburb, Town, Rural)*

District Median Income

School K-3 Student Population

School Number of Cohort Students in Analysis Sample

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

Table 2: Propensity Score Matching Controls

Attendance and Discipline

MATCHING VARIABLE (MEASURED 2012-13)

ATTENDANCE

DISCIPLINE

School % Black, Hispanic, White, Other Race/Ethnicity*

✓

✓

School % Free/Reduced-price Lunch

✓

✓

Locale Description (City, Suburb, Town, Rural)*

✓

✓

Charter School Indicator

✓

✓

District Median Income

✓

✓

School K-3 Student Population

✓

✓

2012-13 School Attendance Rate

✓

2012-13 School Suspension Rate

✓

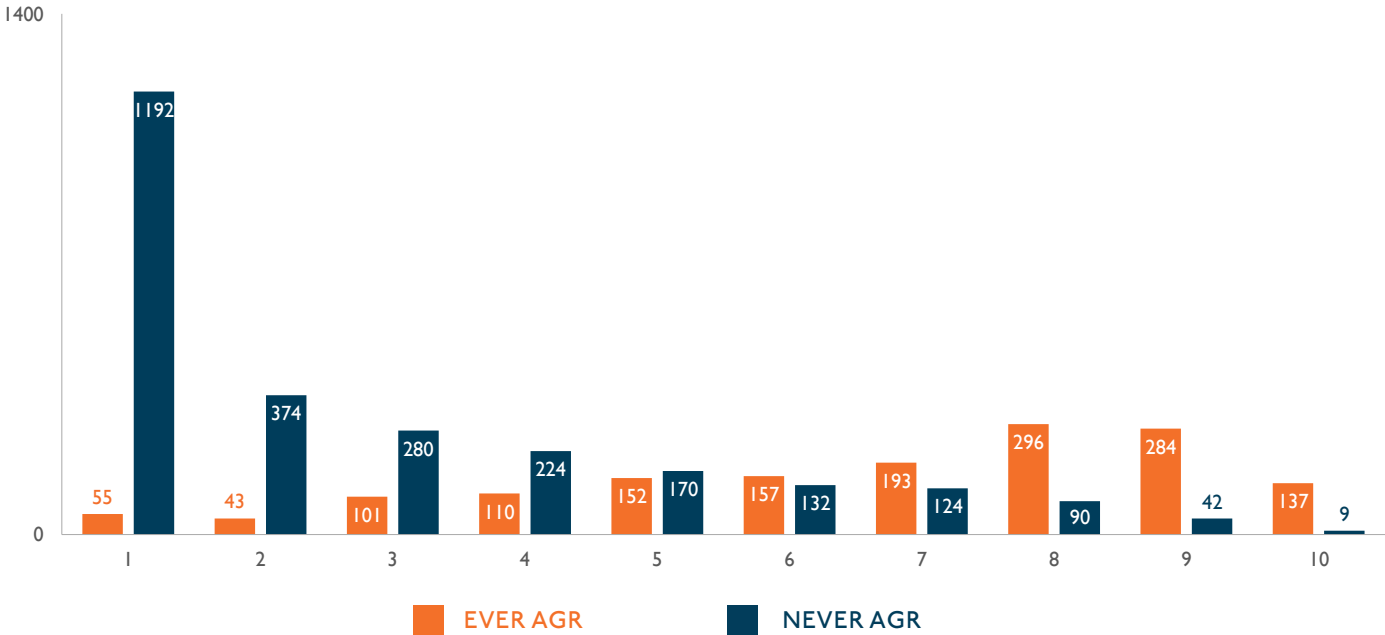
* Due to collinearity, we omit one Race/Ethnicity category and one Locale Description category from the model.

Matching for attendance and discipline outcomes follows a different strategy, matching at the school level based on K-3 demographic characteristics in 2012-13, the first year in our data. The matching controls appear in Table 2.¹⁹ To improve the quality of school matches, we include district-level median income. Absence and discipline models omit 2020-21 data; due to the COVID-19 pandemic, absence and discipline fluctuated substantially relative to previous years.

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

¹⁹ In the 2022 evaluation, we reported results from two separate models, with and without controls for previous school attendance and discipline rates, to test for bias resulting from previous SAGE impacts on these outcomes. Results from these models were qualitatively similar. As a result, this year's evaluation only reports results from models that match on 2012-13 attendance and out-of-school suspension rates.

Figure 3: Common Support for Matching
Third Grade Forward Reading



Note: Matching occurs at the school level within cohort. Figure 3 includes all cohorts, and as a result, there are multiple observations for each AGR school.

Analysis

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

Table 3: Analysis Model Controls

CONTROL VARIABLE	GROWTH	ATTENDANCE	DISCIPLINE
Student Demographics Race/Ethnicity*, 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	✓		
Average Teacher Compensation		✓	✓
District Median Income	✓	✓	✓
District % Under the Poverty Line		✓	✓
School K-3 Student Population	✓	✓	✓
School Number of Cohort Students in Analysis Sample	✓		
2012-13 School Attendance Rate		✓	
2012-13 School Suspension Rate			✓
Cohort Indicators**	✓		
Grade-by-year 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.

*** Indicators for each grade-year combination equal one if a student's grade and year in school match the indicator variable, zero otherwise.

Limitations

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

The primary limitation stems from PSM's assumption that schools matched on observable characteristics, such as test scores and demographics, are also matched on unobserved characteristics, such as schools' ability to properly implement AGR strategies or instructor quality in the local hiring market. If unobserved characteristics are not balanced between AGR and comparison schools and are related to both outcomes and AGR participation, estimates of AGR impacts will be biased. In particular, if AGR schools are systematically more (less) effective than schools in the matched comparison group, impact estimates will be biased upward (downward).

The second limitation occurs due to the phase out of PALS testing. Through 2015-16, Wisconsin mandated PALS for kindergarten students. When that mandate ended prior to the 2016-17 school year, PALS usage dropped 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.

Finally, readers should note that the COVID-19 pandemic's impact on student attendance, learning models, and testing may impact results from 2019-20 and 2020-21. For example, the 2020-21 and 2021-22 samples of AGR students and schools included in Forward analyses is 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.

Section 4

AGR Demographics

AGR Demographics

We begin with information on the characteristics of AGR students and schools. Table 4 shows the number of AGR schools for each of the six years of the program. The first AGR cohort started in 2015-16 with 96 schools, followed by the second cohort in 2016-17, which brought the total to 408 schools. After a small increase in subsequent years, the overall number of AGR schools dropped slightly. In 2021-22, there were 404 AGR schools.

The numbers of students in AGR schools from 2015-16 to 2021-22, overall and by grade, are presented in Table 5. The first cohort of AGR schools included approximately 18,000 students, while the addition of the second cohort in 2016-17 brought the total to over 77,000 students. Since, student participation has declined, decreasing to 70,754 in 2021-22.

Table 4: Number of AGR Schools

By Grade and Year

GRADE	2016	2017	2018	2019	2020	2021	2022
0	88	393	391	395	396	391	386
1	91	398	397	403	400	395	391
2	91	398	398	403	400	395	391
3	88	391	391	395	394	389	384
Overall (K-3)	96	408	408	413	412	408	404

Table 5: Number of AGR Students

By Grade and Year

GRADE	2016	2017	2018	2019	2020	2021	2022
0	4,138	18,382	18,262	18,406	18,430	17,659	17,739
1	4,570	19,294	18,813	18,786	18,539	17,950	17,749
2	4,682	20,053	19,158	18,959	18,594	18,002	17,826
3	4,544	19,506	19,227	18,532	18,093	17,493	17,440
Overall (K-3)	17,934	77,235	75,460	74,683	73,656	71,104	70,754

Figure 4 and Figure 5 compare the demographic characteristics of AGR students to all K-3 Wisconsin students (including AGR students) in 2021-22. 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
2021-22

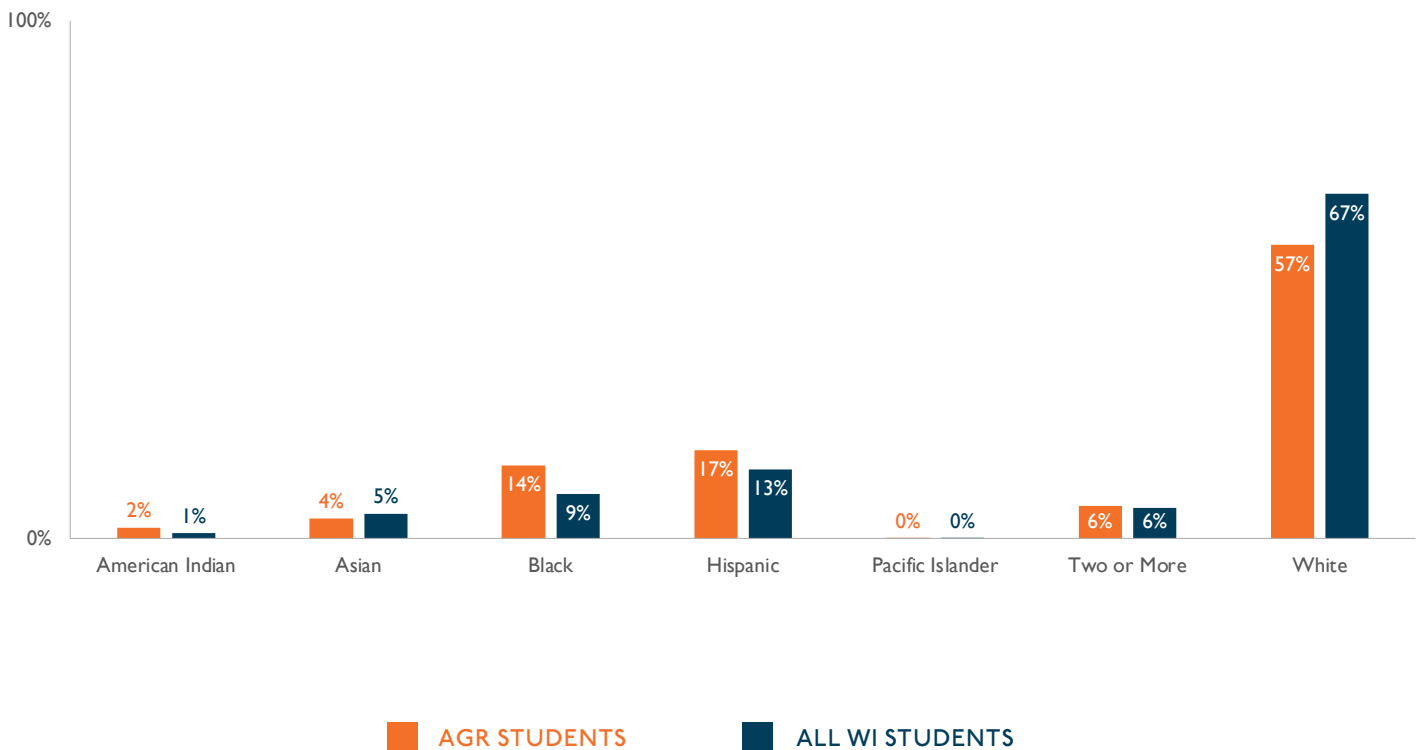


Figure 5: Percentage of AGR and WI Students who were English Learners, Eligible for FRL, and in Special Education

2021-22

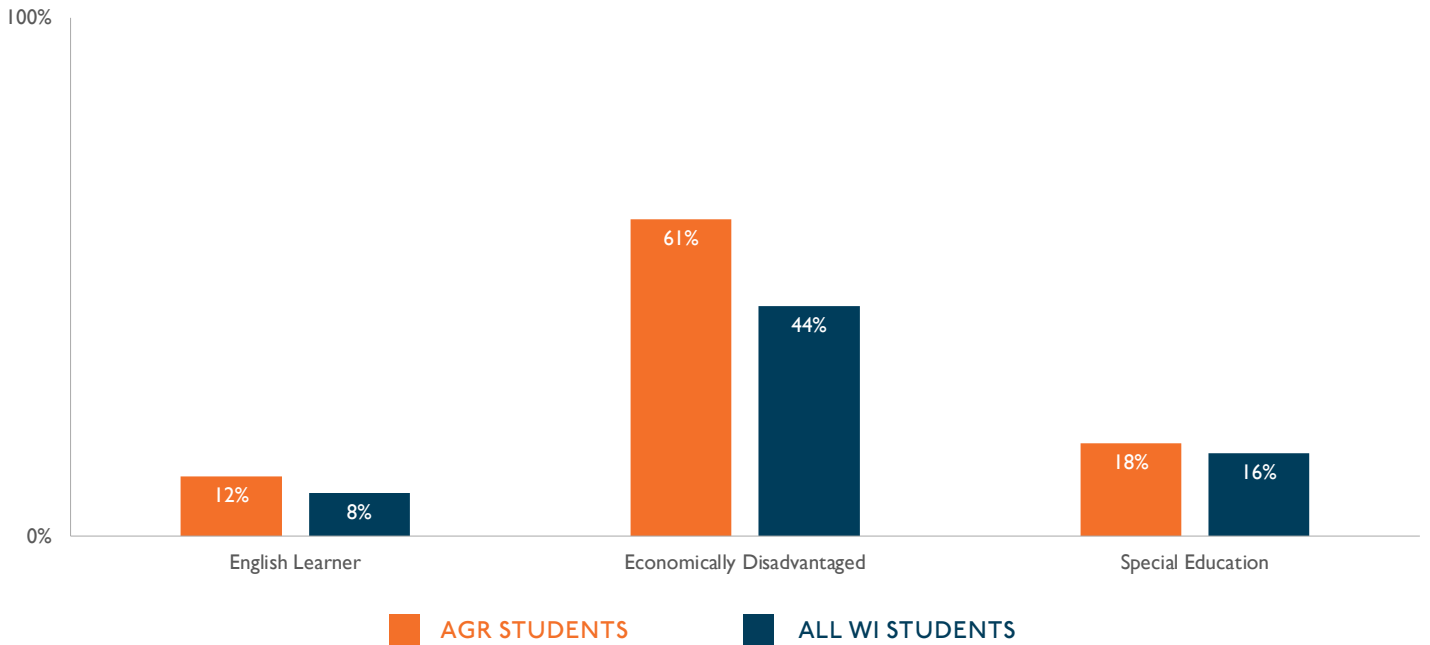
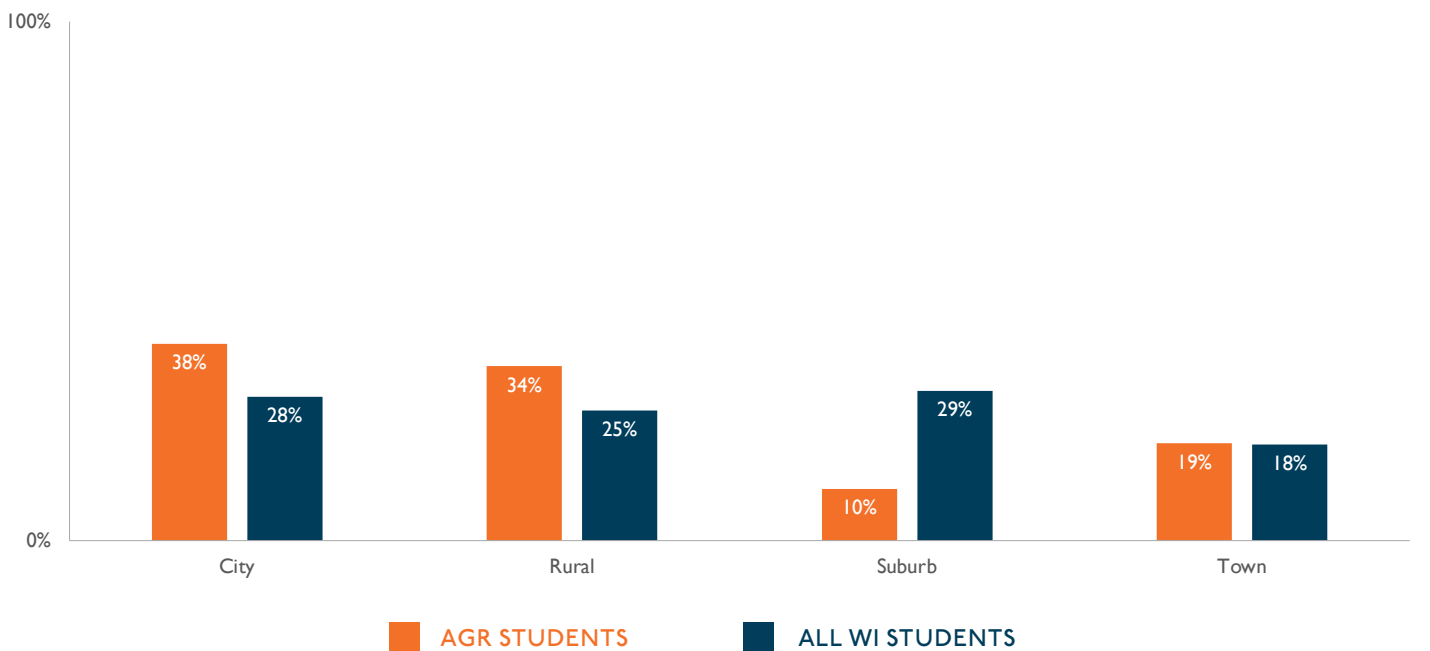


Figure 6: Locale Description of AGR and WI Students

2021-22



Third Grade Forward Analysis Sample

Examining third grade Forward requires several population restrictions to create a sample of students for analysis. First, we include only students with third grade Forward results as well as students with information on each of the controls used in the analysis (Table 3). As noted above, this results in dropping the cohort of students starting kindergarten in 2016-17, who would have taken the Forward exam in 2019-20 had it not been cancelled due to the COVID-19 pandemic. Since PALS was the state-mandated reading assessment for kindergarten from 2012-13 through 2015-16, nearly all students in the first four cohorts have Fall kindergarten PALS information, but after 2015-16, participation in PALS dropped to less than 50 percent of kindergarteners 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 cohort of students starting kindergarten in 2017-18 and attending third grade in 2020-21, we omit any schools that were not in person at least 75 percent of the days between September and April, as noted above. The resulting sample of analysis students compared to the total number of AGR students is presented in Table 6.

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

	2013	2014	2015	2016	2017	2018	2019	ALL YEARS
All AGR Students	22,215	21,304	20,750	19,646	18,994	18,818	18,543	140,270
Analysis AGR Students	12,622	12,711	12,271	11,751	0	2,413	1,625	53,393
Percentage in Analysis	56.8%	59.7%	59.1%	59.8%	0.0%	12.8%	8.8%	38.1%

Because not all AGR students are included in the Forward analysis sample, the results may not apply to all AGR students. 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 2018-19 kindergarten cohort the differences are larger. Analysis students in the 2018-19 cohort are less likely to be Black, Hispanic, English learners, and eligible for FRL, and more likely to be White (Figure 10 and Figure 11). The schools included in the Forward analysis in the 2018-19 cohort are more likely rural and far less likely urban (Figure 12).²⁰The 2017-18 cohort is similar to the 2018-19 cohort – for more information see the [2020-21 AGR evaluation report](#). Based on the data in Figures 10-12 and in last year's report, results from the 2017-18 and 2018-19 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 and 2018-19 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.

²⁰ The lack of students in the city locale in the 2018-19 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

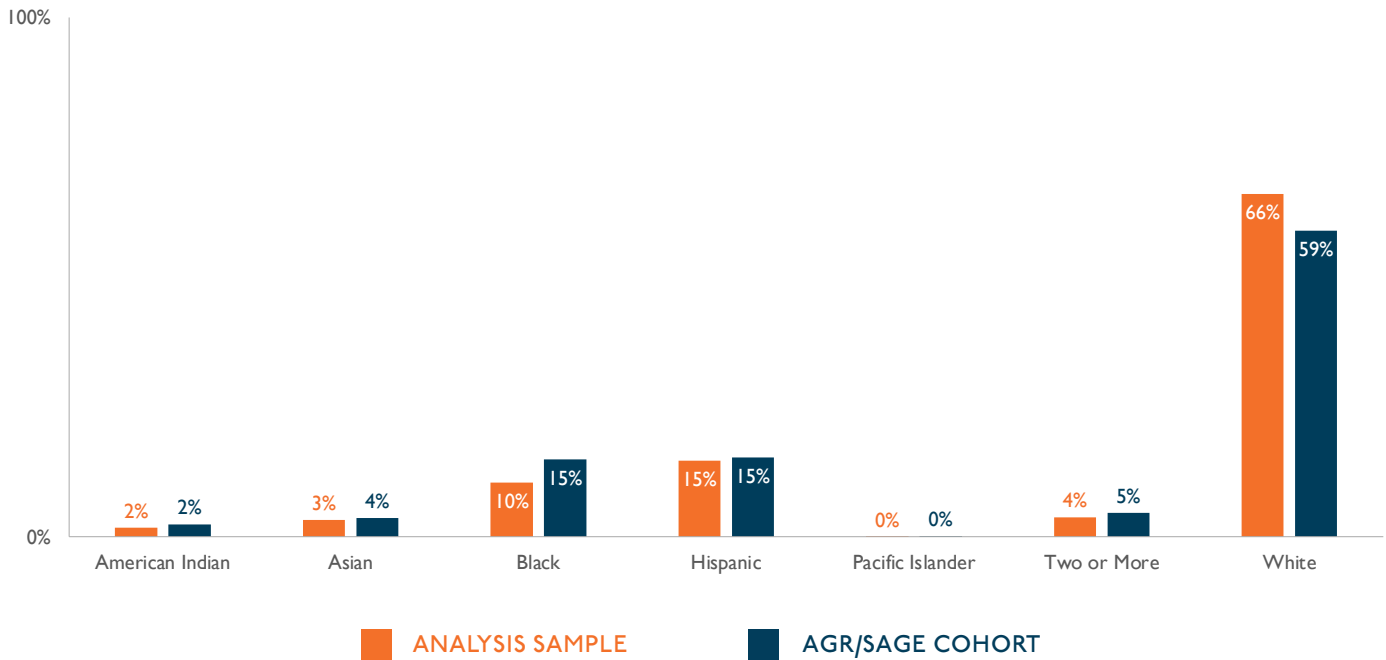


Figure 8: AGR Forward Analysis Students and AGR Students Overall who were English Learners, Eligible for FRL, and in Special Education

All Cohorts

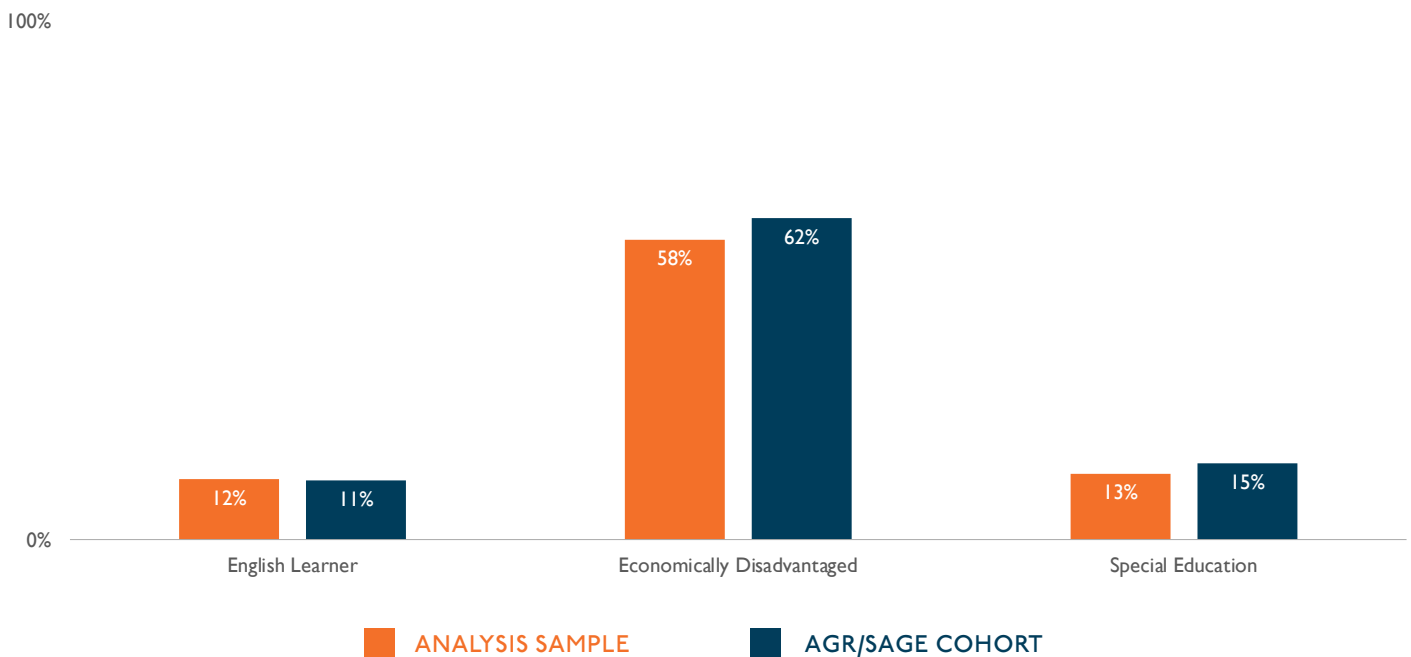


Figure 9: Locale Description of AGR Forward Analysis Students and AGR Students Overall

All Cohorts

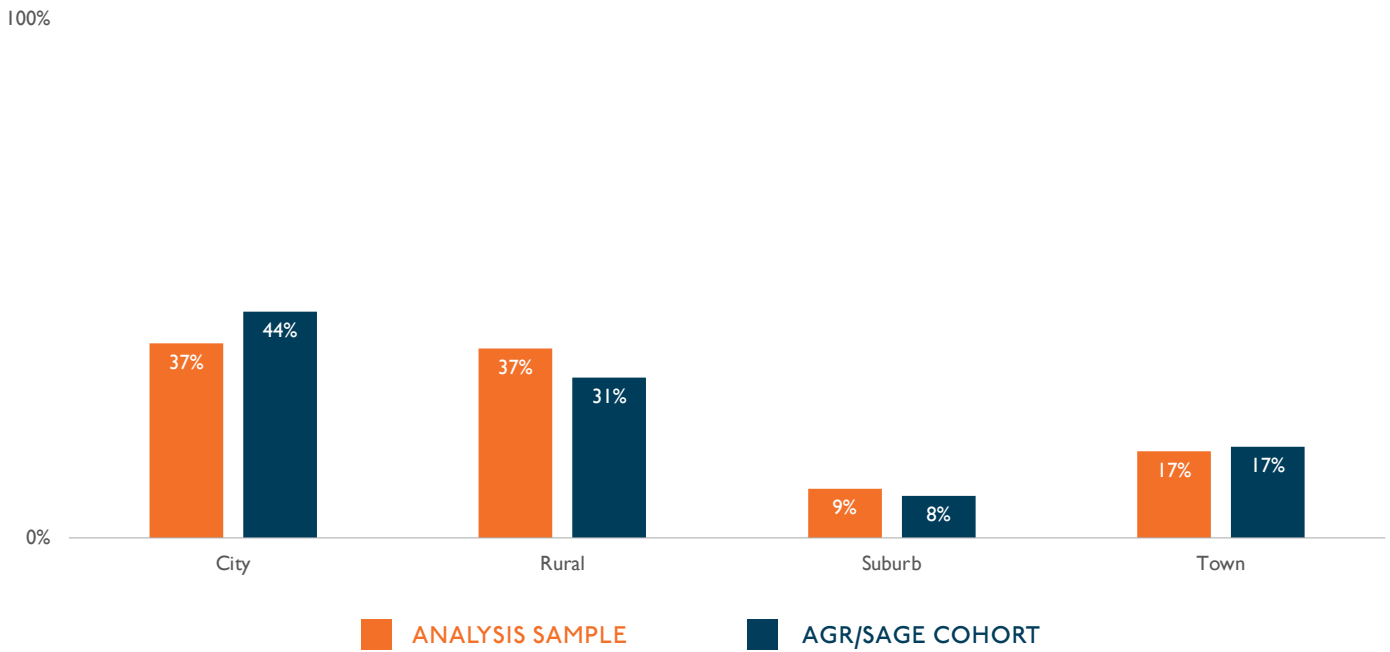


Figure 10: Race/Ethnicity of AGR Forward Analysis Students and AGR Students Overall

2018-19 Cohort

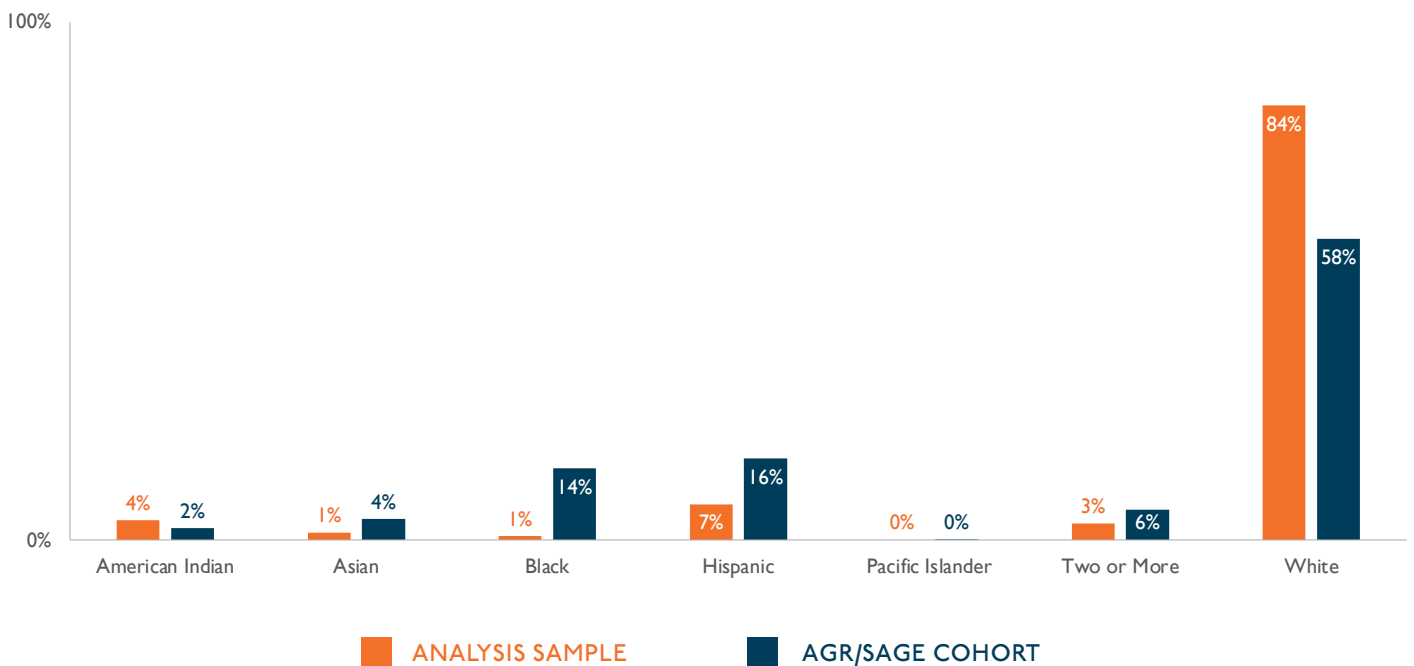


Figure II: AGR Forward Analysis Students and AGR Students Overall who were English Learners, Eligible for FRL, and in Special Education

2018-19 Cohort

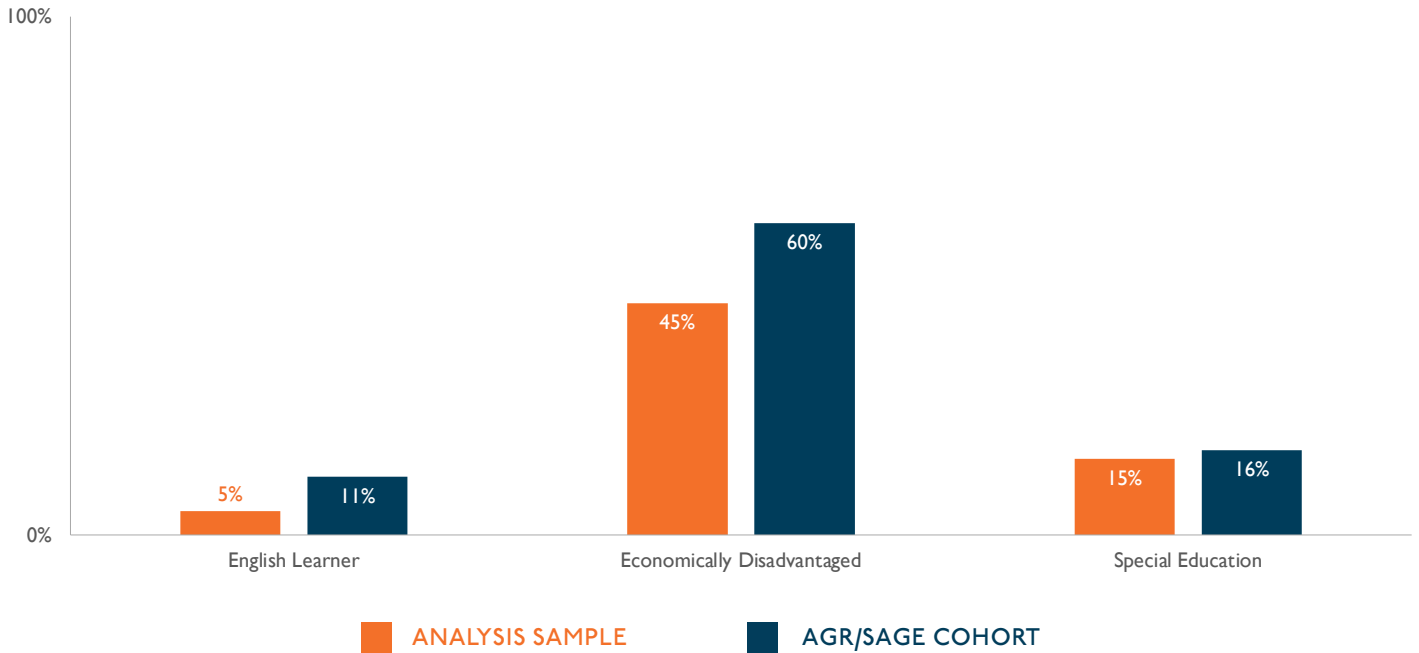
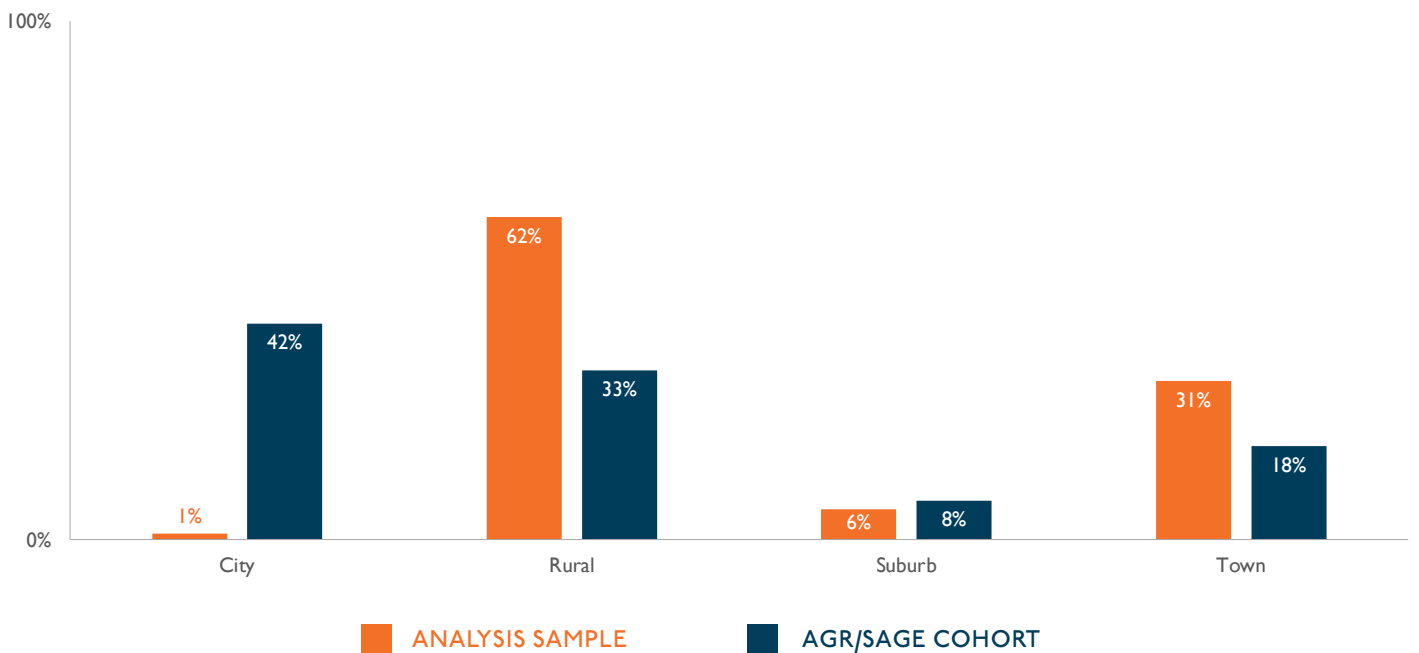


Figure I2: Locale Description of AGR Forward Analysis Students and AGR Students Overall

2018-19 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. As noted previously, the three strategies include:

Provide professional development related to small group instruction and reduce the class size to one of the following:

- No more than 18.
- No more than 30 in a combined classroom having at least 2 regular classroom teachers.

Provide data-driven instructional coaching for the class teachers.

Provide data-informed, one-to-one tutoring to pupils in the class who are struggling with reading or mathematics.

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 2021-22. These figures also provide information on the number and percentage of students affected by each strategy combination. Schools most frequently used class size reduction and coaching combined, all three strategies, class size reduction only, and coaching only. Only three schools used only tutoring as a strategy.

Figure 13: Implementation of AGR Strategies 2021-22

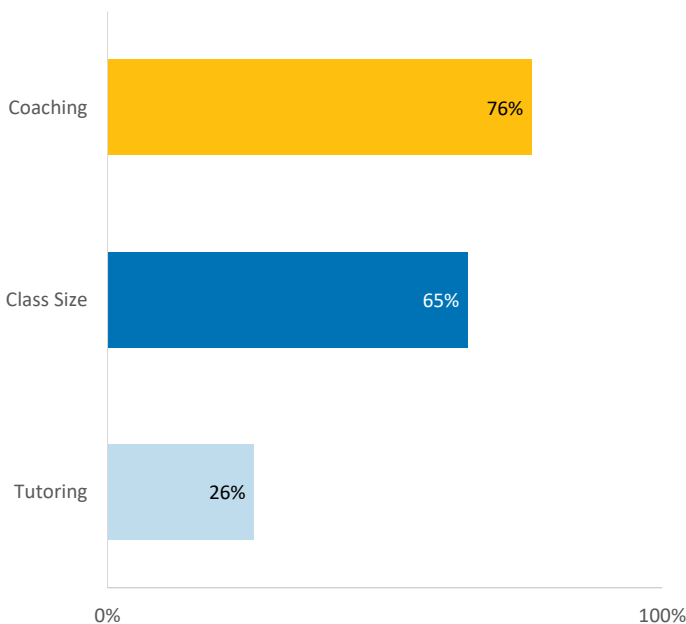
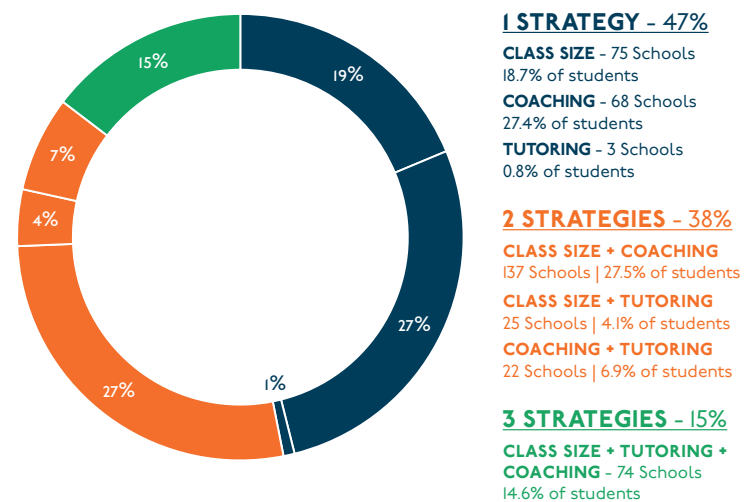


Figure 14: Implementation of AGR Strategies by Count of Strategies, 2021-22



Section 6

AGR Impacts

AGR Impacts

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

To answer these questions, we begin by providing results of AGR's impacts by comparing outcomes of AGR students relative to students in similar, non-AGR schools. We provide AGR impact results statewide and for various subpopulations of students, including low-income students. We also present AGR impacts relative to impacts from previous SAGE implementation. 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, absences, and OSS. 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. All p-values presented in this report are corrected to account for multiple estimates (see the Technical Appendix for details).

21 Data Recognition Corporation. (2022). *Wisconsin Forward Exam Spring 2022 Technical Report*. https://dpi.wi.gov/sites/default/files/imce/assessment/pdf/2022_Technical_Report_Final.pdf. See Table 10-I.

22 See Richardson, J., & Sim, G. 2020. [Evaluation of the Achievement Gap Reduction Program: 2015-16 through 2018-19](#).

Third Grade Forward Impacts

Table 7 and Table 8 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, four years of SAGE, or some combination of the two. Results are broken down by student demographics.

As shown in Table 7, AGR impacts on third grade Forward reading are generally positive but small and not statistically different from zero, both for all students and for most subsets of students (the notable exception, Black students in non-urban [town, suburb, and rural] schools, is a very small demographic group). For context, the impact from four years of AGR (0.97 scale score points) represents 0.02 standard deviations of Forward growth.²¹ Table 7 also shows that impacts from four years of SAGE and impacts from having a mix of both AGR and SAGE are both 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 1-3.²² Taken together, the Forward, PALS, and MAP/STAR results indicate that either (a) kindergarten AGR impacts fade out by the time students reach third grade, a common phenomenon among education interventions, and/or (b) improvements in PALS skills do not translate to improvements in the skills tested on Forward.

AGR impacts on third grade Forward math, as shown in Table 8, are similar to those of reading. The impact of four years of AGR (1.58 scale score points) is equal to 0.03 standard deviations.²³ 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 or for students with four years of SAGE.

23 Data Recognition Corporation. (2022). *Wisconsin Forward Exam Spring 2022 Technical Report*. https://dpi.wi.gov/sites/default/files/imce/assessment/pdf/2022_Technical_Report_Final.pdf. See Table 10-I.

Table 7: 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	FORWARD SCORE IMPACT (4 YEARS OF SAGE)	P-VALUE
ALL	0.97	0.85	-0.08	0.99	1.54	0.54
FRL	1.83	0.63	-0.77	0.78	1.16	0.73
Non-Urban	1.81	0.73	0.68	0.86	2.32	0.49
FRL & Non-Urban	1.74	0.72	-0.36	0.93	2.11	0.62
Non-Urban and Black	-6.25	0.51	1.59	0.85	-4.07	0.69
Non-Urban & Hispanic	6.01	0.24	-1.35	0.83	1.08	0.89
Non-Urban & White	1.63	0.78	0.87	0.83	2.54	0.47
Non-Urban & Other Race	1.57	0.89	-0.47	0.98	2.32	0.83

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

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

SAMPLE	FORWARD SCORE IMPACT (4 YEARS OF AGR)	P-VALUE	FORWARD SCORE IMPACT (MIX OF AGR/SAGE)	P-VALUE	FORWARD SCORE IMPACT (4 YEARS OF SAGE)	P-VALUE
ALL	1.58	0.73	-0.40	0.89	1.31	0.68
FRL	0.36	0.95	-1.04	0.72	2.41	0.40
Non-Urban	1.30	0.83	-0.19	0.97	0.42	0.92
FRL & Non-Urban	-1.94	0.73	-0.66	0.87	1.12	0.83
Non-Urban & Black	-0.50	0.99	0.91	0.91	-3.51	0.79
Non-Urban & Hispanic	5.47	0.36	-2.73	0.55	2.05	0.79
Non-Urban & White	1.52	0.79	0.20	0.97	0.65	0.90
Non-Urban & Other Race	-7.04	0.51	-1.89	0.83	-1.65	0.89

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

Fourth Grade Forward Impacts

Table 9 and Table 10 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, four years of SAGE, or some combination of the two. Results are broken down by student demographics.

As shown in Table 9, AGR impacts on fourth grade Forward reading are negative but small and not statistically different from zero, both for all students and for the subset of students who receive FRL (with the exception of AGR students who were Black and in non-urban schools, which is a very small demographic group). For context, the impact of four years of AGR (-4.11 scale score points) represents 0.08 standard deviations.²⁴ Table 9 also shows that the impacts of four years of SAGE or a mix of AGR and SAGE are not statistically different from zero.

AGR impacts on fourth grade Forward math, as shown in Table 10, are slightly different from reading impacts. The overall impact of AGR is small, positive, and not statistically significant. The impacts of AGR on various subgroups of students show mixed results, but all are also not statistically different from zero. The impact of four years of AGR (0.67 scale score points) is equal to 0.01 standard deviations.²⁵ There are no statistically significant impacts on Forward math for students with four years of SAGE or a mix of AGR and SAGE.

²⁴ Ibid.

²⁵ Ibid.

Table 9: 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	FORWARD SCORE IMPACT (4 YEARS OF SAGE)	P-VALUE
ALL	-4.11	0.21	-1.68	0.36	-2.38	0.35
FRL	-2.49	0.52	-2.53	0.23	-3.74	0.17
Non-Urban	-3.99	0.21	-2.06	0.37	-1.93	0.59
FRL & Non-Urban	-3.59	0.29	-3.04	0.22	-3.14	0.36
Non-Urban and Black	8.66	0.37	2.85	0.79	-6.64	0.56
Non-Urban & Hispanic	-1.64	0.87	-3.89	0.36	-1.98	0.83
Non-Urban & White	-4.75	0.13	-2.29	0.33	-1.88	0.59
Non-Urban & Other Race	2.24	0.86	2.31	0.86	-3.88	0.77

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

Table 10: 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	FORWARD SCORE IMPACT (4 YEARS OF SAGE)	P-VALUE
ALL	0.67	0.91	-2.67	0.14	-2.68	0.27
FRL	-0.28	0.99	-3.37	0.10	-4.00	0.13
Non-Urban	0.74	0.92	-3.53	0.09	-2.56	0.36
FRL & Non-Urban	-1.24	0.86	-4.29	0.06	-3.32	0.32
Non-Urban and Black	8.72	0.60	-1.24	0.91	-1.84	0.92
Non-Urban & Hispanic	5.03	0.51	-4.63	0.30	-3.57	0.60
Non-Urban & White	0.41	0.96	-3.48	0.11	-2.07	0.52
Non-Urban & Other Race	-3.07	0.79	-3.33	0.73	-11.10	0.21

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

During educational change, such as the transition from SAGE to AGR, outcomes can suffer an “implementation dip” as teachers and administrators create new systems, learn new skills, and integrate new programs into existing school structures.²⁶ When schools transitioned from SAGE, which only allowed for class size reduction, to AGR, most schools chose to change their implementation to include instructional coaching and/or tutoring (see Table I7). If an “implementation dip” occurred, we would expect to observe higher test score growth for students who experienced four years of AGR or SAGE than students with a mix of AGR and SAGE, those who experienced AGR as it was first implemented. Indeed, Tables 7 through I0 show some evidence of an “implementation dip” with the results for a mix of AGR and SAGE being somewhat lower than for four years of AGR or for four years of SAGE. With limited samples of students experiencing four years of AGR and generally high p-values, however, the results are inconclusive.

AGR Impacts on Absences and OSS

Table II and Table I2 present the impacts of AGR on absences and OSS, respectively. Results from these tables show the average difference in outcomes between AGR students and non-AGR students (or SAGE students). Results include the sample of all students and multiple subgroups.

In each of these tables, negative impacts are desirable, reflecting that AGR reduces absences or OSS.

Note that, across all results, OSS models are statistically noisy – available student and school-level controls are unable to explain much of the variance in the outcome. Readers should interpret results with caution.

Table II shows AGR impacts on absences, which are presented both in terms of percentage point differences and the implied difference in absence days during a typical school year. For all AGR students and most student subgroups, the AGR impact, relative to non-AGR, non-SAGE students, is small and not statistically significant. Comparing students in urban schools, AGR students have significantly greater absences than those in the non-AGR comparison group. Relative to SAGE students, AGR students in the non-urban, White, and White students eligible for FRL subgroups have significantly fewer absences. However, none of these results are larger than the equivalent of one day of instruction.

For OSS (Table I2), there are no statistically significant differences between AGR schools and non-AGR schools, or AGR schools and SAGE schools overall and for most subgroups of students. There are statistically significant, positive results showing that AGR schools have fewer OSS than non-AGR schools for EL students and Black students in non-urban schools (town, suburb, and rural), although this demographic group is small.

²⁶ Fullan, M. (2001). *Leading in a Culture of Change*. Jossey-Bass.

Table II: Impact of AGR on Absences

	AGR VS NON-AGR/NON-SAGE			AGR VS SAGE		
	IMPACT (PERCENTAGE SAMPLE POINTS)	IMPACT (DAYS)	P-VALUE	IMPACT (PERCENTAGE SAMPLE POINTS)	IMPACT (DAYS)	P-VALUE
All Students	0.52	0.9	0.17	-0.34	-0.6	0.29
Free & Reduced Lunch	0.51	0.9	0.21	-0.17	-0.3	0.69
English Language Learners	0.52	0.9	0.30	0.12	0.2	0.69
Urban	0.73	1.3	0.02	0.21	0.4	0.60
Non-Urban	0.40	0.7	0.51	-0.73	-1.3	0.01
Black	0.88	1.5	0.05	0.52	0.9	0.24
Hispanic	0.52	1.0	0.14	0.27	0.5	0.45
Other	0.51	1.4	0.34	-0.19	-0.3	0.69
White	0.42	0.7	0.38	-0.66	-1.2	0.01
Free & Reduced Lunch and Urban	0.70	1.2	0.09	0.35	0.6	0.36
Free & Reduced Lunch and Non-Urban	0.36	0.6	0.66	-0.76	-1.3	0.02
Free & Reduced Lunch and Black	0.83	1.5	0.12	0.72	1.3	0.13
Free & Reduced Lunch and Hispanic	0.49	0.9	0.31	0.28	0.5	0.50
Free & Reduced Lunch and Other	0.99	1.7	0.34	-0.08	-0.1	0.90
Free & Reduced Lunch and White	0.37	0.7	0.58	-0.69	-1.2	0.02
Urban and Black	1.02	1.8	0.06	0.62	1.1	0.21
Urban and Hispanic	0.80	1.4	0.12	0.57	1.0	0.12
Urban and Other	1.06	1.9	0.27	0.10	0.2	0.87
Urban and White	0.52	0.9	0.10	-0.27	-0.5	0.41
Non-Urban and Black	0.20	0.3	0.83	-0.22	-0.4	0.76
Non-Urban and Hispanic	0.17	0.3	0.79	-0.41	-0.7	0.34
Non-Urban and Other	0.55	1.0	0.73	-0.69	-1.2	0.18
Non-Urban and White	0.39	0.7	0.52	-0.74	-1.3	0.01

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

Table 12: Impact of AGR on OSS

	AGR VS NON-AGR/NON-SAGE		AGR VS SAGE	
	IMPACT (PERCENTAGE SAMPLE POINTS)	P-VALUE	IMPACT (PERCENTAGE SAMPLE POINTS)	P-VALUE
All Students	-0.26	0.55	0.08	0.87
Free & Reduced Lunch	-0.36	0.44	0.06	0.91
English Language Learners	-0.61*	0.02	-0.08	0.69
Urban	-0.22	0.76	0.06	0.91
Non-Urban	-0.42	0.36	0.16	0.69
Black	-0.40	0.69	-0.04	0.99
Hispanic	-0.41	0.27	-0.01	0.99
Other	-0.44	0.73	-0.01	0.98
White	-0.01	0.99	0.26	0.34
Free & Reduced Lunch and Urban	-0.34	0.62	0.02	0.98
Free & Reduced Lunch and Non-Urban	-0.57	0.30	0.19	0.72
Free & Reduced Lunch and Black	-0.56	0.60	-0.02	0.987
Free & Reduced Lunch and Hispanic	-0.54	0.23	-0.08	0.86
Free & Reduced Lunch and Other	-0.36	0.77	-0.04	0.98
Free & Reduced Lunch and White	-0.06	0.91	0.30	0.39
Urban and Black	-0.28	0.83	0.00	1.00
Urban and Hispanic	-0.40	0.47	0.01	0.99
Urban and Other	0.15	0.89	0.35	0.52
Urban and White	-0.03	0.98	0.25	0.69
Non-Urban and Black	-2.37*	0.02	-0.47	0.68
Non-Urban and Hispanic	-0.67	0.08	-0.07	0.90
Non-Urban and Other	-1.54	0.37	-0.48	0.63
Non-Urban and White	0.01	0.98	0.24	0.30

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

Differences in Outcomes by AGR Strategies

This set of student performance results examines how schools' AGR strategies impact outcomes. For schools that changed their AGR strategies over time, we estimate how those strategy changes impact outcomes. For Tables I3-I6, the impact is the difference in the outcome between AGR students in schools before and after a strategy is employed. For instance, if a school switched from not using class size reduction to using class size reduction, the resulting estimated change in outcome is the impact presented below. Due to the number of combinations of possible strategies used, the estimates provided only examine the extent to which each individual type of strategy (class size reduction, coaching, and tutoring) changed, regardless of the other strategies used within a school. The analysis uses annual MAP and STAR reading and math data to capture year-to-year changes in test score growth over time in grades 1-3, which would not be possible using Forward, which is first administered in third grade.

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

Table I3 shows the impact of a strategy change for class size, coaching, and tutoring on reading growth, and Table I4 shows the impact on math growth. For each subject, none of the strategies are associated with a statistically significant impact on academic growth. Similarly, Table I5 shows no statistically significant associated impact on absence rates for changing to any particular AGR strategy. For discipline, however, Table I6 highlights that changing to using class size reduction as an AGR strategy is associated with reducing OSS, a statistically significant impact. This was the case overall and for low-income and Black students.

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

SAMPLE	STRATEGY	IMPACT (STANDARDIZED)	IMPACT (APPROX. MAP SCALE)	P-VALUE
All Students	Class Size	0.01	0.16	0.82
	Coaching	0.01	0.16	0.85
	Tutoring	-0.05	-0.71	0.58
FRL Students	Class Size	0.02	0.22	0.73
	Coaching	0.02	0.23	0.79
	Tutoring	-0.05	-0.77	0.57
Black Students	Class Size	0.01	0.14	0.90
	Coaching	0.05	0.68	0.67
	Tutoring	0.02	0.32	0.85
Hispanic Students	Class Size	0.01	0.13	0.90
	Coaching	-0.03	-0.47	0.65
	Tutoring	-0.07	-1.04	0.45
White Students	Class Size	0.01	0.19	0.79
	Coaching	0.01	0.17	0.86
	Tutoring	-0.06	-0.90	0.44


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

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

SAMPLE	STRATEGY	IMPACT (STANDARDIZED)	IMPACT (APPROX. MAP SCALE)	P-VALUE
All Students	Class Size	-0.02	-0.23	0.79
	Coaching	-0.03	-0.45	0.56
	Tutoring	0.01	0.10	0.93
FRL Students	Class Size	-0.02	-0.22	0.79
	Coaching	-0.03	-0.36	0.69
	Tutoring	0.00	-0.02	0.99
Black Students	Class Size	-0.04	-0.55	0.51
	Coaching	-0.02	-0.27	0.88
	Tutoring	0.07	-0.99	0.40
Hispanic Students	Class Size	-0.05	-0.67	0.49
	Coaching	-0.11	-1.43	0.10
	Tutoring	0.01	0.17	0.91
White Students	Class Size	0.00	-0.03	0.99
	Coaching	-0.03	-0.34	0.68
	Tutoring	0.00	0.04	0.98

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

Table 15: Impact of AGR Strategies on Absences

SAMPLE	STRATEGY	IMPACT (STANDARDIZED)	IMPACT (APPROX. MAP SCALE)	IMPACT (DAYS)	P-VALUE
All Students	Class Size	0.00	0.0	0.0	0.99
	Coaching	0.01	0.1	0.1	0.91
	Tutoring	0.02	0.2	0.3	0.59
FRL Students	Class Size	0.00	-0.1	-0.1	0.94
	Coaching	0.02	0.3	0.3	0.83
	Tutoring	0.01	0.2	0.2	0.73
Black Students	Class Size	-0.02	-0.2	-0.2	0.69
	Coaching	0.00	0.0	0.0	0.99
	Tutoring	0.02	0.3	0.3	0.79
Hispanic Students	Class Size	0.01	0.1	0.1	0.90
	Coaching	0.04	0.7	0.7	0.46
	Tutoring	0.02	0.3	0.3	0.69
White Students	Class Size	-0.01	-0.1	-0.1	0.94
	Coaching	0.00	0.0	0.0	1.00
	Tutoring	0.02	0.3	0.3	0.52

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

Table 16: Impact of AGR Strategies on OSS

SAMPLE	STRATEGY	IMPACT (STANDARDIZED)	IMPACT (APPROX. MAP SCALE)	P-VALUE
All Students	Class Size	-0.01	-0.6	0.02
	Coaching	0.00	0.1	0.87
	Tutoring	0.00	-0.1	0.91
FRL Students	Class Size	-0.01	-0.8	0.01
	Coaching	0.00	-0.1	0.91
	Tutoring	0.00	0.1	0.91
Black Students	Class Size	-0.02	-1.9	0.00
	Coaching	0.00	-0.3	0.86
	Tutoring	0.00	-0.2	0.90
Hispanic Students	Class Size	-0.01	-0.5	0.24
	Coaching	0.00	0.2	0.84
	Tutoring	0.00	0.1	0.90
White Students	Class Size	0.00	0.3	0.46
	Coaching	0.00	0.2	0.69
	Tutoring	0.00	-0.1	0.83

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

Section 7

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

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

At the conclusion of each school year, DPI surveys AGR schools to produce an EOY 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 five years of EOY data collected from 2017-18 through 2021-22, and three years of school board report data, collected from 2017-18 through 2019-20. Due to the pandemic and the pressures it put on schools,

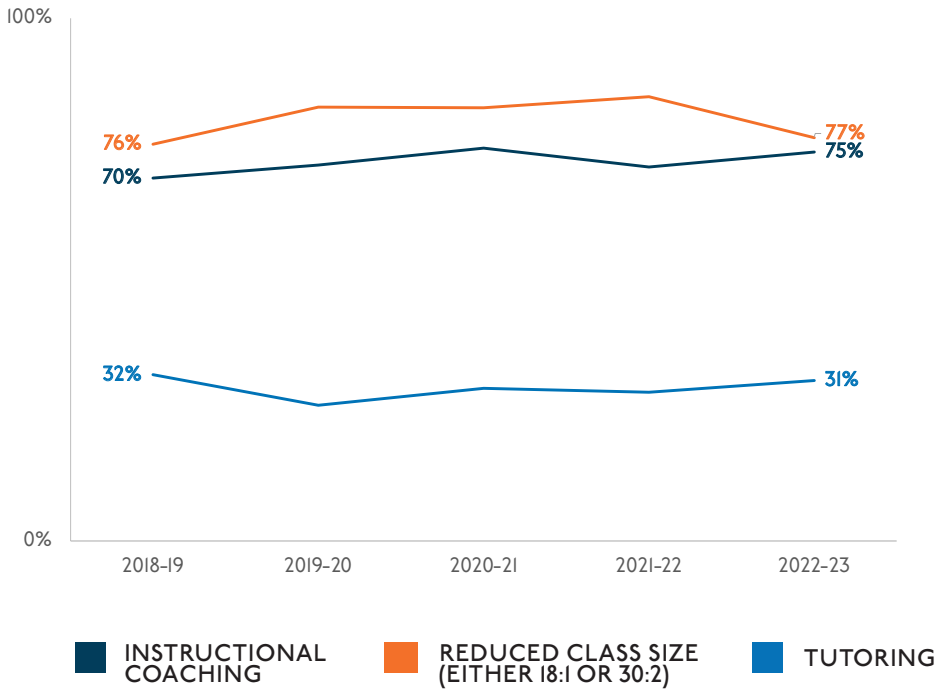
DPI did not require schools to submit their 2020-21 school board reports to DPI. 2021-22 school board reports are also excluded from this analysis but will be included in next year's report.

Table 17 shows AGR strategy combinations schools chose during each of the five sample years. Schools are listed as employing a strategy if they indicate it on either their EOY or school board reports. As can be seen in Table 17, strategy choices have been stable over time. One exception is that schools combining class size reduction and instructional coaching increased by from 29 percent to 40 percent from 2017-18 to 2019-20 but decreased to 34 percent by 2021-22.

Table 17: School-level AGR Strategies

STRATEGY	2017-18	2018-19	2019-20	2020-21	2021-22
Coaching Only	19%	12%	11%	12%	17%
Class Size Only	17%	19%	16%	22%	19%
Tutoring Only	1%	0%	1%	0%	0%
Coaching and Class Size	29%	38%	40%	38%	34%
Coaching and Tutoring	4%	3%	3%	3%	5%
Class Size and Tutoring	9%	7%	7%	7%	6%
All Three	22%	22%	23%	18%	18%

Figure 15: Schools Using Each Strategy in Any Grade



Reduced class size has remained AGR's most used strategy since the program's beginning, when its predecessor program SAGE required that all AGR schools reduce class sizes in grades K-3. As seen in Figure 15, about 77 percent of schools use class size reduction in at least one grade, slightly higher than the number that use instructional coaching. Throughout the sample, less than 40 percent of schools have used tutoring in any grade, despite the consensus in the research literature demonstrating the benefits of intensive tutoring.

AGR statewide policy allows schools to choose how to distribute each strategy across grades and across classrooms within grades. As a result, some schools choose to implement particular strategies in just some grades. Figure 16 shows that schools using class size reduction do not always use that strategy uniformly across grades K-3. Class size reduction is more common in kindergarten than grades 1-3. In kindergarten, over 75 percent of schools using class size reduction do 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

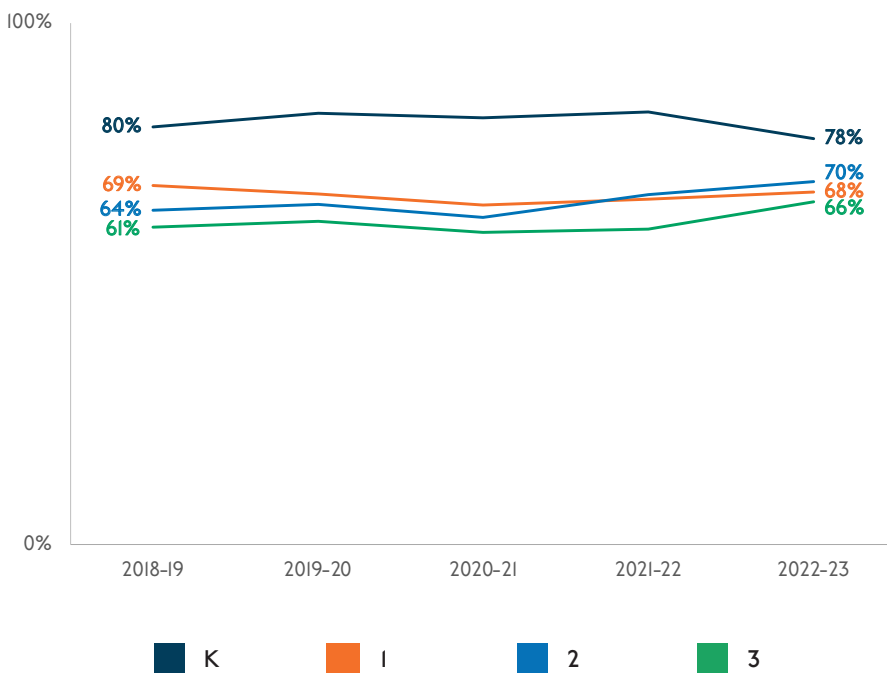
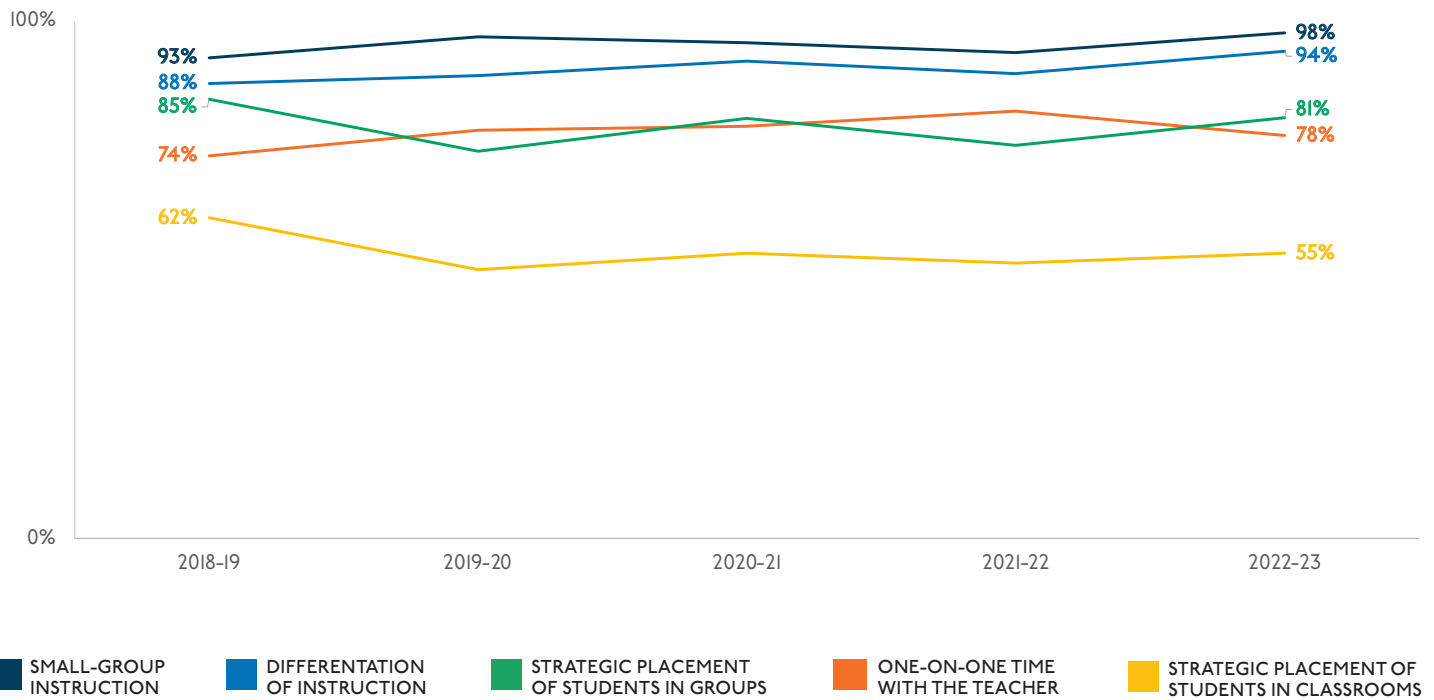


Figure 17: Instructional Strategies in Schools with Class Size Reduction



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

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 five 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 nearly 60 percent of schools. Only 2-4 percent of schools reported that they use no additional instructional strategies due to reduced class sizes (not shown).

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

By Grade

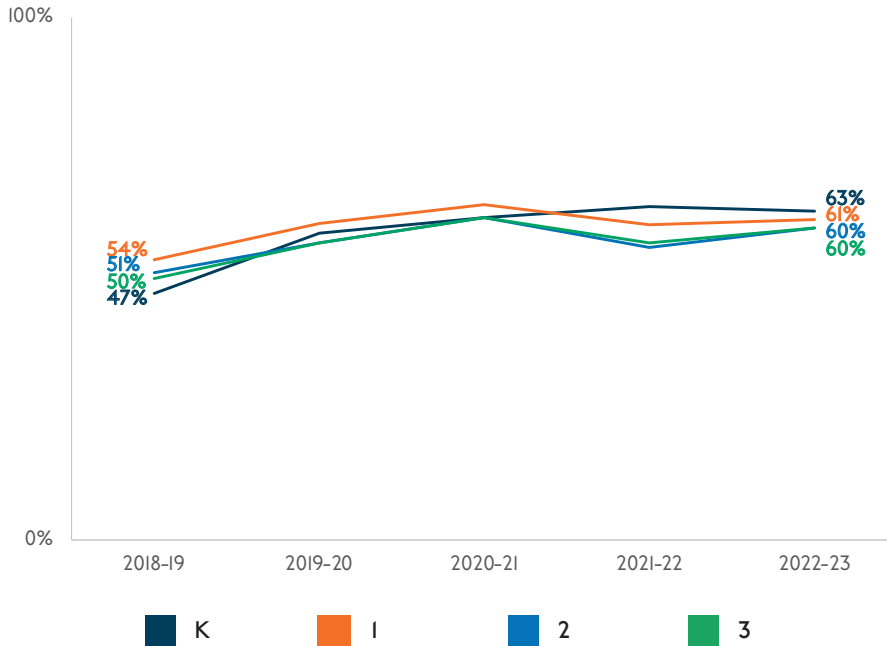
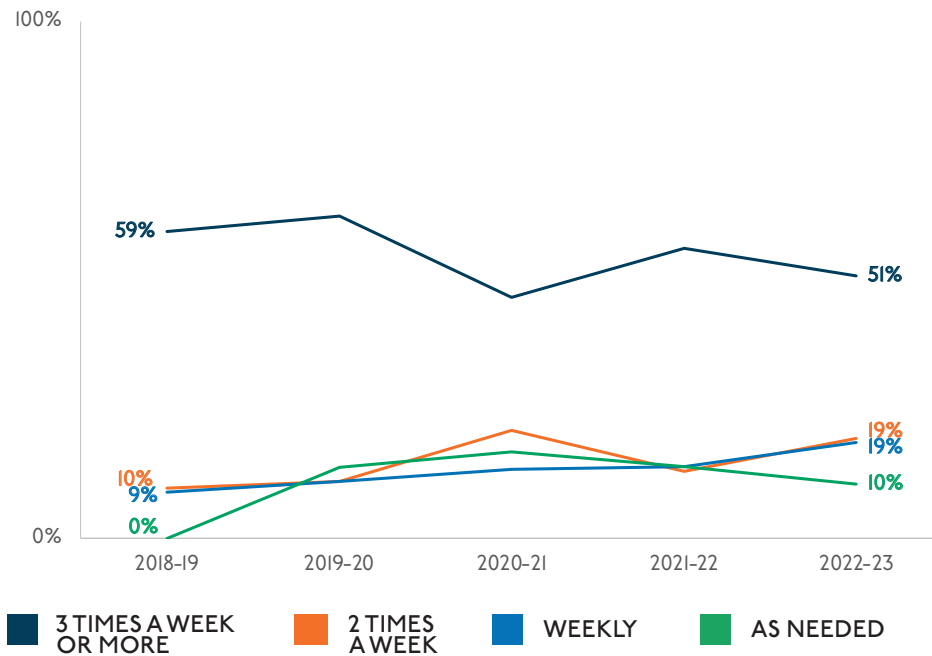


Figure 18 shows that, unlike schools' distribution of class size reduction, tutoring is equally common in all grades. Within grades, since 2017-18, schools that use tutoring have been increasingly likely to use tutoring in at least 75 percent of classrooms. Figure 18 also reflects that schools implementing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

Studies show that more frequent tutoring results in larger impacts on standardized exam scores.²⁷ Most schools implementing AGR tutoring do so frequently, as seen in Figure 19. A little more than 50 percent of schools provide tutoring at least three times per week, and a little less than 20 percent provide tutoring two times per week. In 2019-20, perhaps due to the COVID-19 pandemic, fewer schools tutored students three times per week, but there was a parallel increase in schools tutoring two times per week.

Figure 19: School-Reported Tutoring Frequency

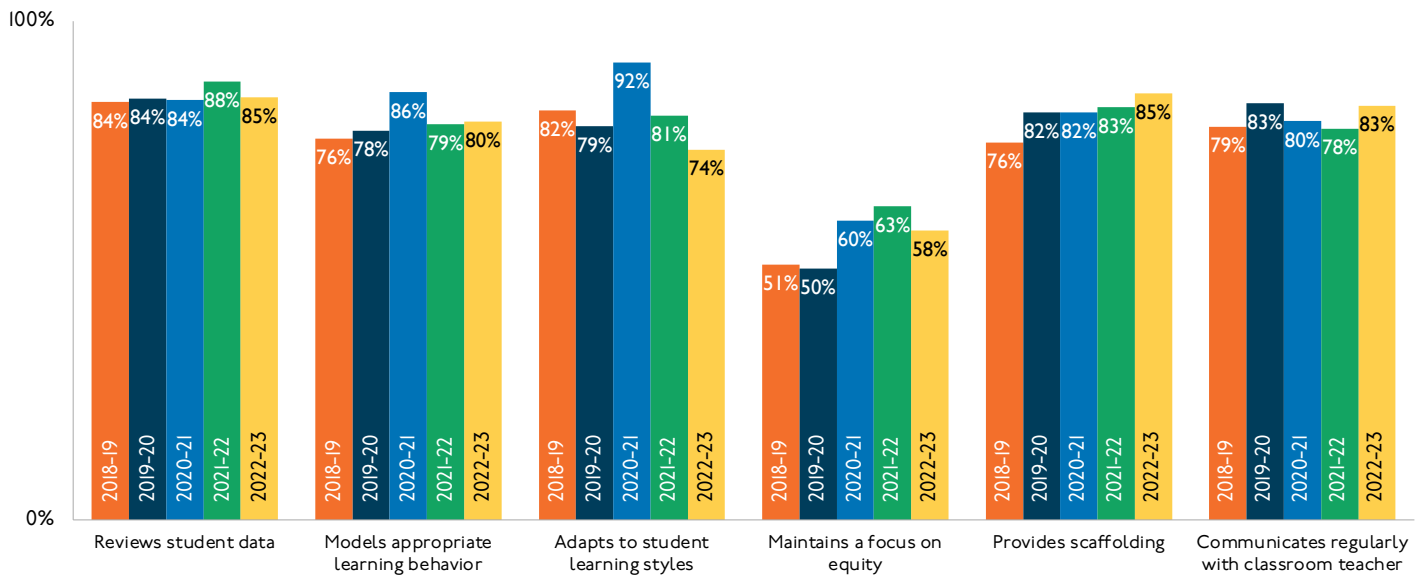


* Note: Not Pictured - No specific instructional strategies, Other, Not Sure/Don't Know

27 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 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 2021-22, 58 percent of schools used tutoring for equity purposes, up from 51 percent in 2017-18.

Figure 20: School-Reported Tutoring Practices



* Note: Not Pictured – No specific instructional strategies, 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 60 percent do so in at least three quarters of classrooms in any particular grade. These patterns have been stable since 2017-18.

Figure 21: Schools Using Instructional Coaching Doing So in at Least 75% of Classrooms By Grade

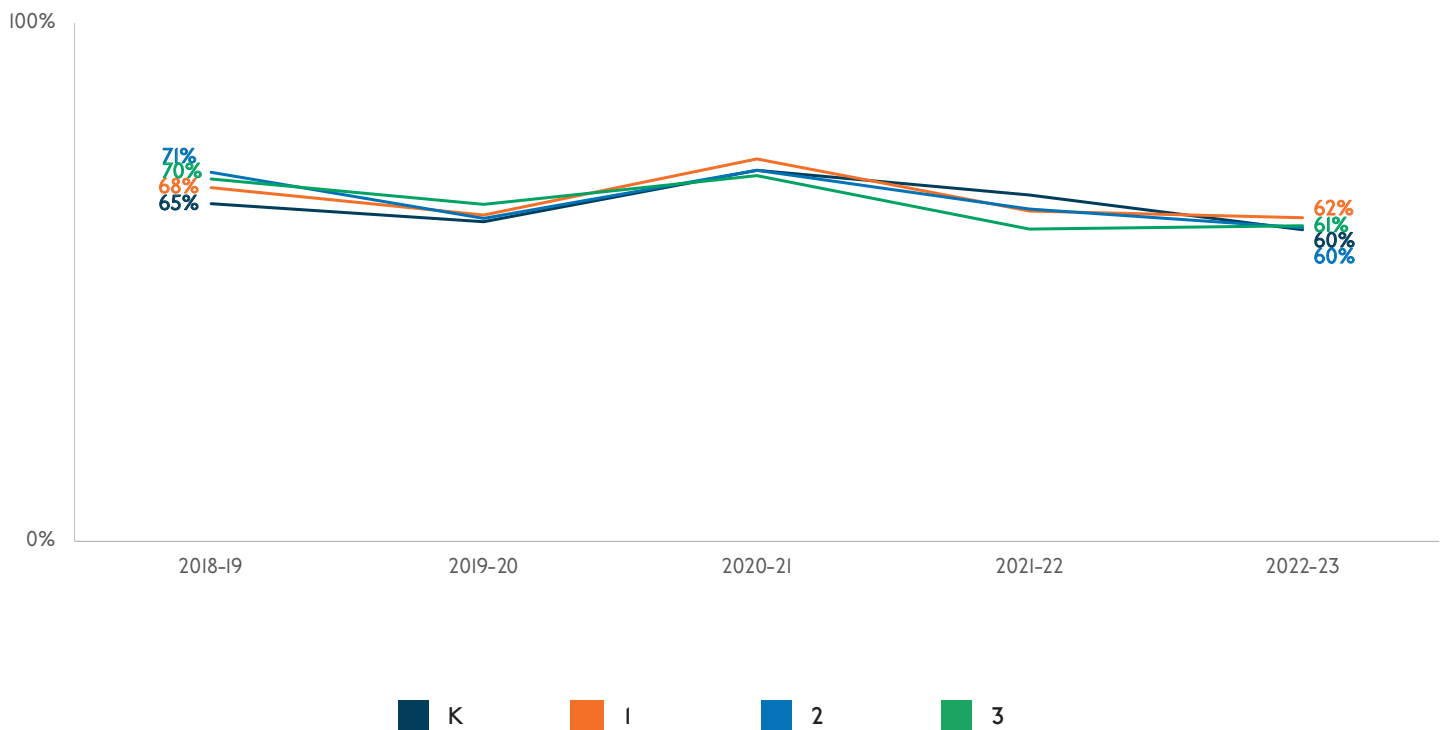
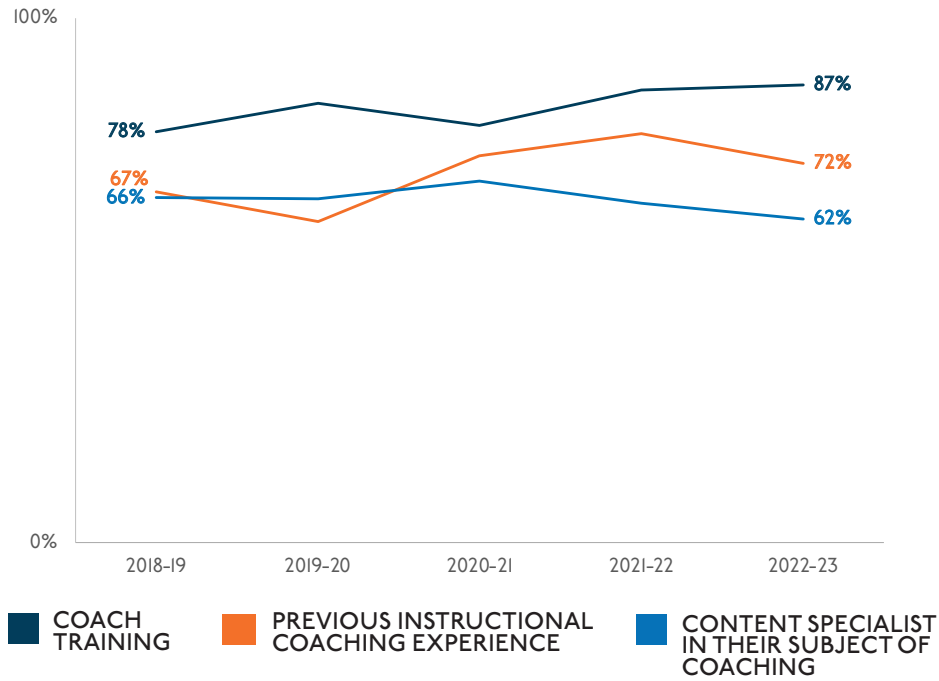


Figure 22: School-Reported Characteristics of Instructional Coaches

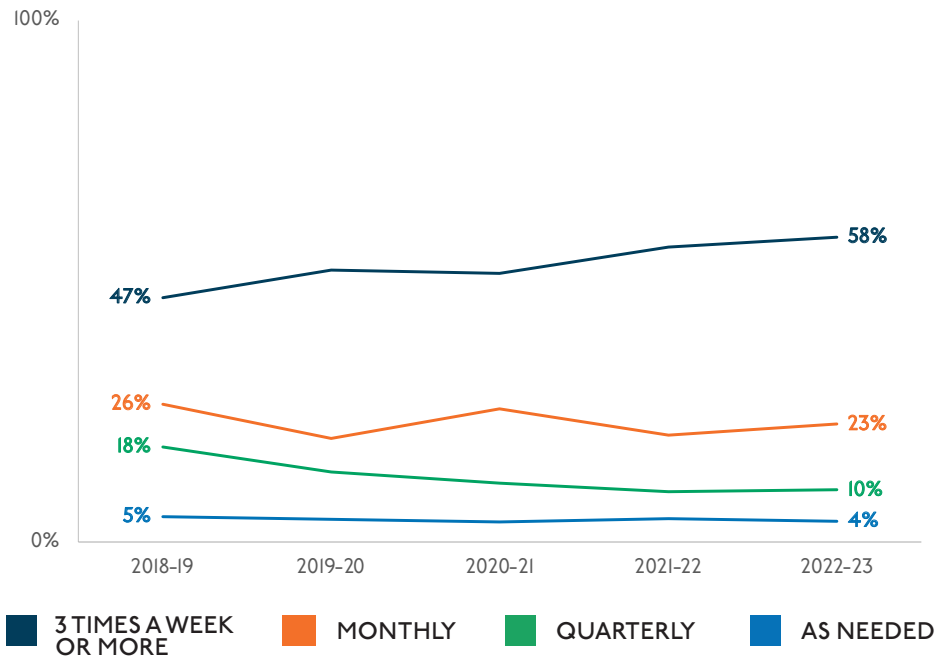


* Note: Not Pictured – No specific instructional strategies, Other, Not Sure/Don't Know

Schools responded affirmatively to almost all listed instructional coaching characteristics (Figure 22). As expected, over the sample period, increasingly more schools reported that their coaches had previous coaching experience, increasing from 61 percent of schools in 2018-19 to 72 percent in 2021-22. The percentage of schools whose coaches had previous training and were content specialists in their subjects of coaching remained static over the sample period.

From 2017-18 to 2021-22, schools reported increasing frequency of instructional coaching (Figure 23). Schools providing weekly coaching increased from 47 percent in 2017-18 to 58 percent in 2021-22, while the fraction of schools providing coaching on an as-needed basis steadily declined.

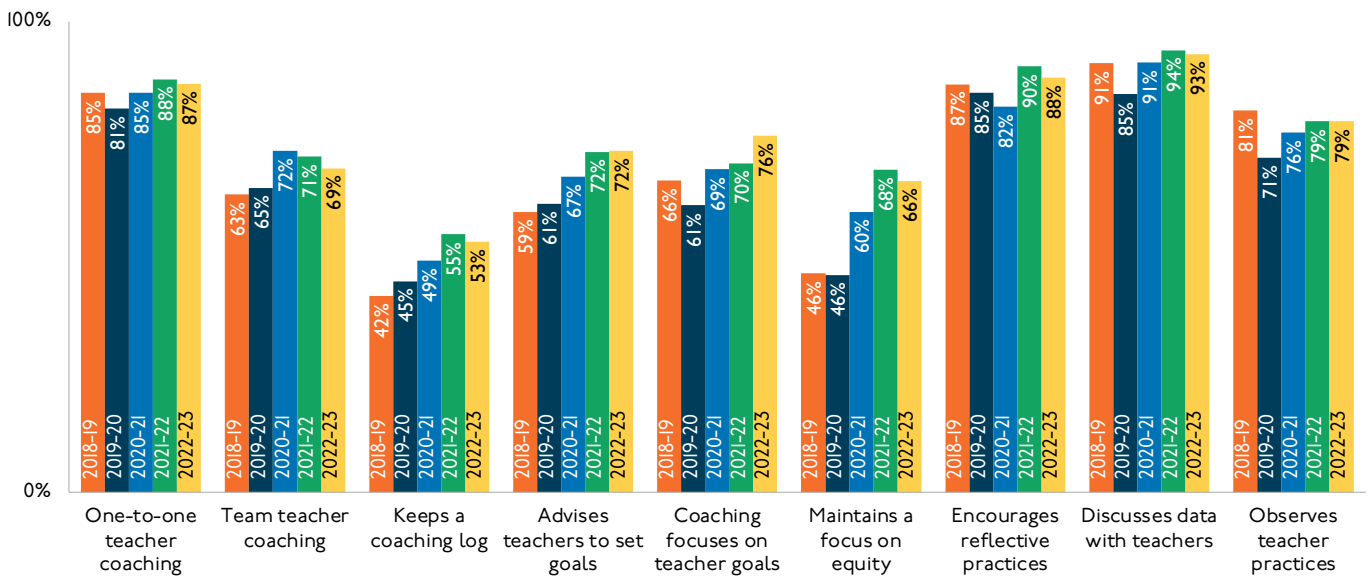
Figure 23: School-Reported Instructional Coaching Practices



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

Figure 24 shows that, during the sample period, schools reported increasing use of most of the instructional coaching practices listed on the EOY survey. Similar to practices associated with one-to-one tutoring, the most substantial growth occurred for maintaining a focus on equity. In 2021-22, 66 percent of respondents cited maintaining a focus on equity, 20 percentage points more than in 2017-18 (46 percent). Keeping a coaching log and advising teachers to set goals also saw substantial growth.

Figure 24: School-Reported Instructional Coaching Practices



Section 8

Future AGR Evaluation

Future AGR Evaluation

Rigorous evaluation of AGR impacts on student testing outcomes requires a pre-program baseline measure of student knowledge. That measure (or measures) should apply to as many students as possible, so that the sample of AGR students used to estimate impacts is a valid representation of the overall AGR student population and key subgroups.

Fortunately, Wisconsin requires that districts administer fall early literacy screeners to kindergarteners, before students have received significant benefits of AGR or other programs. PALS was Wisconsin’s mandated literacy screener through 2015-16. Afterwards, the state continued to require that districts administer an early literacy screener but could choose assessments other than PALS.²⁸ As a result, PALS usage has decreased since 2015-16 (Table 18). For the kindergarten cohort of 2018-19, the most recent cohort included in this report, about 40 percent took PALS. As can

be seen in Table 18, only 12 percent of the 2020-21 and 2021-22 AGR cohorts took PALS. Both Milwaukee and Madison, the state’s two largest districts representing over 20 percent of AGR schools, have opted for screeners other than PALS.

For the AGR evaluation, DPI has made data available for three assessments – PALS, STAR, and MAP. However, these three assessments have decreased in popularity. As a result, evaluation of AGR’s impacts increasingly relies on a sample that does not accurately represent AGR students and schools (as seen earlier in Figure 10 – Figure 12). As districts continue to phase out PALS, evaluators will need access to data from other kindergarten assessments. Based on information collected in the End-of-Year Report, most AGR schools use one of a handful of kindergarten screeners. Having access to data from these screeners is essential for the continued rigor of any AGR evaluation.

28 Wisconsin Department of Public Instruction. *PALS Early Literacy Screener*. <https://dpi.wi.gov/assessment/historical/pals>

Table 18: Students with PALS Scores

KINDERGARTEN YEAR	AGR/SAGE SCHOOLS	ALL SCHOOLS
2012-13	94%	96%
2013-14	95%	96%
2014-15	95%	96%
2015-16	95%	95%
2016-17	54%	56%
2017-18	46%	49%
2018-19	40%	43%
2019-20	37%	37%
2020-21	12%	13%
2021-22	12%	14%

Section 9

Summary and Conclusions

Summary and Conclusions

The socioeconomic achievement gap is wide and has remained relatively unchanged over the past fifty years.²⁹ Researchers and policymakers have hypothesized dozens of causes, including funding deficits for districts located in high-poverty areas.³⁰ The AGR program seeks to reduce the socioeconomic achievement gap by providing additional funding to districts with large proportions of economically disadvantaged students.

This report provides evidence regarding program impacts on math and reading growth, student attendance, and OSS. These results are presented at the state level and disaggregated for 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 preliminary evidence on the relative effectiveness of the three AGR strategies.

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 and fourth 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 2019, AGR impacts on third grade Forward reading and math are small and not statistically different from zero, for both the statewide sample and for many student subgroups, including students receiving FRL. Impacts for fourth 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. 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.

The report also estimates impacts for non-testing outcomes through the 2021-22 academic year, omitting 2020-21 due to changes in absence and suspension rates due to the COVID-19 pandemic. We find no consistent evidence of AGR impacts on absences or OSS.

29 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

30 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>

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 75 percent of schools use reduced class sizes in at least one grade. Despite strong evidence in the research literature that tutoring can positively impact achievement score growth, relatively few AGR schools implemented tutoring, either alone or in combination with other strategies. Examinations of schools that change strategies over time continues to find evidence that class size reduction reduces OSS rates for some subgroups of students by a small amount.

As in any observational study, this evaluation has several limitations. The PSM methodology matches schools on observable characteristics, but comparison schools may not match AGR schools on unobserved characteristics such as schools' ability to properly implement AGR or instructor quality in the local labor market. The long history of SAGE, AGR's precursor program that provided funding for reduced class sizes only, also limits the study. For absences and OSS, previous school outcomes used for matching likely include SAGE impacts as well, which would bias AGR impacts toward zero. Inconsistent testing patterns in grades K-3, including diminishing use of PALS, a lack of testing in spring of 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.

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

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

Comparison schools with high percentages of students eligible for FRL are not AGR participants for two primary reasons. First, poverty in those schools may have increased since the last SAGE eligibility period. Those schools currently would be eligible for AGR based on poverty thresholds but are ineligible because they did not participate in SAGE. 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 rests on the strong assumption that balancing AGR and non-AGR schools on observed characteristics also balances those schools on unobserved characteristics. The most typical method of addressing bias from fixed, unobserved characteristics would be to include school fixed effects in the estimation. For the AGR analysis, however, including school fixed effects would only allow comparisons of AGR to SAGE because all AGR schools previously participated in SAGE. 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, the methodology for estimating impacts of individual AGR strategies, multiple comparisons corrections of p-values, and an analysis of AGR implementation changes during COVID and the relationship to school learning models.

Propensity Score Matching

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

Equation (1)

$$\ln \left[\frac{\Pr(\text{EverAGR}_s)}{1 - \Pr(\text{EverAGR}_s)} \right] = \alpha + \beta X_s + \epsilon_s$$

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

Specifying the Propensity Score Model

To determine which variables to include in the propensity score matching model, we tested the influence of many demographic and academic variables. The final list of covariates appears in Table 1 and Table 2 of the main report. For each of the models, the most important matching variables measure the average outcome in a previous time period (pretests), such as the school's average test scores from the previous time period. The choice of pretest was complicated by both the level of matching (school or school-grade-year) and by the fact that AGR schools previously participated in SAGE. To the greatest extent possible, we aimed to remove previous program impacts from the matching model.

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. We used two strategies to address matching on post-program outcomes. For Forward, we matched separately for each school-cohort, using pre-program kindergarten fall PALS measures along with other, school-level controls (see Table 1). For the attendance and discipline models, we included average attendance rate (or suspension rate), limiting the effect of including a post-program outcome to just one year (see Table 2). The inclusion of school-level outcome controls may cause bias, however, depending on whether SAGE, which was in place in 2012-13, impacted those 2012-13 outcomes. All AGR schools previously participated in SAGE. As a consequence, matching AGR schools to non-AGR schools based on post-SAGE outcomes risks biasing the results toward zero (toward estimating smaller impacts), because schools would be matched on previous-period outcomes that already include the treatment impact (in this case, the SAGE program is similar enough to AGR to raise similar concerns). If SAGE had no impact on 2012-13 outcomes, however, omitting the outcomes from matching risks bias from poor matches. With available data it is not possible to know whether SAGE impacted 2012-13 attendance and OSS outcomes.

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.³¹

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 (2016) kindergarten cohort, who did not take third (fourth) grade Forward due to COVID-19 shutdowns, students from the 2018 and 2019 (2017 and 2018) kindergarten cohorts whose schools did not host at least 75 percent of class days in person 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 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.

³¹ We use Stata's `kmatch` package with an Epanechnikov Kernel and allow Stata to select the optimal bandwidth based on the outcome.

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.³² 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.

32 What Works Clearinghouse. (2022). *Procedures and Standards Handbook, Version 5.0*. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5_0-0-508.pdf

Table A-I: Matching and Analysis Strategies and Samples

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

Impact Analysis

After matching, we model impact estimates separately for each outcome. The analysis model for Forward differs from the model for absences and OSS, although in both cases we model impact estimates for SAGE and AGR in the same regression. All models include weights generated by the kernel PSM procedure. Standard errors are clustered at the school level. Models for Forward and absence rate use Weighted Least Squares, and the suspension rate model, where the outcome is an indicator of whether a student received at least one suspension during the year, uses a logit specification. To account for the non-linearity of absence rate as an outcome, we first convert absence rates onto the standard normal distribution using a probit transformation. To provide meaningful results, we then use an inverse transformation of the raw impact estimates before reporting. Outcome statistics are available upon request.

Forward Reading & Math

Equation (2)

$$\text{Forward}_{i,g3,s,c} = \alpha + \lambda \text{PALS}_{i,gK} + \beta_1 \text{FullAGR}_i + \beta_2 \text{FullSAGE}_i + \beta_3 \text{AGR/SAGEMix}_i + \gamma X_i + \theta Z_s + \delta_c + \varepsilon_{i,g3,s,c}$$

where $\text{Forward}_{i,g3,s,c}$ is an outcome for student i in grade 3 (or 4), school s , and year y . $\text{PALS}_{i,gK}$ is student i 's fall kindergarten PALS score, which acts as a pretest for both reading and math. FullAGRs , FullSAGE_i , and AGR/SAGEMix_i are indicators of treatment type. X_i represents a vector of student-level covariates, and Z_s 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 3 in the main report. As described above, the models include all school-level variables from the PSM procedure.

Absences & OSS

Equation (3)

$$Y_{i,s,g,y} = \alpha + \beta_1 \text{SAGE}_{s,y} + \beta_2 \text{AGR}_{s,y} + \gamma X_{i,y} + \theta Z_{s,y} + \delta_{g,y} + \varepsilon_{i,s,g,y}$$

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

Alternative Forward Reading and Math Analysis

For this year's evaluation, we tested models of third and fourth grade Forward scores that include 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. One benefit of this alternative specification is that the generalizability of results improves as fewer schools are removed from the analysis sample due to employing a virtual learning model during 2020-21.

Tables A-2 through A-5 show results from these analyses overall and for subgroups of students for third grade reading, third grade math, fourth grade reading, and fourth grade math respectively. With the addition of schools that changed to a virtual learning environment for 2020-21 for the most recent cohorts, estimated overall impacts of four years of AGR were worse across the four outcomes, though nearly all results remained not statistically different from zero. The only exception was third grade Forward reading results for Hispanic and urban (city) students with four years of AGR where the impact was negative and significant.

Table A-2: Alternative 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	FORWARD SCORE IMPACT (4 YEARS OF SAGE)	P-VALUE
ALL	0.07	0.99	-0.07	0.99	1.61	0.51
FRL	0.08	0.98	-0.72	0.79	1.24	0.69
ELL	-2.58	0.67	-0.01	1.00	5.55	0.14
Urban	-4.5	0.17	-0.96	0.79	0.67	0.89
Non-Urban	2.34	0.68	0.51	0.89	2.23	0.51
Black	0.71	0.79	0.89	0.59	3.17	0.73
Hispanic	-4.23	0.35	-2.49	0.37	0.43	0.95
White	1.16	0.85	0.78	0.79	2.3	0.39
Other Race	0.71	0.92	0.89	0.86	3.17	0.41
FRL & Urban	-3.63	0.36	-1.77	0.59	0.44	0.93
FRL & Non-Urban	2.17	0.66	0.01	0.99	1.85	0.69
FRL & Black	-0.99	0.90	-2.44	0.67	-1.84	0.81
FRL & Hispanic	-4.39	0.36	-2.77	0.36	0.43	0.95
FRL & White	1.82	0.69	0.2	0.96	2.22	0.51
FRL & Other Race	0.89	0.90	1.12	0.83	2.93	0.50
Urban & Black	-2.84	0.69	-3.25	0.55	-1.89	0.82
Urban & Hispanic	-9.54	0.02	-2.87	0.39	0.64	0.93
Urban & White	-3.25	0.50	1.1	0.79	1.31	0.79
Urban & Other Race	0.19	0.99	1.61	0.74	3.84	0.41
Non-Urban & Black	-3.93	0.73	1.34	0.87	-4.2	0.69
Non-Urban & Hispanic	3.85	0.52	-2.36	0.69	0.18	0.99
Non-Urban & White	2.61	0.65	0.77	0.85	2.57	0.46
Non-Urban & Other Race	-1.01	0.93	-0.68	0.95	2.14	0.84

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

Table A-3: Alternative Impact of AGR on Third Grade Forward Math Growth

SAMPLE	FORWARD SCORE IMPACT (4 YEARS OF AGR)		FORWARD SCORE IMPACT (MIX OF AGR/SAGE)		FORWARD SCORE IMPACT (4 YEARS OF SAGE)	
		P-VALUE		P-VALUE		P-VALUE
ALL	-1.44	0.69	-0.46	0.86	1.42	0.63
FRL	-4.1	0.19	-0.8	0.81	2.79	0.34
ELL	-1.85	0.86	0.19	0.99	6.97	0.15
Urban	-6.42	0.18	-0.16	0.99	3.15	0.37
Non-Urban	0.72	0.89	-0.73	0.83	0.11	0.99
Black	-3.27	0.35	0.68	0.99	4.17	0.52
Hispanic	-6.16	0.35	-1.58	0.74	3.39	0.50
White	0.55	0.91	-0.47	0.88	-0.03	1.00
Other Race	-3.27	0.62	0.68	0.90	4.17	0.34
FRL & Urban	-7.51	0.17	-0.66	0.90	4.47	0.24
FRL & Non-Urban	-2.51	0.52	-1.22	0.73	0.79	0.89
FRL & Black	-6.85	0.37	0.56	0.96	5.39	0.45
FRL & Hispanic	-7.62	0.29	-2.06	0.68	3.91	0.44
FRL & White	-1.89	0.63	-1.1	0.72	-1.89	0.86
FRL & Other Race	-5.28	0.38	0.23	0.99	4.2	0.39
Urban & Black	-8.28	0.30	-0.2	0.99	5.76	0.44
Urban & Hispanic	-11.66	0.13	-1.02	0.89	4.28	0.50
Urban & White	-0.84	0.91	-0.44	0.91	-0.56	0.91
Urban & Other Race	0.23	0.99	2.33	0.69	7.32	0.12
Non-Urban & Black	-8.5	0.93	-1.76	0.88	-1.01	0.83
Non-Urban & Hispanic	2.7	0.73	-3.04	0.57	1.95	0.83
Non-Urban & White	1.32	0.79	-0.42	0.92	1.32	0.98
Non-Urban & Other Race	-8.5	0.35	-1.76	0.83	-1.01	0.92

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

Table A-4: Alternative Impact of AGR on Fourth Grade Forward Reading Growth

SAMPLE	FORWARD SCORE IMPACT (4 YEARS OF AGR)		FORWARD SCORE IMPACT (MIX OF AGR/SAGE)		FORWARD SCORE IMPACT (4 YEARS OF SAGE)	
		P-VALUE		P-VALUE		P-VALUE
ALL	-2.44	0.37	-1.51	0.40	-2.36	0.35
FRL	-1.98	0.54	-2.25	0.31	-3.62	0.18
ELL	-9.52	0.11	-0.45	0.93	1.42	0.87
Urban	-2.71	0.63	-1.14	0.79	-2.99	0.43
Non-Urban	-2.36	0.46	-1.86	0.41	-1.98	0.56
Black	1.97	0.79	-4.43	0.36	-9.15	0.08
Hispanic	-5.96	0.32	-1.69	0.69	-0.74	0.92
White	-2.92	0.35	-1.32	0.51	-1.60	0.59
Other Race	-1.55	0.87	0.38	0.95	-2.30	0.70
FRL & Urban	-3.31	0.51	-2.27	0.57	-4.42	0.28
FRL & Non-Urban	-2.07	0.63	-2.79	0.26	-3.11	0.37
FRL & Black	2.32	0.73	-4.65	0.37	-9.29	0.08
FRL & Hispanic	-7.24	0.21	-1.90	0.69	-1.25	0.86
FRL & White	-2.07	0.62	-2.48	0.30	-2.66	0.41
FRL & Other Race	-1.48	0.89	0.24	0.99	-4.05	0.37
Urban & Black	0.89	0.91	-5.37	0.34	-9.15	0.11
Urban & Hispanic	-12.10	0.08	-1.28	0.86	-0.41	0.98
Urban & White	-1.45	0.69	1.79	0.69	-1.27	0.87
Urban & Other Race	-2.57	0.73	-0.67	0.93	-1.45	0.87
Non-Urban & Black	1.49	0.93	3.59	0.69	-5.88	0.63
Non-Urban & Hispanic	-0.03	1.00	-3.24	0.41	-1.46	0.87
Non-Urban & White	-2.99	0.35	-2.15	0.35	-1.91	0.57
Non-Urban & Other Race	-0.96	0.97	2.60	0.84	-3.68	0.77

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

Table A-5: Alternative Impact of AGR on Fourth Grade Forward Math Growth

SAMPLE	FORWARD SCORE IMPACT (4 YEARS OF AGR)		FORWARD SCORE IMPACT (MIX OF AGR/SAGE)		FORWARD SCORE IMPACT (4 YEARS OF SAGE)	
		P-VALUE		P-VALUE		P-VALUE
ALL	-1.53	0.73	-2.46	0.19	-2.50	0.31
FRL	-3.27	0.39	-2.71	0.20	-3.42	0.21
ELL	-6.69	0.37	-1.27	0.85	-1.79	0.83
Urban	-2.37	0.36	-0.75	0.87	-5.08	0.59
Non-Urban	-0.10	0.98	-3.69	0.08	-2.76	0.35
Black	-8.74	0.17	-4.29	0.47	-8.17	0.22
Hispanic	-6.46	0.35	-1.26	0.85	-1.42	0.86
White	0.53	0.92	-2.69	0.17	-1.70	0.55
Other Race	-2.19	0.85	-1.41	0.81	-4.01	0.42
FRL & Urban	-6.99	0.17	-1.62	0.72	-3.75	0.36
FRL & Non-Urban	-1.42	0.83	-4.15	0.07	-3.24	0.33
FRL & Black	-9.03	0.17	-4.48	0.43	-8.25	0.21
FRL & Hispanic	-8.05	0.20	-1.70	0.79	-1.50	0.87
FRL & White	-0.35	0.96	-3.23	0.14	-2.29	0.50
FRL & Other Race	-1.31	0.91	-1.31	0.84	-5.35	0.32
Urban & Black	-9.89	0.14	-4.88	0.45	-8.63	0.23
Urban & Hispanic	-13.20	0.06	-0.20	0.99	-0.72	0.95
Urban & White	9.22	0.10	1.24	0.82	-0.20	0.99
Urban & Other Race	-1.58	0.86	-0.94	0.90	-0.95	0.92
Non-Urban & Black	-4.85	0.87	0.06	1.00	-0.40	0.99
Non-Urban & Hispanic	1.67	0.87	-3.81	0.40	-2.63	0.73
Non-Urban & White	-0.11	0.99	-3.85	0.08	-2.45	0.41
Non-Urban & Other Race	-4.01	0.83	-2.23	0.83	-10.18	0.23

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

Strategy Analysis

This analysis only includes AGR schools with no comparison schools. Each school serves as its own control observation – the analysis compares outcomes before the strategy change to outcomes after the change. The analysis includes the years 2017-18 through 2019-20. In the years prior to 2017-18, high quality information on strategy use was not available. For the 2020-21 school year, the pandemic resulted in lower MAP and STAR participation as well as fluctuations in attendance and discipline rates. As a result, this analysis does not include the 2020-21 school year. Since this evaluation measures year-to-year changes in strategies, it requires at least two consecutive years of data. As a result, the analysis also omits 2021-22.

Because the analysis identifies strategies' impacts on year-to-year changes in outcomes, reading and math growth outcomes use assessment data from two local assessments, MAP and STAR, which many schools administer in fall and spring each year.³³ While some readers may be familiar with these assessments' scaling and typical growth, other readers may lack this familiarity. To provide interpretability of results across assessments, we standardize assessment scores to have a mean of zero and standard deviation of one. For each test window, subject, and grade combination (e.g., Fall 2nd grade MAP), we take each individual's score, subtract the mean test score across all test takers, then divide by the standard deviation of the score across all test takers.

For the MAP assessment, we use 2015 national norms available from the vendor, NWEA.³⁴ These national norms include means and standard deviations for each subject, grade, and test window combination. NWEA created these 2015 norms from a national sample of data from fall of 2011 through spring of 2014. NWEA calculated 2020 norms from a national sample of data from fall of 2015 through spring of 2018.³⁵ While the 2020 norms provide a more recent estimate of means and standard deviations throughout the tested MAP population, we continue to use 2015 norms for two main reasons. First, while Wisconsin represents only a fraction of the NWEA testing population, making it unlikely that any impacts of AGR may be present in the national sample of MAP norms, we want to exclude any possibility. Second, we used 2015 norms in prior years of the evaluation. Since 2020 norms are different from 2015 norms, we select the 2015 norms to retain consistency in our estimates of AGR impacts.

³³ We use both MAP and STAR to increase the number of schools in the analysis.

³⁴ Thum, Y. M., & Hauser, C.H. (2015). *NWEA 2015 MAP Norms for Student and School Achievement Status and Growth*. Northwest Evaluation Association. https://www.elmbrookschoools.org/uploaded/SSMigration/data/files/gallery/ContentGallery/NWEA_2015_Full_Norming_Study.pdf

³⁵ Thum, Y. M., & Kuhfeld, M. (2020). *NWEA 2020 MAP Growth: Achievement Status and Growth Norms for Students and Schools*. NWEA. <https://teach.mapnwea.org/impl/normsResearchStudy.pdf>

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

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

Equation (4)

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

where Y is the standardized STAR or MAP spring test score in either subject for student i in school s , grade g , and academic year y .³⁶ On the right side of Equation (4), the superscripts f , w , m , and r refer to fall, winter, math, and reading, respectively, so that, for example, f,m denotes the fall math test score. $X_{i,y}$ represents a vector of student-level covariates, and $Z_{s,y}$ represents a vector of school-level covariates. Student-level covariates include gender, race/ethnicity, FRL status, English learner status, and special education status. School-level covariates include percentages of the student-level covariates, school population, and average fall test scores. We include grade-by-year fixed effects $\delta_{g,y}$ to control for unobserved, statewide effects that vary by grade and/ or academic year. For the final predictions of Spring 2020 assessment scores used in the analysis models, we use both Fall 2019 and Winter 2020 assessments to capture actual test score growth that occurred before the COVID-19 school closures and use that actual growth trajectory to predict the growth trajectory that is most likely to have occurred between Winter and Spring 2020.

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

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

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

Equation (5)

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

where $Y_{i,s,g,y}$ is an outcome for student i in grade g , school s , and year y . $\text{class size}_{s,g,y}$, $\text{coachings}_{s,g,y}$, and $\text{tutorings}_{s,g,y}$ are indicators for whether a school employed that particular strategy, in each grade and year. X_{iy} represents a vector of student-level covariates, including lagged values of the outcome Y . Grade-by-year fixed effects, $\delta_{g,y}$, are included to control for any unobserved, statewide effects that vary by grade and/or time. School fixed effects, π_s , are included to control for any unobserved effects that vary by school. Standard errors are clustered at the school level. Models for MAP/STAR math and reading and absence rate use a linear regression, and the suspension rate model, where the outcome is an indicator of whether a student received at least one suspension during the year, uses a logit specification. To account for the non-linearity of absence rate as an outcome, we first converted absence rates onto the standard normal distribution using a probit transformation, then after estimation used an inverse transformation for interpretable results.

Multiple Comparisons Analysis

Estimating multiple impact models, as this report does, increases the likelihood for false positives – results that are statistically significant due to random chance rather than actual program impacts. For example, a 0.05 significance level implies that 5 percent of statistically significant estimates are produced by random chance. To adjust for potential false positives, we apply the Benjamini-Hochberg procedure, a common method of correcting for multiple comparisons by accounting for the total number of statistical tests as well as the strength of the estimates.³⁷ 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.

³⁷ 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>

Analysis of AGR Implementation Changes During COVID and their Relationship to School Learning Models

This report also examined whether schools changed AGR strategies in response to the COVID-19 pandemic, and if so, whether strategy changes were related to choices of virtual, in-person, and hybrid learning models during 2020-21. We also explored whether schools returned to pre-COVID AGR strategies during 2021-22. To answer these questions, this section analyzed school learning model data from 2020-21 as well as AGR school strategy data from 2019-20 to 2021-22.

Schools' chosen AGR strategies remained relatively stable over the sample period. Approximately 17 percent of schools from 2019-20 to 2020-21 and 22 percent of schools from 2020-21 to 2021-22 changed their AGR strategy, while about 45 percent of schools did not change their strategy at all (Table A-6).

AGR school strategies over the sample period were examined by school learning mode during the 2020-21 school year. Schools that reported having less than 30 percent of their school days in-person were labeled as "mostly virtual," schools with more than 90 percent of their school days reported as in-person were labeled as "mostly in-person," and schools that fell in-between were labeled as "mixed."

Table A-6: Percentage of AGR Schools that Changed Strategies during COVID

CHANGE	PERCENTAGE OF SCHOOLS
2019-20 to 2020-21	16.8%
2020-21 to 2021-22	22.4%
Changed Both Years	15.5%
No Change	45.3%

Tables A-7 and A-8 show that, during the sample period, schools that were mostly virtual did not report any substantial changes in their AGR strategy choices. Tables A-9 and A-10, showing data from mostly in-person schools, and Tables A-11 and A-12, showing data for mixed schools, reveal similar results. Overall, the analysis showed no substantial results, meaning that there is no evidence that the pandemic significantly impacted AGR schools' strategy choices.

Table A-7: Mostly Virtual Schools' Strategy Changes From 2019-20 to 2020-21

STRATEGY FOR 2019-20	2020-21 STRATEGY							
	ALL	COACHING	CLASS SIZE	TUTORING	CLASS SIZE & COACHING	CLASS SIZE & TUTORING	COACHING & TUTORING	TOTAL
All	9	0	0	0	1	0	0	10
Coaching	0	2	0	0	3	0	0	5
Class Size	0	0	2	0	2	0	0	4
Tutoring	0	0	0	0	0	0	0	0
Class Size & Coaching	0	4	3	0	80	0	0	87
Class Size & Tutoring	0	0	0	0	0	0	0	0
Coaching & Tutoring	0	0	0	0	0	0	0	0
Total	9	6	5	0	86	0	0	106

Table A-8: Mostly Virtual Schools' Strategy Changes From 2020-21 to 2021-22

STRATEGY FOR 2020-21	STRATEGY FOR 2021-22							
	ALL	COACHING	CLASS SIZE	TUTORING	CLASS SIZE & COACHING	CLASS SIZE & TUTORING	COACHING & TUTORING	TOTAL
All	8	0	0	0	1	0	0	9
Coaching	0	4	0	0	2	0	0	6
Class Size	0	0	3	0	2	0	0	5
Tutoring	0	0	0	0	0	0	0	0
Class Size & Coaching	2	28	2	0	54	0	0	86
Class Size & Tutoring	0	0	0	0	0	0	0	0
Coaching & Tutoring	0	0	0	0	0	0	0	0
Total	10	32	5	0	59	0	0	106

Table A-9: Mostly In-Person Schools' Strategy Changes From 2019-20 to 2020-21

STRATEGY FOR 2020-21								
STRATEGY FOR 2019-20	ALL	COACHING	CLASS SIZE	TUTORING	CLASS SIZE & COACHING	CLASS SIZE & TUTORING	COACHING & TUTORING	TOTAL
All	25	2	4	0	7	2	3	43
Coaching	0	10	1	0	0	0	0	11
Class Size	0	1	30	0	1	6	0	38
Tutoring	0	1	0	0	1	0	1	3
Class Size & Coaching	4	1	7	0	19	0	0	31
Class Size & Tutoring	5	0	8	0	0	5	0	18
Coaching & Tutoring	1	0	0	1	0	0	5	7
Total	35	15	50	1	28	13	5	151

Table A-10: Mostly In-Person Schools' Strategy Changes From 2020-21 to 2021-22

STRATEGY FOR 2021-22								
STRATEGY FOR 2020-21	ALL	COACHING	CLASS SIZE	TUTORING	CLASS SIZE & COACHING	CLASS SIZE & TUTORING	COACHING & TUTORING	TOTAL
All	19	0	3	0	2	4	7	35
Coaching	1	7	0	0	7	0	0	15
Class Size	4	0	33	0	8	5	0	50
Tutoring	0	0	0	1	0	0	0	1
Class Size & Coaching	6	4	3	0	14	0	1	28
Class Size & Tutoring	1	0	6	0	2	4	0	13
Coaching & Tutoring	4	0	0	1	0	0	4	9
Total	35	11	45	2	33	13	12	151

Table A-II: Mixed Schools' Strategy Changes From 2019-20 to 2020-21

STRATEGY FOR 2020 - 21								
STRATEGY FOR 2019-20	ALL	COACHING	CLASS SIZE	TUTORING	CLASS SIZE & COACHING	CLASS SIZE & TUTORING	COACHING & TUTORING	TOTAL
All	17	3	2	0	9	6	1	38
Coaching	0	19	0	0	6	1	1	27
Class Size	2	0	15	0	0	0	0	17
Tutoring	0	0	0	0	0	0	0	0
Class Size & Coaching	5	3	7	0	24	2	0	41
Class Size & Tutoring	0	0	4	0	0	4	0	8
Coaching & Tutoring	5	0	0	0	0	0	0	5
Total	29	25	28	0	39	13	2	136

Table A-I2: Mixed Schools' Strategy Changes From 2020-21 to 2021-22

STRATEGY FOR 2021-22								
STRATEGY FOR 2020-21	ALL	COACHING	CLASS SIZE	TUTORING	CLASS SIZE & COACHING	CLASS SIZE & TUTORING	COACHING & TUTORING	TOTAL
All	18	0	1	0	3	0	7	29
Coaching	0	21	0	0	3	0	1	25
Class Size	3	1	16	0	4	4	0	28
Tutoring	0	0	0	0	0	0	0	0
Class Size & Coaching	2	1	2	0	32	2	0	39
Class Size & Tutoring	3	0	2	0	2	6	0	13
Coaching & Tutoring	1	1	0	0	0	0	0	2
Total	27	24	21	0	44	12	8	136



Wisconsin Evaluation Collaborative