Evaluation of the Alabama Accountability Act: Academic Achievement Test Outcomes of Scholarship Recipients through 2020-2021

The Institute for Social Science Research
The University of Alabama

Erika Steele, Ph.D.
Joan M. Barth, Ph.D.

September 1, 2022

Executive Summary

This report fulfills the Alabama Accountability Act (AAA) evaluation requirements by examining the academic achievement of scholarship recipients through the 2020-2021 academic year.

The report has three objectives:

1. Describe the academic achievement of students in the scholarship program.
2. Compare scholarship recipients to Alabama public school students.
3. Assess changes in achievement across time.

COVID-19 impacted academic achievement for students throughout the country due to disruptions to normal schooling. The results of this report must be considered in the context of the pandemic.

Scholarship Granting Organizations (SGOs) provided demographic information and achievement test scores for scholarship recipients. Achievement test score information for Alabama public school students was retrieved from the Alabama State Department of Education (ALSDE) website, the Public Affairs Research Council of Alabama, the ACT Inc, and the College Board.

- **The SGOs provided information on 2968 student scholarship recipients for 2020-2021.**
- **2,203 recipients were in grades 2-8, 10, and 11 and required to submit test scores.**

Methodological Limitations

- The lack of a uniform standardized test among schools continued to constrain the accurate assessment of scholarship recipients' academic achievement and the comparisons that could be made to Alabama public school students.
  - Norm-referenced tests and criterion-referenced tests are based on different standards and cannot be directly compared.
  - Schools using the same test often reported scores based on different national norms and these cannot be combined.
  - Some achievement tests were used by only one school or included only a small number of students, making analyses unreliable.
  - ALSDE used a new test in 2020-2021 for grades 2 through 8, the Alabama Comprehensive Assessment Program (ACAP). There are no longitudinal data for ACAP, and because only a few AAA students took this test, no direct comparisons could be made for these grades.
- Inconsistencies in test score reporting from schools and missing test data limited the number of students who could be included in the evaluation sample.

After accounting for these issues, the evaluation was based on 1,658 scholarship recipients attending 90 schools in 36 counties. This represented 75% of the scholarship recipients in the grades for which testing was required. Students varied in their demographic characteristics:

- Number of years receiving a scholarship:
  - 4% were first-time scholarship recipients.
  - 29% were in their 7th year or more as a scholarship recipient.
  - 4.4 was the average number of years of being in the scholarship program.
- 94% were eligible for free/reduced lunch subsidies.
- 60% were Black/African American (AA), 19% were White, and 19% were Hispanic.

Objective 1: Describe the academic achievement of students in the scholarship program.

- On norm-referenced tests, scholarship students as a group did not perform as well as students in the U.S. taking the same test.
  - Typically, the mean percentile scores across tests were significantly below the 50th percentile.
  - There were some exceptions in which mean scores were not below the 50th percentile for small numbers of students on specific tests. Variability among the grades, subject areas, and demographic groups in which these scores occurred revealed no reliable pattern.
- On criterion-referenced tests:
  - For students in grades 3-8 who took the ACT Aspire, the majority of scholarship recipients met the proficiency benchmarks for English but failed to meet proficiency benchmarks for Reading and Math.
  - For 10th graders who took high school college entrance exams, the majority met benchmarks for Reading, English or Reading-Writing assessments, but most failed to make benchmarks in Math.
  - For 11th graders who took high school college entrance exams, the findings were mixed: On the SAT and ACT, the majority of students did not meet benchmarks in each subject area; on the PSAT/NMSQT, 51% met benchmarks for Reading-Writing, but the majority failed to meet benchmarks in Math.
- Generally, when comparisons could be made among race/ethnicity groups, outcomes were poorer for Black/AA participants compared to Hispanic and White students.

Objective 2: Compare the learning achievement of scholarship recipients to students attending public schools.

- No comparisons could be made for grades 2-8 due to a lack of a common test.
- For 11th graders taking the ACT, scholarship recipients generally performed similarly to economically disadvantaged public school students.
  - Mean Reading scores were higher for scholarship students than economically disadvantaged public school children, but the percentage of students meeting benchmarks in Reading was similar.
- For 10th and 11th graders taking the PSAT/NMSQT, Reading-Writing and Math performance was poorer than comparable economically disadvantaged public school children in Alabama.
  - Mean Reading-Writing scores for both grades were above benchmarks.

Findings for Objective 3: Assess changes in achievement across time.

- On average, the number of years a student participated in the scholarship program was not strongly associated with significant improvement on standardized test scores.
- The analyses of the ACT mean scores and proficiency rates did not show a consistent pattern of improvement over time, especially in more recent years.

Table of Contents

List of Charts and Tables

Page

List of Abbreviations

AAA	Alabama Accountability Act
AA	African American
ACAP	Alabama Comprehensive Assessment Program
AL	Alabama
ALSDE	Alabama State Department of Education
CAT	Computer adaptive test
Econ. Dis/Disadv	Economically Disadvantaged
ELA	English Language Arts
EBRW	Evidenced Based Reading-Writing
FERPA	Federal Education Rights and Privacy Act
ISSR	Institute for Social Science Research
K-12	Kindergarten through 12th grade
N	Number of people in a group
n	Number of people in a subgroup
NA	Not applicable
NAEP	National Assessment of Educational Progress
PARCA	Public Affairs Research Council of Alabama
PDF	Portable Document Format
PSAT/NMSQT	The Preliminary SAT/National Merit Scholarship Qualifying Test
r	Correlation coefficient
S	Spring norms, for example Iowa S2011 means Iowa Spring 2011 norms
SAT	Scholastic Aptitude Test
SAT-10	Stanford Achievement Test-10
SGO	Scholarship Granting Organization

Evaluation of the Alabama Accountability Act: Academic Achievement Test Outcomes of Scholarship Recipients through 2020-2021

Introduction

This report fulfills the state-mandated evaluation of the academic outcomes of students receiving scholarships under the Alabama Accountability Act (AAA) as set forth in the AAA legislation. It follows a series of reports starting in 2016 authored by the Institute for Social Science Research (ISSR) at the University of Alabama. These biannual reports described the achievement test results from the 2014-2015 through the 2018-2109 academic years, compared the outcomes to students attending public schools in Alabama, and examined changes in scholarship recipients' achievement test scores over time. The 2022 report examines these same issues for the 2020-2021 academic year.

This report first provides an overview of the pertinent AAA legislation. The methodology is described next, which includes a description of the 2020-2021 sample and the achievement tests that are part of this report. The findings are organized around three objectives: 1) describe the academic achievement of students receiving tuition scholarships in the 2020-2021 academic year, 2) compare their performance to public school children, and 3) examine changes in achievement over time. The conclusion of the report summarizes the overall impact of the AAA scholarship program on student academic achievement.

Overview of AAA

The Alabama Accountability Act (AAA), passed by the legislature in 2013 and amended in 2015, established a statewide scholarship program for low-income students to attend public or private schools. The scholarship program is funded by a tax credit program, and the scholarship awards are managed by Scholarship Granting Organizations (SGOs), which must comply with standards set by the AAA. All students receiving scholarships must meet family income eligibility requirements. Priority is given to students who are zoned to attend a failing public school as designated by the Alabama State Department of Education (ALSDE). However, students meeting AAA income requirements who attend non-failing public schools may receive scholarships if additional funds are available. Scholarships are awarded from the SGO to the student to attend a school that must meet standards set forth in the AAA. Scholarships may cover all or part of tuition and mandatory fees for one academic year. In 2015, the legislature amended the AAA to place limits on the amount that could be awarded to a student depending on the grade level (elementary, middle, or high school). The Alabama State Department of Revenue oversees the implementation of the AAA. This report fulfills the evaluation component of the 2013 Alabama Accountability Act by providing evidence for the academic achievement of scholarship recipients in the 2020-2021 academic year.

Scholarship Recipient Testing Requirements

The academic accountability standards require the SGOs to ensure that schools accepting scholarship students "annually administer either the state achievement tests or nationally recognized norm-referenced tests that measure learning gains in math and language arts to all students receiving an educational scholarship in grades that require testing under the accountability testing laws of the state for public schools." The purpose of these tests is to assess the learning gains for scholarship recipients and to provide a means of comparing scholarship recipients to students who attend Alabama public schools.

Evaluation Reporting Requirements

The AAA states that the evaluation shall include the following:

- The learning achievements of students receiving educational scholarships, aggregated by grade level, gender, family income level, number of years of participation in the tax credit scholarship program, and race of the student receiving an educational scholarship.
- A comparison of the learning gains of students participating in the tax credit scholarship program to the statewide learning gains of public school students with socioeconomic and educational backgrounds similar to those students participating in the tax credit scholarship program.
- A report to be made every two years, starting in 2016.

Following these requirements, this report has three objectives: a) describe the academic achievement of students in the scholarship program for the 2020-2021 school year, b) make comparisons between the level of achievement of the scholarship recipients and comparable students attending public schools for the 2020-2021 school year, and c) measure the achievement gains of students in the scholarship program over time.

Alabama State-Mandated Testing in Public Schools 2020-2021 Academic Year

The Alabama State Department of Education (ALSDE) assesses children in grades 2 through 8 using the Alabama Comprehensive Assessment Program (ACAP). ACAP is an online assessment designed to provide state stakeholders with information regarding student progress toward mastery of the Alabama Course of Study Standards. Students in grades 2 through 8 take assessments covering English Language Arts and Math. Alabama tenth graders took the PreACT and eleventh graders were required to take the ACT college entrance exam. Tests are typically administered during the spring semester in March and April.

Impact of COVID-19

The COVID-19 global pandemic had an enormous impact on education in Alabama and the entire U.S. during the 2019-2020 school year. The State of Alabama closed all public schools for in-person instruction in March 2020. The remainder of the spring semester was conveyed in a variety of formats throughout the state, with some districts attempting to teach virtually, others relying on school work being delivered to children in a paper format, and some using a combination. On March 27, 2020, the

U.S. Department of Education approved the state's request to waive federally required student assessments and other measures of student achievement for students in grades K-12. Thus, standardized testing, including college achievement and entrance exams, that typically occurs in the latter half of the spring term was cancelled. The majority of private schools attended by scholarship recipients were also closed during this time and consequently did not test students. The lack of test data for the 2019-2020 academic year impacts Objective 3 of this report, which examines the change in scholarship students' academic achievement over time. In the past, through various approaches, we tried to track change in student scores by comparing test performance over consecutive years. The missing 2019-2020 data created a serious obstacle for this approach, and so alternative methods were used to address this objective.

During the 2020-2021 academic year, schools throughout the state continued to react to health and safety concerns as the pandemic impacted the lives of students, educators, and families. A variety of educational approaches was employed throughout Alabama, including virtual learning, staggered in-person schooling, hybrids of virtual and in-person learning, and regular in-person learning. Moreover, the approach changed throughout the year resulting in rolling shutdowns of in-person schooling and other adjustments as health concerns waxed and waned. Nationally, the negative impact of the disruption to in-person learning on academic achievement continued to be evident in the 2020-2021 academic year, especially for children in lower grades (Allen, 2021) and students who come from economically disadvantaged families (Kuhfeld et al. 2022). Beyond instruction, schools faced higher rates of misbehavior, violence, and mental health issues, which may have also contributed to poorer achievement (Kuhfeld et al. 2022).

As findings are presented throughout this report, it is important to keep in mind that all results and conclusions must be considered in the context of the pandemic. In subsequent reports, we may be better able to understand the impact of COVID-19 on the students in the AAA program.

Method

The methodology for the 2022 report follows that of previous years, and similarly, the conclusions that can be drawn from this report are limited in several ways by the nature of the testing data that are reported to the evaluation team. These are briefly discussed as they remain largely unchanged from previous reports. A major limitation in the reporting of the results is the lack of a uniform achievement test among schools, which constrains the conclusions that can be made about student achievement outcomes and the types of comparisons that can be made to students attending public schools. Schools provided scores from 21 standardized tests. Comparisons across tests are invalid because tests vary in their content and are designed for unique purposes. *Norm-referenced tests*, such as the Iowa Assessment and the Stanford Achievement Test, and *criterion-referenced* tests, such as the ACT Aspire and the ACT college entrance exam, are based on different standards and cannot be directly compared. Norm-referenced tests are designed to compare student achievement relative to others at a particular grade level and distinguish between high and low achievers. For example, students scoring at the 70th percentile on a norm-referenced test achieved a score that was better than or equal to 70 percent of students in the nation at their grade level taking the same test. Interpreted alone, the percentile scores do not indicate if a child has acquired the academic skills and content that are appropriate for his or her age group. In contrast, criterion-referenced test scores typically describe student success in terms of meeting achievement readiness benchmarks that indicate if the student is on track to meeting a long-term academic goal, such as admission to college. In theory, 100% of students could achieve these criterion benchmarks. In criterion-

referenced tests, the emphasis is on achieving scores that meet benchmarks, and consequently, percentile scores are less meaningful with respect to achievement. Even tests within the same broad categories of norm- or criterion-referenced cannot be combined for analyses since each test has unique content and unique scoring systems.

Additionally, some tests were used by only one school or taken only by a small number of students. Small numbers for some grade levels and demographic groups also make comparisons potentially unreliable. Guidance from ACT Inc. recommends a sample of at least 25 students, and this standard was adopted in this report.

Even when the same test is used across schools, students at different schools at times have scores that are based on different norms. Some schools report achievement scores using outdated norms when more up-to-date norms are available. For example, one school may report test scores for the Stanford Achievement Test based on 2018 norms, while another school may report scores based on 2002 norms. These are not comparable because the older tests are not based on the Common Core, the current national standards for children in grades K-12. Using these older norms makes it difficult to know accurately how well today's students are performing. The 2022 report focuses on test score data based on the most recent norms used by the schools so that a more accurate assessment of scholarship students' academic performance can be given.

Every year the evaluation team communicates with the SGOs about the specific test scores that should be reported in the test reports, including the subject areas (Reading, Language/English, and Math) and types of scores (national percentiles and scale scores). These expectations are communicated to the schools. School adherence to these guidelines has improved over time, but missing data continues to compromise the integrity of the findings, sometimes resulting in entire schools being excluded from the report.

With these challenges noted, the remainder of the report describes outcomes for the 2020-2021 academic year.

Data Sources

The following data sources were used to evaluate the academic achievement of the 2020-2021 scholarship recipients:

- Demographic reports from each year of the program from eight SGOs: Scholarships for Kids, C2 Opportunity Scholarship Fund, Bama Works Fund, Academics Plus, Alabama Opportunity Scholarship Fund, Rocket City Scholarship Granting Organization, Children's Tuition Fund, and Renaissance Scholarship Fund.
- Test reports collected by the SGOs from participating schools and shared with ISSR. Test scores were received as PDFs.
- 2020-2021 Alabama Comprehensive Assessment Program results available from the ALSDE website.
- Eleventh (11^{th}) grade ACT scores for public school students in Alabama retrieved from the Public Affairs Research Council of Alabama (PARCA) report available on their website.

- National and state scores on the Practice SAT-National Merit Scholarship Qualifying Test (PSAT/NMSQT) scores available from the College Board website.

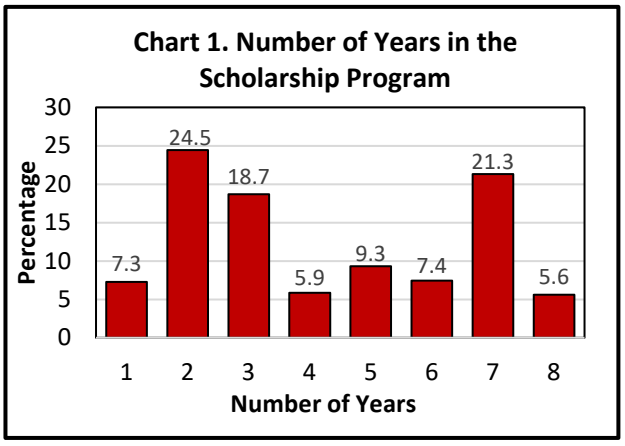
Statistical Analyses

Statistical comparisons were conducted throughout the report to aid in drawing conclusions. These statistical tests consider the sample size and the variation in the data to inform us of the likelihood of a reliable difference. As is customary in educational research, a probability value (*p)* of $\leq .05$ was used as the criterion to determine significance.

- T-tests were used to compare mean scholarship student test scores to established benchmarks, to compare genders, or to compare racial/ethnic groups of scholarship students.
- Analysis of Variance (ANOVA) was used to compare the mean scores of multiple groups, such as racial groups.
- Chi-Square analyses were used to compare demographic groups on the percentages of students meeting a benchmark score.
- Z-tests were used to compare the percentages of scholarship students meeting benchmarks to comparable indicators of public school students.
- Correlations (*r*) assessed the relation between achievement test scores and the number of years of participation in the AAA scholarship program.

2020-2021 Sample

The first part of this report focuses on the new data from the 2020-2021 academic year, as earlier reports analyzed the previous academic years. The SGOs provided information on 2968 students who had received scholarships during the 2020-2021 academic year. This group of students was 50% female and in kindergarten through 12th grade. Chart 1 graphically illustrates the number of years participants had been in the AAA scholarship program. Only 7% were in their first year, indicating that the majority of the students (93%) had received at least one previous scholarship. Half of the students were in their 4th year or more of receiving a scholarship, and 27% had participated for seven or more years. The average number of years in the program was 4.2. As in previous years, the scholarship recipients primarily represented three racial/ethnic groups, Black/African American (Black/AA; 64%), White (19%), and Hispanic (15%). The remaining 2% of the sample were another race or no information was provided. Ninety-five percent (95%) were free/reduced lunch eligible. Students resided in 42 counties in the state and attended 121 different schools.



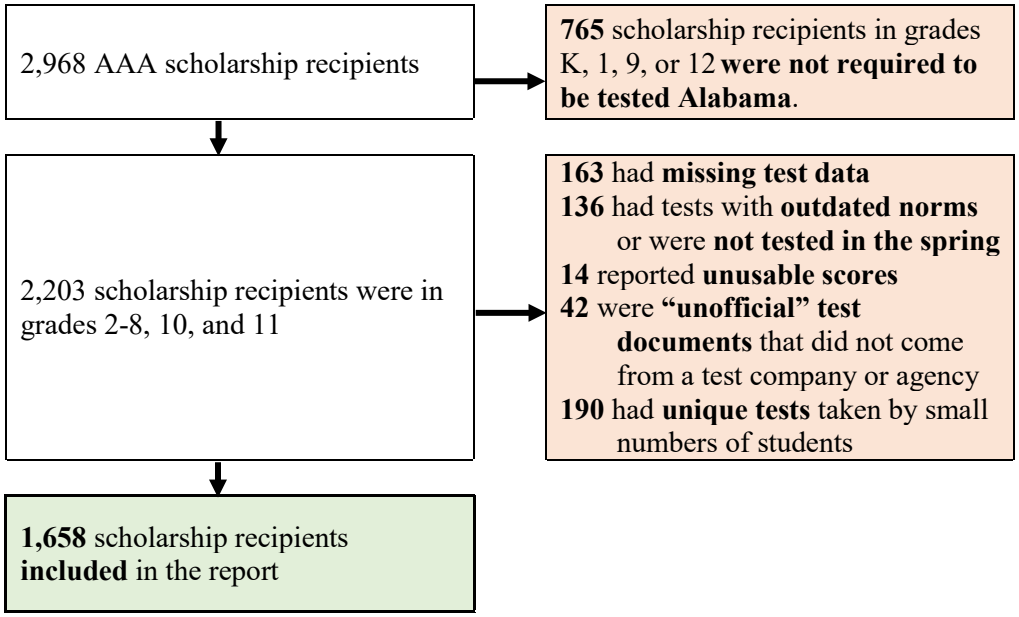
Achievement Test Data for 2020-2021 Scholarship Participants

A total of 2,203 scholarship students were in grades 2 through 8, 10, and 11, which are the grades the state requires to be tested for Language Arts and Math. These grades are the focus of this report. Students in grades kindergarten, first, ninth, and twelfth comprised 26% (n = 765) of scholarship recipients and were not required to be tested according to the AAA.

Data for 545 students in grades required to test were not included in the report for several reasons. The flowchart below summarizes factors affecting the 2021-2022 sample size.

1. Test data were missing for 163 students (7% of those required to test) for a number of reasons: the student withdrew before testing, the school did not test the student, the student was absent for testing, the school did not submit scores to the SGO, or there was no explanation for the missing test. The most common explanation, accounting for 53% of the missing tests, was a student withdrawing from the school before the tests were administered. It should be noted that the number of missing tests has decreased over time. ISSR will continue to work with the SGOs to ensure that all students who are in grades that are tested in the State of Alabama take a standardized test or the appropriate alternate assessment.
2. Some schools administered tests during the fall or winter or used outdated norms when more recent norms were available (e.g., using Stanford-10 2002 instead of Stanford-10 2018 norms). Fall and winter scores assess achievement partway through the school year and are not comparable to those taken at the end of the school year in the spring. The ALSDE conducts testing in April for grades 2 through 8, and so the more appropriate comparison group of AAA scholarship recipients are students tested in the spring. This excluded 136 students from the analyses, approximately 6% of students required to test.
3. Further attrition occurred because schools did not include scores for Reading, Math, or English/Language Arts (n = 11) or did not include percentile scores (n = 3) where these were needed to interpret the test results.
4. In addition, several schools provided student data that were "unofficial" reports from the schools that did not come directly from the standardized testing company (e.g., Scantron, Iowa, MAP Growth, ACT) or a testing administration company (e.g., Seton, Abeka) that administers and assesses standardized tests for home- and private-schools (n = 42). These issues accounted for 3% of students required to test. ISSR will communicate to the schools through the SGOs the expectation that schools administer tests in the spring, use the most recent norms available for a test, and provide test scores that can be verified.
5. Finally, 21 different standardized tests were given by 130 different schools, and as with previous reports, some schools used tests that few schools or no other school used. These schools typically had few scholarship recipients. Making these test results public is not desirable for two reasons: a) Schools and individual children could be identifiable, a violation of FERPA; and b) Small samples, as noted earlier, are not likely to be representative of the full group of scholarship recipients. Therefore, 190 students (9% of those required to test) attending these schools were excluded from this evaluation.

Flowchart: Student Inclusion Process



A total of 1,658 students or 75% of students for whom testing was required according to the AAA had potentially reportable test data from eleven standardized tests. Table 1 indicates the number of students who took each test and the number of schools represented by each test. Collectively, students in this group attended 90 unique schools. The discrepancy between this total and the numbers listed in Table 1 is due to some schools giving more than one test (e.g., a K-12 school might give the ACT Aspire for grades 3 through 8, the PSAT/NMSQT for grade 10, and the ACT for grade 11).

Table 1. Tests Included in the Evaluation for Grades 3 through 8, 10, and 11

Test	Number of Students	Number of Schools
1. ACT	110	31
2. ACT Aspire	130	11
3. Iowa Assessment 2017 Norms	662	36
4. Measures of Academic Progress (MAP) Growth	195	17
5. PreACT	72	10
6. PSAT/NMSQT	84	13
7. Scantron	113	2
8. Scholastic Aptitude Test (SAT)	32	4
9. Stanford Achievement Test 10 2018 Norms	143	17
10. STAR	77	3
11. TerraNova 3	40	2
Total	**1658**	

Description of Tests

Nearly all of the achievement tests purport to base their test questions on nationally recognized educational standards, such as those of the National Assessment of Educational Progress (NAEP). They provide a score, such as a national percentile, that can be used to evaluate student performance relative to other students in the U.S. A child who scores at the 50th percentile is performing as well as or better than half of the students in the nation who are at the same grade level. Scale scores are derived from the number of items answered correctly and are often used to determine if students are meeting grade-level benchmarks on criterion-referenced tests or to track progress over time. Generally, scores on these tests are used to assess whether students or school systems have met requirements set by national or state standards and consequently, meet the testing requirement put forward in AAA. A brief description of each of the 11 tests taken by AAA scholarship recipients follows. Additionally, although only one school with too few students (n < 10) took the ACAP, a description is provided because this is the test that Alabama public school children in grades 2 through 8 took.

Many of the tests are computer adaptive (CATs) in which students are given harder or easier questions as they proceed through the test based on whether their answers are correct (resulting in harder questions) or wrong (resulting in easier questions). In comparison to fixed form tests (in which all students are shown the same questions), CATs may be better at providing insight into student achievement by reducing the impact of test anxiety, boredom, and guessing. CATs can be criterion-referenced (ACAP), norm-referenced (Scantron), or have features of both (MAP Growth and STAR).

Criterion-Referenced Tests

- The *ACT Aspire* assesses progress toward college and career readiness. Benchmarks are used to evaluate if a student is on track to succeed in college. Scale scores are used to assess students' performance against a set of learning standards for each grade level. As such, ACT Aspire scores are labeled criterion-referenced, and it is possible for every child to get a score that meets the benchmark. The ACT Aspire includes test scale scores for Reading, English, and Math, in addition to other areas. National percentile scores are also provided that are used to compare a student's performance on the test to similar students. The scale score places students into one of four readiness levels that are used to predict the likelihood a student will succeed in future college courses: 1) *Exceeding*, 2) *Ready*, 3) *Close*, or 4) *In Need of Support*.
- The *PreACT* is used to prepare high school students to take the ACT college entrance exam. The scores can be used to predict how well a student might perform on the ACT college entrance exam. Reports include an estimated ACT score (1-36) and a national percentile score. Subscales are provided for Reading, English, and Math. Proficiency benchmarks are provided by ACT Inc. for both 10th and 11th grades to assess college readiness. Student performance on the PreACT can be categorized into one of three readiness levels, 1) *On target*, 2) *On the cusp*, or 3) *In need of intervention*.
- The *ACT* is a nationally normed college entrance exam, usually taken by high school juniors and seniors to predict college readiness. Reports include an ACT score (1-36), which can be used to determine college readiness (criterion-referenced score), and a national percentile

score. Subscale scores are provided for Reading, English, and Math. ACT Inc provides college readiness benchmarks and has set proficiency benchmarks for high school students.

- The *Alabama Comprehensive Assessment Program (ACAP)* is a criterion-referenced assessment designed to measure grade level performance in English Language Arts and Math based on standards set by the Alabama Course of Study for students in grades 2 through 8. Students in grades 4, 6, and 8 are also assessed in science. Scaled scores on the ACAP range from 250-800 and are used to determine grade level-proficiency.
- The *Practice SAT-National Merit Scholarship Qualifying Test (PSAT/NMSQT)* is used to prepare students to take the Scholastic Aptitude Test (SAT) college entrance exam and is usually taken in the 10th and 11th grades of high school. The scores include a composite score that aligns with a predicted SAT score. The composite score is the sum of the Math and Evidence-Based Reading and Writing (EBRW) scores. Scores on the Math and EBRW sections range from 160-760. Benchmarks are provided to assess students' college readiness. In addition, national percentile scores that allow students to be compared to similar students in the nation are provided for all subject areas.
- The *Scholastic Aptitude Test (SAT)* is a criterion-referenced assessment of a student's readiness for college. The SAT includes a composite score, and subscale scores range from 200 to 800 for Math and Evidence-Based Reading and Writing. Grade level and college readiness benchmarks are provided by the College Board and test developers. For this report, we compared student performance to grade level benchmarks.

Norm-Referenced Tests

- *Iowa Assessment (previously Iowa Test of Basic Skills)* was developed by the Education Department at the University of Iowa and is a norm-referenced test. Test items were developed to align with the Iowa Core of State Educational Standards. The test has been validated at the national level, and it provides national percentile scores for Reading, Language, and Math. The scale scores can be used to track a student's progress over time, but do not indicate whether a student is performing at grade-level. The Iowa Assessment provides Achievement Levels based on the student's scale score. This report includes test results and interpretations based on national norms developed in 2017, the most recent available at the time of testing.
- *Measures of Academic Progress (MAP) Growth* is a computer adaptive test developed by the Northwest Evaluation Association (NWEA). MAP Growth has features of both norm- and criterion-referenced tests. The Achievement Percentile Ranks allow student performance to be compared to students in a norm group for the fall, winter, and spring terms. The **R**asch Un**IT** (RIT) Scores (100 to 350) are used to determine student proficiency levels based on cut scores set by individual states (based on state and Common Core State Standards) or default cut-scores for U.S. states and international schools that have not set benchmarks. Because Alabama does not have benchmarks for this test, national Achievement Percentile Ranks were used to assess student performance in this report. Scores are provided for Reading, Language, and Math.
- *Scantron Performance Series* is a norm-referenced computer adaptive test developed to provide a longitudinal view of student growth in various subject areas. National percentile scores are used to compare student performance to similar students in the norm group. Scale

scores can range from 1300-3700 and are used to estimate student ability and to determine student growth in grades 2 through 12. Scores are provided for Reading, Language, and Math.

- The *Stanford Achievement Test, 10th Edition (SAT-10)* is a norm-referenced test developed by Pearson Assessment. The SAT-10 was developed to compare a child's academic achievement relative to others in the nation based on a national percentile score. The SAT-10 provides national percentile scores in Language Arts, Reading, and Math for students in grades kindergarten through 12. Pearson identifies a percentile score of 24 or greater as performing in an "Average" or better Performance Cluster. National percentile rank and Performance Clusters do not indicate whether a child is performing at grade level; instead, they are indicators of relative performance compared to other students. This report includes SAT-10 percentile scores based on 2018 norms.
- *The Standardized Testing and Reporting Assessment (STAR)* was developed by Renaissance. The STAR assessment is a computer adaptive test that has qualities of both norm- and criterion-referenced tests. Like the other norm-referenced tests, national percentile rankings compare student performance to similar students in the nation. Grade-level benchmarks based on the scale scores are provided by Renaissance for students in grades 1 through 12 in Reading.
- The *TerraNova 3rd edition* is also a norm-referenced test. The test content aligns with the framework of the NAEP. The national percentile scores indicate how well a child compares to other students at the same grade level, similar to the SAT-10 and Iowa. Included in the report are scores for Language Arts, Reading, and Math based on 2017 norms.

Demographic Information for Scholarship Recipients Included in the Evaluation

Based on demographic information provided by the SGOs, the 1,658 scholarship recipients with usable test scores were similar to the larger sample from which they were drawn. The racial/ethnic make-up of the sample was predominantly from three groups, Black/AA (60%), White/Caucasian (19%), and Hispanic (19%), and the remaining 2% of students were either another race, more than one race, or no race was designated. Half (50%) of the students in this group were female, and 94% were free/reduced lunch eligible. Students represented 36 counties in the state and attended 90 different schools, and these numbers are lower than the sample as a whole. Similar to the larger sample, the majority of students in this subgroup had received at least one previous scholarship, but this percentage was slightly higher than for the sample as a whole (96% vs. 93%), and the percentage of students having been in the program 7 or more years was also slightly higher in this subsample (29%) than the sample as a whole (27%). However, the average number of years that a student had been in the AAA program was similar (4.4 vs. 4.2). These small discrepancies are likely due to the exclusion of students in the youngest grades who were not required to test and were more likely to be first-time scholarship students because they were just starting school.

Findings for the 2020-2021 Academic Year

Objective 1: Describe the Academic Achievement of Scholarship Recipients

In this section, outcomes are described for each of the 11 tests for the 2020-2021 academic year. For each test, a brief description of the student demographics is provided, and additional details relevant to understanding the test scores are given. When possible, test scores disaggregated by grade, race/ethnicity, and gender are presented. Statistical tests comparing scores among racial/ethnic groups and between genders were conducted when there were sufficient numbers of students in these groups ($n \geq 25$). National percentile scores are included for norm-referenced tests, and scale scores are generally used for criterion-referenced tests to assess grade-level benchmarks. Due to rounding, sometimes percentages in a table or chart sum to a number slightly greater or less than 100%.

The presentation of the results is organized by the type of test, norm- or criterion-referenced since the tests within each type measure achievement in similar ways. The first six tests, Iowa, MAP Growth, Scantron, SAT-10, STAR, and TerraNova 3, are norm-referenced tests. The five criterion-referenced tests, ACT Aspire, PreACT, ACT, PSAT/NMSQT, and the SAT, are summarized next. The AAA legislation asks for test scores for Math and Language Arts subject areas. For some tests, English scores were provided rather than Language Arts, but the content of these subjects is similar. Furthermore, because the State of Alabama has used Reading scores to evaluate public school students in the past, Reading scores are included in this report as well. Due to the low representation of other races/ethnicities (typically 1.5% or less), descriptive information is only provided for Black/AA, White, and Hispanic groups.

Norm-Referenced Test Results

It is important to recall that the scores for norm-referenced tests do not indicate if a child has acquired the knowledge and skills expected for their grade. Rather, these tests focus on percentile scores that assess students' performance relative to other children at the same grade level in the country. As has been noted in previous reports, the 50th percentile is often used as the yardstick for evaluating performance, but it is not a good indicator of whether a child or a group of children have mastered grade-level material. As a marker for performance, however, the average scholarship recipients' scores should be close to the 50th percentile if as a group they are achieving at levels similar to others in the U.S. Generally, meeting or exceeding this standard would be considered a positive outcome. Statistical comparisons to the 50th percentile were made separately for each of the norm-referenced tests to determine if the scholarship recipients achieved at a level comparable to students in the U.S.

Iowa Assessment

Results for the Iowa Assessment-Spring 2017 Norms were available for 662 students in grades 2 through 8, 10, and 11. Females comprised 52% of the test takers. The racial/ethnic make-up was 49% Black/AA, 16% White, 31% Hispanic, and 4% of another race or unknown. The number of years that students had received a scholarship varied considerably: 4% were in their 1st year, and 32% were in their 7th year or more, with a mean of five years. The vast majority were free/reduced lunch eligible (91%). Table 2 presents the test results for each grade level disaggregated by gender and race/ethnicity when the sample size was sufficient. There were less than 25 students in 10th

grade (n = 11) and 11th grade (n = 15), so those scores are not reported here. The numbers of Black/AA and Hispanic students were sufficient in grades 3 through 8 to report their scores, but no other racial/ethnic group had 25 or more students at a grade level.

Table 2 reveals that nearly all mean percentile scores were statistically significantly below the 50th percentile (designated by *). However, there were exceptions where scores were not significantly different from the 50th percentile, primarily for grades 3 (all subjects and all groups) and 7 (Reading and Language for all groups except Black/AA). Hispanic and female students also had several scores that did not significantly differ from the 50th percentile in several grades.

Table 2: Mean Iowa Assessment Percentile Scores and Achievement Levels for Grades 2 – 8 (Spring 2017 Norms)

		Reading		Language	Math	
Grade	Group (N)	Mean Percentile	% at Achievement Level	Mean Percentile	Mean Percentile	% at Achievement Level
2	All (39)	46*	NA	39*	40*	NA
	Black/AA (21)	---	---	---	---	---
	Hispanic (11)	---	---	---	---	---
	Females (19)	---	---	---	---	---
	Males (20)	---	---	---	---	---
3	All (66-67)	45	39	49	48	30
	Black/AA (28-29)	44	17	52	47	21
	Hispanic (25)	46	48	46	54	40
	Females (41)	45	37	48	47	27
	Males (25-26)	45	42	50	50	35
4	All (94-95)	44*	39	47	41*	34
	Black/AA (46-47)	39*	34	43	32*	22
	Hispanic (31)	50	52	49	51	52
	Females (44-45)	46	47	52	39*	39
	Males (50)	43*	32	42*	42	30
5	All (91-92)	40*	34	44*	39*	40
	Black/AA (35-36)	29*	19	33*	26*	19
	Hispanic(37)	43*	38	46	41*	46
	Females (48)	43*	37	48	39*	39
	Males (43-44)	37*	32	41*	38*	41
6	All (110-112)	38*	35	43*	32*	30
	Black/AA (45-46)	28*	20	29*	19*	20
	Hispanic (38-39)	40*	44	47	36*	31
	Females (66-67)	40*	42	44	31*	33
	Males (44-45)	36*	24	40*	33*	27
7	All (130-132)	48	34	52	38*	35
	Black/AA (70-71)	40*	25	42*	26*	24

Table 2: Mean Iowa Assessment Percentile Scores and Achievement Levels for Grades 2 – 8 (Spring 2017 Norms)

		Reading		Language	Math	
Grade	Group (N)	Mean Percentile	% at Achievement Level	Mean Percentile	Mean Percentile	% at Achievement Level
	Hispanic (30)	53	55	56	52	61
	Females (69-71)	49	32	56	38*	33
	Males (61)	46	36	47	37*	38
8	All (97-99)	40*	24	40*	30*	17
	Black/AA (52)	34*	28	32*	21*	11
	Hispanic (27-28)	41*	14	44	39*	32
	Females (41)	44	26	47	31*	19
	Males (56-58)	37*	22	35*	29*	16

--- Indicates an insufficient number of students in the group (< 25) for reporting.
* Mean score is significantly below the 50th percentile.
Mean scores without a * designation are not significantly different from the 50th percentile.
NA indicates that achievement levels were not available for 2nd grade.

Comparisons between scores for males and females at each grade level were not significant. However, comparisons between Hispanic and Black/AA students indicated that Hispanic students scored significantly higher than Black/AA students in Math for grades 4 through 8. Hispanic students were also significantly higher than Black/AA students in Reading and Language for grades 5, 6, and 7.

The Iowa Assessment also provides "Achievement Levels" to help interpret test results for Reading and Math. Table 2 indicates the percentage of students who reached a minimum level of achievement designated as "Proficient" or higher. Generally, the majority of students at each grade level failed to meet this standard, although rates above 50% were observed for Hispanic students in grades 4 and 7.

Measures of Academic Progress (MAP) Growth

Results for the MAP Growth test are reported for 195 students in grades 2 through 8, 10, and 11. Over half (53%) of the students were female. As with the larger population of AAA scholarship recipients, the racial/ethnic background of the students who took the MAP Growth test was majority Black/AA (62%), followed by 22% White and 14% Hispanic. Twelve percent (12%) of students who took the MAP Growth Test received their first scholarship during the 2020-2021 school year. The remaining students had received a scholarship for two or more years, with three years being the average. Eleven percent (11%) of students had received a scholarship for seven or more years. Nearly all (93%) of the students were eligible for free/reduced lunch.

Table 3 presents the mean percentile scores in each subject area for all grade levels combined and separately for grades 2, 3, and 8 where there were a sufficient number of students ($n \geq 25$). Mean percentile scores for race and gender are reported separately for the sample as a whole (all grades

and demographic groups combined). The mean percentile scores were significantly below the 50th percentile for the sample as a whole and for Black/AA (all grades combined) and male students (all grades combined). White students scored significantly *above* the 50th percentile for Reading and Language and did not significantly differ from the 50th percentile for Math. Hispanic and female students' scores for Reading and Language also did not differ from the 50th percentile. Additionally, for all subjects in grade 2 and for Reading and Language in grade 3 the mean percentile scores were not significantly different from the 50th percentile.

Statistical tests indicated that White students performed significantly better than Black/AA students in all subject areas and better than Hispanic students in Reading and Math. Black/AA students also performed more poorly than Hispanic students in Math and Language. Females and males were not statistically different in their performance in Reading, Language, or Math when all grade levels were combined.

Table 3: Mean MAP Growth Percentile Scores for Grades 2 – 8, 10 and 11 (Spring 2020 Norms)

		Reading	Language	Math
Grade	Group (N)	Mean Percentile	Mean Percentile	Mean Percentile
2-8, 10 & 11	All (184-195)	45*	44*	32*
	Black/AA (112-120)	39*	36*	24*
	White (40-43)	59#	60#	50
	Hispanic (27)	46	51	38*
	Female (97-103)	47	47	30*
	Male (87-92)	43*	41*	34*
2	All (40-44)	54	52	43
3	All (25)	42	44	31*
8	All (29)	40*	40*	32*

* Mean score is significantly below the 50th percentile.
# Mean score is significantly above the 50th percentile.
Mean scores without a * or # are not significantly different from the 50th percentile.

Stanford Achievement Test 10 (SAT-10)

Findings for the SAT-10 (Spring 2018 norms) are reported for 143 students in grades 2 through 8, 10, and 11. Half of the test takers (50%) were female. Students who took the SAT-10 were predominantly Black/AA (57%), followed by 40% White and 4% Hispanic. All students had received two or more scholarships with the average being five. Forty percent (40%) of the students in this group had received a scholarship for seven or more years. As with the larger sample, nearly all students were free/reduced lunch eligible (94%). There were sufficient numbers of students in the 7th ($n = 32$) and 8th ($n = 27$) grades to report these grade-level scores, but there were not enough students to provide scores based on race/ethnicity or gender for these grades. No other grade had more than 25 students. Table 4 reports the combined scores for all grades and separately for grades 7 and 8.

Table 4 reveals that the mean percentile scores for all subject areas, grades, and demographic groups were significantly below the 50th percentile. The SAT-10 Spring 2018 norms identifies a percentile score of 24 or greater as performing in an "Average" or "Above Average" Performance Cluster. The percentage of students meeting the minimum standard for "Average" is indicated in Table 4. If students as a group were performing at the level of most students in the U.S., then it would be expected that 77% of students should be in the Average cluster or higher. Generally, more than half of the students performed in the Average range or higher for Reading and English but below the expected 77% mark. The majority of students were below the Average Performance mark in Math. Comparisons among racial groups indicated that White students performed significantly higher than Black/AA students in Math. Females performed significantly higher than males in Reading and Language.

Table 4: Mean SAT-10 Percentile Scores and Performance Clusters for Grades 2 – 8, 10, and 11 (Spring 2018 Norms)

		Reading		Language		Math	
Grades	Group (N)	Mean Percentile	% Perf. Cluster	Mean Percentile	% Perf. Cluster	Mean Percentile	% Perf. Cluster
2-8, 10 & 11	All (137-143)	37*	63	31*	56	25*	39
	Black/AA (78-81)	33*	56	29*	54	18*	27
	White (54-57)	41*	74	33*	59	32*	51
	Female (69-71)	41*	72	38*	70	26*	42
	Male (68-72)	31*	54	23*	43	24*	33
7	All (30-32)	38*	60	31*	53	20*	31
8	All (27)	34*	59	24*	48	21*	22

% Perf. Cluster = Percentage of students that meet or exceed the Average Performance Cluster standard of a percentile score $\geq$ 24%.
* Mean score is significantly below the 50th percentile.

Scantron

The results for the Scantron test are reported for 113 students in grades 2, 5 through 8, 10, and 11. The sample was 89% male, and the racial/ethnic make-up was 99% Black/AA and 1% Hispanic. The average number of years a student had participated in the scholarship program was three; 12% were first-time scholarship recipients, and 8% had received a scholarship for seven years. Nearly all (99%) were free/reduced lunch eligible. There was a sufficient number of students to report scores separately for grades 7, 8, and 11 for some subject areas. Across grade levels, only one student was not Black/AA and there were very few female students (two to five per grade). Consequently, scores were not disaggregated by race or gender, but it should be recognized that the scores represent primarily Black/AA males. Additionally, for Language, 62 students were missing this subject area completely or the percentile score was not provided. Table 5 reveals that the mean scores for all but 11th grade Reading were significantly below the 50th percentile.

Table 5: Mean Scantron Percentile Scores for Grades 2, 5 – 8, 10, and 11				
		Reading	Language	Math
Grade	Group (N)	Mean Percentile	Mean Percentile	Mean Percentile
2, 5-8, 10 & 11	All (51-105)	36*	25*	25*
7	All (21-27)	---	---	21*
8	All (22-29)	27*	---	16*
11	All (20-28)	44	---	---
--- Indicates an insufficient number of students in the group (< 25) for reporting. * Mean score is significantly below the 50th percentile. Mean scores without a * designation are not significantly different from the 50th percentile.				

The Standardized Testing and Reporting Assessment (STAR)

STAR test results are reported for 77 students in grades 2 through 8, 10, and 11 who took the test during the spring semester of the 2020-2021 school year. Female students comprised 52% of the sample, which was predominantly Black/AA (95%). All students had received two or more scholarships, and 48 % had received a AAA scholarship for seven or more years. The average number of years for having received a scholarship was five. Nearly all (99%) were free/reduced lunch eligible.

STAR provides scores for Reading and Math, and these are presented in Table 6. Because only three students were not Black/AA, the test results were not reported by race. Thus, the results in Table 6 primarily reflect a Black/AA racial group. There were sufficient numbers of males and females to report their scores separately. The mean percentile scores ranged from 32 to 38 across all groups and subject areas and were statistically significantly below the 50th percentile. The mean scores for males and females were not significantly different.

Table 6: Mean STAR Percentile Scores for Grades 2 – 8, 10, and 11			
		Reading	Math
Grade	Group (N)	Mean Percentile	Mean Percentile
2-8, 10 & 11	All (76-77)	36*	35*
	Female (39-40)	36*	32*
	Male (37)	36*	38*
* Mean score is significantly below the 50th percentile.			

TerraNova 3

Results for the TerraNova 3 (2017 Norms) are reported for 40 students in grades 2 through 8. There were slightly more female students (55%) than male students. The racial/ethnic make-up was 50% White, 47% Black/AA, and 3% Hispanic. All students had received two or more scholarships with the average number of years of participation in AAA being four. Approximately a third (35%) had received a scholarship for seven or more years. Nearly all students (98%) were eligible for free/reduced lunch. There were not enough students to disaggregate student scores by race, gender,

or grade. As shown in Table 7, the mean percentile scores were between 44 and 53 and were not significantly different from the 50th percentile.

Table 7: Mean TerraNova 3 Percentile Scores for Grades 2 – 8 (2017 Norms)

		Reading	Language	Math
Grade	Group (N)	Mean Percentile	Mean Percentile	Mean Percentile
2-8	All (39-40)	53	44	44
Mean percentile scores are not significantly different from the 50th percentile.				

Summary for Norm-Referenced Test Results

Findings for the norm-referenced tests are summarized in the chart below. A review of the scores from the six tests indicates that a majority of the average percentile scores were significantly below the 50th percentile, similar to previous evaluations of the AAA. For five of the six norm-referenced achievement tests, the majority of statistical comparisons for mean scores to the 50th percentile in each subject area were below the 50th percentile. The TerraNova 3 results are the exception in this year's report, as they were in previous years, and it is not clear why this is such a persistent finding. The content on the TerraNova 3 might be different than other tests, or the results could be due to differences in the schools that use these tests, or other factors related to students in these schools. Additionally, although performance at some grade levels and subject areas was not statistically below the 50th percentile, no discernable pattern emerged. For example, better performance was seen on the Iowa Assessment for 3rd graders (all subjects) and 7th graders (Reading and Language) and for the MAP Growth test for 2nd graders (all subjects) and 3rd graders (Reading and Language). This finding follows previous reports, where some grade levels and subject areas were above the 50th percentile, but no pattern emerged. When Performance Clusters (Stanford) or Achievement Levels (Iowa) were available, the percentage of students making standards for their grades was below expectations based on national norms.

With respect to the performance of different racial groups, with only a few exceptions, the performance of Black/AA students was generally below the 50th percentile as evident in the findings for the Iowa Assessment and the MAP Growth when scores were disaggregated by race and on the Scantron and STAR tests where the test takers were nearly 100% Black/AA. In comparison, the performance of Hispanic students was relatively better. For example, on the Iowa Assessment none of the Language scores for Hispanic students was significantly below the 50th percentile, and in grades 3, 4, and 7 the mean scores for Reading and Math were also not significantly below the 50th percentile. Furthermore, on the MAP Growth test for all grades combined, Hispanic students' mean scores for Reading and Language were not significantly below the 50th percentile. Results for White students were only available for the MAP Growth and the SAT-10, and the findings were mixed. On the MAP Growth, White students' scores were significantly *above* the 50th percentile in all subject areas; whereas for the SAT-10 their mean scores were significantly below the 50th percentile. When statistical comparisons could be made between racial groups, Black/AA students generally performed more poorly than Hispanic and White students, although there were exceptions for some subject areas on some tests (e.g., Reading and Language on the SAT-10).

Summary for Norm-Referenced Test Results

Tests included: Iowa, MAP-Growth, SAT-10, Scantron, STAR, and TerraNova 3.

Scholarship students as a group did not perform better than other students in the U.S.

- It was most typical for the mean percentile scores across tests to be **significantly below the 50th percentile.**
- **When Performance Clusters (Stanford) or Achievement Levels (Iowa) were available**, the percentage of students meeting standards for their grades was **below expectations based on national norms.**

There were anomalous findings to this generalization for specific grades and standardized tests.

- The variability in findings across the tests suggests there may be unmeasured factors associated with the schools using particular tests that could explain these results.
- Small sample sizes impact the reliability of some findings.

Race/ethnicity comparisons indicated that outcomes were poorer for Black/AA students compared to other racial groups.

- Black/AA students generally performed below the 50th percentile in all subjects.
- Where comparisons could be made, Hispanic and White students often had statistically significantly higher scores than Black/AA students.

Gender comparisons generally suggest that males and females performed similarly.

- There were several instances on the Iowa and MAP Growth tests where **female students' scores were not significantly below the 50th percentile for Reading and/or Language,** but male scores were below this mark.

With respect to gender, when the comparisons could be made, there were few significant differences in the mean achievement test scores. Significant differences were only found on the SAT-10 test when all grades were combined, indicating that Reading and Language scores were significantly higher for females than males. There were several instances where female students' scores were not significantly below the 50th percentile for Reading or Language when scores were reported for the sample as a whole or grade level: On the Iowa Assessment in grades 4 through 7 and for the MAP Growth test for all grades combined.

To summarize, although there are some exceptions, as reported above, the findings for norm-referenced tests suggest that the scholarship recipients generally performed below national norms for their grade levels. This is consistent with the performance described in the 2020 report. Compared to previous years, in the current report the number of norm-referenced tests more than doubled, but in some cases the number of students taking a particular test was small (e.g., STAR and TerraNova 3). When sample sizes are small, there is an increased probability of variation among individuals within a comparison group (e.g., grade, race/ethnicity, gender) that may give the appearance of some groups performing better than others. Aggregating over grade levels as was necessary for all but the Iowa Assessment, potentially obscures differences at earlier and more advanced grades. Larger samples allow us to be sure that performance on a test is related to student academic achievement and not unrelated factors such as a single student in the group having a good testing day or better resources for test preparation at a single school, among others.

By focusing on tests with the most recent norms taken in the spring, the report attempted to address variability among the different tests. However, the fluctuations in findings across the tests suggest unmeasured factors associated with the schools using particular tests that could explain these

results (e.g., school resources, class sizes, availability of help for struggling students, accommodations for special needs students). Even though all the tests included in this report are based on current standards outlined by NAEP or Common Core State Standards, we cannot rule out test design or psychometrics as also playing a part in the variation in scores. More modern tests such as the STAR and MAP Growth are computer adaptive tests and adjust the questions based on student ability, whereas at the time of the report, the Iowa Assessment (2017 norms) and the SAT-10 (2018 norms) do not.

Criterion-Referenced Test Results

ACT Aspire

Results for the ACT Aspire test are reported for 130 students in grades 3 through 8 and 10. This group was 57% female, and the racial make-up included 66% Black/AA, 22% Hispanic, and 23% White. The eligibility rate for free/reduced lunch was 95%. The average number of years that a student had received a scholarship was four. Only one student was a first-time scholarship recipient, and 15% were in their 7th or 8th year in the scholarship program. There were not enough students to report scores separately for any grade level, so the results are presented for all grades combined.

Because the ACT Aspire is a criterion-referenced test, percentile scores are less meaningful for evaluating students' academic performance. Instead, ACT Aspire has set proficiency benchmark scores based on the scale scores for each grade level and subject area. Table 8 presents the mean percentile score and the percentage of students who reached proficiency for all grade levels combined. Scores are also disaggregated by race and gender. For Reading and Math, the percent proficient ranged from 27% to 39%, indicating that the majority of students did not meet the benchmarks for their grade level. Results for English were better, with 71% to 78% of students meeting the benchmarks for this subject area. This pattern of performance across the different subject areas was evident in the earlier reports on AAA, but it is not clear why English scores are consistently higher. Statistical comparisons among the racial groups and genders were not significant.

Table 8: Mean ACT Aspire Percentile Scores and Proficiency Rates for Grades 3 – 8 and 10

		Reading		English		Math	
Grade	Group (N)	Mean Percentile	% Prof.	Mean Percentile	% Prof.	Mean Percentile	% Prof.
3-8 & 10	All (127-129)	44	31%	45	74%	41	36%
	Black/AA (70-71)	43	29%	44	74%	44	39%
	Hispanic (27-28)	46	36%	47	78%	36	30%
	White (25-29)	45	31%	46	72%	40	37%
	Girls (96-99)	45	34%	48	76%	40	34%
	Boys (54-55)	43	27%	41	71%	43	39%

% Prof. = percent meeting the proficiency benchmark.

PreACT Test

PreACT test scores are included for 72 students in grades 10 and 11. The racial/ethnic make-up of this group of students was 72% Black/AA, 17% White, and 10% Hispanic. With respect to gender, 46% were female. Most students were free/reduced lunch eligible (97%). More than half of the students (53%) had received a scholarship for seven or more years, only 4% were first-time scholarship recipients, and the average number of years in the scholarship program was five. There were 62 students in 10th grade and ten in 11th grade. Due to the small number of 11th graders, the summary statistics focus on 10th graders.

For the PreACT, the critical scores are the scale scores (range 1-36) that correspond to the ACT college entrance exam scores, rather than percentile scores. Benchmark scores are provided to indicate college readiness. Specifically, according to the PreACT Technical Bulletin, these benchmarks indicate "the level of achievement required for students to have a 50% chance of obtaining a B or higher or about a 75% chance of receiving a C or higher in corresponding credit-bearing first-year college courses." Because the ACT is normally taken in the 11th grade, additional college readiness indicators are provided for 10th graders to account for the fact that 10th grade students will continue to gain skills and knowledge over the course of the year. As a result, these indicators can be used to make predictions as to the likelihood of meeting the benchmark scores in 11th grade. Three benchmark levels for 10th grade are defined for each subject area: *In need of intervention, On the cusp,* and *On target.*

Table 9: Mean PreACT Scale Scores and Readiness Indicators for Grade 10

		Reading		English		Math	
Grade	Group (N)	Mean Scale Score	Readiness Indicator	Mean Scale Score	Readiness Indicator	Mean Scale Score	Readiness Indicator[1]
10	All (62)	20	On Target	16	On Target	16	Intervention
	Black/AA (44)	18	On Cusp	13	On Cusp	15	Intervention
	Female (30)	20	On Target	15	On Target	16	Intervention
	Male (32)	20	On Target	16	On Target	17	On Cusp

[1] Readiness indicators are for 10th grade students.

Table 9 presents the mean scale scores for 10th grade students and provides the corresponding college readiness indicator level for 10th graders. There was a sufficient number of students to report scores for Black/AA students and for male and female students. The mean scores indicate that students were generally *On target* for Reading and English but were *In need of intervention* for Math. Scores appear to be slightly lower for Black/AA students compared to the sample as a whole. Statistical comparisons between the two genders were not significant.

The percentages of 10th grade students who fell into each of the three readiness categories were calculated, and the results are presented in Table 10. These results show that more than half the students were *On target* to meet the ACT readiness benchmarks for English (57%), but for Reading and Math, the majority of students failed to meet benchmarks. The results for Math are especially of concern as 71% were *In need of intervention*. In interpreting the two sets of results presented in Tables 9 and 10, it is important to consider that for criterion-referenced tests, the goal is that 100%

of students should meet benchmarks. With that standard as an ideal, the scores for the PreACT fall short in all three subject areas.

Table 10: PreACT Percentage of Students in Grade 10 within Each Readiness Category

	Reading			English			Math		
Grade (N)	Inter-vention	On Cusp	On Target	Inter-vention	On Cusp	On Target	Inter-vention	On Cusp	On Target
10 (62)	36%	21%	44%	32%	11%	57%	71%	13%	16%

ACT College Entrance Exam

ACT scores are reported for 110 students in 10th and 11th grade. The majority of this sample was Black/AA (63%), followed by 26% White and 11% Hispanic. Ninety-eight percent (98%) of the students were eligible for free/reduced lunch and 52% were female. There were only two first-year scholarship recipients (2%). The average number of years that a student had received a scholarship was five, and 36% had received a scholarship for seven or more years. Only the 11th grade had a sufficient number of students (97) to report scores, and there were enough students to break out scores by gender, Black/AA, and White students. See Table 11.

Table 11: Mean ACT Scores and Proficiency Rates for Grade 11

		Reading		English		Math	
Grade	Group (N)	Mean Scale Score	% Prof.	Mean Scale Score	% Prof	Mean Scale Score	% Prof
11	All (97)	19	25%	17	37%	17	11%
	Black/AA (60)	17	17%	15	27%	16	3%
	White (27)	21	41%	19	59%	18	19%
	Females (49)	18	14%	16	37%	16	4%
	Males (48)	19	35%	17	38%	17	19%

The benchmark scores for 11th grade ACT scores are 22 for Reading, 18 for English, and 22 for Math.

Similar to the PreACT, the relevant scores for the ACT are the scale scores (range from 1 to 36), which align with proficiency benchmarks for each grade level. The benchmark scores are similar to those for the PreACT and are interpreted the same way. The benchmark scores for 11th grade ACT scores are 22 for Reading, 18 for English, and 22 for Math. The average ACT scale scores were statistically significantly below benchmarks for college preparedness for all subjects for all groups represented in Table 11. An examination of the percentage of students who reached proficiency reveals a similar pattern of poor performance. With the exception of White students' English scores, the majority of students failed to reach proficiency.

Comparisons between genders did not reveal statistically significant differences between the mean scores, but the proficiency rates for males were higher than proficiency rates for women for Reading and Math. The comparisons between Black/AA and White students indicated that White students' mean scores and proficiency rates were significantly higher than Black students' scores in all subject areas.

PSAT/NMSQT

The PSAT/NMSQT results are reported for 84 scholarship recipients in the 10th and 11th grades. There were more female students (60%) than male students. The racial/ethnic make-up was 56% Black/AA, 13% White, 24% Hispanic, and 7% other ethnicities. All students had received two or more scholarships, and the average was five scholarships. Thirty-nine percent (39%) had received a scholarship for seven or more years. Most students (89%) were eligible for free/reduced lunch. Scores could not be reported by racial/ethnic demographic groups because of the low number of students. There were enough female students to report their scores separately.

The PSAT/NMSQT combines Reading, Writing, and Language scores into an "evidenced-based Reading-Writing score." As a result, the combined scores are presented in Table 12. The Reading-Writing and Math scores are aligned with benchmarks used to predict college readiness. The benchmark scores correspond to a 75% likelihood of achieving a grade of "C" or better in the first semester of college for courses in related areas. The benchmark for Reading-Writing corresponds to a scale score of 430 for 10th grade and 460 for 11th grade. The mean scale scores for Reading-Writing were not statistically different from the benchmarks for each grade. However, less than 50% of students in 10th grade (46%) and just over half of the students in 11th grade (51%) made the benchmark. The majority of female 10th graders (56%) made the Reading-Writing benchmark. Performance was poorer in Math. The average scale scores for Math fell statistically significantly below the benchmark. Only 15% of 10th graders met the benchmark score of 480, while only 11% of 11th graders met the benchmark of 510.

Table 12: Mean PSAT/NMSQT Scores and Percent Meeting Benchmarks for Grades 10 and 11

		Reading-Writing		Math	
Grade	Group (N)	Mean Scale Score	% Meets Benchmark	Mean Scale Score	% Meets Benchmark
10	All (39)	434	46%	406	15%
	Female (27)	447	56%	401	15%
11	All (45)	465	51%	436	11%

Reading-Writing benchmarks: 430 for 10th grade and 460 for 11th grade.
Math benchmarks: 480 for 10th grade and 510 for 11th grade.

Scholastic Aptitude Test (SAT)

The SAT was administered to 32 students in the 11th grade. There were more female students (59%) than male students. The racial/ethnic make-up was 72% Black/AA, 25% Hispanic, and 3% White. All students had received two or more scholarships, with an average of four years of participation in the AAA program. Sixteen percent had received a scholarship for seven or more years. The majority of students in this group (88%) were eligible for free/reduced lunch.

The SAT aligns with the PSAT/NMSQT and reports results using similar subject areas (Reading-Writing and Math) and benchmark scores. Scale scores align with benchmark scores to determine if a student is "college ready." Benchmark scores for Reading-Writing indicate that a student has a 75% chance of earning a "C" or better in first-semester college courses in history, literature, social sciences, or writing. Similarly, meeting the benchmark score for Math indicates that a student has a 75% chance of earning a "C" or better in first-semester college courses in algebra,

statistics, pre-calculus, or calculus. For 11th grade students, the Reading-Writing benchmark score is 460, and the Math benchmark score is 510.

Table 13 shows the mean scale score and the percentage of students who met the benchmark scores for each subject area. The mean scale scores fell significantly below the benchmark scores, and only 31% and 3% of students met the benchmarks for Reading-Writing and Math, respectively. It should be noted, that in Math, only one student made the benchmark.

Table 13: Mean SAT Scores and Percent Meeting Benchmarks for Grade 11

		Reading-Writing		Math	
Grade	Group (N)	Mean Scale Score	% Meets Benchmark	Mean Scale Score	% Meets Benchmark
11	All (32)	428	31%	389	3%
Reading-Writing benchmarks: 460; Math benchmarks: 510					

Summary for Criterion-Referenced Test Results

The key performance indicator for students taking criterion-referenced tests is the percentage of students meeting benchmarks on each of the tests. Generally, most students did not meet benchmarks, but the findings vary among tests and grade levels. The summary graphic below presents the key findings.

The ACT Aspire is the only test that included students in 8th grade or younger, as well as high school students. Combined results across grade levels, race/ethnicity, and gender indicated that the majority of students were not meeting benchmarks for their grade in Reading (69%) and Math (64%). In contrast, the majority of students (74%) did meet benchmark scores in English. Results were similar across racial/ethnic groups and gender, and there were no significant differences among any of these groups.

The PSAT/NMSQT and PreACT both included scores for 10th graders, and the findings were similar across the two tests. Reviewing scores across all demographic groups combined revealed that on the PreACT the mean scores for Reading and English met the *On target* benchmark score, and 44% and 57% of students met the standards for Reading and English, respectively. Math performance was worse, with the mean score indicating that intervention was needed, and only 16% of students were *On target*. The results for the PSAT/NMSQT were similar for 10th graders in that the average scale score for Reading-Writing was above the benchmark, and 46% met or exceeded this score. In contrast for Math, the mean scale score was below the benchmark and only 15% of students met the benchmark. Together these results suggest a relatively more positive performance for 10th graders in Language Arts and Reading compared to Math. It should be noted that this pattern of results is very similar to those reported in 2020.

Summary for Criterion-Referenced Test Results
Students in grades 3-8 took the ACT Aspire:
⊙ The **majority failed** to meet grade level benchmarks for **Reading** and **Math**. ⊙ For **English** the majority of scholarship recipients **met** or e**xceeded** the benchmarks.
PSAT/NMSQT and PreACT scores were available for grade 10:
⊙ Average performance on Reading, English, and writing met benchmarks. • On the **Pre-ACT,** mean scores for **Reading and English made the *On target* benchmark and 44% and 57% of students were at or above the benchmarks, respectively** • On the **PSAT/NMSQT**, **mean Reading-Writing scores exceeded the benchmark, and** 46% of students were at or above the benchmark. ⊙ **Performance in Math was below the benchmarks for the majority of students on both tests.**
Students in grade 11 took the PSAT/NMSQT, ACT, or SAT:
⊙ **For the SAT and ACT**, the majority of 11th grade students did **not meet benchmarks** in M**ath, Reading,** or **English, and mean scores were below benchmarks.** ⊙ On the **PSAT/NMSQT** for the combined **Reading-Writing assessment** the mean score **exceeded the benchmark** and 51% of students **met** or **exceeded** the benchmark. • Performance in **Math was below the benchmark.**

Eleventh grade students were represented in three standardized tests: ACT, PSAT/NMSQT, and SAT. Across the three tests, there was a great deal of consistency in Math performance in that only a very small percentage of students (range from 3% to 19%) made benchmarks across all demographic groups combined. Performance for Reading, Reading-Writing, and English was better, but varied among the tests. On the PSAT/NMSQT the mean Reading-Writing score was above the benchmark, and 51% of students met or exceeded the benchmark score; however, only 31% met the Reading-Writing benchmark on the SAT, and the mean score was below the benchmark. For the ACT, 25% and 37% of students for all demographic groups combined met benchmarks for Reading and English, respectively, and the mean scores were below benchmarks. For the ACT, scores were reported separately for racial/ethnic groups and gender. Similar to other results, White students' scores were significantly higher than Black/AA students' scores in all subject areas. Males had higher proficiency rates than females for Reading and Math.

Taken together, the general pattern of results suggests that on the criterion-referenced tests, most of the scholarship students failed to make benchmark scores, although performance was relatively better in Reading, Reading-Writing, and English compared to Math.

Objective 1 Conclusion

Variability in test type (norm- or criterion-referenced), test delivery (computer or paper and pencil), and whether or not the test is adaptive (adjusts for student ability) or fixed form (all students get the same questions) makes it difficult to draw general conclusions about the academic performance of scholarship recipients. On the norm-referenced test results, most AAA students performed significantly below the 50th percentile. Similar to previous years, for criterion-referenced tests the majority of students failed to meet benchmarks. An exception to this generalization is that 10th graders, on average, were meeting benchmarks in Reading-Writing, Reading, and English on the PreACT and the PSAT/NMSQT. Similar results were also found for 2018-2019 scholarship recipients in the 2020 report. An interesting finding across both norm- and criterion-referenced tests was the relatively poorer performance in Math compared to other subjects. In addition to factors typically associated with poor performance (e.g., race, poverty) that vary among the group of students taking any given test, differences in student performance were likely due to factors that vary by school, such as curriculum, pedagogy, and teacher quality.

As with previous reports, the information presented so far does not indicate whether the scholarship recipients' academic achievement represents an improvement, decline, or no change over time as a result of the AAA, nor does it indicate how these students directly compare to public school children in the State of Alabama. The next section of the report provides some insights into these issues.

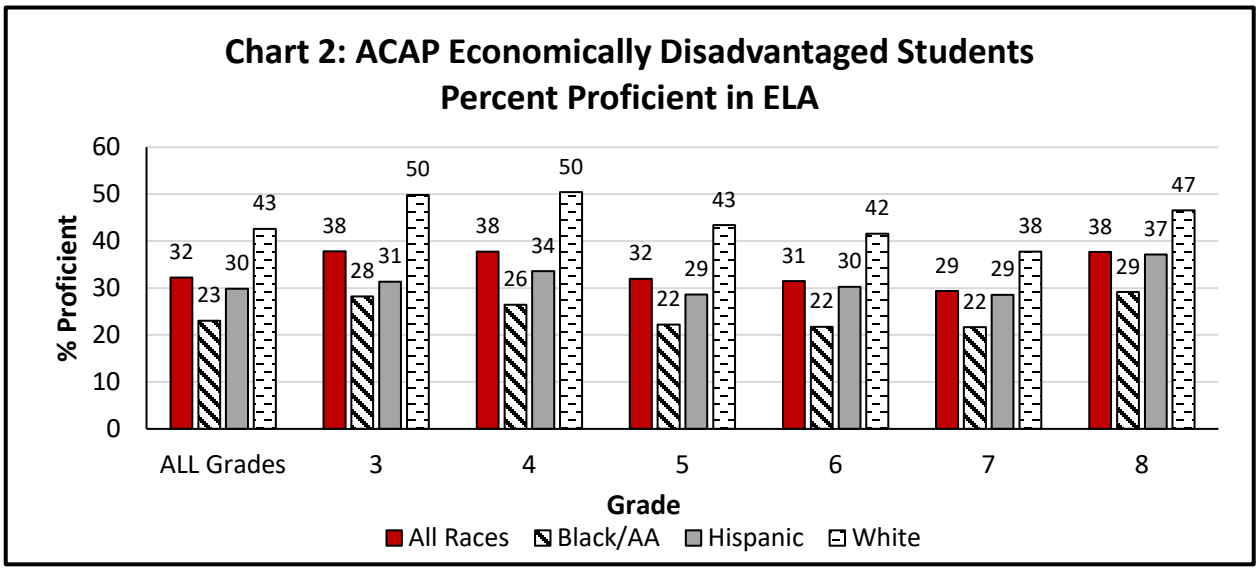
Objective 2: Compare Scholarship Recipients to Alabama Public School Students

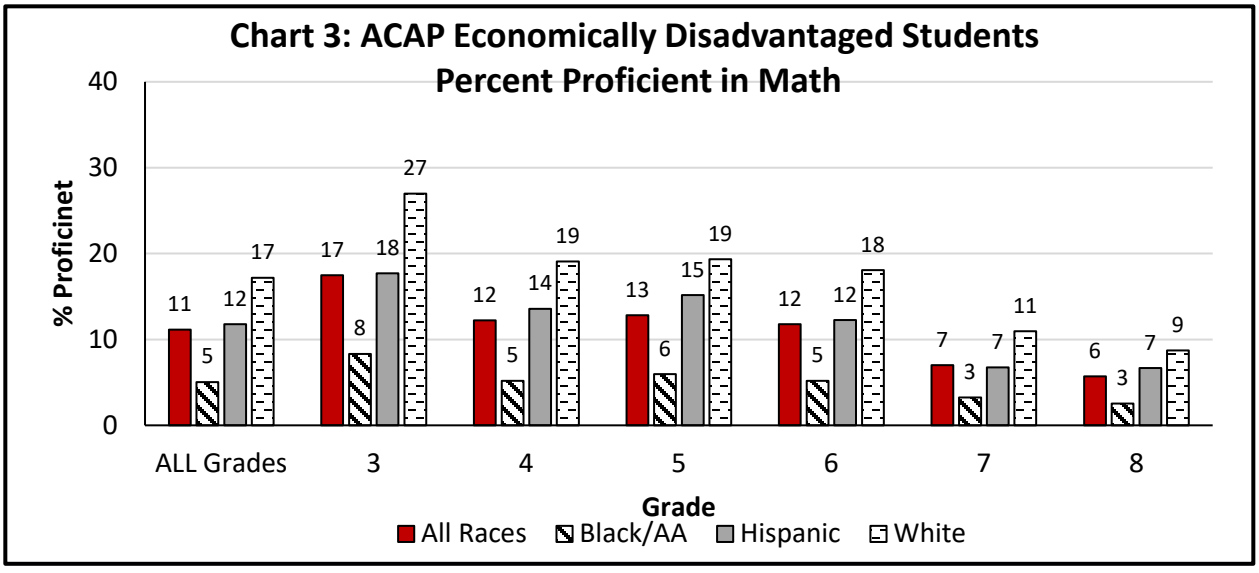
For the 2020-2021 academic year, students attending public schools in Alabama in grades 2 through 8 took the Alabama Comprehensive Assessment Program (ACAP) test, those in grade 10 took the PreACT, and those in grade 11 took the ACT college entrance exam. The spring of 2021 was the first year the ACAP was administered. Previously, ALSDE used the Scantron test. Among the scholarship recipients, only one school with two students gave the ACAP test, and thus no direct comparisons can be made. Additionally, test results were not reported for 2nd graders on the ALSDE website. Nevertheless, with the objective of providing some information on the performance of public school students in grades 3 through 8, the results for the ACAP are reported. Additionally, test results were not available for the PreACT for Alabama public school children. ACT data were available to make comparisons between the AAA scholarship students and Alabama public school children in 11th grade. State and national data were available for the PSAT/NMSQT for both 10th and 11th grades, and these results are reported so that more scholarship students are represented in the analysis of Objective 2. However, it should be kept in mind that the PSAT/NMSQT is not the state-mandated test for high school students.

For Objective 2, economically disadvantaged public school students are the appropriate comparison group for scholarship students, since 94% of the AAA scholarship students were eligible for free/reduced lunch. Scores for Black/AA, Hispanic, and White students are also reported since sometimes performance among the AAA scholarship students varied by race.

ACAP

The ALSDE provided proficiency rates for two subject areas, English Language Arts (ELA) and Math. The proficiency scores for each grade level align with performance expectations for that grade. Charts 2 and 3 present the percentage of students from economically disadvantaged backgrounds (i.e., free/reduced lunch eligible) that met benchmarks for ELA and Math, respectively. These charts indicate the majority of economically disadvantaged public school children failed to meet benchmarks for their grade level in both subject areas. For all demographic groups combined, the percentage of students who were proficient in ELA ranged from 29% (7th grade) to 38% (grades 3, 4, and 8) across grade levels. Black/AA and Hispanic students performed more poorly than White students, and Black/AA students performed slightly lower than Hispanic students. The percentage of students proficient in Math was much lower compared to ELA. For all demographic groups combined, the percentage of students who were proficient in Math ranged from 6% (8th grade) to 17% (3rd grade) across grade levels. The pattern of performance in Math for Black/AA, Hispanic, and White students was similar to that for ELA.

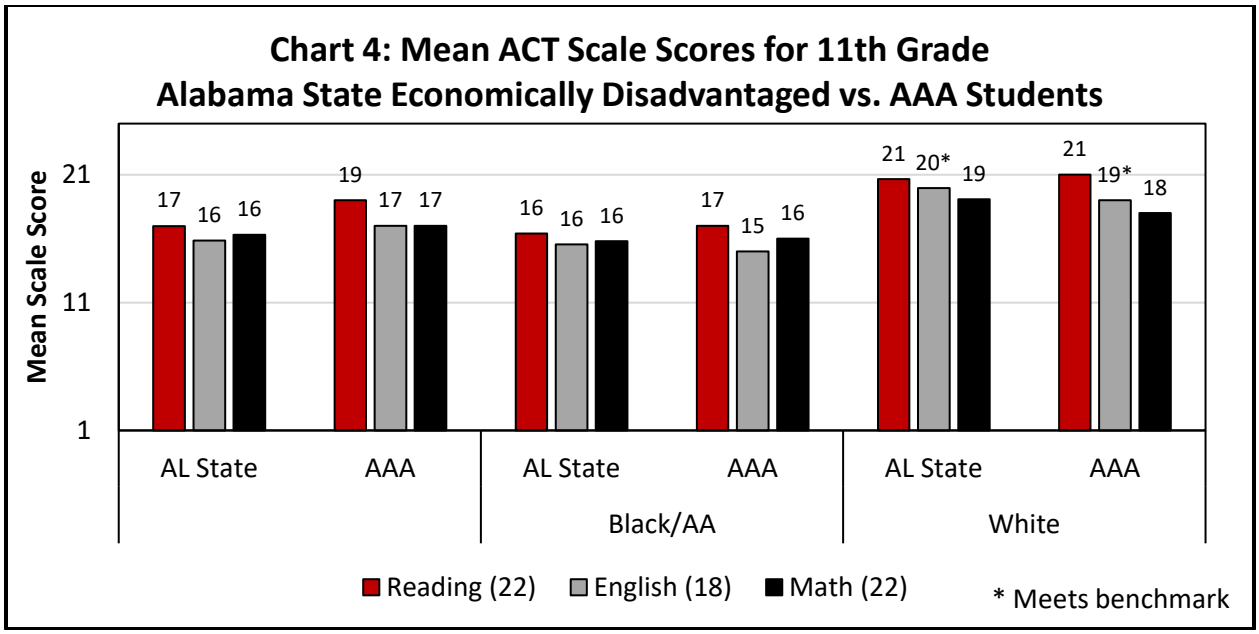


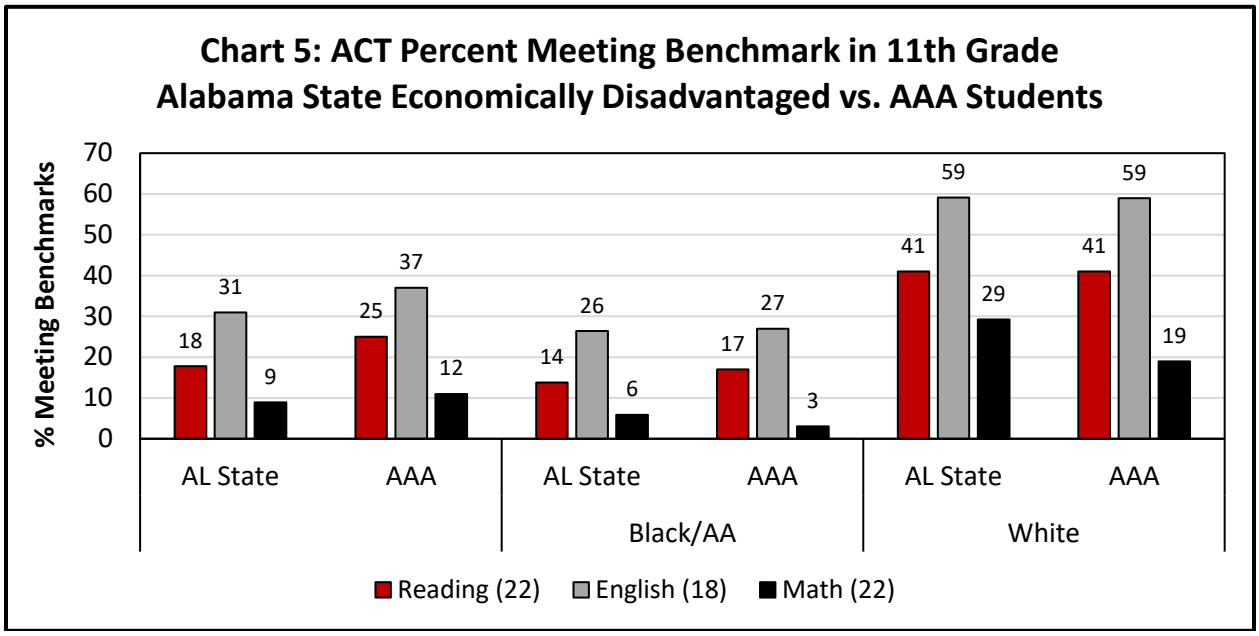


ACT

The Public Affairs Research Council of Alabama (PARCA) published the percentage of Alabama public school children in 11th grade who met the ACT college readiness benchmarks in each subject for the 2020-2021 academic year. As a reminder, the benchmark scale scores are 22 for Reading and Math and 18 for English, with a maximum score of 36. Chart 4 compares the mean scale scores for economically disadvantaged 11th graders attending public schools to those of the AAA scholarship students. Mean scale scores are also presented for Black/AA and White students because there was a sufficient number of scholarship students in these groups ($n \geq 25$). Generally, for both groups of students, the mean scores were below the benchmarks, with the exception of English for White students. Statistical comparisons between the mean scores for comparable groups indicated that the mean Reading score for AAA students (19) was statistically higher than the mean Reading score for disadvantaged public school students (17). No other comparisons were significant.

To further compare the performance of these two groups, Chart 5 displays the percentage of students making the benchmark score in each subject area. The percentages were very similar between the two groups and statistical comparisons yielded no significant differences. With the exception of White students' scores for English, the percentage of students meeting benchmarks for all students was below 50%. Taken together, these analyses indicate that 11th grade scholarship recipients collectively performed similarly to their public school counterparts.

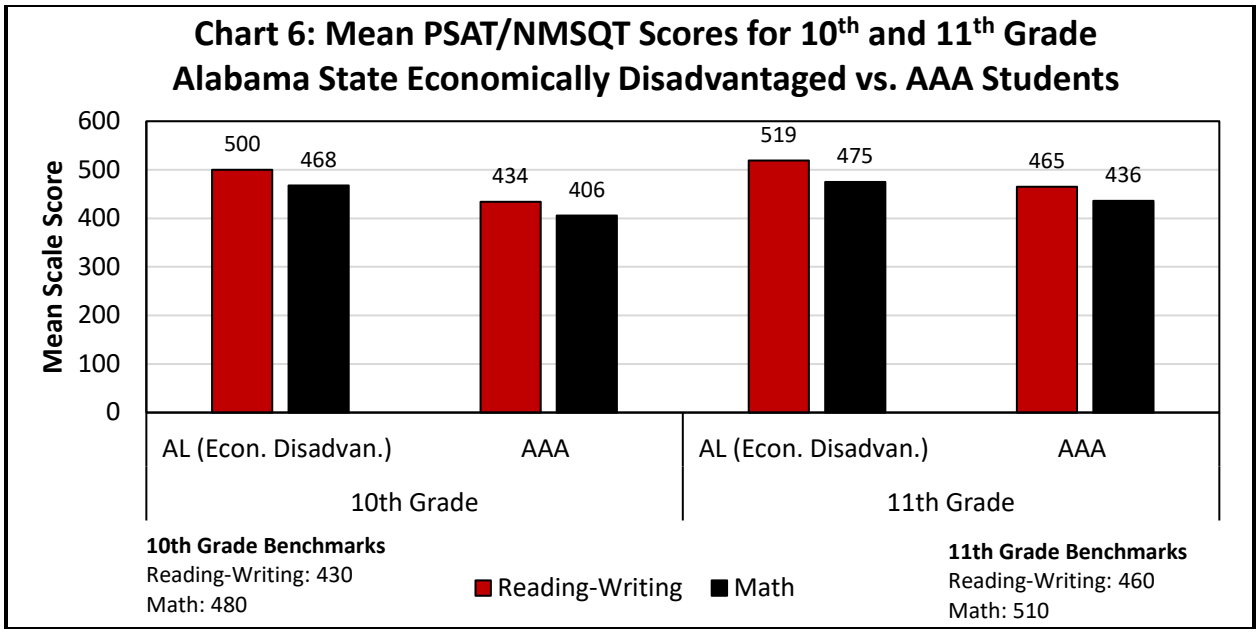


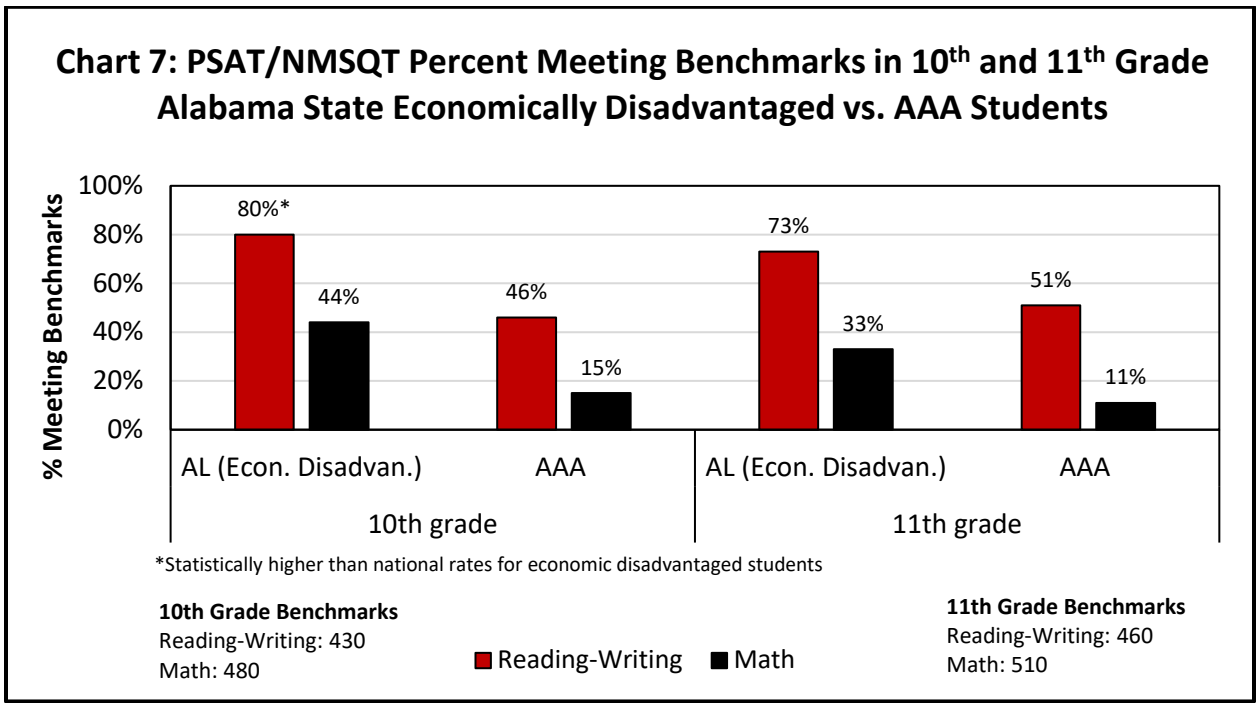


PSAT/NMSQT

The College Board reports the percentage of Alabama public school children who met the PSAT/NMSQT college readiness benchmarks for the 2020-2021 academic year. Chart 6 compares the mean scale scores of AAA students to economically disadvantaged students in Alabama. The Reading-Writing benchmarks are 430 and 460 for 10th and 11th grade, respectively, and the Math benchmarks are 480 and 510 for 10th and 11th grade, respectively. Chart 6 reveals that scores for 10th and 11th graders in both groups were above benchmarks for Reading-Writing, but below benchmarks for Math. Statistical comparisons between the scores for comparable groups indicated that the mean scores for AAA students in grades 10 and 11 were statistically lower than the scores of comparable Alabama public school students in both Reading-Writing and Math.

To further compare the performance of these two groups, Chart 7 displays the percentage of students making the benchmark score in each subject area. The percentages of AAA scholarship recipients in grades 10 and 11 who met the benchmarks for Reading-Writing and Math were significantly lower than similar economically disadvantaged students in Alabama. Additional data from the College Board indicated that Alabama public school students performed better than economically disadvantaged students in the nation in Reading-Writing for the 10th grade and the same as economically disadvantaged students in the nation for 10th grade Math and 11th grade Reading-Wring and Math.





Objective 2 Conclusion

With the data available, stronger conclusions can be made for 11th grade students compared to the other grades required to take standardized tests. The results for the ACT indicate few significant differences in performance between the AAA scholarship recipients and economically disadvantaged public school children. Additionally, 11th grade students are near the end of their state mandated education, and the majority of scholarship students who took the ACT had received a scholarship for five years or more. Thus, this group might represent the cumulative effects of receiving a scholarship. The findings for this report suggest that economically disadvantaged 11th

graders performed similarly on the ACT, regardless of their participation in AAA. In contrast, AAA scholarship recipients did not do as well as economically disadvantaged students attending public schools in Alabama on the PSAT/NMSQT. It is important to point out that students in Alabama are required to take the ACT in grade 11, and thus the AAA students are being compared to similar students in Alabama. In contrast, students in Alabama who took the PSAT/NMSQT might represent a subset of high-achieving college bound students who self-selected to take this test to compete for the National Merit Scholarship. Similar to past reports, because these comparisons include just a small percentage of the scholarship students, some caution must be taken in generalizing them to the larger group of scholarship students. Finally, the majority of public school children in grades 3 through 8 failed to reach proficiency benchmarks. Although no direct comparisons can be made to AAA scholarship students, both groups of students appear to struggle to meet national standards on the tests that they were administered.

Before moving to the next objective, it is important to consider that the public school comparison groups did not represent the average student in the state; rather they represented students from economically disadvantaged homes. When all Alabama public school students' scores are considered (regardless of economic status) the proficiency rates are higher than those represented in the charts for economically disadvantaged students. Thus, the differences between the AAA scholarship students and the average student in the state strongly favors the public school children.

Summary for Objective 2: Scholarship Recipients vs. Alabama Public School Students

- **Due to the lack of appropriate comparative data**, **strong conclusions cannot be made** for the relative performance of the scholarship recipients and the scholarship recipients.
- Six years after the first AAA report, there **is little evidence** that the scholarship program has resulted in academic achievement that **is superior to that of comparable Alabama public schools.**

ACAP findings for 3rd through 8th grade

- Only **a small percentage of economically disadvantaged public school children met proficiency standards.**
- Performance in **Math was much lower than in Language Arts**.
- **No comparisons could be made between Alabama public school students and the scholarship recipients** because only a few scholarship students took the ACAP.

ACT findings for 11th graders

- Based on their mean scores, **scholarship recipients collectively** performed **better than economically disadvantaged public school students** in **Reading,** but were **similar** to public school children in **Math and English**.
- The **proficiency rates in each of the three subject areas were comparable** for economically disadvantaged public school students and scholarship recipients.

PSAT/NMSQT for 10th and 11th graders

- Average performance for **both grades met national benchmarks in Reading-Writing, but not in Math.**
- AAA students performed statistically **lower than the mean Reading-Writing and Math** scores for economically disadvantaged students in Alabama.
- For both grades, **the percentage of AAA students meeting benchmarks for Reading-Writing and Math was significantly lower** than economically disadvantaged students in Alabama.

Objective 3: Changes in Achievement across Time

The third objective of this report examines changes in scholarship students' performance over time. This objective explores if greater participation in the AAA program is related to higher standardized achievement scores. Several challenges were faced in meeting this objective:

- Ideally, such an analysis would calculate the average change in national percentile scores or proficiency groups over time for scholarship students and compare it to comparable changes for public school students. A significant obstacle to this approach is the missing test data from 2019-2020 for all students. Additionally, change in scholarship students' performance from one year to the next is difficult to assess because many students do not take the same test each year due to schools changing tests, students changing schools (especially from 8th grade into high school), or no test data being available (because a student was not required to test due to his or her grade or the test report was not submitted). Thus, a large percentage of students would be excluded from this longitudinal analysis. Furthermore, as has been noted previously, ALSDE changed the required achievement test for grades 2 through 8 from the Scantron to the ACAP. As a result, for these grades, there is no appropriate longitudinal data for the ACAP.
- Second, as noted throughout this report, without a common test across the two groups of students, limited comparisons can be made. Test results were available for Alabama State and AAA scholarship students for the ACT for 11th graders. Thus, a direct comparison in performance over time could be made for this test.

With these limitations in mind, two approaches were taken to examine change over time. The first approach examined the relationship between the number of years a student had received a scholarship and their achievement test scores for the 2020-2021 academic year. Several independent correlation analyses were conducted between test performance and years in the scholarship program using the test data included in Objective 1. These correlation analyses include the greatest number of scholarship students and test types, but they do not reveal the amount of change over time, only the direction of change. Second, because the ACT has been consistently administered over the years, performance was compared between scholarship and public school students over five years.

Correlations between 2020-2021 Test Performance and Number of Years Receiving a Scholarship

Correlation analyses were used to infer a relationship between performance on the 2020-2021 achievement tests and the number of years a student was in the scholarship program. Correlations can be positive, negative, or not significant, and they can range from -1 to +1. A significant positive correlation would indicate that the longer a student was in the scholarship program, the better they performed on the achievement tests. A significant negative correlation would imply a relationship between increased years in the program and lower performance. Non-significant correlations would suggest that there is no relationship between achievement test scores and the number of years a student had received a scholarship. Finally, it should be noted that significant correlations cannot be interpreted as participation *causing* scores to change; rather they can only suggest that the two are related.

Similar to making comparisons based on mean scores or proficiency groups, a minimum sample size is necessary to detect a reliable correlation. A minimum sample size of 60 was set, which is the sample size necessary to detect a moderate relationship between test performance and the number of years receiving a scholarship. Additionally, these analyses only included students in grade 6 or higher because grades lower than that had a more restricted range for the number of years they could have received a scholarship. For example, a student in 2nd grade could have at most three years of participation in the AAA program (kindergarten, first, and second grade). A restricted range can cause correlations to be attenuated. As a result, correlation analyses were not conducted for the PSAT/NMSQT, SAT, STAR, and TerraNova 3 due to an insufficient number of students in 6th grade or higher.

First, correlations were calculated between number of years a student had received a scholarship (one to eight years) and their percentile scores in Reading, English/Language, and Math for five tests: Iowa Assessment, SAT-10, MAP Growth, Scantron, and ACT Aspire. Only three correlations were significant, each showing positive correlations:

- Iowa Assessment Language: $n = 362$, $r = .158$, $p = .002$
- Iowa Assessment Math: $n = 363$, $r = .188$, $p < .001$
- Scantron Math: $n = 96$, $r = .235$, $p = .02$

Next correlations were calculated between the number of years a student had received a scholarship and the scale scores for 10th graders on PreACT and 11th graders on the ACT. None of these correlations was significant.

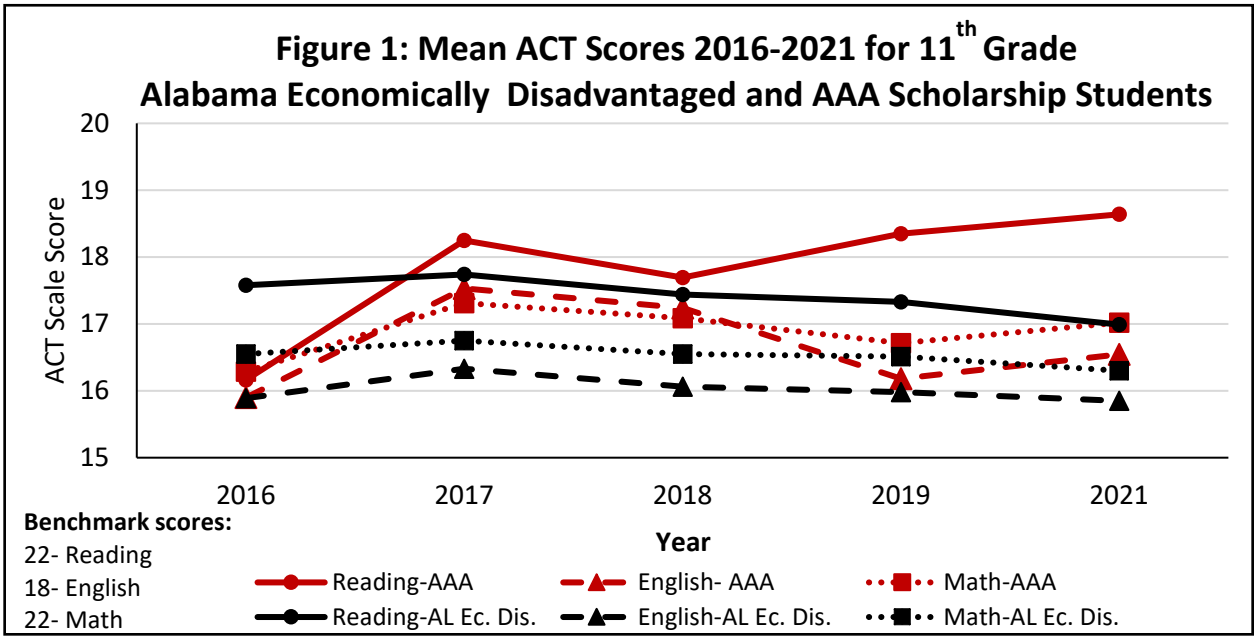
Considering these results together, out of the 20 correlations calculated, only three were significant. These positive correlations suggest that some students may improve the longer they participate in the program, but for the majority of students and the majority of tests, there was no relationship between years of participation and academic achievement. The significant correlations are also relatively small (possible range -1 to +1), which indicates that they are not strongly related to national percentile scores.

Comparison of Students in Grade 11 over Time

Compared to earlier time points, the 2021 cohort of AAA students had relatively more years attending an alternative school to their assigned public school. If the AAA program is having a positive impact on achievement, then it might be expected that scholarship recipients in more recent years should have higher scores and rates of proficiency compared to earlier cohorts. Moreover, the same level of improvement should not be evident for the economically disadvantaged public school children.

To examine change over time, mean ACT scores for 11th grade were gathered for the scholarship students starting in the 2015-2016 academic year through 2018-2019 and 2020-2021. Scores are not included for 2019-2020 because many students were not tested due to the COVID-19 pandemic. Comparable data were available from PARCA for economically disadvantaged public school children in Alabama. Before proceeding, it should be noted that often seemingly large changes in proficiency rates and scores in Figures 1 and 2 are not statistically significant. The non-significant statistical tests tell us that despite their size these are probably not reliable differences.

Figure 1 plots the mean ACT scale score for Reading, English, and Math for each group of students. Scholarship students' scores are represented in red and public school students are represented in black. The mean scores for the disadvantaged public school children varied only slightly over time, not more than one scale score point over the five years plotted in Figure 1. Statistical analyses for each subject area examined if the mean scores for AAA Scholarship students are improving over time. Results of these tests indicated that, as a group, ACT scores did not change over time. Follow-up analyses indicated that the mean Reading scores for 2019 and 2021 were significantly higher than those for 2016, but not higher than the intervening years (2017 and 2018). Because this trend is based on a small sample, it might not be reliable, so it must be viewed cautiously.



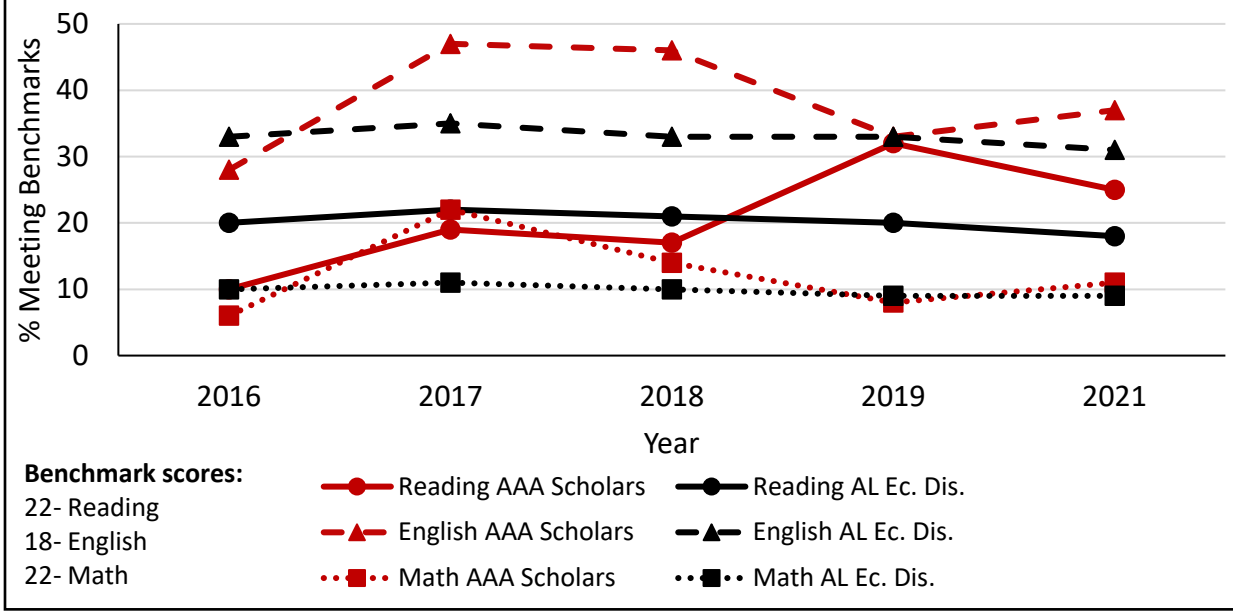
Comparisons between public school children and AAA scholarship recipients for each subject area at each time point revealed only two significant differences out of 15 comparisons. For Reading, the mean score for public school students was significantly higher than AAA students in 2016, but significantly lower than AAA students in 2021. It is also important to point out that for both groups of students, mean scale scores fell below benchmarks (22 for Reading and Math; 18 for English).

Further insight into the question of whether there are positive changes in ACT achievement scores over time can be gained by comparing the percentage of students who met benchmarks across years. Figure 2 plots the percentage of students meeting benchmark scores in Reading, English, and Math for scholarship students and economically disadvantaged Alabama public school students from 2016 through 2019 and 2021. It should first be noted that in any given year, less than 50% of the students met the benchmarks for Reading, English, or Math for both groups of students. The results revealed very little change in the proficiency rates for disadvantaged public school children. For the scholarship students the critical comparisons are for the 2019 and 2021 results compared to the earlier time points. For Reading, proficiency rates were significantly lower in 2016 compared to 2019 and 2021, suggesting some improvement over time and consistent with the findings for the mean scores. For English proficiency, rates were significantly lower in 2016

compared to 2018, but there were no other significant changes between years. This suggests that these gains were not maintained over time. Finally, for Math, the 2017 proficiency rates were higher than those for 2016 and 2019. The 2021 Math proficiency rates were not significantly different from any other year. Together this indicates that there was no consistent pattern of improvement for Math.

Additional comparisons were made between the proficiency rates for the scholarship students and the economically disadvantaged public school students at each time point for each subject area. Across 15 comparisons, with three exceptions, the proficiency rates for the scholarship students were no different from those of the Alabama economically disadvantaged group. In each of these exceptions, the scholarship students' rates were significantly higher: a) 2019 Reading, b) 2018 English, and c) 2017 Math.

Figure 2: ACT Percent Meeting Benchmarks 2016-2021 for 11th Grade Alabama Economically Disadvantaged and AAA Scholarship Students



Objective 3 Conclusion

Together, the data on scholarship students does not reveal a consistent pattern of improvement or decline over time. The correlation analyses indicated that the number of years that a student had been in the AAA program was not strongly related to achievement test scores, and this is similar to the findings reported for the 2018-2019 students in the 2020 report. The results for the 11th grade ACT scores are important because in 2021 they included students who had been in the scholarship program for an average of five years. However, there was little indication in the analyses of the mean scores or proficiency rates that the ACT scores were consistently improving over time. For Reading, analyses for both proficiency rates and mean scores showed that there was improvement in 2021 compared to 2016, but there were no significant changes among the scores from 2017 through 2021. Comparative data for economically disadvantaged public school children indicated little variation in these students' performance over the years. Overall, the 11th grade scholarship

students performed similarly to the economically disadvantaged public school children. There were a few exceptions to this generalization indicting that scholarship students performed better than public school students, but these differences did not extend from one year to the next, suggesting a lack of reliability. Together these results do not indicate strong advantages or disadvantages to being in the scholarship program. It will be important to attend to the Reading scores in future reports as there are some indications that scholarship students may be improving; however, due to fluctuations that have been observed in previous years, it is too early to tell if the current findings are reliable.

It is likely that disruptions due to the COVID-19 pandemic affected student performance in 2021 (Allen, 2021). The data in Figures 1 and 2 for the ACT for economically disadvantaged public school children show that performance in 2021 was close to pre-pandemic levels, but often slightly lower. Nevertheless, for both groups of students at all grade levels, performance generally fell below national standards. The conclusion that test scores as a group are not changing over time should not be interpreted as suggesting that individual children do not improve over time. It is possible that many students are improving, but these are offset by other students who are declining or stable over time. Also, it is important to note that a nationally representative longitudinal study of academic performance conducted by the NAEP revealed that achievement tests scores in Reading and Math have been generally stagnant or slightly declining since 2012.

Summary for Objective 3: Changes in Achievement across Time

- **The number of years of participation in the scholarship program was not strongly associated with significant improvement** on standardized tests scores.
- The lack of consistent improvement over time **followed the same pattern seen in public school students** in Alabama.

- The **number of years** that a student participated in the scholarship program was generally **not strongly correlated with higher** achievement test scores.
- On the **ACT, 11th grade** students' scores **did not show a significant long-term consistent pattern of gain or decline.**
  - Some analyses suggest that **greater participation in the scholarship program is associated with higher Reading scores, but it is too early to know if this pattern is reliable.**
  - Overall, the 11th grade **scholarship students performed similarly to the economically disadvantaged public school children,** but when there were differences, they usually favored scholarship students.

General Conclusion

This report provides the most recent assessment of how the scholarship program enacted through the AAA affects the academic achievement of scholarship recipients. The academic performance of scholarship recipients was analyzed by utilizing the demographic and test score data provided annually by the SGOs and schools that enroll students with scholarships. Many factors that impact the reliability and validity of the findings were noted throughout the report, and these are nearly all linked to the lack of a common test among schools. Within these limitations, the evaluation addressed three objectives:

- Objective 1 described the achievement test performance of the 2020-2021 scholarship recipients. Scholarship students generally performed near or below the mean percentile scores for students in the U.S. on norm-referenced tests. The findings were mixed for the proficiency rates on criterion-referenced tests. Students performed better in English/Language and Reading in which there were several instances where either the mean scale score met a benchmark or the percentage of students making the benchmark was above 50%. Math performance was more consistent, with the majority of students not meeting benchmark scores on any of the tests. The inconsistency in results between norm- and criterion-referenced tests is not easy to resolve and underscores the need for a common test across schools.
- Scholarship students were compared to economically disadvantaged Alabama public school students for Objective 2. No comparisons could be made for elementary and middle school students due to lack of a common test. On the ACT, 11th grade scholarship students' performance was comparable to economically disadvantaged public school students in all subject areas. However, comparisons on the PSAT/NMSQT indicated that the scholarship recipients performed more poorly than public school children in all subject areas. Only a small percentage of scholarship students took the ACT or the PSAT/NMSQT, which hampers the ability of this report to draw definitive conclusions.
- The third objective assessed if scholarship recipients' achievement scores improved over time. The number of years that a student participated in the program was weakly, but inconsistently correlated with test performance. The lack of a reliable relationship across tests, and the lack of replication from the 2020 report does not give us strong confidence in this result. Additionally, students who took the ACT in 11th grade had the highest rate of participation in AAA, and performance on this test has not reliably improved over time, similar to their public school counterparts.

Overall, the results for the three objectives replicate the pattern from previous reports in that the majority of AAA scholarship students performed similarly to their peers in public schools and often fell below national expectations for their grades. Performance was generally better in Reading and English/Language than Math, similar to public school children. The objectives of the AAA program are to improve the learning outcomes of students zoned to attend failing public schools in Alabama, and the results from this report indicate that this goal has yet to be realized. In the 2020 report, we noted some optimism for the future as school accreditation requirements included in the AAA legislation took effect. However, improvements that might have resulted from higher quality schools could have been obscured by the adverse impacts of the COVID-19 pandemic on daily school activities. The next report, due in 2024, should include three years of post-pandemic outcomes, which may provide more consistent and conclusive findings.

Limitations

The reporting requirements set forth in the AAA legislation do not permit a rigorous, tightly controlled evaluation of the program. Since the initial report in 2016, we have noted that the lack of a common assessment severely limits our capacity to draw strong conclusions regarding the academic achievement of scholarship recipients relative to students attending public schools. In 2020-2021 only a few students took the ACAP, and although more scholarship students could be directly compared to public school children on the ACT (97 11th graders), they only represented 4% of the AAA students who were required to test. An accurate model of the effects of the scholarship program would require statewide student-level assessments that use the same standardized test and link test scores to student demographic information.

This report made the best use of the data available, but there are challenges inherent in working with the data, which are noted here and in previous reports. For example, results for a particular test (e.g., the TerraNova 3) are confounded by idiosyncratic characteristics of the schools that use that test, such as the composition of race, household income, or number of years receiving a scholarship. These confounding factors cannot be readily accounted for in the evaluation. Small sample sizes for some results also impact the statistical reliability of the report. Additionally, missing data from students who were required to test (see Flowchart p. 7) may decrease the validity of the findings. The most meaningful comparison between scholarship recipients and public school students would contrast scholarship students' performance to the performance of students in the public school for which they were zoned, rather than aggregating across all schools in the state. However, lack of a common test among schools and small sample sizes made these comparisons impossible to conduct.

Some schools opted to evaluate student performance using tests with outdated national norms, and several schools provided results for tests taken in the fall. Using older tests may save money for a school, but the value of this practice for evaluating student learning is questionable. As noted in Objective 1, tests taken in the fall do not represent the learning a student has achieved during the year.

Finally, due to COVID, many students were not tested during the 2019-2020 school year, making the measurement of the change in student performance more complex. Due to the variability in tests used by students in grades 2 through 8, we could not measure how these students performed over time or how their scores were affected by the pandemic with accuracy. Although results were reported for the ACT and PSAT/NMSQT, it should be kept in mind that the comparisons involve different cohorts of students, rather than tracking the same students over time. The students in the 2020-2021 cohort of 10th and 11th graders may differ from previous cohorts in terms of the private schools attended, household income, and other factors associated with performance on standardized tests.

In closing, it is important to recall that the AAA scholarship program targets low-income students and has been utilized by families belonging to demographic groups (e.g., racial minorities) that have historically lagged behind others in the state and the U.S. in academic achievement. This report, along with state and national data, make it evident that sustained and lasting improvement for low income students is difficult to achieve.

Glossary of Terms

Common core. The Common Core is a set of academic standards for what every student is expected to learn in each grade level, from kindergarten through high school in the U.S. They cover math and English language arts.

Computer adaptive test. A test delivered through a computer terminal in which students are given harder or easier questions as they proceed through the test based on whether their answers are correct (resulting in harder questions) or wrong (resulting in easier questions).

Criterion-referenced test. These tests assess students' learning against a fixed set of predetermined learning standards that are specific for their grade level. In an ideal school, every student would meet the criterion score for their grade level.

Economically disadvantaged student. An ALSDE designation applied to public school children who qualify for free or reduced lunch subsidies.

Fixed form test. A test in which students are shown the same questions regardless of ability or performance.

Mean. A mean test score is calculated by adding together every score in a group and dividing by the number of people in the group. It is one way to represent the score of a typical person in the group.

National percentile. National percentile scores can range from 1 - 99. The percentile rank indicates the percent of students nationwide who scored lower than a particular raw score on the same test at the time the norms were compiled.

Norm-referenced test. These tests are designed to compare student achievement relative to others at a particular grade level with the goal of distinguishing between high and low achievers. National percentile scores are commonly used as a reference point for these tests, with the 50th percentile indicating the score achieved by the average student in the U.S.

Proficiency Scores/Groups. Proficiency groups provide an assessment of student achievement based on a set of criteria, such as national educational standards or college readiness.

Raw score. A raw score is the number of items that a student answered correctly on a test.

Scale(d) score. A scaled score is a mathematical transformation of a raw score. Scaling provides a continuous metric across the different forms and levels of a test (such as tests for different grade levels). Higher scale scores indicate higher levels of academic achievement.

Scholarship Granting Organization (SGO). An organization that provides educational scholarships to eligible students attending qualifying schools. SGOs receive donations from individuals and corporations (subject to limitations imposed by the Alabama Accountability Act), which are then distributed in the form of scholarships to eligible students. Donations by taxpayers cannot be restricted or conditional with respect to how the donation is applied to scholarship recipients or schools.

Statistically significant difference. The difference between two or more scores is considered significantly different when there is a low probability (usually 5% or less) that the difference could occur by chance. When a statistically significant difference is observed between the mean scores of two groups of students, it suggests that the difference is likely to be a "real" difference.

References

ACT Aspire. (2021). Interpretive guide for ACT Aspire summative reports. Retrieved from: https://success.act.org/s/article/ACT-Aspire-Interpretive-Guide-for-Summative-Reports

ACT. (2020). The PreACT technical manual (fall 2020 version 1). Retrieved from: https://success.act.org/s/article/PreACT-Technical-Manual

Allen, J. (2021). Have ACT Scores Declined During the COVID-19 Pandemic? An Examination of Fall State and District Testing Data. Retrieved from: https://www.act.org/content/act/en/research/pdfs/COVID-Impact-Fall-ACT-State-and-District-2021-5.html

College Board. (2021). PSAT/NMSQT understanding scores. Retrieved from: https://satsuite.collegeboard.org/media/pdf/psat-nmsqt-understanding-scores.pdf

College Board. (2021). SAT suite of assessments annual report Alabama. Retrieved from: https://reports.collegeboard.org/sat-suite-program-results

College Board. (2021). SAT understanding scores. Retrieved from: https://satsuite.collegeboard.org/media/pdf/understanding-sat-scores.pdf

He, W., & Meyer, J. (2021). MAP Growth universal screening benchmarks: Establishing MAP Growth as an effective universal screener. NWEA. Retrieved from: https://www.nwea.org/content/uploads/2021/05/MAP-Growth-Universal-Screening-Benchmarks-2021-03-12_NWEA_report.pdf

Kuhfeld, M., Soland, J., Lewis, K., & Emily Morton, E. (2022). The pandemic has had devastating impacts on learning. What will it take to help students catch up. *Brown Center Chalkboard*, March, 3. https://www.brookings.edu/blog/brown-center-chalkboard/2022/03/03/the-pandemic-has-had-devastating-impacts-on-learning-what-will-it-take-to-help-students-catch-up/

Oladele, J. I., & Ndlovu, M. (2021). A review of standardized assessment development procedure and algorithms for computer adaptive testing: applications and relevance for fourth industrial revolution. International journal of learning, teaching and educational research, 20(5).

Renaissance. (2022). Renaissance star assessments administration manual: instructions for administering star early literacy, star reading, and star math assessments. Retrieved from: https://doc.renlearn.com/KMNet/R61649.pdf