Evaluation of the Alabama Accountability Act: Academic Achievement Test Outcomes of Scholarship Recipients through 2018-2019

The Institute for Social Science Research
The University of Alabama

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

September 1, 2020

Executive Summary

This report fulfills the 2013 Alabama Accountability Act evaluation requirements by examining the academic achievement of scholarship recipients through the 2018-2019 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.

Scholarship Granting Organizations 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 (PARCA), and the ACT Inc.

Some challenges were encountered in conducting the evaluation:

- The lack of a uniform achievement test among schools continued to constrain the accurate description 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.
- Inconsistencies in test score reporting from schools and missing test data limited the number of students who could be included in the evaluation sample.
- Test score information from ALSDE for grades 3-8 included only the percentage of students in proficiency groups, limiting the types of analyses that could be conducted.

The evaluation was based upon test scores from 1,929 scholarship recipients attending 105 schools in 43 counties. This represented 80% of the scholarship recipients in the grades for which testing was required. These students varied in their demographic characteristics:

- Number of years receiving a scholarship:
  - 18% were first time scholarship recipients.
  - 17% were two- or three-time scholarship recipients.
  - 65% were in their fourth year or more of receiving a scholarship.
- 90% were eligible for free/reduced lunch subsidies.
- 30% were zoned to attend a failing school.
- 63% were Black/African American (AA), 17% were White/Caucasian, and 15% were Hispanic.

Continues

Although this report can show trends for this subsample of scholarship recipients, due to the necessity of excluding a significant proportion of scholarship recipients (20%) from analyses, findings may not be representative of all of the scholarship recipients.

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

- On norm-referenced tests, scholarship recipients' performance was mixed:
  - For the Stanford Achievement Test (2018 Norms), the scholarship recipients were below the national average for English, reading, and math.
  - For the TerraNova (2017 Norms), scholarship recipients did not differ from the national average. However, the average scores for Black/AA students was below the 50th percentile.
  - For the Iowa Assessments (S2017 Norms), scholarship recipients in grades 5-7 and Black/AA in all grades performed below the national average. The average scores for students in grades 3 and 4 were either not different from the 50th percentile or above it.
- On criterion-referenced tests:
  - The majority of scholarship recipients met the proficiency benchmarks for English but failed to meet proficiency benchmarks for reading and math.
  - Outcomes were poorer for Black/AA participants who made up the majority (67%) of scholarship recipients.
- In contrast to previous reports in which the majority of scholarship recipients fell below national norms and benchmarks, the pattern this year is mixed, with results indicating that performance was often on par with these standards for English on criterion-referenced tests.

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

- In grades 4-8 scholarship students' rates of academic achievement proficiency were lower than economically disadvantaged public school students for math but were not different for reading.
- Eleventh grade scholarship students' proficiency rates for English and math were comparable to economically disadvantaged public school students and higher than this group for reading, although Black/AA scholarship students performed more poorly in all subject areas.
- Six years after the passage of the AAA, there is no evidence that the scholarship program has resulted in academic achievement that is superior to Alabama public schools, and majorities in both groups fail to meet academic benchmarks.

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

- On average, over time, participating in the scholarship program was not associated with significant improvement on standardized tests scores.
- The lack of change over time followed the same pattern seen in public school students in Alabama and is likely not attributable to participation in the scholarship program.

Table of Contents

List of Charts and Tables

Page

List of Charts, Tables, and Figures-Continued

List of Abbreviations

AAA	Alabama Accountability Act
AA	African American
AL	Alabama
ALSDE	Alabama State Department of Education
Econ. Dis/Disadv	Economically Disadvantaged
FERPA	Federal Education Rights and Privacy Act
ISSR	Institute for Social Science Research
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
SGO	Scholarship Granting Organization

Evaluation of the Alabama Accountability Act: Academic Achievement Test Outcomes of Scholarship Recipients through 2018-2019

Introduction

In September, 2016, the Institute for Social Science Research (ISSR) at the University of Alabama completed the first state-mandated evaluation of the academic outcomes of students receiving scholarships under the Alabama Accountability Act (AAA) as set forth in the AAA legislation. Thus far, in four previous reports, ISSR has described the achievement test results from the 2014-2015 through the 2017-2108 academic years, compared the outcomes to students attending public schools in Alabama, and examined changes in scholarship recipients' achievement test scores over time in comparison to comparable children attending public schools in Alabama. The current report follows a similar approach with the 2018-2019 achievement test results.

This report first provides an overview of the pertinent AAA legislation. The methodology is described, next, which includes a description of the 2018-2019 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 2018-2019 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

This report fulfills the evaluation component of the 2013 Alabama Accountability Act by providing evidence for the academic achievement of scholarship recipients in the 2018-2019 academic year. The Alabama Accountability Act (AAA), passed by the legislature in 2013 and amended in 2015, established a 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 the standards set by the AAA. The AAA places restrictions on who can receive scholarships based on family income. 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 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 implementation of the AAA.

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.

Thus, the current 2020 report has three major objectives: a) describe the academic achievement of students in the scholarship program for the 2018-2019 school year, b) make comparisons between the learning achievement of the scholarship recipients and comparable students attending public schools for the 2018-2019 school year, and c) measure the achievement gains of students in the scholarship program over time.

Alabama State-Mandated Testing in Public Schools 2018-2019 Academic Year

Students attending public schools in Alabama during the 2018-2019 academic year were tested in March and April. Math and reading were assessed with the Scantron Performance Series for students in grades 3-8. Alabama tenth graders took the PreACT and eleventh graders were required to take the ACT college entrance exam.

Method

As in previous years, several challenges to meeting the evaluation objectives set forth in the AAA were encountered. Primary among these is the lack of a uniform achievement test among schools, which limits the conclusions that can be made about student learning gains. Schools provided scores from a total of 19 unique 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 Assessments and the Stanford Achievement Test, and criterion-referenced tests, such as the ACT

Aspire and Scantron Performance series, are based on different standards and cannot be directly compared. 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 entrance to college. In theory, 100% of students could achieve these criterion benchmarks. In contrast, 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, a student 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 his or her grade level taking the same test. 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. Further, even when the same test is used across schools, students at different schools often have scores that are based on different norms. For example, one school may report test scores based 2017 norms while another school may report test scores based on 2011 norms, which often not comparable.

Although improvements have been made, the schools continue to be inconsistent in providing test reports with national percentile and scale scores for math, reading, and language arts/English. This missing data compromises the integrity of the report findings, and ISSR continues to work with the SGOs to ensure that the schools provide appropriate test reports. 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.

An improvement to the 2020 report is that greater attention is given to the norms used in calculating percentile or proficiency scores so that a more accurate assessment of scholarship students' academic performance can be given. For example, the Iowa test results could have been reported using norms developed in 2017, 2011, or 2005, and the Stanford Achievement Test scores included four sets of norms, 2018, 2007, 2005, and 2002; whereas the TerraNova 3 used 2017 norms. Norms dated more than five years ago are not comparable to newer norms. The older tests are not based on the Common Core, the current national standards for children in grades K-12. Given this variability, descriptive statistics are provided for each test and norm, but when drawing conclusions about the overall performance of scholarship recipients, the report focuses on test score data based on the most recent norms.

With these challenges noted, the remainder of the report describes outcomes for the 2018-2019 academic year. Statistical comparisons were conducted throughout the report to aid in drawing conclusions. T-tests were used to compare the average 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 scores of multiples groups, such as changes in scores over multiple years. Z-tests were used to compare the percentages of scholarship students meeting benchmarks to comparable indicators of public school students. Correlations were used to assess the relation between achievement test scores and the number of years of participation in the AAA scholarship program. 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.

Data Sources

The following data sources were used to evaluate the academic achievement of the 2018-2019 scholarship recipients:

- Demographic reports from each year of the program from eight SGOs: Scholarships for Kids, AAA Scholarship Foundation, Alabama Opportunity Scholarship Fund, Rocket City Scholarship Granting Organization, Children's Tuition Fund, 100 Black Men of Mobile, and Renaissance Scholarship Fund.
- Test reports collected by the SGOs from participating schools and shared with ISSR. Test scores were received as PDFs or hard copies (2014-2015 through 2018-2019).
- 2017-2018 and 2018-2019 Alabama State Scantron Performance Series results available from the ALSDE website.
- The 10th grade Pre-ACT and the 11th grade ACT results for all students in Alabama available online from the ACT Inc. website.
- The 11th grade ACT scores for public school students in Alabama retrieved from the Public Affairs Research Council of Alabama (PARCA) report available on their website.

2018-2019 Sample

The first part of this report focuses on the new data from the 2018-2019 academic year, as earlier reports have analyzed the previous academic years. The SGOs reported that a total of 3,683 students (50% female) in kindergarten through 12th grade had received scholarships during the 2018-2019 academic year. The majority of the students (71%) had received at least one previous scholarship: 9% had received one previous award, 13% had received two previous awards, and 49% had received three or more previous awards. Nearly all students were free/reduced lunch eligible (91%). The scholarship recipients primarily represented three racial/ethnic groups, Black/African American (Black/AA; 68%), White/Caucasian (16%), and Hispanic (12%). Four percent (4%) were another race or no information was provided. Students resided in 62 counties in the state, with approximately 32% zoned to attend a failing school based on the SGOs' reports.

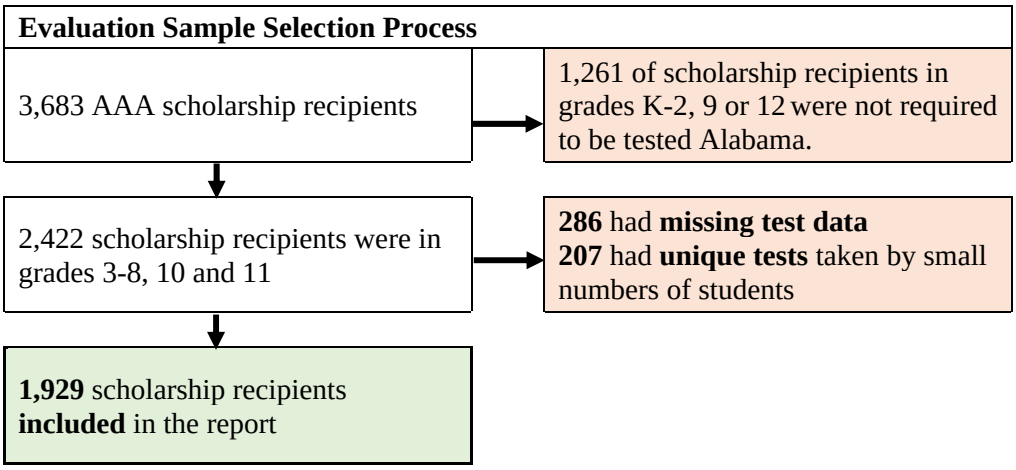
Achievement Test Data for 2018-2019 Scholarship Participants

A total of 2,422 scholarship students were in grades 3-8, 10, and 11, which are the grades the state requires to be tested for reading and math. These grades are the focus of this report. Students in grades kindergarten through second grade and grades 9 and 12 comprised 1,261 (34%) of scholarship recipients and were not required to be tested according to the AAA.

Data for 493 students in the grades required to test were not included in the report for two reasons: 1) the test data were missing or 2) the school used a unique test.

1. Test score reports were provided for 2,136 (88%) students but were missing for 286 (12%) students. Test scores were missing 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. In the cases where the school did not test a child who was enrolled in their school, the most common explanations were that the school did not test a particular grade, or the child had a disability (although the ALSDE requires such students to be tested). 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. Nineteen different standardized tests were given by 130 different schools, and unfortunately, some schools used tests that few schools or no other school used. These schools typically had a low number of scholarship recipients. Making these test results public (especially when disaggregated by grade, race, or gender) would lead to undesirable results: a) Schools and individual children could be identifiable; the latter is a violation of FERPA; and b) Small samples, as noted earlier, are not likely to be representative of the full group of scholarship recipients. For these reasons, results from these schools would not contribute meaningfully to the AAA evaluation, and therefore, the 207 students (less than 9% of those required to test) attending these schools were excluded from this evaluation. Figure 1 provides a flow chart that summarizes factors affecting the 2018-2019 sample size.



A total of 1,929 students or 80% of students for whom testing was required according to the AAA had potentially reportable test data from eight standardized tests: 1) ACT Aspire, 2) Scantron Performance Series (also used by ALSDE), 3) The Stanford Achievement Test 10, 4) TerraNova 3, 5) The Iowa Assessment, 6) The PreACT (practice college entrance exam), 7) The ACT (college entrance exam), and 8) The Practice SAT-National Merit Scholarship Qualifying Test (PSAT/NMSQT).

The 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 105 unique schools. The discrepancy between this total and the numbers listed in the table is due to some schools giving more than one test (e.g., a K-12 school might give the ACT Aspire for grades 3-8, the

PSAT/NMSQT for grade 10, and the ACT for grade 11). Further attrition occurred because schools might not have included a particular subject area in their reports, did not report usable scores (e.g., number correct) or individual students may not have tested in a subject area. These instances are described as the results for each test are presented.

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

Test	Number of Students	Number of Schools
ACT	81	22
ACT Aspire	152	18
Iowa Assessments	974	49
PreACT	111	8
PSAT/NMSQT	174	18
Scantron	91	3
Stanford Achievement Test 10	178	20
TerraNova	168	9
Total	**1929**	

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 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 eight tests follows.

- *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. ACT Inc provides college readiness benchmarks, and ALSDE has set proficiency benchmarks for high school students. Subscale scores are provided for reading, English, and math.
- 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 scores for reading, English, and mathematics, in addition to other areas. National percentile scores are also provided that are interpreted similar to those in other tests.

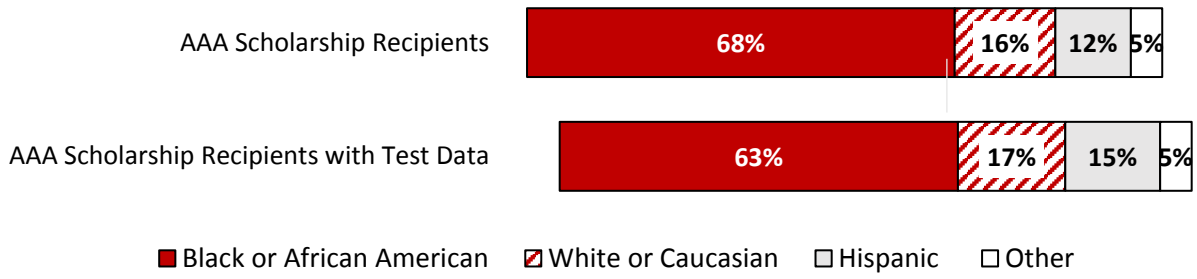
- *Iowa Assessments (previously Iowa Test of Basic Skills)* was developed by the Education Department at the University of Iowa and is also a norm-referenced test. Test items were originally developed to align with the Iowa Core of State Educational Standards. The test has been validated at the national level, and it provides national percentile scores for reading, English, and math. 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. In contrast to criterion-referenced benchmarks, 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. This report includes test results based on national norms developed in 2017, 2011, and 2005.
- *The Practice SAT-National Merit Scholarship Qualifying Test* (PSAT/NMSQT) is used to prepare students to take the 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, as well as a subscale score in math and a combined reading and writing subscale score. National percentile scores are provided for all subject areas.
- *The PreACT* is used to prepare high school students to take the college ACT. 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. Proficiency benchmarks are provided by ACT Inc. for both 10th and 11th grades to assess college readiness. ALSDE has set proficiency benchmarks for high school students. Subscale scores are provided for reading, English, and math. High school students commonly take this test their first and second years of high school.
- *Scantron Performance Series* is a criterion-referenced computer-adaptive test developed to provide a longitudinal view of student growth in various subject areas. In addition to scale scores and national percentiles, the Scantron provides benchmarking (Placement Indicator Quartiles/ Performance Bands). The ALSDE phased out the ACT-Aspire in 2017 and used the Scantron for testing students in grades 3-8 for the 2017-2018 and 2018-2019 school years. ALSDE has adapted the benchmarking used by Scantron to assess students at four levels; 1) *Emerging Learner*, 2) *Developing Learner*, 3) *Proficient Learne*r, and 4) *Distinguished Learner.* This report includes reading and math scores.
- The *Stanford Achievement Test 10* is a norm-referenced test and was developed, among other reasons, to compare a child's academic achievement relative to others in the nation. The scale scores follow a bell curve, or normal distribution. The Stanford provides achievement/ability scores in language arts, reading, and math. This report includes scores based on 2002, 2007, and 2018 norms.
- *TerraNova, 3rd edition* is a norm-referenced test similar to the Stanford Achievement Test and Iowa Assessments. 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 Stanford Achievement Test. Included in the report are scores for language arts, reading, and math. The scores reported in the report are based on 2017 norms.

Demographic Information for Scholarship Recipients Included in the Evaluation

Based on information provided by the SGOs, the 1,929 scholarship recipients with usable test scores differed somewhat from the larger group in their previous enrollment in the scholarship

program. For both sets of students the majority had previously received a scholarship, but there was more than a 10% difference between the two percentages: 71% for all recipients vs. 82% for those included in the evaluation. This discrepancy is 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. In the evaluation sample (18%) were first time scholarship recipients, 7% were two time scholarship recipients, 11% were three time recipients, and 65% were in their fourth or higher year. The two samples were similar in other demographic characteristics. Nearly all of the students in evaluation sample were eligible for free or reduced lunch (90%). The SGOs reported that 30% of the scholarship recipients were zoned to attend a public school that was designated as failing by the ALSDE. As with the larger sample (Chart 1), the racial/ethnic make-up of the sample was predominantly from three groups, Black/AA (63%), White/Caucasian (17%), and Hispanic (15%), and the remaining 5% of students were either another race, more than one race, or no race was designated. About half (51%) of the students in this group were female. Students represented 42 counties in the state and attended 105 different schools.

Chart 1
Racial Demograhics of the Evaluation Sample Compared to the Total Population of Scholarship Recipients



Findings for the 2018-2019 Academic Year

Objective 1: Describe the Academic Achievement of Scholarship Recipients

In this section, outcomes are described for each of the eight tests for the 2018-2019 academic year. For each test a brief description of the student demographics is provided, and additional test details relevant for 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 most tests. When relevant, scale scores were reported to aid in interpreting the test score information, usually to describe outcomes related to benchmarks or proficiency groups associated with criterion-referenced tests 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 three tests, Stanford Achievement Test 10, TerraNova, and Iowa Assessments are norm-referenced tests. The criterion-referenced ACT Aspire, Scantron, PreACT, ACT, and PSAT/NMSQT are summarized next. The AAA 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 uses reading scores to evaluate public school students, 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/Caucasian, and Hispanic groups when enough student data was available.

Norm-Referenced Test Results

Stanford Achievement Test 10

The Stanford Achievement Test 10 was given to 178 students in grades 3 through 8, 10, and 11. School test reports included four different sets of norms (2002, 2005, 2007, and 2018) and the reports from one school did not include the norm information (n = 12). In past reports, scores using the 2002 norms were converted to the 2007 norms so that more student data could be included. As noted in the introduction, an improvement made this year is to report scores separately for each norm year. Students whose scores were reported using the 2005 norms (n = 9) and those students for whom no norm information was available are excluded from this report. This resulted in a sample of 157 students. Among these students, 83% were repeat scholarship recipients. Nearly two-thirds (65%) were in their 4th year or more of being a scholarship student, 18% were in their second or third year, and 17% were first-time scholarship students. The free and reduced lunch rate was 87%. The racial/ethnic make-up was 82% Black/AA, 13% White/Caucasian, and the remaining students (5%) either had no information on race or were classified as another racial group. There were slightly more females (52%) than males.

Results for Stanford 2018 Norms. There were 74 students in grades 3-8, 10, and 11 whose scores were reported using the 2018 norms. Unfortunately, no grade level had the minimum of 25 students. Consequently, all grade levels were combined (Table 2). There were enough Black/AA students to report their scores separately, but not for any other racial group. Statistical comparisons between males and females indicated that female students had statistically significantly higher scores for language arts. Reviewing the data in Table 2, it appears that the average percentile scores fell near the bottom third of test takers nationally.

Table 2
Stanford 2018 Norms: Mean National Percentile Test Scores for All Grades

Grades	Group (N)	Reading	Language	Math
3-8, 10	All (71-74)	31	31	19
	Black/AA (52-55)	27	28	17
	Female (33-35)	34	38	19
	Male (38-39)	28	25	19

Results for Stanford 2007 Norms. Thirty-two (32) students in grades 4-8 took a test that reported 2007 norms. No grade level met the minimum standard of 25 students, so the scores presented in

Table 3 were averaged across grades 4-8. The only racial group to meet the sample size minimum was Black/AA, and there were not enough scores to report results separately by gender. Statistical comparisons could not be made among racial/ethnic groups due to the lack of a comparison racial group with a sufficient sample size (25 or more). The scores presented in Table 3 were between the 31st and 39th percentile nationally, suggesting the students were performing more poorly than the majority of students in the country.

Table 3
Stanford 2007 Norms: Mean National Percentile Test Scores for All Grades

Grades	Group (N)	Reading	Language	Math
4-8	All (32)	39	36	31
	Black/AA (27)	39	36	32

Results for Stanford 2002 Norms. This test was taken by 51 students in grades 3-8 and 10. No grade level met the minimum standard of 25 students, so the scores presented in Table 4 were averaged across grades. The only racial group to meet the sample size minimum was Black/AA, and there were enough scores to report results separately only for males. Statistical comparisons could not be made among racial/ethnic groups or gender due to the lack of an appropriate comparison group with a sufficient sample size (25 or more). The scores presented in Table 4 were between the 57th and 61st percentile nationally, suggesting the students on average were performing better than the majority of students in the country in 2002.

Table 4
Stanford 2002 Norms: Mean National Percentile Test Scores for All Grades

Grades	Group (N)	Reading	Language	Math
3-8, 10	All (51)	59	57	57
	Black/AA (46)	58	57	56
	Males (29)	59	57	61

TerraNova 3

TerraNova test scores (2017 norms) were submitted for 168 students. However, one school representing 103 students failed to provide composite scores for reading, language, or math. Additional schools failed to provide usable data for five students, resulting in a final sample of 60 students whose data could be included in the report. These students were in grades 3-8 and 10. This group of students was 45% male, 57% Black/AA, 40% White/Caucasian, and 3% all remaining racial groups. First time scholarship recipients comprised 27% of the students, 20% had received a scholarship for two to three years, and 53% had been in the scholarship program for four years or more. The vast majority (79%) were free/reduced lunch eligible. Due to the reduction in the number of students with usable test scores, scores are reported across all grades combined. Scores are reported separately for Black/AA students and by gender. Statistical comparisons between males and females indicated no significant differences between the two genders for any subject area. The scores presented in Table 5 are between the 39th and 51st percentile.

Table 5
TerraNova: Mean National Percentile Test Scores for All Grades

Grades	Group (N)	Reading	Language	Math
3-8, 10	All (60)	47	46	44
	Black (33-34)	42	39	39
	Female (32-33)	51	51	44
	Male (27)	42	39	44

Iowa Assessment

The Iowa was administered to 974 students in grades 3-8, 10, and 11. The racial/ethnic make-up consisted of 57% Black/AA, 16% White/Caucasian, and 21% Hispanic students. First time scholarship recipients comprised 15% of the Iowa test takers, 13% had received a scholarship for two or three years, and 72% were in their fourth year or higher of receiving a scholarship. The vast majority were free/reduced lunch eligible (90%), and 54% of the test takers were female.

Students were given versions of the Iowa test based on Spring 2005 (*n* = 180), Fall 2011 (*n* = 13), Spring 2011 (*n* = 202), and Spring 2017 (*n* = 579) norms. The number of students in grades 10 (*n* = 10) and 11 (*n* = 3) were too small to provide reliable results. For the 2011 norms, only tests taken in the spring were included because this corresponds to when Alabama public schools administer the test. Thus 961 student scores were included in the analysis. Statistics were calculated separately for the S2017, S2011, and S2005 norms.

Results for Iowa S2017 Norms (*n* = 579). Examining the results for all students at each grade level revealed that average scores ranged from 36% (6th grade math) to 56% (3rd grade English), that (Chart 2).

Chart 2
Iowa Test S2017 Norms: Mean National Percentile Scores for Grades 3-8

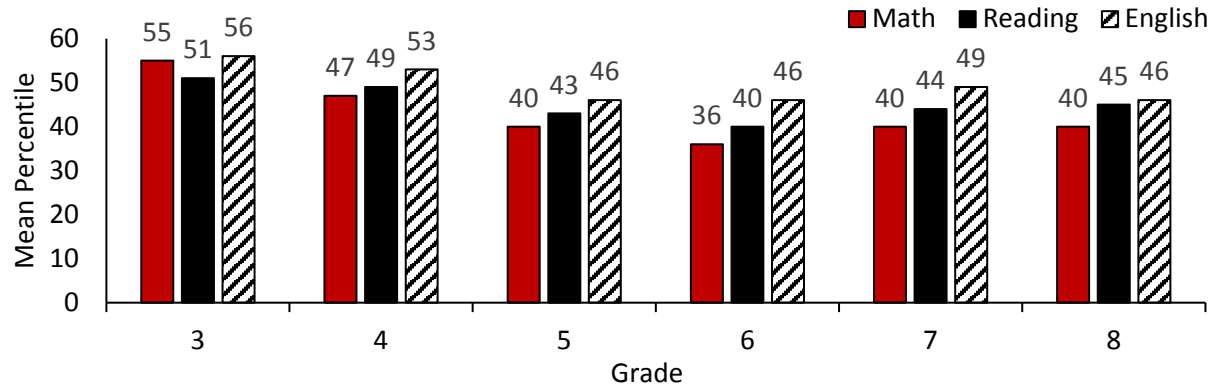
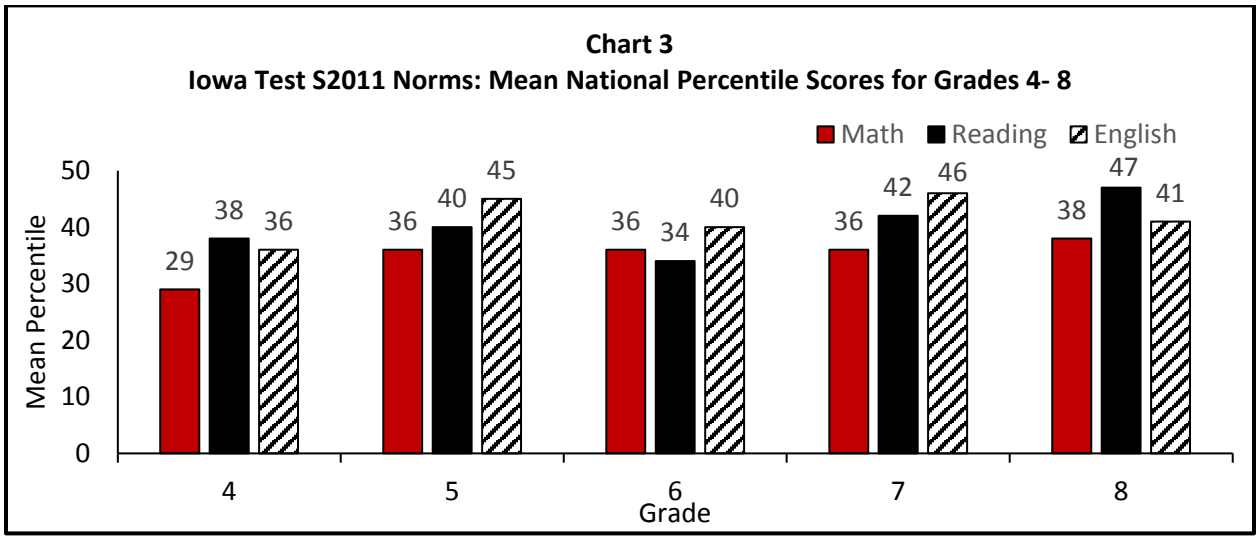


Table A1 in the Appendix provides the detailed data for each grade level disaggregated by gender and race when appropriate. There were enough students to compare Black/AA and Hispanic students in grades 3-6. Students performed similarly on all subjects in the 3rd grade regardless of

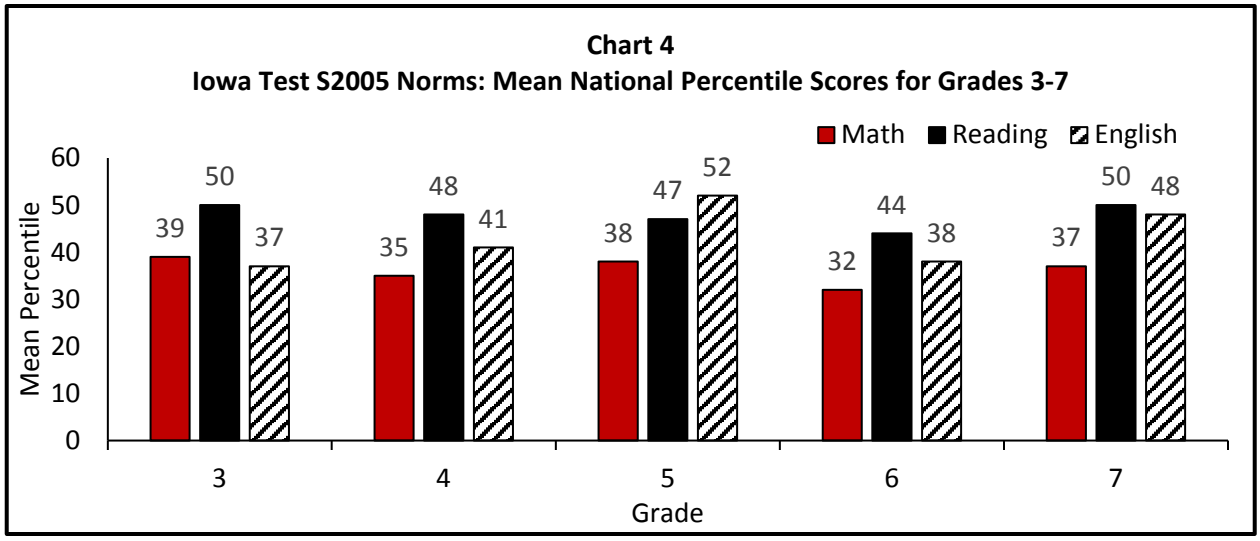
race. Black/AA students performed significantly lower than Hispanic students in several of the comparisons that were made: in grades 4-6 for math, grades 4 and 6 for English, and in grade 6 for reading. See Table A1 for the means.

Comparisons between genders for each grade level yielded three significant effects: In the 3rd grade, girls scored significantly higher than boys in reading (Mean percentiles 57 vs. 44) and English (Mean percentiles 62 vs. 49). In the 4th grade, girls scored significantly higher than boys in English (Mean percentiles 58 vs. 46).

Results for Iowa S2011 Norms (*n* = 202). Students who took the Iowa test using S2011 norms had mean percentile scores that ranged from a high of 47 (8th grade English) to a low of 29 (3rd grade math; Chart 3). For most grades, there were not enough students to disaggregate scores by race or gender, but in grade 6, there were enough Black/AA and male students to disaggregate the data (Table A2 in the Appendix). No statistical comparisons could be made due to the lack of a comparison group with a sufficient number of students.



Results for Iowa S2005 Norms (*n* = 180). Chart 4 presents the mean percentile scores for students in grades 3-7. The highest mean percentile score was 52 (5th grade English) and the lowest was 32 (6th grade math), all below the national median. There were not enough students to disaggregate test scores by gender, but in some grades (grades 3 and 5-7), there were enough to report means for Black/AA students. (Table A3 in the Appendix). No statistical comparisons could be made due to the lack of a comparison group with a sufficient number of students.



Summary for Norm-Referenced Test Results

In considering the findings across the three norm-referenced tests it is important to recall that the percentile scores are an assessment of students' performance relative to other children at the same grade level in the country. By themselves, the scores do not indicate if a child has acquired the knowledge and skills expected for their grade. As has been noted in previous reports, although the 50th percentile is often used as the yardstick for evaluating performance, 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. A review of the scores from the three tests indicates that nearly all of the average percentile scores were below the 50th percentile—a finding consistent with previous evaluations of the AAA. Statistical comparisons to the 50th percentile were made separately for each of the three tests, utilizing data from only the most recently normed edition. As noted earlier, these norms provide the best indicators for how students are performing using the most contemporary educational standards.

Considering first the Stanford Achievement Test (2018 Norms; Table 2), the average scores for the combined grade levels, Black/AA students, and each gender were less than the 50th percentile at statistically significant levels for all subject areas.

For the TerraNova (2017 norms; Table 5), three scores out of the nine presented in Table 5 were significantly below the 50th percentile: reading and math scores for Black/AA students and reading scores for males. The remaining scores were not significantly lower or higher than the 50th percentile, suggesting these students performed near the median of U.S. students taking this test.

The relevant scores for the Iowa test are based on the 2017 norms and appear in Chart 2 and Table A1. Statistical comparisons revealed a mixed pattern of findings. The results for grade 3 were the most positive, with both English and math scores testing significantly higher than the 50th percentile, but reading scores showing no statistical difference. For 4th grade there were no significant differences. Grades 5-7 showed a similar pattern indicating that math and reading

scores were significantly below the 50th percentile, but English scores were not statistically lower or higher than that 50th percentile. In 8th grade, math scores were also significantly below the 50th percentile, but reading and English were not statistically higher or lower than the 50th percentile.

Considering different racial groups on the Iowa test, Black/AA students' scores were below the 50th percentile in grades 4-8. Hispanic students in grades 3-6 were generally not statistically different from the 50th percentile, similar to the sample as a whole. The only exception was that Hispanic students in 6th grade performed significantly below the mean for English.

<table>
<tr><th colspan="3">Summary for Norm-Referenced Test Results</th></tr>
<tr><td colspan="3">Scholarship students as a group did not perform better than other students in the U.S.
It was most typical for students to perform near or below the 50th percentile compared to the nation, but this pattern varied depending on the standardized test.
Black/AA students generally performed below the 50th percentile on tests.
Comparisons among racial groups were limited to only the Iowa Assessment, and as a result, no conclusions can be drawn.
There are 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 and outdated test norms adversely impact the reliability of some findings.</td></tr>
<tr><th>Stanford Achievement Test
(2018 Norms)</th><th>TerraNova
(2017 Norms)</th><th>Iowa Test
(2017 Norms)</th></tr>
<tr><td>The average scores for each grade level and subject area were below the 50th percentile for tests adopting the 2018 norms.
Results by racial group:
⊙ Similar to the sample as a whole, Black/AA students typically performed below the 50th percentile.</td><td>The mean percentile scores combined across all grade levels were not significantly different from the 50th percentile.
Results by racial group:
⊙ Black/AA students performed below the 50th percentile in reading and math.</td><td>The average scores were significantly below the 50th percentile in grades 5-7 for reading and math and 8th grade math but were above or no different from the 50th percentile in all subjects for 3rd and 4th grades and for English in grades 5-8.
Results by racial group:
⊙ Black/AA students scored significantly lower than the 50th percentile in grades 4-8.
⊙ Hispanic students' scores were not significantly above or below the 50th percentile.
⊙ Hispanic students performed better than Black/AA students in all subjects for grade 6, in math for grades 4 and 5, and in English in grade 4.</td></tr>
</table>

The summary graphic provides the key findings for the norm-referenced tests. It is difficult to draw strong conclusions when results vary across tests, grade levels, and race. With the exception of the third graders on the Iowa test, the findings generally indicate performances near national norms or below them. Disregarding racial groups, half of the comparisons (12/24) were not statistically different from the 50th percentile scores, and in comparison to previous reports, this might suggest a positive trend. When race of the student is considered, the results for Black/AA students indicate consistent performances below the national norm; whereas Hispanic students generally did not perform statistically above or below the 50th percentile. The variability in findings across the tests suggests that there may be 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). With the relatively small sample sizes, there is an increased probability of variation among individuals within a grade that may result in some grades performing better than others. This could be due to any number of performance related factors, such as ability, having a good testing day, or differences in teacher quality, among others. By focusing on tests with the most recent norms, the report attempted to address one source of variability among the different tests. As more data accumulate over time, the pattern will become clearer.

Criterion Referenced Test Results

ACT Aspire

The ACT Aspire was administered to 152 students in grades 3-8 and 10. Thirty-seven percent (37%) of these students were first time scholarship recipients, 30% were second or third year scholarship recipients, and 34% had received a scholarship for four or more years. Females comprised 52% of the group. Nearly all of the ACT Aspire test takers were eligible for free/reduced lunch subsidies (92%). The students who took the ACT Aspire were 53% Black/AA, 21% White/Caucasian, 12% Hispanic, 11% indicated more than one racial group, and the remainder were another racial group or unspecified.

In addition to percentile scores, the ACT Aspire provides four proficiency benchmarks based on scale scores that classify students as 1) *In need of support*, 2) *Close*, 3) *Ready*, and 4) *Exceeding*. Students who are at or above the *Ready* benchmark are considered on track to be college ready by the time they are in 11th grade. Table 6 below includes the average percentile and scale scores. The proficiency group corresponding to the average scale score is also provided. Only grades 8 and 10 met the minimum of 25 students for reporting. There were insufficient numbers of students in these grade levels to disaggregate scores by race or gender. The two grade levels showed a similar pattern of results indicating that on average students were below proficiency for reading but met proficiency for English. Math scores could only be reported for 10th grade, and on average these were below proficiency. The percentages of students proficient at each grade level and for each subject were also calculated. For 8th grade the percentages proficient were 44% for reading and 76% for English. For 10th grade, the percentages proficient were 20% for reading, 53% for English, and 21% for math.

Table 6
ACT Aspire: Mean Scale Scores, National Percentiles, and Corresponding Proficiency Levels for Grades 8 and 10

	Reading			English			Math		
Grade (N)	Scale Score	Prof. Level	Percentile	Scale Score	Prof. Level	Percentile	Scale Score	Prof. Level	Percentile
8 (24-25)	422	2	51	427	3	48	*	*	*
10 (29-30)	422	2	43	429	3	45	422	1	37

Proficiency Levels: 1 = In need of support, 2 = Close, 3 = Ready, 4 = Exceeding
* Insufficient number of students in the category, n < 25.

To capture a greater portion of the ACT Aspire test results, Table 7 reports the average percentile scores and the percentage of students who met the proficiency benchmarks across all grades. There were enough students to report scores for Black/AA and White/Caucasian students and for each gender. Across all groups, there were higher percentile scores and greater rates of proficiency for English (71%) compared to reading (33%) and math (41%). White/Caucasian students showed higher rates of proficiency than the other demographic groups across all subjects, and their percentile scores were significantly higher than Black/AA students. There were no significant differences in the test scores for males and females.

Table 7
ACT Aspire: Mean National Percentiles and Corresponding Proficiency Levels for Grades 3-8 and 10 Combined

Group (N)	Reading		English		Math	
	Percent Proficient	Percentile	Percent Proficient	Percentile	Percent Proficient	Percentile
All (150-152)	33%	49	71%	49	41%	47
Black /AA (80-81)	19%	42	61%	42	31%	40
White (32)	66%	63	84%	65	53%	54
Female (77-79)	37%	51	73%	52	38%	44
Male (73)	29%	48	69%	46	44%	49

Scantron Performance Series

The Scantron, which is the test adopted by ALSDE, was administered to 91 scholarship students in grades 4-8, 10, and 11. Language arts scores were not available for Alabama public school children and none of the scholarship students had language arts scores. Students taking the Scantron were 88% Black/AA, 6% White/Caucasian, and the rest were another race or unspecified. Additionally, the Scantron test takers were 82% male and 95% were eligible for free/reduced lunch. Forty percent (40%) were first-time scholarship recipients, 28% had received a scholarship for two or three years, and 33% had been a scholarship recipient for four or more years. Due to the small sample sizes, scores were not reported by grade level. Because the Scantron is used by ALSDE for grades 3-8, and so that direct comparisons could be made to the appropriate scores in

Objective 2, the scores presented in Table 8 are reported for grades 4-8 combined (there are no 3rd graders who took this test) and then for all grades that took the Scantron (4-8, 10, and 11). As shown in Table 8 students were at or below the 40th percentile in both math and reading across all demographic groups and grade levels.

Table 8
Scantron: Mean National Percentile Test Scores for Grades 4-8, 10, and 11

Grade	Group (N)	Reading Percentile	Math Percentile
4-8	All (72)	39	30
	Black/AA (61)	39	30
	Male (63)	40	31
4-8, 10, 11	All (81-86)	35	28
	Black/AA (70-76)	34	28
	Male (67-71)	37	30

In addition, ALSDE has identified four proficiency groups similar to those described by Scantron. The four proficiency groups defined by ALSDE (Scantron) are: 1) *Emerging Learner* (Far Below), 2) *Developing Learner* (Below), 3) *Proficient Learner* (Above), and 4) *Distinguished Learner* (Far Above). As shown in Table 9, the majority of students in grades 4-8 performed below grade level in both reading and math. Across these combined grade levels, 12% of the scholarship recipients were proficient in math and 35% in reading. Proficiency scores were not reported for 10th and 11th grade due to the small sample sizes.

Table 9
Scantron: Proficiency Groups for Grade 4-8

Grade	Group (N)	Reading Proficiency Groups				Math Proficiency Groups			
		1	2	3	4	1	2	3	4
4-8	All (72)	32%	33%	22%	13%	39%	49%	8%	4%
	Black/AA (61)	31%	34%	23%	12%	39%	46%	10%	5%
	Male (63)	30%	25%	24%	11%	36%	49%	10%	5%

Proficiency Levels: 1 = *Emerging Learner (Far Below)*, 2 = *Developing Learner (Below)*, 3 = *Proficient Learner (Above)*, 4 = *Distinguished Learner (Far Above)*

PSAT/NMSQT

The PSAT/NMSQT was administered to 174 students in grades 7, 8, 10, and 11. There were only 5 students in each of the 7th and 8th grades who took this test, so the report focused on grades 10 and 11 (n = 164). Of these students, 8% were first time scholarship recipients, 15% had received a scholarship for two or three years, and nearly 77% had received a scholarship for four years or more. Eighty-eight percent (88%) of the students were eligible for free/reduced lunch subsidies. The racial/ethnic make-up was 53% Black/AA, 20% White/Caucasian, and 22% Hispanic. The remaining 5% of students were either from another racial/ethnic group or had no race information. The students taking the PSAT/NMSQT were 56% female. There were sufficient numbers to report scores separately for Black/AA students and by gender. The PSAT/NMSQT combines reading, writing, and language scores into *an evidenced-based reading and writing score*. As a result, the combined percentile scores are presented in Table 10.

The reading-writing and the 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. Scoring for the PSAT/NMSQT places students' scores into one of three categories: *Need to strengthen skills, Approaching benchmark, or Met or exceeded benchmark.* In Table 10 all of the mean reading-writing scores met the grade level benchmark, but none of the math scores did, all falling into the *Need to strengthen skills* category. Examining the full distribution of scores for each grade level indicated that among the 10th grade students, 29% met the benchmark for math and 62% met the benchmark for reading-writing. For 11th graders the percentages making benchmarks for math and reading-writing were 26% and 61%, respectively. Comparisons between male and female students revealed no significant differences. Together these results suggest that a majority of scholarship recipients are meeting the benchmarks for reading-writing but not for math.

Table 10
PSAT/NMSQT: Mean National Percentile Scores and Corresponding Proficiency Group for Grades 10 and 11

Grade	Group (N)	Reading-Writing		Math	
		Percentile	Proficiency Group	Percentile	Proficiency Group
10	All (82)	50	3	39	1
	Black /AA (41)	39	3	27	1
	Female (45)	48	3	33	1
	Male (37)	52	3	47	2
11	All (81)	52	3	39	1
	Black/AA (45)	45	3	27	1
	Female (45)	56	3	38	1
	Male (36)	47	3	40	1

Proficiency groups: 1 = *Need to strengthen skills*, 2 = *Approaching benchmark*, 3 = *Met or exceeded benchmark.*

PreACT

The PreACT was administered to 111 students in grades 10 and 11. The racial/ethnic make-up of this group of students was 77% Black/AA, 14% White/Caucasian, 1% Hispanic, and the rest were another race. Half of the students who took the PreACT were male. Most were free/reduced lunch eligible (94%). Only 5% were first-time scholarship recipients, 23% had received a scholarship for 2-3 years, and 73% had received a scholarship for four or more years.

For the PreACT, the critical scores are the scale scores (possible 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." The college readiness benchmarks set by ACT Inc. for 11th graders are reading-22, math-22, and English-18, but these differ from the benchmarks set by the State of Alabama for these subjects: reading-19, math-19, and English-8. Because the ACT is

normally taken in the 11th grade, additional college readiness indicators are provided for 10th graders. The rationale behind the additional indicators is 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. The three benchmark levels for 10th grade are defined for each subject area: *In need of intervention, On the cusp,* and *On target.*

Table 11 presents the mean scale scores for 10th and 11th grade students and provides the corresponding college readiness indicator level for 10th graders. There were a sufficient number of students to report scores for Black/AA students for both grades and for male and female students for 10th grade.

Table 11
PreACT: Mean Scale Scores and Readiness Indicators for Grades 10 and 11

Grade	Group (N)	Reading		English		Math	
		Scale Score	Readiness Indicator[1]	Scale Score	Readiness Indicator[1]	Scale Score	Readiness Indicator[1]
10	All (71)	19	On Cusp	16	On Target	16	Intervention
	Black/AA (54)	18	On Cusp	15	On Target	15	Intervention
	Female (36)	19	On Cusp	16	On Target	16	Intervention
	Male (35)	19	On Cusp	15	On Target	16	Intervention
11[2]	All (40)	19	NA	17	NA	17	NA
	Black/AA (31)	18	NA	15	NA	16	NA

[1] Readiness indicators are for 10th grade students only. NA = not applicable
[2] 11th grade college benchmark scores set by the State of Alabama are reading-19, English-18, and math-19. The benchmarks set by ACT are reading-22, English-18, and math-22.

With the exception of English, the 10th grade scores generally did not meet the readiness benchmarks. The percentages of 10th grade students who fell into each of the three readiness categories were calculated, and the results are presented in Table 12. These results show that more than half the students were *On target* to meet the ACT college readiness benchmarks for reading (53%) and English (64%), but over 75% were *In need of intervention* for math. Statistical comparisons between male and female students revealed no significant differences.

Table 12
PreACT: Percentage of Students in Grade 10 within each Readiness Category

Grade (N)	Reading			English			Math		
	Inter-vention	On Cusp	On Target	Inter-vention	On Cusp	On Target	Inter-vention	On Cusp	On Target
10 (71)	37%	10%	53%	18%	18%	64%	76%	14%	10%

For 11th graders, the mean scale scores met the State of Alabama's benchmark for reading but fell below the English and math benchmarks (Table 11). None of the mean scores met the benchmarks set by ACT for college readiness. To further investigate, the percentages of students who met or

exceeded the State benchmarks were calculated: reading-48%, English-45%, and math-23%. The corresponding benchmark percentages for the ACT college readiness standards are: reading-33%, English-45%, and math-15%. Regardless of the standard used, the majority of students did not meet performance benchmarks. Statistical comparisons between male and female students revealed no significant differences.

Together the data suggest a different pattern for 10th and 11th graders. Whereas the majority of 10th graders were *On target* for reading and English, less than half of 11th graders made benchmarks for their grade level. However, the two grade levels were similar for math, in that only a small percentage of students (10% to 23%) were *On target* or made the benchmark.

ACT

The ACT was administered to 81 students in grades 10 (n = 14) and 11 (n = 67). Only the 11th grade had a sufficient number of students to report scores out separately. The majority of the 11th grade sample was Black/AA (67%), followed by White/Caucasian (30%), and Hispanic (3%). About half of the students (48%) were female. All but eight students (88%) were eligible for free/reduced lunch. Similar to the other tests, only a small percentage (10%) of the students were first year scholarship recipients, 24% had received a scholarship for two to three years, and 66% were in their fourth year or more of receiving a scholarship. There were enough students to report scores for Black/AA students and by gender. Both percentile scores and scale scores are presented in Table 13. The ACT scale scores have a possible range from 1 to 36. A statistical comparison of the two genders indicated that female students scored statistically higher than males in English.

Table 13
ACT: Mean National Percentile and Scale Scores for Grade 11

Grade	Group (N)	Reading		English		Math	
		Scale Score	Percent meeting benchmark	Scale Score	Percent meeting benchmark	Scale Score	Percent meeting benchmark
11	All (67)	18	32%	16	33%	17	8%
	Black (45)	17	22%	15	29%	16	7%
	Female (32)	19	41%	17	44%	17	6%
	Male (34-35)	17	24%	15	24%	17	9%

The ACT benchmark scales scores for 11th grade are reading-22, English-18, and math-22. ALSDE benchmarks for 11th grade are reading-19, English-18, and math-17.

Two sets of benchmarks are available for the ACT. First, similar to the PreACT, the ACT testing program provides college readiness benchmarks, which are 22 for reading, 18 for English, and 22 for math. The average ACT scale scores for all 11th graders and the subgroups (Table 13) fell below benchmark scores for college preparedness for reading, math, and English. Less than half of the students met the benchmarks in any subject, with about a third meeting the benchmark for reading and English and 8% meeting the benchmark for math.

ALSDE has set the following benchmark scores for 11th grade: reading-19, English-18, and math-17. The percentages of 11th grade students who were proficient based on ALSDE standards were 42% for reading, 33% for English, and 41% for math. Together, these results suggest that the

scholarship recipients who took the ACT generally failed to meet national standards predictive of college achievement and also failed to meet Alabama State standards, although performance was better in English relative to other subjects.

Summary for Criterion-Referenced Test Results

The key performance indicator for students taking criterion-referenced tests is the percentage of students making benchmarks on each of the tests. The summary graphic below presents the principle findings. It is important to note that comparisons could not be made across racial groups due to insufficient representation of any racial group except Black/AA on most tests.

For students in grades 3 through 8, the ACT Aspire and Scantron findings are applicable, and results varied depending on the subject area and test. A little over a third of students on both tests were proficient in reading. English scores were only available for the ACT Aspire, and these scores are a bright spot in the findings, with 71% of the students meeting or exceeding the benchmark. Although the results from both tests indicate that the majority of students were below the proficiency benchmark for math, the percentages varied dramatically between the tests, with 41% of students meeting or exceeding the math benchmark on the ACT Aspire, but only 12% did so on the Scantron. Together these results indicate that the majority of scholarship recipients in grades 3-8 failed to meet grade level benchmarks for reading and math, but a large majority did so for English.

Tenth graders were represented in three different tests: ACT Aspire, PSAT/NMSQT, and PreACT. The findings, again, depend on the test and subject area. Similar to the students in younger grades, 10th grade students performed best in English, with the majority meeting or exceeding the English benchmarks on the PreACT and ACT Aspire. The PSAT/NMSQT combines reading and language arts into a single score, and the majority of scholarship recipients met that benchmark as well (62%). The results for reading were not consistent across tests, with the majority meeting the benchmark on the PreACT (53%), but only 20% meeting the comparable benchmark on the ACT Aspire. Across the three tests, the majority of scholarship recipients failed to meet benchmarks for math.

Eleventh grade students were also represented in three standardized tests: PSAT/NMSQT, PreACT and ACT. With one exception, results were similar across the three tests in that the majority of 11th grade students did not meet benchmark scores in math, reading, or English. The exception is for the combined reading-writing scores on the PSAT/NMSQT in which over 60% of the students met the benchmark.

Taken together, the pattern of results suggests that many of the scholarship students are making benchmarks for English but fail to do so for math. Over 60 percent of students in grades 10 and 11 met or exceeded benchmarks on the combined reading-writing assessment for the PSAT/NMSQT. Reading proficiency based on a stand-alone test showed a mixed pattern that varied by grade level and test. The majority of scholarship recipients in grade 10 met the reading benchmark on the PreACT; however, the majority of those in grades 3-8 and 11 failed to do so on the ACT Aspire, Scantron, or ACT. Thus, the overall pattern for English and math is clearer than that for reading. It is not clear why English scores are generally better than math and reading, but it is a bright spot in this report.

Summary for Criterion-Referenced Test Results
Students in grades 3-8 took either the ACT Aspire or Scantron:
⊙ The **majority failed** to meet grade level benchmarks for **reading** and **math**. ⊙ For **English and language arts the** majority of scholarship recipients **met** or e**xceeded** the benchmarks.
Students in grade 10 took the ACT Aspire, PSAT/NMSQT, or PreACT:
⊙ The **majority failed** to meet benchmarks for **math.** ⊙ For **English and language arts the** majority of students **met** or e**xceeded** the benchmarks. ⊙ The results for **reading** were inconsistent across tests.
Students in grade 11 took the PSAT/NMSQT, PreACT, or ACT:
⊙ Generally, the majority of 11th grade students did **not meet benchmarks** in **math, reading,** or **English.** ⊙ The exception is on the PSAT/NMSQT combined **reading-writing assessment** on which over 60% of students met or exceeded the benchmark.

Objective 1 Conclusion

It is difficult to draw general conclusions about the academic performance of the 2018-2019 scholarship recipients due to the variability in performance between the type of test (norm- or criterion-referenced), among the tests within each type, and grade levels. Based on the norm-referenced test results, which mostly comprised students in grades 3-8, there were many instances where average scores were near the 50th percentile or better on the TerraNova and Iowa Assessments, suggesting that some scholarship students, as a group are performing similar to other students in the U.S. Yet, a significant number of the results suggest the opposite, indicating other scholarship students are not performing as well as other students in the nation. In addition to factors typically associated with poor performance (e.g., race, poverty, attending a failing school), the difference in student performance is likely due to factors that may vary by school, such as curriculum, pedagogy, and teacher quality.

The pattern for students taking the criterion-referenced tests is more interpretable, with the majority of students making benchmarks in English, failing to make benchmarks in math, and providing a mixed pattern for reading.

In previous reports, performance of the scholarship students was clearly below national norms and standards, but the same generalization cannot be made for this report. The pattern this year is mixed, with results indicating performance was often on par with these standards for English on criterion-referenced tests. As will be noted later, one reason for the seeming improvement may be that some schools included in the previous reports have been eliminated from the AAA scholarship program because they have failed to meet state standards outlined in the Act. Eliminating these substandard schools could be responsible for the better performance. 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 on these issues.

Objective 2: Compare Scholarship Recipients to Alabama Public School Students

For the 2018-2019 academic year, the relevant comparison data are the test scores on the Scantron (grades 4-8) and the ACT college entrance exam (11th grade). Scholarship students were compared to the performance of Alabama public school students in general and, when possible, to Alabama public school students labeled as *economically disadvantaged* (free/reduced lunch eligible) by ALSDE and Black/AA students.

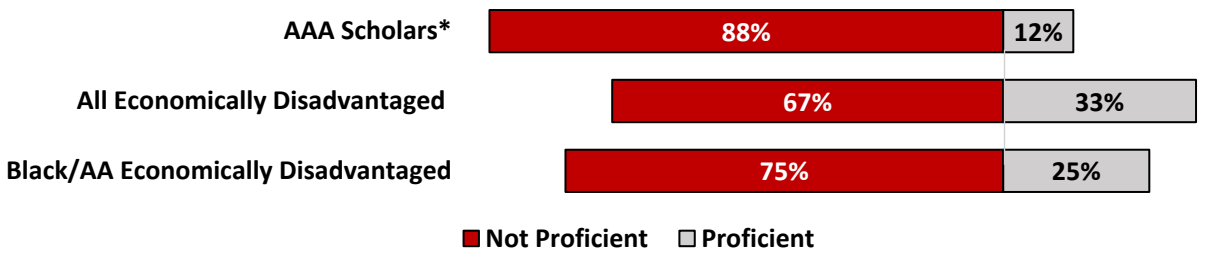
Before presenting the comparative data, there are some significant limitations to the interpretation of the results that must be noted. The scholarship student group represents a very small subsample of all scholarship students, approximately 7% of scholarship students in the grades required to be tested and may not be representative of all of the participants in the AAA program. The relatively small number of scholarship students with Scantron scores (72 in grades 4-8; 4% of all students in these grade levels) and ACT college entrance exam scores (67 in 11th grade; 25% of all 11th graders) collectively represents only 23 (22%) of the 105 schools that students attended. There may be factors associated with the schools that used the Scantron and the ACT (as opposed to other tests, such as the Iowa) that make these schools unrepresentative of the rest of the schools with scholarship recipients (e.g., demographic characteristics of students, class sizes, teacher quality, and pedagogical approaches). With these limitations in mind, the comparisons that are set forth in the evaluation requirements for the AAA were made.

Scantron

The proficiency rates of scholarship recipients in grades 4-8 were compared to economically disadvantaged (free/reduced lunch eligible) Alabama public school children (all students and for the subgroup of Black/AA students) in the same grades. These comparison groups are appropriate because 85% of the scholarship recipients who took the Scantron were Black/AA and nearly all (96%) were eligible for free or reduced lunch. Results are presented in Charts 5 and 6. Statistical comparisons indicated that scholarship students performed more poorly in math in comparison to both the public school groups, but there were no statistically significant differences in reading. Regardless of the group of students being examined, Charts 5 and 6 clearly indicate that most students are not meeting proficiency standards in math and reading.

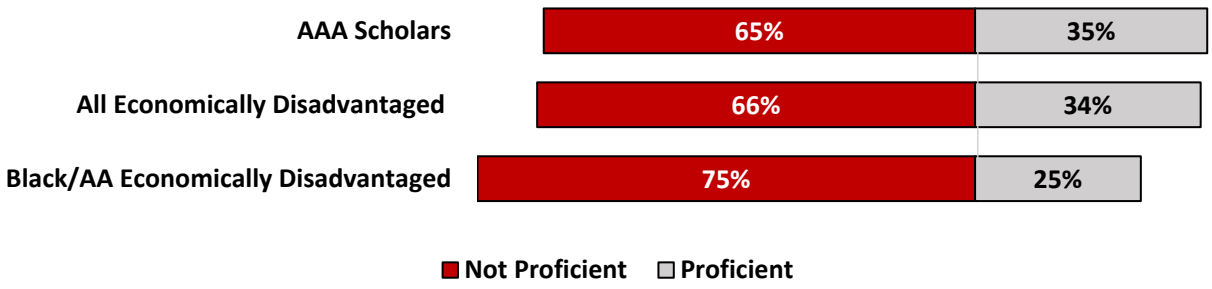
The 72 students in this group differed from the larger group of scholarship recipients on several demographic characteristics: 85% of the Scantron test-takers were Black/AA compared to 67% in the larger sample, 89% were male compared to 50%, 47% were first-time scholarship recipients compared to 29%, and 48% would have attended a failing public school compared to 32% in the larger sample. Most of these elevated characteristics are associated with poorer standardized test performance among students in the U.S., suggesting that lower scores for the scholarship recipients may be in part due to other factors associated with these demographic characteristics, such as less access to resources associated with better educational outcomes. Given these factors, how well these findings generalize to the scholarship students as a whole is unclear.

Chart 5
Scantron: Grades 4-8 Percent Proficient in Math
Scholarship Recipients and Economically Disadvantaged Alabama Public School Students



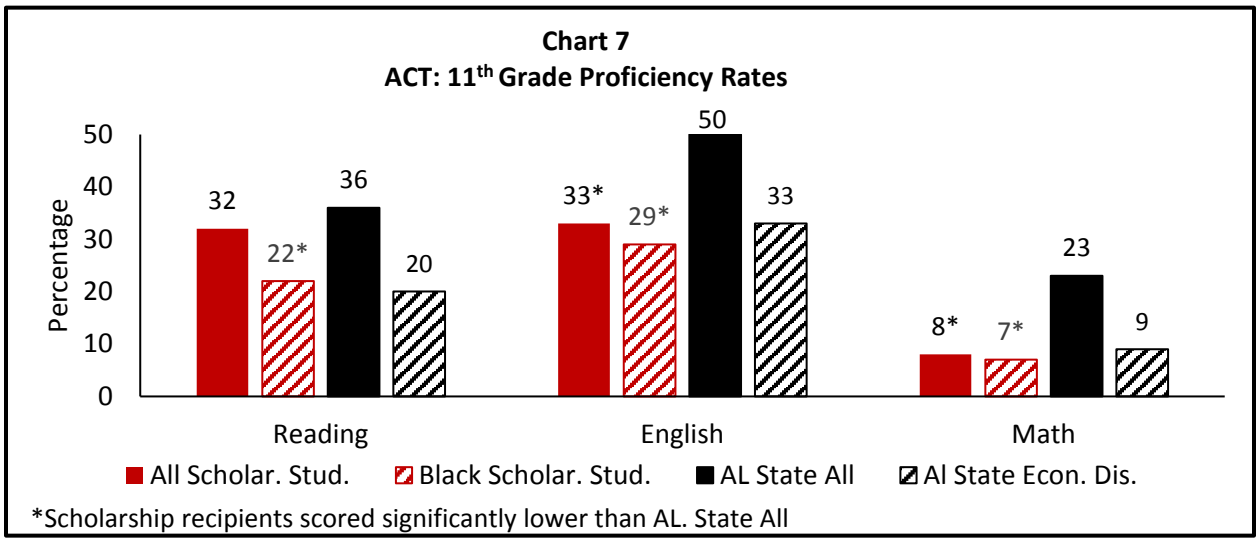
*Scholarship recipients scored significantly lower than both groups

Chart 6
Scantron: Grades 4-8 Percent Proficient in Reading
Scholarship Recipients and Economically Disadvantaged Alabama Public School Students



ACT

The Public Affairs Research Council of Alabama (PARCA) published the percentage of Alabama public school children who met the ACT college readiness benchmarks in each subject. Chart 7 compares the state rates for all students and economically disadvantaged students (in black) to those of the scholarship students (in red). Statistical comparisons indicated that the percentages of scholarship students (the group as a whole and Black/AA students) meeting benchmarks were lower than the State percentages for all students, with one exception: The proficiency rate for scholarship students in reading (32%) was not significantly lower than the proficiency rate for the State of Alabama (36%). Comparisons between scholarship students and economically disadvantaged public school children were not significantly different, with one exception: The reading proficiency rate for all scholarship students (32%) was significantly higher than that for economically disadvantaged public school children (20%). For comparative purposes, it should be noted that for all students in the U.S., the percentages making benchmarks were 45% for reading, 59% for English, and 39% for math.



Additional data available from PARCA provided the average ACT scores for Alabama public school children: 19.8 for reading, 18.8 for English, and 18.4 for math. Statistical analyses comparing these scores to the averages for the scholarship students (18 for reading, 16 for English, and 17 for math) indicated that the scholarship students' scores for English and math were significantly lower. The difference between the two groups of students for reading was not statistically different. Repeating these comparisons for Black/AA scholarship recipients indicated that their average scores (17 for reading, 15 for English, and 16 for math) were significantly lower than the Alabama public school averages in all subject areas.

Together these analyses indicate that 11th grade scholarship recipients collectively performed more poorly than their public school counterparts as a whole in English and math but were not different from them in reading. The performance of Black/AA scholarship students was lower in all subject areas. However, comparisons made to Alabama State economically disadvantaged students generally revealed no significant differences. Similar to the Scantron results, the small number of scholarship students in these comparisons raises concerns about how representative the findings are for the scholarship students as a whole.

Objective 2 Conclusion

For both the Scantron and the ACT, there were comparisons between scholarship students and Alabama public school students that suggest that the two groups are performing comparably and others that suggest the scholarship students are performing more poorly. Generally, scholarship students in grades 4-8 performed lower in math than their counterparts in public schools (all students and economically disadvantaged students), but they were not different in reading. Similarly, the ACT the proficiency rates for scholarship students and public school students in 11th grade did not differ from each other in reading (all students and economically disadvantaged students). For English and math, scholarship students had similar proficiency rates on the ACT as economically disadvantaged public school students but performed more poorly than public school students as a whole.

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. However, these results replicate the pattern from previous reports in that the majority of AAA scholarship students fail to meet benchmarks on standardized tests. Although in some cases their performance is comparable to public school children, the objectives of the AAA program are not to replicate the performance of Alabama public schools, put to exceed it. The results from this report suggest that this is an objective yet to be realized even after six years of the AAA. Furthermore, the overall poor performance of all Alabama students indicates a need for improvement across the board.

Summary for Objective 2: Scholarship Recipients vs. Alabama Public School Students
• Generally, scholarship students' rates of **academic achievement proficiency were lower than those of students attending public schools in math and English.** • Scholarship students' **proficiency rates for reading were generally comparable to their peers in public schools**, although Black/African American scholarship students performed more poorly. • Six years after the passage of the AAA, there **is no evidence** that the scholarship program has resulted in academic achievement that **is superior to that of Alabama public schools.**
Scantron findings for grades 4-8 ⊙ For **math**, scholarship students in grades 4-8 performed **more poorly** than students attending public schools, including those who were economically disadvantaged. ⊙ For **reading**, scholarship students in grades 4-8 achieved **at similar levels** to students attending public schools.
ACT findings for 11th graders ⊙ Scholarship recipients collectively performed **better than economically disadvantaged public school students** in **reading.** ⊙ Scholarship students' proficiency rates for **English and math** were **comparable** to **economically disadvantaged** public school students, but **below** 11th grade public school **students as a whole.**

Objective 3: Changes in Achievement across Time

The third objective of this report addresses whether participation in the scholarship program over time results in achievement score changes that meet, exceed, or fall below those of students attending public schools. Ideally, such an analysis would calculate the average change in national percentile scores or proficiency groups over time for scholarship students and public school students, and then comparisons would be made between the two groups of students. This approach met with four obstacles.

- First, as has been noted previously, without a common test across the two groups of students, limited comparisons can be made. Only a small group of scholarship students took the same tests as those administered by the ALSDE (Scantron and ACT). Analyses based on only a small group of children are likely to be unrepresentative of the scholarship students as a whole and may not be reliable.

- Second, ALSDE changed to the Scantron Performance Series in the 2017-2018 academic year, so there are only two years of new public school data available for grades 3-8 to address this objective. Additionally, due to the small number of scholarship students (*n's* < 25 taking the Scantron for two consecutive years, a direct comparison between the two groups of students over two years cannot be made).

- A third issue is that changes can only be observed as state-wide gains or losses in proficiency groups for public school children in grades 3-8, which may obscure the actual amount of change occurring for individual students. For example, if proficiency rates remain constant from year to year, it is not clear whether this is due to there being no changes in individual student scores or if instead that the percentage of students who gained in proficiency was off-set by a similar percentage who dropped in proficiency.

- Finally, with respect to scholarship students, an individual student would need to have taken the same standardized test the first year of their scholarship and the current academic year to make longitudinal comparisons. 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 their age or test data were not submitted), a large percentage of students would be excluded from this longitudinal analysis.

With these limitations in mind, three 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 2018-2019 academic year. This correlation analysis includes the greatest number of scholarship students and test types, but it does not reveal the amount of change over time, only the direction of change. The second approach focused on students in grades 3-8. Over the years the test most frequently taken by scholarship students in grades 3-8 has been the Iowa test using the 2011 norms. There were sufficient data to examine the average percentile scores in the three subject areas over four academic years, starting with the 2015-2016 academic year and ending with the current 2018-2019 academic year. These results are interpreted in the context of changes in public school children's scores over time. Finally, performance for 11th graders on the ACT was compared between scholarship and public school students over four years.

Correlations between 2018-2019 Test Performance and Number of Years Receiving a Scholarship

Correlation analysis was used to infer a relationship between performance on the 2018-2019 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 their test performance. A significant negative correlation would imply a relationship between increased years in the program and lower performance. Non-significant correlations would suggest 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 longer 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. For the correlation analyses, a minimum sample size of 60 was set, which was the sample size necessary to detect a moderate relationship between tests scores and the number of years receiving a scholarship. Thus, this analysis included all students who took one of nine tests for which there were at least 60 scholarship students: ACT, ACT Aspire, PreACT, PSAT/NMSQT, Iowa Assessments (Normed for 2017, 2011, and 2005), Stanford (2018 norms), and Scantron. Correlations were calculated between number of years a student had received a scholarship (one to six years) and their percentile scores in reading, English/language arts, and math as indicated in the tables and charts in Objective 1. Out of the 26 correlations calculated, only four were significant, each indicating a small positive relation between the number of years receiving a scholarship and test scores:

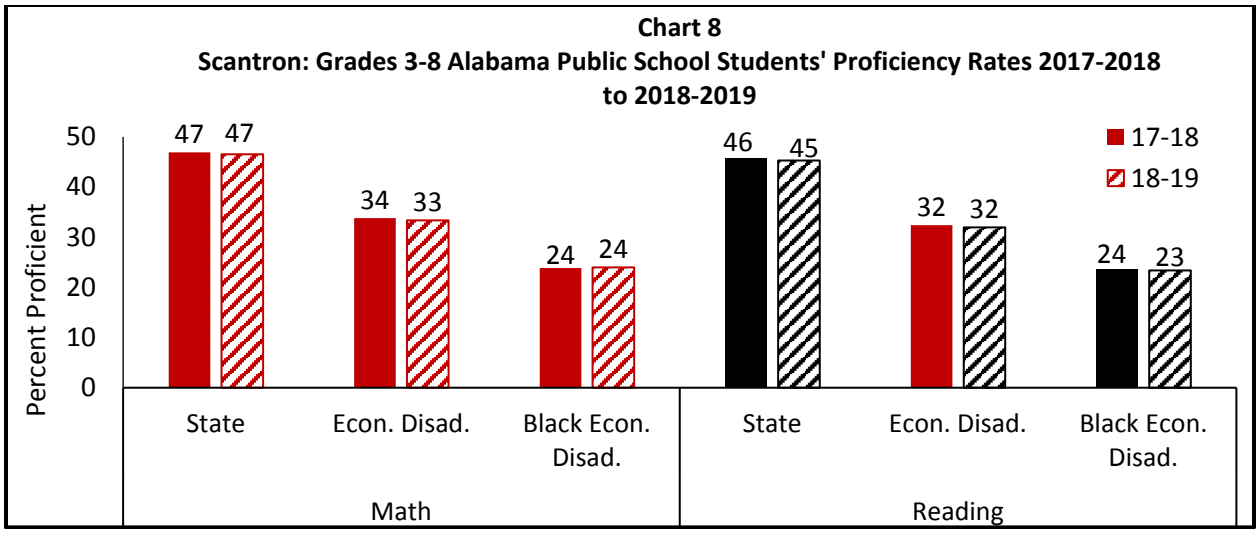
- For the Iowa S2011 norms (n = 201) reading (r = .15) and English (r = .17) were significant;
- For the Iowa S2005 norms (n = 178), reading (r =.17) was significant
- For the ACT (n = 80), English was significant (r = .24).

These four 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 four correlations are also relatively small, suggesting that the number of years in the program is likely not a strong predictor of performance. Results also suggest that similar to the public school children in Alabama, on average, scholarship student performance has been generally stable over time.

Comparison of Students in Grades 3-8

Public School Students' Performance

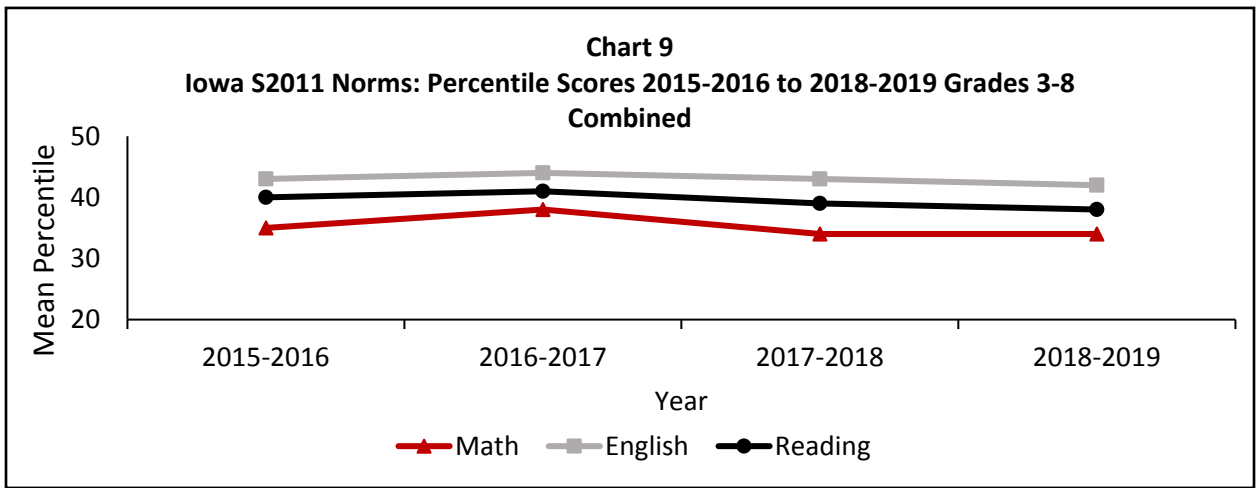
Before presenting the results for the scholarship students, the change in achievement scores for public school children is described for comparative purposes. For the test currently in use, the Scantron, ALSDE has only two years of test score data available starting in the 2017-2018 academic year. Reading and math proficiency rates were compared between the 2017-2018 and 2018-2019 academic years for the state as a whole and for two subpopulations: Economically disadvantaged (all races), and Black/AA economically disadvantaged. As can be seen in Chart 8 there is virtually no difference in proficiency rates across the two years for the state as a whole and for the subpopulations. A similar analysis of three years of ACT Aspire scores, the test mandated by ALSDE from 2014-2015 to 2016-2017, also showed very little change over time (see the 2018 evaluation report for details).



Scholarship Students' Performance on the Iowa Assessments 2014-2015 to 2018-2019

The Iowa test scores with Spring 2011 norms were used to examine grade level improvements over the course of four years. This test was chosen because it included the largest percentage of the scholarship students compared to the other tests available, and as a result, it would be more representative of the scholarship recipients as a group. Although it would have been better to examine results for the most recent Iowa Assessments norms (2017), there were not enough students to make this comparison over time. This report focuses on the Iowa test for the Spring 2011 norms to provide a consistent comparison across the years and grade levels and is an improvement from past reporting. Change over time is examined for all of the scholarship students as a group first, followed by a breakdown by grade level.

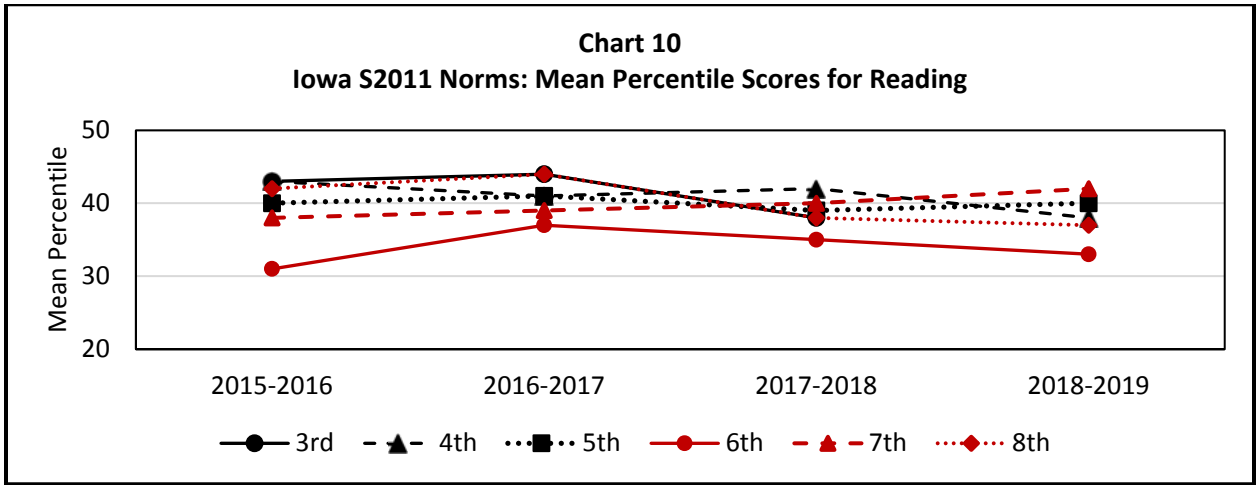
Chart 9 presents the mean percentile scores collapsed across grades 3-8. There were no significant differences in mean percentile scores between any of the four academic years. Students performed similarly on the Iowa test every year in each subject area. The highest mean percentile score was in English (44%) for the 2016-2017 school year, indicating that throughout the years, on average, students who took the Iowa test scored below the 50th percentile.

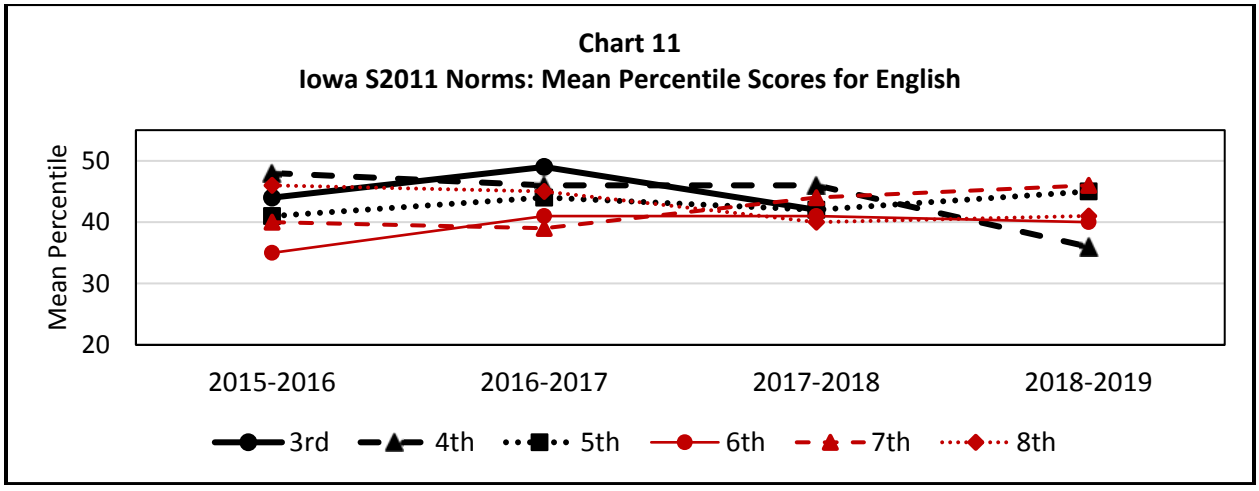


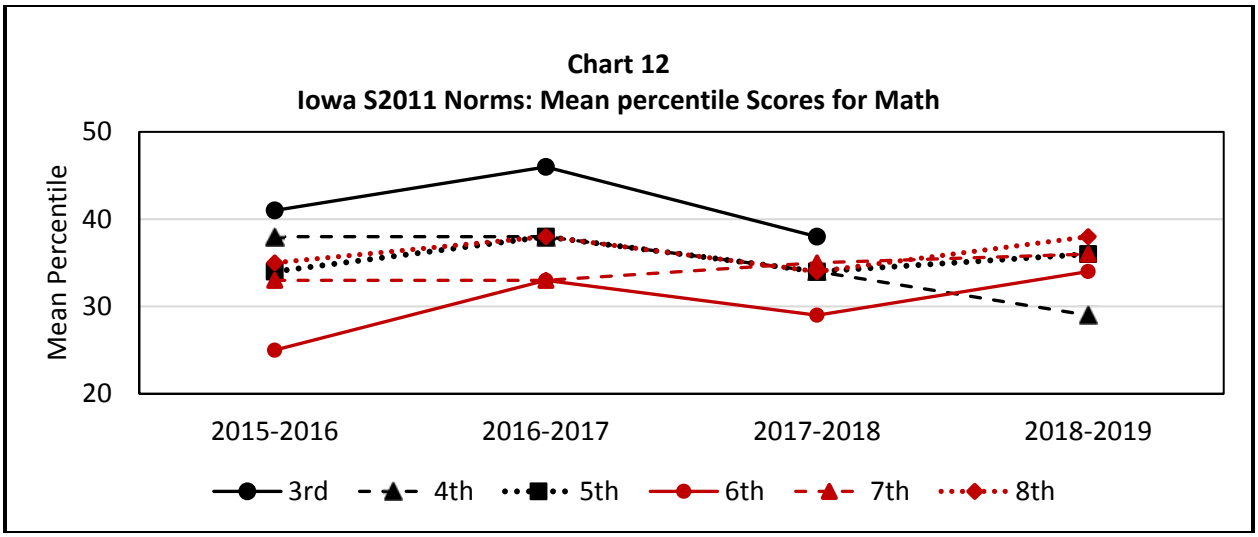
Next, the mean percentile scores by grade level from 2015-2016 to 2018-2019 were examined. Charts 10-12 present the results for reading, English, and math, respectively. Students consistently performed below the 50th percentile for all grades and all subjects. This suggests that as a group, AAA scholarship students generally perform below the 50th percentile and continue to do so throughout out their elementary and middle school years. (See Table 4A for details.)

Statistical analyses (ANOVA) were used to assess the differences in mean scores on the Iowa test at each grade over the four years. In some places in Charts 10-12, a relatively big change in scores is not statistically significant. This is likely due to the small sample size and variability within the sample, which are taken into account in the statistical tests, and indicate that such changes are likely not reliable. When comparing grade levels from year to year, with few exceptions, there were no significant differences in mean percentile score. The exceptions were:

- Third grade math, reading and English scores declined from 2016-2017 to 2017-2018. (There were not enough 3rd grade students in 2018-2019 to report their scores in this chart.)
- Sixth grade was the only grade that saw improvements. Math scores in 2016-2017 were higher than those in 2015-2016.
- Eighth grade reading scores declined from 2016-2017 to 2017-2018.



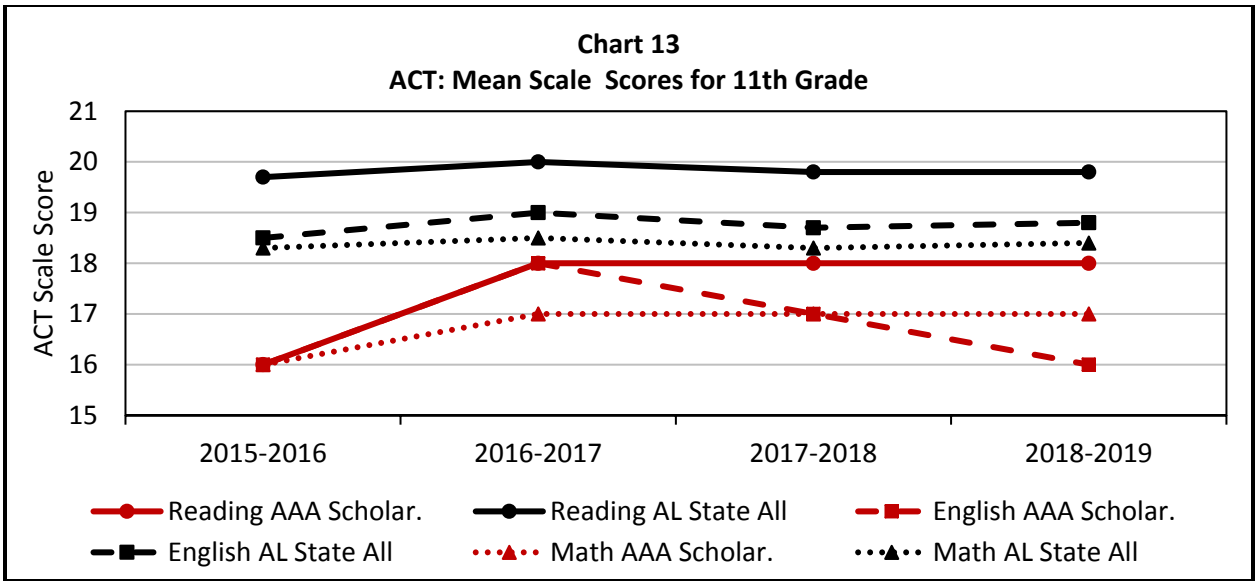




To summarize, for three grades (4th, 5th, and 7th) the Iowa test scores did not show a significant improvement or decline, similar to the public school data. Two grades, 3rd and 8th, showed some decline in scores over time, although not consistently so. One grade, 6th, showed some improvement over two years, but these improvements were also not consistent over the four-year time span. Together, these results do not present a clear pattern of change on the Iowa test using the Spring 2011 norms. Given the small sample sizes for some grades and the variability in the scores, we cannot be certain that the significant differences are not an artifact of the samples, rather than a true difference that would generalize to all scholarship students. Any differences found are likely attributable to factors not examined in this report, such as quality of a particular school, curriculum, or teaching methods. Regardless, neither the Alabama public school students nor the scholarship students in grades 3-8 generally improved over time.

Comparison of Students in Grade 11

Mean ACT scores for 11th grade were available for the scholarship students starting in the 2015-2016 academic year through the 2018-2019 academic year. Comparable date were available from PARCA for public school children in Alabama. Chart 13 plots the mean ACT scale score for reading, English and math for each group of students. Scholarship student scores are represented in red and public school students are represented in black. Although there was a rise in scholarship student scores from the 2015-2016 to the 2016-2017 academic year, the more recent pattern suggests stability or slight decline (for English). Public school students similarly showed very little change on average over time.



Proficiency rates were available for the ACT college readiness standards for scholarship students and Alabama public school students. ACT scores for the 11th grade scholarship students were examined over the same four years as the Iowa test (Charts 14-16). Similar to the results for the Iowa test, throughout the years, scholarship students' performance on the ACT has been poor. In any given year, less than 50% of the scholars met the benchmarks for reading, math, or English (Charts 14-16). Statistical comparisons for change in proficiency rates over time for each subject area revealed no significant changes over the years for reading and English. For math there was a significant improvement from the 2015-2016 academic year to the 2016-2017 academic year. However, this was offset by a significant decline in proficiency rates from the 2016-21017 academic year to 2018-2019. It should be noted that often seemingly large changes in proficiency rates in Charts 14-16 are not statistically significant. The non-significant statistical tests tell us that despite their size these are likely not reliable differences.

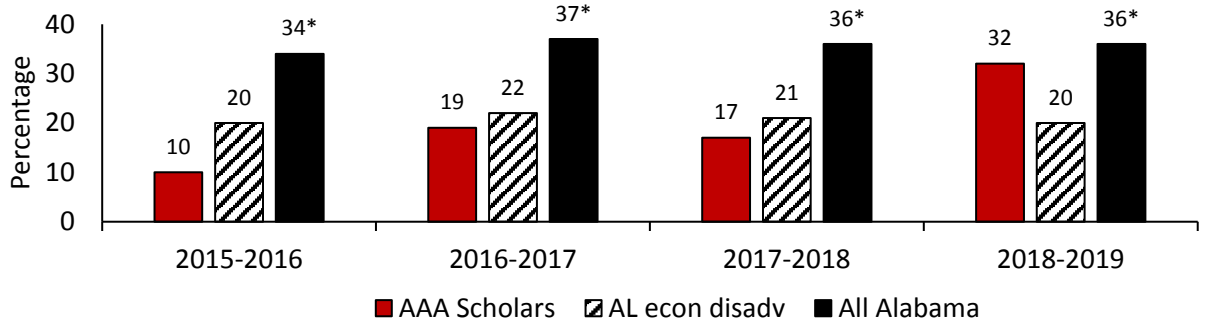
Charts 14-16 also show comparative proficiency rates for 11th graders attending Alabama public schools (all students and economically disadvantaged students). The charts reveal that as a group, the public school students' proficiency rates showed little change over time.

Additional comparisons were made between the proficiency rates for the scholarship students and the two public school groups at each time point for each subject area. With four exceptions, the proficiency rates for the scholarship students were no different from those of the Alabama economically disadvantaged group, but significantly lower than those for all Alabama public school children combined. The exceptions were:

- The English 2016-2017 proficiency rate for scholarship students was not significantly lower than the rate for all Alabama public school children.
- The English 2017-2018 proficiency rate for the scholarships students was significantly greater than the rate for the economically disadvantaged group and no different from that rate for all Alabama public school children.

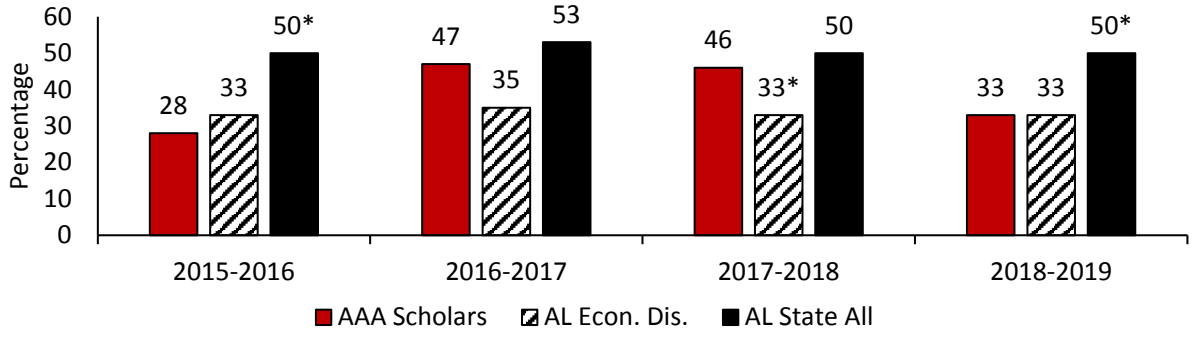
- The Math 2016-2017 proficiency rate for the scholarships students was significantly greater than those for the economically disadvantaged group and no different from those for all Alabama public school children.
- The Math 2017-2018 proficiency rate for scholarship students were not significantly lower than those for all Alabama public school children.

Chart 14
Percent Meeting College Readiness Standards for Reading on the ACT



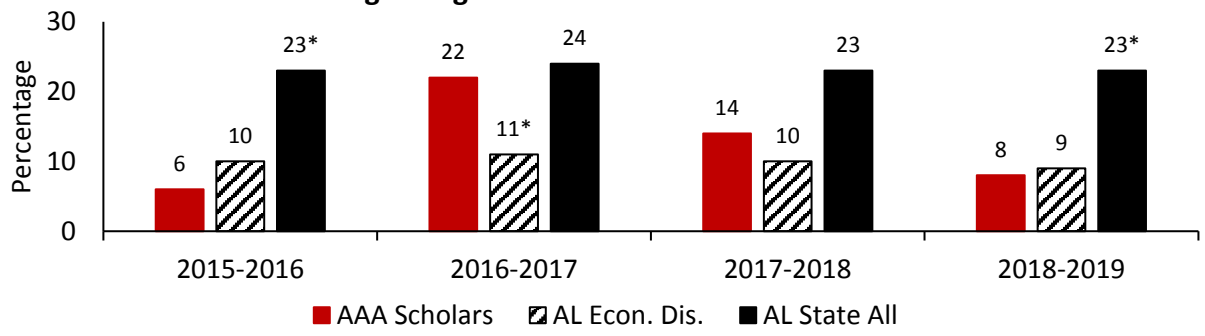
* Significantly different from AAA students the same year.

Chart 15
Percent Meeting College Readiness Standards for English on the ACT



*Significantly different from AAA students in the same year.

Chart 16
Percent Meeting College Readiness Standards for Math on the ACT



*Significantly different from AAA students in the same year.

Objective 3 Conclusion

Together, the data on scholarship students does not reveal a consistent pattern of improvement or decline over time. However, the analyses presented above indicate that the same is likely true of students attending public schools in Alabama. Moreover, for both groups of students, performance generally falls below national standards. Despite the low scores overall, the 11th grade scholarship students consistently performed similarly to the economically disadvantaged public school children. 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 likely offset by other students who are declining or stable over time.

Summary for Objective 3: Changes in Achievement across Time

- On average, over time, **participating in the scholarship program was not associated with significant improvement** on standardized tests scores.
- The lack of change over time **followed the same pattern seen in public school students** in Alabama and is likely **not attributable to participation in the scholarship program**.

- The **number of years** that a student participated in the scholarship program was generally **not correlated with higher** achievement test scores.
- For the majority of scholarship recipients in **grades 3-8 there** was **no consistently observed gain or decline in mean percentile scores** on the **Iowa S2011 norms.**
- On the **ACT, 11th grade** students scores **did not show a consistent pattern of gain or decline**. These scores were generally **no different** from **economically disadvantaged** Alabama public school students, but **lower** than that for 11th graders in the state as a whole.

General Conclusion

The purpose of this evaluation is to assess how the scholarship program enacted through the AAA affects the academic achievement of students in the program. Throughout the report, many concerns have been voiced about the reliability and validity of the findings due to unknown factors associated with missing achievement tests and due to issues related to subsamples included in specific comparisons, such as whether a subsample of students accurately represented the larger group of scholarship students. Within these limitations, the report made use of the available information to describe the performance of scholarship recipients based on the most recent data available (2018-2019 academic year). The evaluation addressed three objectives to reach this goal:

- In the first objective, the achievement test performance of the scholarship recipients was described. On norm-referenced tests, scholarship students generally performed near or below the median scores for students in the U.S. On criterion-referenced tests, proficiency rates indicated an uneven performance. There were several instances when more than half the scholarship students met the benchmarks for English, but there was a less consistent pattern for reading. Math performance showed a reliable pattern, with the majority of students not meeting benchmark scores. Taken together, the results suggest that, with the exception of English, the majority of 2018-2019 scholarship students performed below national norms or standards.
- Objective 2 compared scholarship students to Alabama public school students on the Scantron and ACT. For grades 3-8, the two groups of students performed comparably on reading, but public school children performed better in math. On the ACT, 11th grade scholarship students' performance was comparable to economically disadvantaged public school students in English and math, but better in reading. However, the majority of students in both groups failed to meet proficiency benchmarks. Only a small percentage of scholarship students took the Scantron or the ACT, which hampers the ability of this report to draw definitive conclusions.
- Finally, the third objective assessed if scholarship recipients' achievement scores improved, declined, or remained the same over time. Similar to their public school counterparts, findings suggest that, on average, performance of the group as a whole has not changed over time.

Limitations

As with previous reports, the analyses found in this report have many shortcomings that are inherent in the data available for the evaluation. For example, there is a litany of possible confounding differences among students and the schools they attend that cannot be accounted for in this work. This includes potential differences in test or grade samples, many of which have already been discussed, such as different compositions of race, household income, or number of years receiving a scholarship. The most meaningful comparison between scholarship recipients and public school students would compare 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. Unfortunately, this type of data is not available.

Creating 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. In all previous reports, we have noted that the lack of a common assessment severely limits the ability to draw strong conclusions regarding the academic achievement of scholarship recipients relative to students attending public schools. Only two years of Scantron data were available for this report, and only a small percentage of the schools receiving scholarship students have adopted this test. The Iowa test was taken by the largest group of scholarship students, and we focused on this test data in Objective 3 because it had the potential to represent the largest portion of students compared to the other tests available.

Unfortunately, many schools opted to evaluate student performance using tests with outdated national norms, which inaccurately reflect the achievement of their students against current educational standards. This is primarily an issue for schools that use the Stanford and the Iowa tests. Although it may save money for a school to use outdated test norms, the value of this practice for evaluating student learning is questionable.

Finally, it is important to reiterate that the use of proficiency scores to discern differences in student performance over time may not be sensitive to meaningful changes in performance. Change in performance is only registered when students transition from one group to the next, but each proficiency group represents a range of scores. From a policy perspective, a considerably smaller change in scores could be considered significant. Additionally, students who are closest to the cutoff scores are more likely to change proficiency groups, entailing the potential for a disproportionate impact of a relatively small number of students. In this report, this was primarily an issue for interpreting the Scantron and ACT results. A better understanding of student academic gains could be achieved by either using student-level testing results, or by knowing the means and other statistical information for test scores across demographic groups**.**

In closing, we note that students in the AAA program belong to demographic groups (low income, racial minorities) that have lagged behind other students in the state and the U.S. in academic achievement. There were several instances in this report where scholarship students performed comparably to economically disadvantaged students in Alabama, suggesting there may be some equivalence in performance. Nevertheless, the ultimate goal of the AAA program is to raise the achievement level of scholarship students above what would be expected if they were attending a public school. To date, the evidence in this report indicates this goal has not been achieved.

Some optimism for a better outcome for the future can be held as school accreditation requirements that were written into the AAA take effect. At the time this report was written, 54 schools that previously received scholarship students are no longer eligible to take them due to the lack of accreditation. If the quality of scholarship schools improves over time, the scholarship students may eventually reach parity with public school children in all subject areas.

Glossary of Terms

Criterion-referenced test. These tests assess students' learning against a fixed set of predetermined learning standards that are set 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.

Mean. A mean test score is calculated by adding together every test 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 criterion, such as national educational standards or college readiness.

Raw score. A raw score is the number of items that a child 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 less than a 5% chance) 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.

Appendix

Table A1
Iowa S2017: Mean National Percentile Scores Grades 3-8

Grade	Group (N)	Math Percentile	Reading Percentile	English Percentile
3	All (96-97)	55	51	56
	Black (32-33)	49	48	51
	Hispanic (38)	53	45	51
	White (<25)	*	*	*
	Male (44)	52	44	49
	Female (52-53)	59	58	62
4	All (114)	49	47	53
	Black (44)	33	38	39
	Hispanic (38)	51	47	53
	White (<25)	*	*	*
	Male (48)	46	53	46
	Female (66)	58	45	58
5	All (113-114)	40	42	46
	Black (59-60)	30	37	40
	Hispanic (32)	48	42	44
	White (<25)	*	*	*
	Male (54)	42	42	43
	Female (59-60)	40	43	49
6	All (91)	36	40	46
	Black (44-45)	28	32	37
	Hispanic (28)	41	46	51
	White (< 25)	*	*	*
	Male (48)	32	38	48
	Female (42)	33	41	44
7	All (71)	40	44	49
	Black (36)	25	34	38
	Hispanic (<25)	*	*	*
	White (< 25)	*	*	*
	Male (29)	43	44	47
	Female (42)	38	44	50
8	All (69-71)	40	45	46
	Black (58)	28	38	37
	Hispanic (<25)	*	*	*
	White (<25)	*	*	*
	Male (38)	33	44	45
	Female (33)	37	47	44

* Indicates an insufficient number of students in the group (< 25) for reporting.

Table A2
Iowa S2011: Mean National Percentile Scores for Grades 3-8

Grade	Group (N)	Math Percentile	Reading Percentile	English Percentile
3	All (<25)	*	*	*
	Black (<25)	*	*	*
	Hispanic (<25)	*	*	*
	White (<25)	*	*	*
	Male (<25)	*	*	*
	Female (<25)	*	*	*
4	All (36-37)	29	38	36
	Black (<25)	*	*	*
	Hispanic (<25)	*	*	*
	White (<25)	*	*	*
	Male (<25)	*	*	*
	Female (<25)	*	*	*
5	All (39)	36	40	45
	Black (<25)	*	*	*
	Hispanic (<25)	*	*	*
	White (<25)	*	*	*
	Male (<25)	*	*	*
	Female (<25)	*	*	*
6	All (44-45)	34	33	40
	Black (28)	24	27	31
	Hispanic (<25)	*	*	*
	White (< 25)	*	*	*
	Male (30-31)	30	33	38
	Female (<25)	*	*	*
7	All (36-37)	36	42	46
	Black (<25)	*	*	*
	Hispanic (<25)	*	*	*
	White (< 25)	*	*	*
	Male (<25)	*	*	*
	Female (<25)	*	*	*
8	All (<25)	38	37	41
	Black (<25)	*	*	*
	Hispanic (<25)	*	*	*
	White (< 25)	*	*	*
	Male (<25)	*	*	*
	Female (<25)	*	*	*

* Indicates an insufficient number of students in the group (< 25) for reporting.

Table A3
Iowa S2005: Mean National Percentile Scores for Grades 3-8

Grade	Group (N)	Math Percentile	Reading Percentile	English Percentile
3	All (24-28)	39	50	37
	Black (24-26)	38	49	*
	Hispanic (<25)	*	*	*
	White (<25)	*	*	*
	Male (<25)	*	*	*
	Female (<25)	*	*	*
4	All (24-26)	35	48	41
	Black (<25)	*	*	*
	Hispanic (<25)	*	*	*
	White (<25)	*	*	*
	Male (<25)	*	*	*
	Female (<25)	*	*	*
5	All (26-29)	38	47	52
	Black (26)	33	44	48
	Hispanic (<25)	*	*	*
	White (<25)	*	*	*
	Male (<25)	*	*	*
	Female (<25)	*	*	*
6	All (36-37)	32	44	38
	Black (31)	30	44	37
	Hispanic (<25)	*	*	*
	White (< 25)	*	*	*
	Male (<25)	*	*	*
	Female (<25)	*	*	*
7	All (33)	37	50	48
	Black (32)	37	49	48
	Hispanic (<25)	*	*	*
	White (< 25)	*	*	*
	Male (<25)	*	*	*
	Female (<25)	*	*	*
8	All (<25)	*	*	*
	Black (<25)	*	*	*
	Hispanic (<25)	*	*	*
	White (<25)	*	*	*
	Male (<25)	*	*	*
	Female (<25)	*	*	*

* Indicates an insufficient number of students in the group (< 25) for reporting.

Table A4
Iowa S2011: Mean National Percentile Scores for Grades 3-8

Year	Group (N)	Math Percentile	Reading Percentile	English Percentile
2015-2016	All (637-729)	35	40	43
	3rd (120-140)	41	43	44
	4th (108-126)	38	43	48
	5th (93-106)	34	40	41
	6th (72-90)	25	31	35
	7th (131-144)	33	38	40
	8th (113-123)	35	42	46
2016-2017	All (901-918)	38	41	44
	3rd (178-181)	46	44	49
	4th (159)	38	41	46
	5th (121-128)	38	41	44
	6th (135-136)	33	37	41
	7th (140-145)	33	39	39
	8th (163-170)	38	44	45
2017-2018	All (800-809)	34	39	43
	3rd (158-161)	38	38	42
	4th (146-156)	34	42	46
	5th (131)	34	39	42
	6th (113)	29	35	41
	7th (117-118)	35	40	44
	8th (127-131)	34	38	40
2018-2019	All (201)	34	48	42
	3rd (n<25)	*	*	*
	4th (36-37)	29	38	36
	5th (39)	36	40	45
	6th (44-45)	34	33	40
	7th (36-37)	36	42	46
	8th (36)	38	37	41

* Indicates an insufficient number of students in the group (< 25) for reporting.