MEMORANDUM

TO: Members, Utah State Board of Education

FROM: Rich Nye, Associate Superintendent
Jo Ellen Shaeffer, Director

DATE: December 3-4, 2015

DISCUSSION: SAGE Summative, Interim, and Formative Assessments

---

Background:
Since the initial implementation of the SAGE assessment system, there have been concerns with various deliverables and contractual agreements. The SAGE Overview document identifies some salient features of the SAGE formative, interim, and summative system and seeks to answer some of the predominant concerns with how the SAGE system is serving our public education interests statewide.

Key Points:
Staff will present the Committee with an overview of the SAGE summative, interim, and formative system to broaden and inform the discussion. Material contained in the Overview document are the result of the efforts and collaboration of USOE staff, Assistant Attorney General Chris Lacombe, and The Center for Assessment.

Anticipated Action:
The Committee will receive the information and consider making additional recommendations regarding the AIR contract novation.

Contact: Rich Nye, 801-538-7550 Jo Ellen Shaeffer, 801-538-7811

SAGE

Student Assessment of Growth and Excellence

SAGE Formative, Interim, & Summative Overview

October 8, 2015

CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Introduction

The Utah State Board of Education approved the Utah Core Standards for English language arts (ELA) and mathematics in 2010. These standards were fully implemented in local education agencies (LEAs) in the spring of 2013 for mathematics and ELA. The science standards were adopted and implemented in 2010. The Utah Core Standards describe the educational targets for students in each content area.

The Utah State Office of Education (USOE) entered into a partnership with American Institutes of Research (AIR) to construct, administer, and validate a formative, interim and summative assessment system for the new standards. During the 2013–2014 school year, the Utah State Office of Education working in collaboration with AIR supplemented an existing general education assessment program that aligned the Student Assessment of Growth and Excellence (SAGE) to the Utah Core Standards and satisfied the federal No Child Left Behind (NCLB) requirements. USOE also involved educators and assessment and curriculum specialists in making decisions about how to measure standards. The 2013-2014 SAGE administration was considered an operational field test for students in grades 3–11 for ELA/writing, grades 3-8 for mathematics, along with end-of-course assessments for high school students taking Secondary Mathematics I–III, and for science in grades 4–8, along with end-of-course assessments for high school students taking Earth Science, Biology, Chemistry, and Physics. The SAGE summative assessment was administered a second time for the 2014-15 school year.

Since the initial implementation of the SAGE assessment system, there have been concerns with varying deliverables and contractual agreements. The pages that follow identify some salient features of the SAGE formative, interim, and summative system and seeks to answer some of the predominant concerns with how the SAGE system is serving our public education interests statewide. It should be noted that the reader consider the two primary purposes of a statewide assessment system and whether or not SAGE is meeting the requirements of those two purposes, which are:

- Assessment data to inform *accountability* (Summative)
    - Provides summary measures of what students know and can do at particular points in their education careers.
    - Accountability at the state, district, school, teacher, and student level.
    - Growth Measures
    - *Assessment of Learning*
- Assessment data used to elucidate what and how students are *learning* (Formative, Interim, Summative)
    - Directly supports instruction by generating information at multiple points about how students are learning and about what misunderstandings or misconceptions might be getting in their way.
    - As*sessment for Learning*

SAGE Formative

SAGE Formative is an optional educational system for LEAs that allows educators to create assessments and assignments for students. Students can use SAGE Formative to complete the assignments that educators give to them in class or at home. After a student completes an assignment or assessment, educators can immediately view student scores. Educators may also view reports that provide information about student performance at the aggregate by class, grade level and school.

The introduction of the SAGE formative system lacked many of the supporting features that LEAs needed to make the transition to utilize the system for their benchmark assessments. Some of the initial limitations include, small item bank, inability to create items, inability to share assessments/assignments among teachers, vague reports, etc.

The SAGE formative system has improved incrementally to include better reports, ability to share assignments, and ability to create items. Despite the initial limitations to the SAGE formative system, several LEAs have utilized this system as part of the teaching and learning process. As the formative system has improved LEA usage has increased.

Formative Assessment Use:

There have been **9,869** teachers who have created and administered assessments in the SAGE formative system. Educators have also accessed the SAGE formative system to guide classroom discussions without actually having students login to the system. The table below identifies the number of tests administered and scored in SAGE formative.

Table 1

Summary of Formative Assessment Use Statewide

Number of Tests Created with Student Responses	
41 Districts	1,385,102 Tests Created
72 Charters + EHS	118,234 Tests Created
113 LEAs + EHS	**1,503,336 Tests Created**

Some LEAs have utilized the SAGE formative system extensively. Granite School District for instance has utilized the SAGE formative system for their benchmark assessments in all grades and several content areas. Other LEAs have contracted with formative assessment providers for their benchmark assessment purposes but still use the SAGE formative system to assist in classroom instruction. Table 2 below highlights the ten highest usage districts of the SAGE Formative system from July 2014-June 2015.

Table 2

Summary of Formative Assessment District Use

10 Highest usage districts

District Name	Number of tests created where more than one student responded
Granite	840,574
Cache	81,922
Weber	42,217
Davis	36,475
Washington	34,599
Ogden	32,703
Alpine	31,926
Jordan	31,099
Canyons	29,920
Nebo	27,117

In addition to the previously identified school districts, several charter schools have also accessed the SAGE Formative system for benchmark assessment administration and guided instruction as aligned to the Utah Core Standards. Table 3 below highlights the top 5 highest use charter schools.

Table 3

Summary of Formative Assessment Charter Use

5 Highest Usage Non Districts

District Name	Number of tests created where more than one student responded
Utah Electronic High School	13,123
Utah Virtual Academy	11,332
Pinnacle Canyon Academy	8,904
Syracuse Arts Academy	8,308
American Preparatory Academy	7,262

SAGE Formative Reports

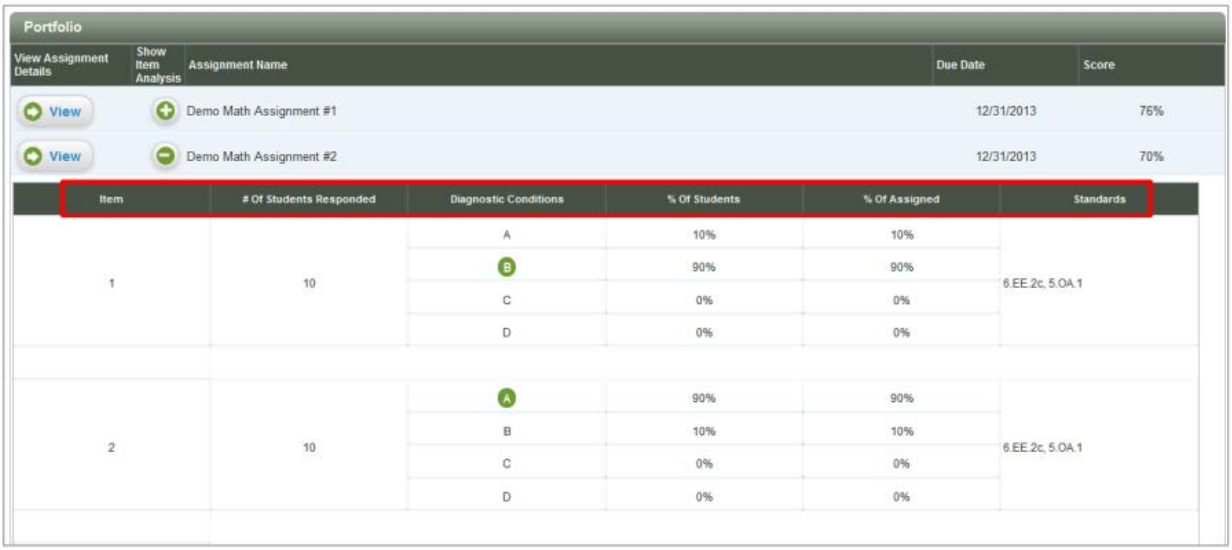
The SAGE formative reporting system was one area that many felt was lacking in terms of what was initially expected. Noting the need for more informative reports, USOE staff have worked with AIR to provide more robust and detailed reports to assist educators in analyzing student performance and adjusting instruction accordingly. Sage formative reporting evolved from the initial implementation and now includes elements of standards level mastery by student,

item analysis, teacher comparisons, school comparisons, and school to school comparisons within the same LEA.

For example, item analysis reporting is done at the standards level and is available to be disaggregated by LEA, school, teacher, and student. This report is beneficial to determine how specific items are functioning and whether or not common misunderstandings are leading students to select various distractors. Teachers and students are also able to identify specific test questions in their analysis and goal setting discussions as it pertains to demonstrating standard mastery.

Figure 1

Exportable report of core standard mastery



Portfolio

View Assignment Details	Show Item Analysis	Assignment Name	Due Date	Score
View		Demo Math Assignment #1	12/31/2013	76%
View		Demo Math Assignment #2	12/31/2013	70%

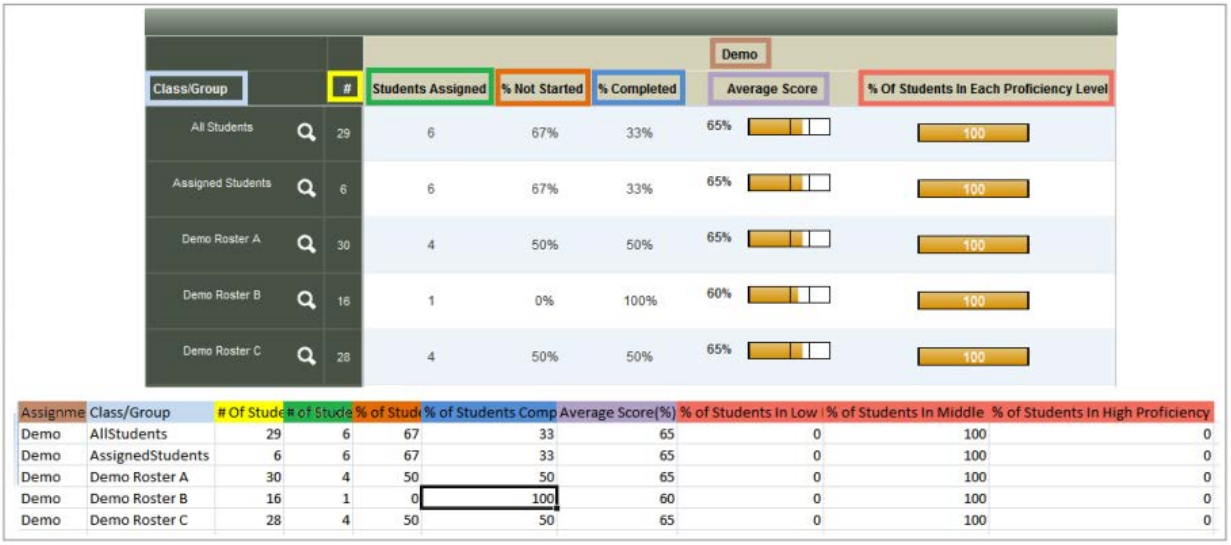
Item	# Of Students Responded	Diagnostic Conditions	% Of Students	% Of Assigned	Standards
1	10	A	10%	10%	6.EE.2c, 5.OA.1
		B	90%	90%	
		C	0%	0%	
		D	0%	0%	
2	10	A	90%	90%	6.EE.2c, 5.OA.1
		B	10%	10%	
		C	0%	0%	
		D	0%	0%	

The exportable report above shows student proficiency and classroom performance by item and standard. This report further allows educators to identify and track trends in mastery and assists in preparing content aligned instruction. This report and others like it are “drillable” to gain greater detail depending on the venue of discussion, whether teacher-student, teacher-teacher, administrator-teacher, etc.

General classroom performance is also an exportable report (Figure 2) that allows educators to determine an overview of classroom understanding. Reports such as the one shown below are critical for understanding the needs of what students may or may not understand and how the instruction to follow is tailored to address areas of misunderstanding.

Figure 2

Exportable report of classroom performance



Demo						
Class/Group	#	Students Assigned	% Not Started	% Completed	Average Score	% Of Students In Each Proficiency Level
All Students	29	6	67%	33%	65%	100
Assigned Students	6	6	67%	33%	65%	100
Demo Roster A	30	4	50%	50%	65%	100
Demo Roster B	16	1	0%	100%	60%	100
Demo Roster C	28	4	50%	50%	65%	100

Assignme	Class/Group	# Of Stude	# of Stude	% of Stud	% of Students Comp	Average Score(%)	% of Students In Low	% of Students In Middle	% of Students In High Proficiency
Demo	AllStudents	29	6	67	33	65	0	100	0
Demo	AssignedStudents	6	6	67	33	65	0	100	0
Demo	Demo Roster A	30	4	50	50	65	0	100	0
Demo	Demo Roster B	16	1	0	100	60	0	100	0
Demo	Demo Roster C	28	4	50	50	65	0	100	0

SAGE Formative Assurances

After careful review of the SAGE Formative system it has been determined that AIR has met, on some level, the services outlined in the AIR proposal and in the novation as currently written (Table 4). Although the assurances appear to have been met, there are instances where what was expected, was not delivered. For example, the SAGE Formative item bank has not been developed to the extent necessary to provide multiple items per standard. There were also expectations that the formative system would provide items that were aligned to subjects other than ELA and Math. Science, social studies, health, etc. have not been developed or broadened in the item bank by AIR. Although educators may create items for all content areas, the initial offering of the items appears to fall short for teachers to maximize assessment *for* instruction.

Table 4

Summary of AIR Formative Assurances

XV Formative Assessment
A. Overview- AIR shall provide an instructional and formative assessment system through it Learning Point Navigator (LPN) system also known as SAGE Formative System. LPN is an online instructional support system which integrates with AIR's interim and summative reporting system, providing resources to support teacher and students in their effort to improve teaching and learning performance.

SAGE Formative Components	Functioning
1. an instructional system where teachers and students can find standards-aligned resources such as assignments, activities, and lessons linked with various learning modalities to enhance student learning	Yes
2. access to lesson plans and meta-instruction/professional development materials to enhance their pedagogical skills	Yes
3. teacher access to materials for students based on individual performance data in order to promote differentiated instruction	Yes
4. student access to performance data and feedback which empower students to manage their own progress, thereby enabling them to guide their own learning by providing access to instructional resources based on areas of strength and weakness	Yes
5. a formative assessment system for both teachers and students by providing access to score reports and feedback	Yes
6. present items in a manner that matches the interim and summative assessment system	Yes
7. the same range of automated scoring options as AIR's test delivery system for the following types of items: graphic response items: propositional responses; equation responses; essay responses	Yes
8. AIR's instructional resource and formative libraries and UTIPS-imported specific content libraries	Yes
9. allow activities to be grouped as a single resource (at publication time) or as assignments (by teacher at assignment time), allowing grouping of activities (such as items) under common stimuli, under common instruction sets, or in any other grouping desired (III-1 to III-4)	Yes

SAGE Interim

The SAGE Interim Assessments are LEA *optional*. Participation is determined locally and is not required by USOE. Student results are provided for LEA and school use, and no interim results are collected by the USOE. These assessments, which can be given twice a year, are computer adaptive to assess the knowledge, skills and abilities described in the Utah Core Standards for English Language Arts (ELA), Math, and Science, so both teachers and students can evaluate their performance according to the reporting categories of the standards.

Students are able to work to improve their mastery of the standards based on results from the interim assessment. A recurring theme from LEAs has been the desire to administer shorter and less time intensive interim assessments. Therefore, new for the 2015-16 school year, SAGE Interim Assessments may be administered as a "Class Period" option which is a truncated version of the SAGE summative assessment. LEAs also have the option to still administer the "Full Reporting" option which mirrors the SAGE Summative assessment in terms of breadth, depth, and length. Table 5 reports the specifics regarding each option.

Table 5

Summary of Content for Interim Options

Subject	Class Period Interim	Full Reporting Interim
ELA: Writing	1 prompt	1 prompt
ELA: Reading & Literacy	30-31 Items	40 Items
Mathematics	28-34 Items	35-42 Items
Science	30-36 Items	38-48 Items 1 Simulation

Similar to the SAGE Formative use by districts, SAGE Interim is also used by districts to set goals and inform instructional decisions (Table 6). The SAGE Interim assessments allow teachers and students to track progress from fall to winter to spring. With the introduction of the class period option there have been **68,510** SAGE Interim Assessments administered from September 8, 2015 to September 27, 2015.

Table 6

Interim Assessment Usage

SAGE Interim Usage between October 15, 2014 and January 26, 2015

SAGE Interim Usage	
41 Districts	217,875 Tests Scored
61 Charters + EHS	39,985 Tests Scored
Statewide	257,860 Tests Scored

The SAGE Interim assessment system could be enhanced further if specific standards were identified for each interim. Potentially assessing all standards that may or may not have been covered in the teaching and learning process may prove to be irrelevant feedback to teacher and student. Rather, providing a testing blueprint or table of specifications that identifies perhaps a third of the standards for the fall interim and another two thirds for the winter interim would provide meaningful guidance for planning instruction. This approach would allow LEAs to target their instruction and make interim reports more meaningful. This approach has been referred to and discussed with AIR and does not appear to be an option in the near future.

SAGE Summative

SAGE Writing

The SAGE writing assessment is a fundamental change in end of level testing as it pertains to grades 3-11. The writing assessment was first administered as part of the operational field testing in February 2014 and again in 2015. The early testing windows were necessary the first two years to allow immediate reporting of the ELA score as students finished the rest of their ELA assessments at the end of the academic year. Many LEAs expressed concern as to the timing of the writing assessment being so early in the academic year and requested a writing window adjustment to be more aligned with the other SAGE assessments. For 2015-16 the writing assessment will occur in alignment with the regular testing window in the spring.

USOE has an obligation to require an online writing assessment for grades 5 and 8 according to Utah Code 53A-1-603 (1) (b). This statutory provision also requires the development of "an assessment method…of students in grades 3 through 11 in mastering basic academic subjects." A "basic academic subject" means "a subject that requires mastery of specific functions, as defined under rules made by the State Board of Education, to include reading, language arts, mathematics, science in grades 4 through 12, and effectiveness of written expression. Utah Code 53A-1-602.

While there is some statutory ambiguity in the extent of assessment required for evaluating the effectiveness of written expression, R277-404-3 A (2) clearly states: "the Board shall maintain a comprehensive assessment system for all students in grades K-12. This assessment system shall include: Online Writing Assessment for grades 3 through 11." Online Writing Assessment means "a Board designated online assessment to measure writing performance for students in grades 3 through 11." R277-404-1. k. Thus, Board rule requires an "online writing assessment for grades 3 to 11.

Besides the Board rule requiring an online writing assessment from grades 3 through 11, writing, as a practical educational matter, is seen as a critical component of the Utah Standards at all levels, and should be assessed completely to measure the depth/breadth of those standards according to developmentally appropriate practices per grade level.

SAGE Writing Times

Over the past two years, USOE and AIR have shortened the stimuli/prompts to respond to time concerns. These efforts have included:

a. Emphasis on both teacher and student expectations for writing lengths and testing times. Student scores tend to max out after approximately 1 hour of time per essay. Students who spend more than 1 hour of time per essay see little if any appreciable increase in scores.

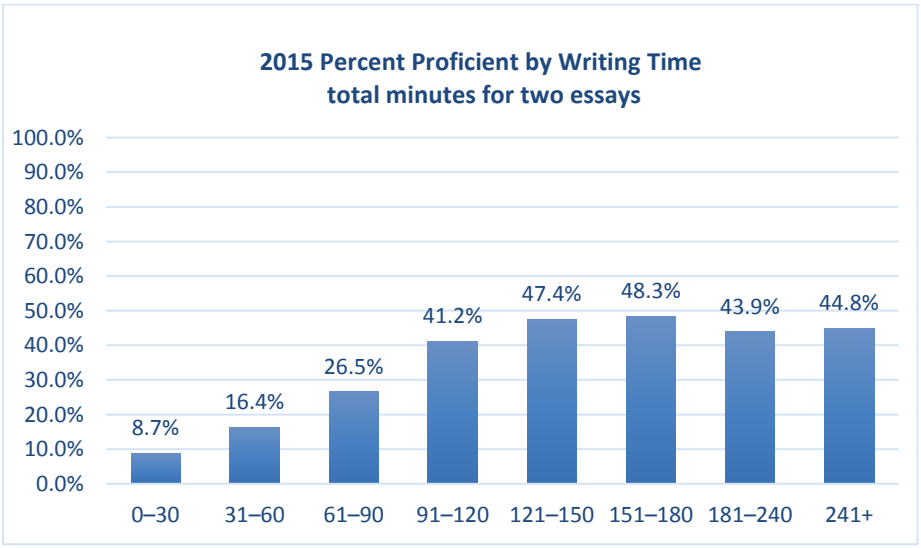
b. USOE and AIR launched a "class period" option for the interim assessment that is shorter for all subjects.

c. In 2015, only 19 LEAs (13% of LEAs) had an average writing time greater than 90 minutes for one or both of the essays. USOE is working directly with these LEAs to determine the cause of these extended testing times.

The following chart depicts the distribution range for the amount of time utilized to answer two essays. This chart suggests the “Law of Diminishing Returns” applies to students who spend more than one hour writing an essay. As a result, AIR and USOE staffs emphasize this data and message to both teachers and students in planning their written responses.

Figure 3

Summary of Percent Proficient by Writing Time



Machine scoring of essays

Machine scoring of essays is about 33% to 50% of the cost of human scoring of essays and has statistically been proven to be a valid and reliable scoring mechanism. During 2013-14 and 2014-15, all essays were human scored. In 2015-16 and 2016-17, all essays will be machine scored. However, 20% of the machine scored essays will be validated by human scoring. In addition, the range finding process is an essential training component of both the machines and the humans who validate the machine scoring of 20% of the essays. If anomalies occur in validating the 20% action will be taken to validate the remaining assessments.

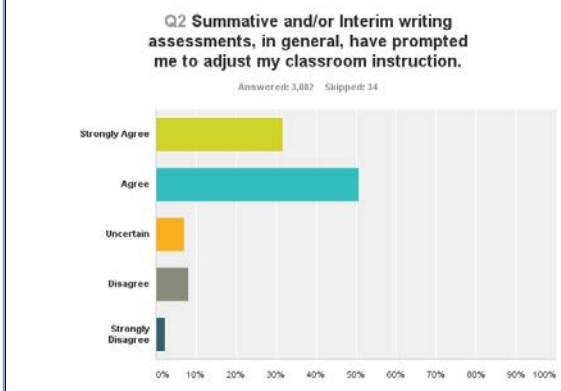
SAGE Informing Writing Instruction

In addition, below is a survey conducted of Utah educators on how the Writing Assessment and feedback received from the writing reports has impacted teaching methods.

Figure 4

Educator Feedback on SAGE Impact on Instruction

2. Summative and/or Interim writing assessments, in general, have caused me to adjust my classroom instruction.



Response Options	*Percentage*	*Responses*
Strongly Agree	31.73%	978
Agree	50.78%	1,565
Uncertain	7.07%	218
Disagree	8.05%	248
Strongly Disagree	2.37%	73

Summative Reports

The first administration of SAGE summative in 2014 required standard setting, data processing, algorithm verification and psychometric auditing procedures. As planned, this process required additional time to make data available to LEAs and occurred in October 2014. During the 2015 administration, LEAs down to the teacher and student level had SAGE data available to them as students finished the assessment.

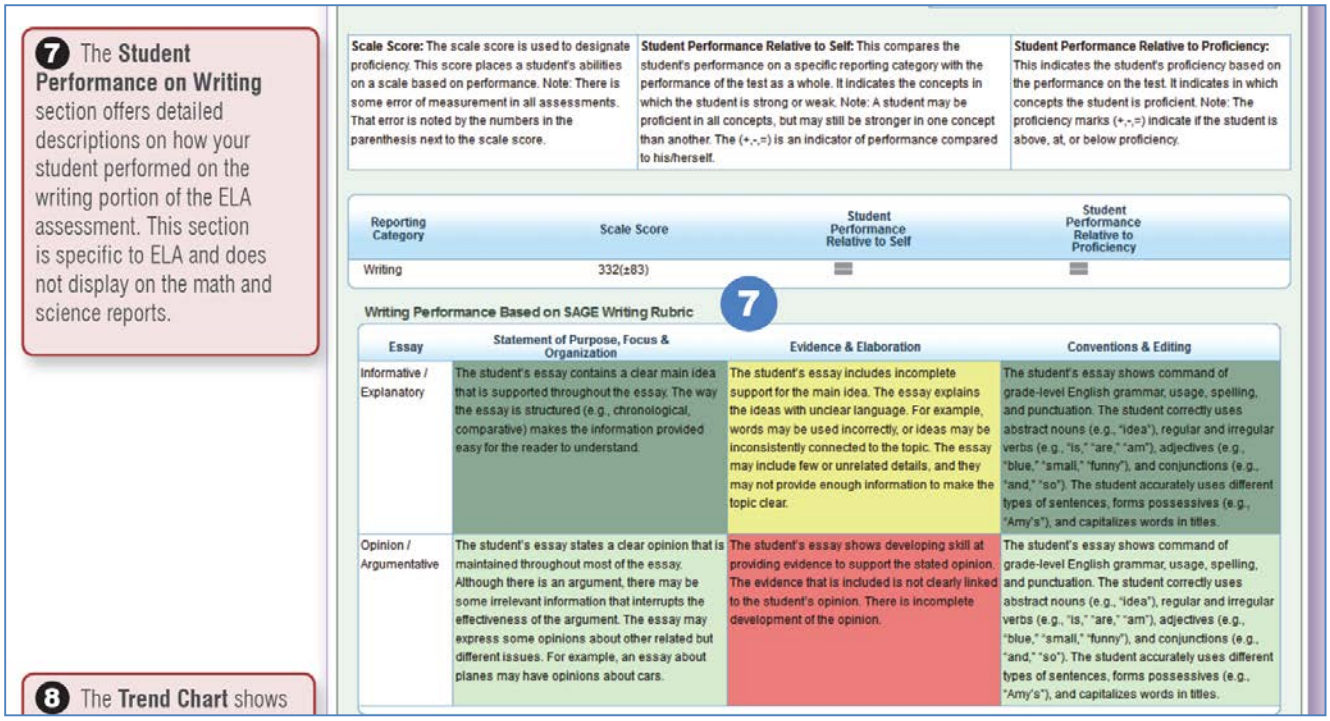
The reports have improved since the initial release and have been modified based on feedback from stakeholders. USOE worked with AIR to enhance the Individual Student Report (ISR) to provide more meaningful data. As schools, districts and the State continue to consume and use SAGE results, additional adjustments to reporting may be made.

During the fall of 2015, USOE is conducting Principal Outreach trainings to assist schools with accessing and using the many SAGE reports available to them through both the SAGE portal and the USOE Data Gateway. Response to these trainings to date has been very positive.

Below is an example of the type of feedback available to teachers and students to help improve student performance:

Figure 5

SAGE Student and Teacher Feedback Example



SAGE Item Bank

An independent review of SAGE test items concluded that **"…**the test items were determined to be error free, unbiased, and were written to support research-based instructional methodology, use student- and grade-appropriate language as well as content standards-based vocabulary, and assess the applicable content standard" (Independent Verification of the Psychometric Validity for the Florida Standards Assessment, 2015)

During Year 1 (2012-13) and Year 2 (2013-14) of the contract, AIR met all of its item development requirements for ELA. It mostly met its item development requirements for science. It did not meet its item development requirements for math. A synopsis of the item development by subject is set forth below in Table 7. A more extensive breakdown of item development by grade level can be provided.

Table 7
SAGE Item Development

Subject	AIR Contract Obligation Year One 2012-13	Items Year One	Diff.	# of Grade Levels met year two	AIR Contract Obligation Year two	Items Year two	Diff.	# of Grade Levels met year two
ELA	1,850	3,382	+1,530	9	450	806	+356	9
Science	450	855	+405	9	450	462	+ 12	6*
Math	1730	951	-779	0	450	437	- 13	3*
Total	**4,030**	**5,188**	**+1,158**		**1,350**	**1,705**	**+355**	

*In each non-compliant grade, at least 45 of 50 items developed.

In 2013, a memorandum was generated by AIR and USOE. This memorandum addressed problems with sufficient numbers of Year 1 ELA items because: 1) a number of Utah's existing CRT ELA items could not be aligned; and 2) AIR encountered difficulties with copyrights of ELA passages. As a result, an agreement was developed where, in Year 1, AIR would develop additional ELA items and decrease the number of math items. In total, AIR was obligated to provide 130 additional items as is set forth in the chart below.

Table 8

Item Development Memorandum

Subject	New Items Contributed to Field Test Pool From New Development Per Grade (Original)	Revised New Items Contributed to the Field Test Pool from New Development Per Grade*	Difference	Actual Year 1 Number on Chart Above
ELA	1,850	2,520	+670	3,382
Math	1,700**	1,160	-540	951
Science	450	450	0	855
Total	**4,000**	**4,130**	**+130**	**5,188**

*A copy of this memorandum is available for review.

** 1,700 and not 1,730 was number of math items identified in memorandum.

The results of the 2015-2016 summative simulations revealed two needed areas of development for the SAGE item banks:

- Secondary Mathematics I, II, and III need additional item development in order to ensure sufficient items across ability groups to increase the flexibility of the pool to deliver items at each student ability level.
- In ELA, there are a number of unused item sets that could be bolstered by additional item development for those passage sets.

Brailling of Items

At the current time, USOE does not have sufficient items in its item bank which meet the criteria to be approved for Brailling. The process for item approval is as follows:

a. all items must go through
    i. Development
    ii. Operational field testing
    iii. Data review/workshops
b. Items that pass through workshops/reviews are eligible for Brailling.
c. Not all items cleared for Brailling can actually be brailed. Certain item types such as grid items (e.g. drag and drop, construction, simulation, etc.) will be excluded from Brailling. Furthermore content-inappropriate items must be excluded from Brailling

Table 9

Summary of Braille Items

No.	Obligation	Source of Obligation	Deadline Date	Cost	Status of Completion
1	Braille 1,350 items	Original Contract and Amendment #1 which extended original contract obligation.	Spring 2014	$676,947	**Completed**, but not in a timely manner.
2	Braille 9,450 items	Amendment #1	Fall 2014		**Partially completed.** 2,677 items have not been Brailed.
3	Braille 1,013 items	Amendment #3	Spring 2015	**$105,703**	**Not Completed**
	Total-11,813 items				As of Spring 2015, 8,123 items brailed. The 2014 Brailling deadlines were to be extended to August 31, 2015 by proposed Amendment #4

SAGE Growth Reports

The AIR proposal establishes an obligation for providing "Trend Reports." In effect, these trend reports provide an overview of student growth by stating that "scores for the state, LEAs, schools, teachers, classes, and students can be plotted on a trend report to illustrate how performance has changed over time."

AIR's reporting tool "emphasizes the context of a student's performance by relating it to aggregate performance and trends over time." In addition to the expected aggregated reports, the reporting tool allows users to create custom rosters of students so that they can be tracked and reported on separately. For example, a teacher may have a class of students, and be tracking the class's test scores; however, the teacher may also have a group of students in the class who are receiving extra help after school. The teacher can create a second roster of only those students to report on their aggregate performance from the group. This may help the teacher begin answering questions about the effectiveness of the extra help or the strengths and weaknesses of this subgroup of students. All aggregated reports are calculated from the student level, so as new students test, reports for the roster, teacher, school, and LEA are updated instantaneously"

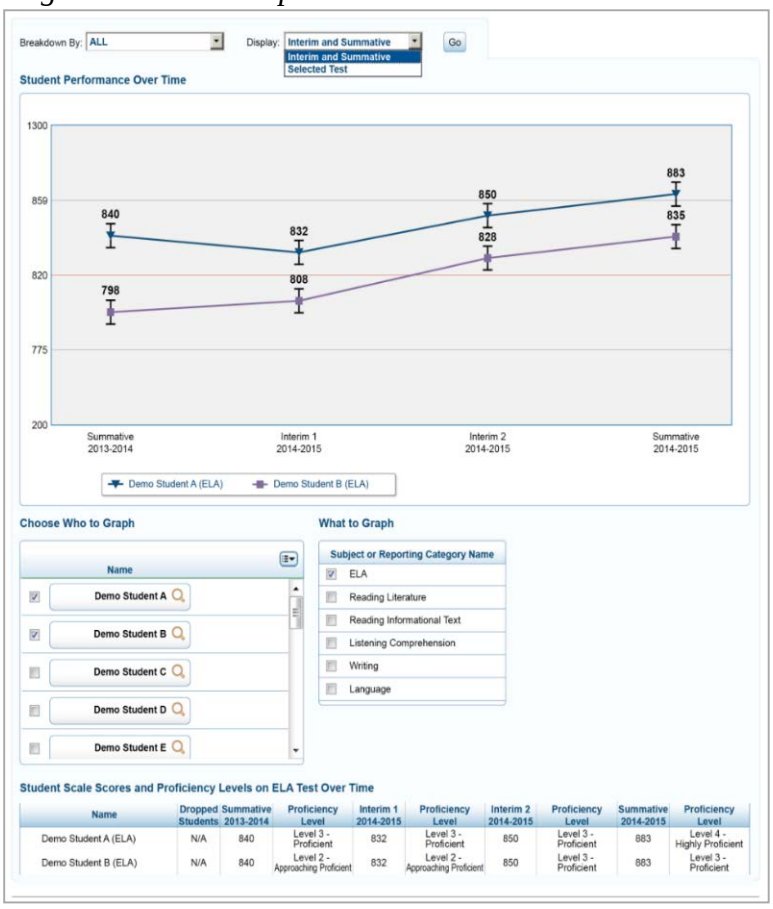
It would appear that AIR's representation in the trend report, interpretation and aggregation of data are in response to the RFP's requirement that "Proposals shall include details and samples of the following reports, which should include, at a minimum, growth (spring to spring for all LEAs and fall to spring optional for LEAs), proficiency, and sub-score information: 1) State summary; 2) LEA summary; 3) School summary; 4) Class summary; and 5) Individual student results".

The SAGE score reports allow authorized users to view scores in the aggregate for LEAs, schools, teachers, and rosters, as well as individual student performance data for previous Summative and Interim ELA, math, and science assessments. Users may take advantage of data analyses to determine what strategies may be needed to be implemented in order to improve teaching and learning. Data can be compared with the overall state and LEA average for the test being analyzed. Additionally, performance trends are viewable and determine whether overall performance is increasing.

The growth reports simply provide some longitudinal data to determine how a student is performing over time. The growth report does not provide analysis to the extent of determining the significance between multiple scores. In other words a scale score of 320 at point 'A' and 340 at point 'B' only indicates an increase. The 20 point difference does not suggest growth above or below what would be considered adequate. An example of a growth report is found in Figure 6.

Figure 6

SAGE Longitudinal Score Report



In addition to the report above, the Individual Student Report (ISR) displays the breakdown of the student's scale score; proficiency level descriptors for the selected subject- or course-based test; and performance at each reporting category. The ISR is made available by LEAs to parents and students to communicate student performance. The report includes average scale scores for the State and LEA for comparison purposes and may also include a graph to display the student's performance on the test over time. Average scores are compiled from the scores of those students who have taken the same test to date; averaged computations do not include those students who have not taken or completed the test.

This trend data is not to be confused with the student growth percentile (SGP) which is an accountability calculation that is provided by USOE through the Data Gateway accessible to teachers and schools. The SGP has been used to determine adequate growth of students and is also reported in the aggregate. Several LEAs utilize the SGP to determine student needs and teacher effectiveness.

SAGE Data Availability

SAGE data results are available at varying times depending on the purpose of the data reporting. There are four main categories of reporting used by the State Office of Education that are reported at different times for different purposes.

1. Raw results
   - Raw test results are immediately available upon student completion through the *Online Reporting System* of the SAGE assessment system. These results are available to students, teachers, schools, and LEAs.

2. Public Data Gateway
   - This data can be viewed publically at the school, district, and state levels only in aggregate form. Test results are provided through the public Data Gateway for all interested parties. These include all tests that students participated in and completed, and include students with allowable accommodations with disabilities. All other tests that were partially completed, or for students that were no longer in a course to take a test, absent, parent opted out, English learners (ELs) who are in the first year of the school system, etc. would not be included in these data. This data is not immediate as it must be verified and merged with student enrollment information through USOE IT. This information is not usually available until July or August following the spring administration.

3. Secure Data Gateway
   - The secure data gateway is for educators, administrators, and others with a vested educational interest to view aggregate test results for subgroup, class, school, district, and state level data. It includes a process to query results for groups of students over time, providing valuable longitudinal data that inform the school improvement process. This data is not immediate as it must be verified and merged with student enrollment information through USOE IT. This information is not available until July or August. Furthermore, users must log into the system to review individual student data.

4. Accountability
   - For school accountability purposes, a different set of inclusion rules apply and be reported at the school level only. Schools must first meet a participation requirement of testing 95% of all students enrolled at their school at the time of test administration. A student is considered to have participated in a test if they were enrolled at the time of the test and were not absent or excused, followed by the specific inclusion rules for students with disabilities and English learners. Students whose parents opt them out of testing will not be counted as a participant in the State's School Grading accountability system; however, they are counted as a participant under the School Federal Accountability Report (SFAR).
   - For academic achievement, students who were present for 160 days of instruction are included. Growth and proficiency are both calculated for the entire school and

again for students who were not yet proficient. In addition, annual measureable objective targets are reported at the subgroup level that includes the following groups with *n* sizes of ten or greater;

i. All students, Asian, African American, American Indian, Caucasian, Hispanic, Pacific Islander, Economically Disadvantaged, English learners, and Students with disabilities.

- This data is not immediate as it must be verified and merged with student enrollment information through USOE IT. It is publically released September 15 annually.

All data reporting is in compliance with The Family Educational Rights and Privacy Act (FERPA) which is a federal law that protects students' privacy by prohibiting disclosure of education records without adult consent.

Validity of SAGE Summative

Validity refers to the degree to which test score interpretations are supported by evidence, and speaks directly to the legitimate use of test scores. Establishing the validity of test score interpretations is thus the most fundamental component of test design and evaluation. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) provide a framework for evaluating whether claims based on test score interpretations are supported by evidence. Within this framework, the Standards describe the range of evidence that may be brought to bear to support the validity of test score interpretations.

The kinds of evidence required to support the validity of test score interpretations depend centrally on the claims made for how test scores may be interpreted. Moreover, the standards make explicit that validity is not an attribute of tests, but rather test score interpretations. Some test score interpretations may be supported by validity evidence, while others are not. Thus, the test itself is not considered valid, but rather the validity of the intended interpretation and use of test scores is evaluated.

Evidence for the validity of test score interpretations is strengthened as evidence supporting test score interpretations accrues. In this sense, the process of seeking and evaluating evidence for the validity of test score interpretation is ongoing. Nevertheless, there currently exists sufficient evidence to support the principle claims for the test scores, including that SAGE test scores indicate the degree to which students have achieved the Utah Core Standards at each grade level, and that students scoring at the proficient level or higher demonstrate levels of achievement consistent with national benchmarks indicating that they are on track to college readiness. These claims are supported by evidence of a test development process that ensures alignment of test content to the Utah Core Standards, a standard setting process that yielded performance standards consistent with those of rigorous, national benchmarks, and evidence that the structural model described by the Utah Core Standards and implemented in the SAGE assessments is sound

In tables 10, 11, and 12 are provided a summary of the bias and average standard errors of the estimated theta by grade and assessment opportunity. In all grades and content areas, the mean bias of the estimated abilities is very small and statistically insignificant, providing the evidence needed to demonstrate that the true score is adequately recovered in the observed score.

The summary statistics of the estimated abilities show that for all examinees in all grades, the item selection algorithm is choosing items that are optimized, conditioned on each examinee's ability. Essentially, this shows that the examinee ability estimates generated on the basis of the items chosen are optimal in the sense that the final score for each examinee always recovers the true score within expected statistical limits. In other words, given that we know the true score for each examinee in a simulation, these data show that the true score is virtually always recovered—an indication that the algorithm is working exactly as expected for a computer-adaptive test.

In addition, the average standard errors are 0.27, 0.30, and 0.31 across all assessment opportunities in reading, mathematics, and science, respectively. Although the item pool is augmented with difficult items to measure the high-performing students' ability more efficiently, it is very challenging to develop item pools that are robust enough to accurately measure students at the extreme levels of knowledge and skills within on-grade-level content standards.

Table 10

Standard Errors of the Estimated Abilities for ELA

Grade	**Average Standard Error**	**SE at 5 Percentile**	**SE at Bottom Quartile**	**SE at Top Quartile**	**SE at 95 Percentile**
3	0.22	0.29	0.21	0.19	0.20
4	0.27	0.35	0.23	0.27	0.29
5	0.25	0.26	0.31	0.25	0.29
6	0.24	0.27	0.23	0.23	0.26
7	0.25	0.28	0.22	0.24	0.24
8	0.26	0.28	0.24	0.26	0.26
9	0.30	0.34	0.31	0.31	0.31
10	0.30	0.37	0.26	0.31	0.31
11	0.32	0.35	0.28	0.32	0.33

Table 11

Standard Errors of the Estimated Abilities for Math

Grade	Average Standard Error	SE at 5 Percentile	SE at Bottom Quartile	SE at Top Quartile	SE at 95 Percentile
3	0.15	0.20	0.16	0.13	0.14
4	0.17	0.26	0.16	0.16	0.18
5	0.19	0.29	0.19	0.16	0.19
6	0.23	0.28	0.21	0.21	0.24
7	0.24	0.36	0.25	0.18	0.19
8	0.29	0.42	0.32	0.21	0.24
SMI	0.37	0.69	0.37	0.27	0.26
SMII	0.52	0.95	0.54	0.38	0.31
SMIII	0.55	1.17	0.55	0.39	0.38

Table 12

Standard Errors of the Estimated Abilities for Science

Grade	Average Standard Error	SE at 5 Percentile	SE at Bottom Quartile	SE at Top Quartile	SE at 95 Percentile
4	0.33	0.34	0.31	0.34	0.38
5	0.37	0.34	0.34	0.35	0.48
6	0.30	0.30	0.24	0.30	0.46
7	0.30	0.32	0.25	0.30	0.44
8	0.32	0.37	0.30	0.33	0.36
Bio	0.30	0.39	0.26	0.29	0.37
Chem	0.27	0.31	0.22	0.24	0.36
ESS	0.33	0.47	0.27	0.29	0.37
Physics	0.31	0.38	0.30	0.29	0.33

Overall, these diagnostics on the item-selection algorithm provide evidence that scores are comparable with respect to the targeted content, and scores at various ranges of the score distribution are measured with good precision. However, it should be noted in Table 11 that there is a greater degree of error for assessing math competencies among lower performing students. There is a need, as mentioned in the item development section and referenced in Table 7 and 8, for more secondary math items at the lower end for determining numeracy.

SAGE Achievement Gap

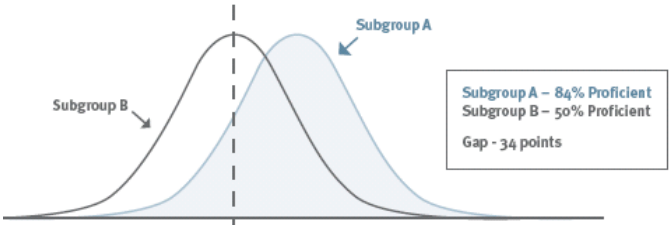
All of the test developers at AIR are trained to write items that are accessible to all students, based on the principles of universal design. In addition, all of AIR's test developers are required to pass a certification examination that certifies their ability to implement AIR's Language

Accessibility Guidelines in the items that they are developing. Each item presented to the Utah review committees is reviewed by three content experts at AIR as well as an editor. At each review level, every item is checked for language accessibility and for adherence to universal design principles.

There is a nationally recognized achievement gap among varying populations and between students of varying demographics. There continues to be an achievement gap in Utah and it would seem that the SAGE summative assessment has exasperated that gap. Typically the achievement gap is viewed by simply looking at the difference between the reported percent proficiency of groups. Unfortunately, this metric is ill suited to report the breadth and depth of the achievement gap. Understanding the achievement gap represents a statistical lens as observed in the figures below:

Figure 7

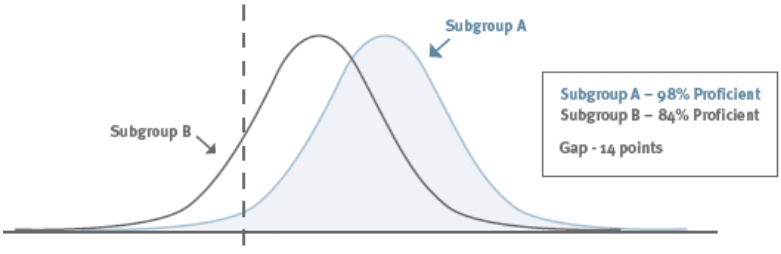
Difference Between Subgroups Time 1



It would appear that there is a 34 point difference or gap between groups. In the figure below both groups progressed to the right and the percent proficient gap would appear to close within 14 points. However, statistically the gap did not close to the extent that the percent proficient metric would suggest.

Figure 8

Difference Between Subgroups Time 2



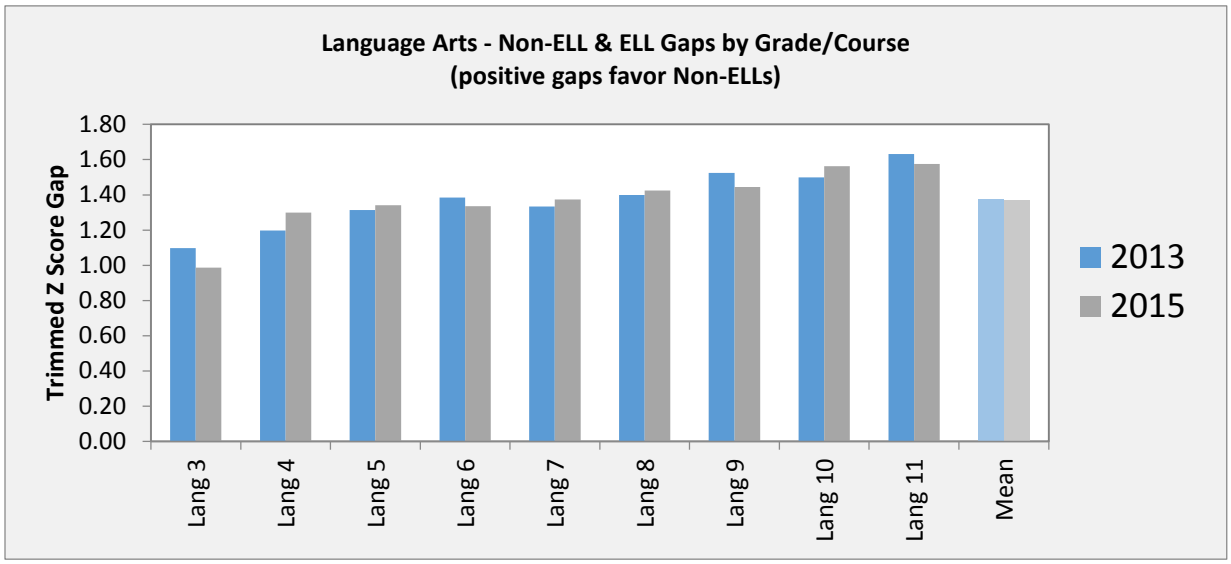
Using effect size is a more robust and statistically accurate alternative to understanding the achievement gap. Effect size is defined by determining the difference between the means of the groups and dividing by the standard deviation of either one of the groups or the pooled

standard deviation of both groups. Effect size has multiple advantages including the use of the full score distribution rather than the single points used in percent proficient to explain achievement gaps. Effect size is also comparable across different assessments which aids in the understanding of subgroup performance between CRT and SAGE.

The graph below illustrates the effect size scores (Z score) between subgroups and between test administrations which is useful in understanding ELA performance on SAGE and CRT.

Figure 9

Achievement Gap on CRT vs. SAGE



The graph above shows the calculated achievement gap between ELL student performance on the 2013 CRT and the 2015 SAGE assessments as compared to Caucasian students. In most cases the effect sizes are negative, meaning the ELL student groups performed worse than the White students from CRT to SAGE but not to the extent that the percent proficient description would indicate between CRT and SAGE. Therefore, although an achievement gap exists between groups, the difference in the gap between CRT to SAGE is not as pronounced as the percent proficiency difference would indicate when considering the full distribution of scores. Effect size differences between SAGE and CRT are less in math and science.

SAGE Access & Accommodations

To help minimize access and language barriers of some populations, several resources are available to all students. In 2014–2015, the available assessment tools included the following: alternate location, assistive communication devices, audio amplification, calculation devices and computation tables, directions signed with certified interpreter, highlight tool, dictionary tool which featured a Thesaurus and Spanish translation options, text to speech, magnification, minimize distractions, scratch paper, spell check, and strike through.

In addition to resources available to all students, there are options available to accommodate students that have been identified with special needs. In the 2014–2015 administration, the available accommodation options included the following: Braille, ASL videos for spring 2015 administration, descriptive audio, printing on request and scribe (non-functional in SAGE systems). There are also accommodations available for English Language Learners. Simply stated an accommodation is a practice or procedure in presentation, response, setting, timing, or scheduling that, when used in testing, provided equal access to all students. State approved accommodations do not compromise the learning expectations, constructs, grade-level standards, or measured outcome of the assessments.

Summary

The SAGE summative, interim, and formative system continues to evolve to meet the diverse needs of LEAs to provide meaningful data to inform the teaching and learning process. Each component of the system has strengths and weaknesses. Where weaknesses persist, LEAs have found ways to supplement shortcomings internally and USOE staff has worked with AIR for improvement. While certainly not meeting all of LEA needs, SAGE represents a system that has provided more resources to impact student achievement than has been previously experienced or provided in any statewide assessment system in Utah.

The SAGE summative system has specifically defined how the standards are assessed and has raised the bar for student learning. Current SAGE results are in line with student performance on ACT, and other external measures. Issues regarding the length of time students are testing have been analyzed and certain steps have been taken to curtail undue time commitments. Access issues for limited English speaking students has also been taken into consideration and accommodations have been provided. The SAGE summative item bank needs to be developed further to address the depth needed for a computer adaptive assessment to adequately address a broader learning/proficiency spectrum.

The SAGE interim system represents the largest shortfall of the SAGE system. SAGE interim has primarily served as an opportunity for students to see the nature of the questions that will be asked on the summative assessment rather than a tool to measure or inform student learning. Students are potentially assessed on material that has not been covered in the classroom. We currently do not have the ability to predetermine which standards will be assessed on each interim. Further work needs to be done to make SAGE interim an improved element to inform learning progressions.

The SAGE formative system as introduced in 2013 largely fell short of general expectations. The formative system has undergone a few meaningful changes that have made it more practical and attractive for LEA use. Currently there are no additional changes being considered to the formative system although several could be made. LEAs have made recommendations and attest to the need of having a substantive formative assessment resource to accomplish the primary purpose of using assessment to drive curricular choices and instructional practices.