Utah State Assessments 2014–2015 Technical Report

Executive Summary

American Institutes for Research

American Institutes for Research

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SPRING 2015 UTAH STUDENT ASSESSMENT OF GROWTH AND EXCELLENCE (SAGE) EXECUTIVE SUMMARY

Validity refers to the degree to which test score interpretations are supported by evidence and speaks directly to the legitimate uses 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, & 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 an attribute not of tests but rather of 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.

There are a number of intended uses for SAGE test scores, including for school accountability, feedback about student and class performance, measurement of student growth over time, evaluation of performance gaps between groups, evaluation of teacher performance, and diagnosis of individual student strengths and weaknesses. Each of these intended uses requires claims to be made about the interpretation of test scores, and the strength of those claims rests on the validity evidence supporting those claims. Some validity evidence will be central to all of the claims, including especially evidence for the alignment of test items and administrations to Utah Core Standards. Other evidence may target more specific claims, such as evidence for measurement of student growth or evaluation of teacher performance. Evaluation of validity evidence should therefore be made with respect to the claim that it is purported to support.

Central to evaluating the validity of test score interpretations is determining whether the test measures the intended construct. Such an evaluation in turn requires a clear definition of the measurement construct. For Utah's SAGE assessments, the definition of the measurement construct is provided by the Utah Core Standards.

The Utah Core Standards specify what students should know and be able to do by the end of each grade level in order to graduate ready for post-secondary education or entry into the workforce. The Utah Core Standards were initially established in 1984 and are regularly revised. The current Utah Core Standards for ELA and mathematics were approved by the Utah State Office of Education in 2010, and these standards were fully implemented in June 2013 for ELA and in April 2013 for mathematics. Utah's science standards were adopted and implemented in 2010. The Utah Core Standards for ELA, mathematics, and science describe the educational targets for students in each subject area.

Because directly measuring student achievement against each benchmark in the Utah Core Standards would result in an impractically long test, each test administration is designed to measure a representative sample of the content domain defined by the Utah Core Standards. To ensure that each student is assessed on the intended breadth and depth of the Utah Core Standards,

item selection in the test delivery system is guided by a set of test specifications, or blueprints, that indicate the number of items that should be sampled from each content strand, standard, and benchmark. Thus, the test blueprints represent a policy statement about the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the Utah Core Standards is evaluated, alignment of test blueprints with the content standards is critical. USOE has published the SAGE test blueprints that specify the distribution of items across reporting strands and depth of knowledge levels.

While the blueprints ensure that the full range of the intended measurement construct is represented in each test administration, tests may also inadvertently measure attributes that are not relevant to the construct of interest. For example, when a high level of English language proficiency is necessary to access content in math and science items, language proficiency may unnecessarily limit the student's ability to demonstrate achievement in those subject areas. Thus, while such tests may measure achievement of relevant math and science content standards, they may also measure construct irrelevant variation in language proficiency, limiting the generalizability of test score interpretations for some student populations.

The principles of universal design of assessments provide guidelines for test design to minimize the impact of construct-irrelevant factors in assessing student achievement. Universal design removes barriers to access for the widest range of students possible. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

Test development specialists receive extensive training on the principles of universal design and apply these principles in the development of all test materials, including items and accompanying stimuli. In the review process, adherence to the principles of universal design is verified.

In addition, the SAGE test delivery system provides a range of accessibility tools and accommodations for reducing construct irrelevant barriers to accessing test content for virtually all students. The range of accommodations provided in the online testing environment far exceed the typical accommodations made available in paper-based test administrations, which were typically limited to large print, Braille, and English and foreign language audio translations. Table 1 lists the accommodations and accessibility supports currently available for the SAGE assessments.

Table 1: Accommodations and Accessibility Supports

Accessibility Feature	Description
Text-to-Speech—Directions, Passages, Items	Computer reads text and graphics aloud on directions, passages, and items. What is read and how it is read is configurable.
Text-to-Speech—Graphic Description	Computer reads graphics and tables aloud.
Magnification Interface	Student can zoom in and zoom out on the entire page. This capability persists throughout the test.
Magnifier	Student can magnify a selected portion of an item.
Variable Font Size	The number of levels (generally, five levels) and rate of increase (generally, 1.25x the previous level) are configurable.
Refreshable Braille/ Tactile With External Embosser Printer	Items can be rendered to desktop embossers that can integrate Braille and tactile graphics. The items will simultaneously render on a reader-accessible screen, and the student will be able to navigate to response spaces to provide answers.
Reverse Contrast	Background turns to black, while text turns to white.
Administrator-Selectable Variable Font and Background Colors	Any foreground and background color can be supported.
Color Overlay	Any color can be laid on the screen. This persists throughout the test.
Increased White Space	This is the streamlined interface.
Sign Language—Directions, Passages, Items	This capability consists of recorded videos on sign language. Avatars are not recommended by hearing-impaired experts since they do not translate well to American Sign Language.
Translations	Versions are available in alternate languages.
Keyword Translation	This enables translators to associate keyword translations.
Glossaries and Dictionaries	These enable content developers to associate additional content with words or phrases. The content can be of multiple types, and the content shown to a student can be controlled by his or her personal profile.
Alternate Language Glossaries and Dictionaries	These enable content developers to associate alternate-language content with words or phrases. The content can comprise multiple types, and the content shown to a student can be controlled by his or her personal profile.
Administrator-Selectable Assistive Devices Integration	Our system has a standard and a streamlined interface. Most assistive devices can work with the former, and an even wider group works with the latter. If the use of the device requires relaxation of certain security features (e.g., if suppression of pop-up windows interferes with on-screen keyboards), the system can be configured to allow the test administrator to select a more permissive mode.
Line Reader	This feature will allow a student to track the line he or she is reading.

Accessibility Feature	Description
Masking	Students can mask extraneous information on the screen.
Speech-to-Text	Speech will be converted to text and then saved in the database. (Available through compatibility with third-party assistive technology.)
Auditory Calming	A tool that plays music or white noise in the background. (Available through third-party software.)
Administrator-Selectable Zoom	Default font size can be set in advance through a file upload or user interface or at the time of testing by the test administrator. Student can zoom in or zoom out at any time.
Administrator-Selectable Large Print Font	Default font size can be set in advance through a file upload or user interface or at the time of testing by the test administrator. Student can zoom in or zoom out at any time.
Administrator-Selectable Screen-Reader	The system supports an integrated screen reader that can be configured to provide a variety of support levels, each selectable by the administrator.
Additional Time	AIR's system currently does not impose a time limit on the test. It is up to the proctor to stop a student's test or stop the entire session. However, if there are unforeseen events, such as a fire alarm, that trigger need for additional testing time, AIR's system can enable a grace period extension (GPE) for a single test opportunity or for multiple test opportunities.
Segment Breaks	AIR's system has the capability of adding test segments within a test. A test segment is made up of multiple item groups and creates a logical break between segments within a test. For example, a segment break might separate a calculator from a non-calculator segment of a test.
Recorded Audio	Computer efficiently delivers recorded audio. We are able to deliver voice-audio using only about 10 Kbps of bandwidth.
Secure Print Facility	A visual accessibility feature, the secure print facility allows the secure printing of items or passages. A student requests that a passage or item be printed; the request is then encrypted and sent securely to the proctor; the proctor needs to approve the request before it is sent to the printer. In addition, this feature also allows for the delivery of real-time paper tests, including large print tests.
Test Pauses and Restarts	An attention accessibility feature, test pauses and restarts, allows the test to be paused at any time and restarted and taken over many days. So that security is not compromised, visibility on past items is not allowed when the test has been paused longer than a specified period of time.
Writing Checklists	An attention accessibility feature generally for essay items, the writing checklist enables a student to check off writing guidelines from a checklist.
Review Test	Students can review the test before ending it.
Area Boundaries	An agility accessibility feature, area boundaries for mouse-clicking multiple-choice options allow students to click anywhere on the selected response text or button.
Language	Any language that is necessary can be supported.
Help Section	A reference feature, the Help Section explains how the system and its tools work.

Accessibility Feature	Description
Performance Report	A reference feature, a performance report is available at the end of the test for the student.

1. Evidence Based on Test Content

Because the SAGE assessments are designed to measure student progress toward achievement of the Utah Core Standards the validity of SAGE test score interpretations critically depend on the degree to which test content is aligned with expectations for student learning specified in the Utah Core Standards.

Alignment of content standards is achieved through a rigorous item development process that proceeds from the content standards and refers back to those standards in a highly iterative item development process that includes the state department of education, test developers, and educator and stakeholder committees. The review process is described in more detail in Section 2.1.1 and is explicitly designed to ensure rigorous alignment of test content to the Utah Core Standards.

Ensuring the alignment of test items to their intended content standards establishes a critical link between the expectations for student achievement articulated in the Utah Core Standards with the SAGE item content. The SAGE test blueprints, in turn, specify the range and depth with which each of the content strands and standards will be covered in each test administration, and thus completes the link between the Utah Core Standards and the SAGE content based test score interpretations.

The test blueprints drive item selection in the adaptive algorithm used to administer SAGE assessments. The adaptive algorithm seeks to meet three objectives: satisfy blueprint constraints, maximize overall test information near the student's ability estimate, and maximize test information within each of the reporting strands as well. Each item satisfies multiple blueprint elements. For example, an item not only measures a particular content standard, but does so at a particular depth of knowledge. As the test progresses, item selection weights increase for blueprint elements that have not been met, while items measuring blueprint elements that have been satisfied are no longer considered. The adaptive algorithm is configured for each assessment to ensure that all critical blueprint elements are satisfied for each test administration.

Moreover, unlike with fixed-form tests in which the same test form is administered to all students statewide, the SAGE assessments are administered adaptively, with students within classrooms and schools administered different samples of items from the subject area pool. Thus, while each student may be administered only one or two items per benchmark, indicators of performance at the classroom and school levels are based on a larger, more representative sample of the content domain than is possible with fixed-form assessments, ensuring that teachers and schools are held accountable for instruction across the full range of the academic content standards.

Details of the procedures used to develop and review the items comprising the SAGE adaptive item pools are provided in Volume 4.

2. Independent Alignment Study

While it is critically important to develop and strictly enforce an item development process that works to ensure alignment of test items to content standards, it is also important to independently

verify the alignment of test items to content standards. USOE has contracted with the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) to conduct an independent alignment study.

The CRESST study is two-pronged, and is designed to evaluate the adequacy of the item pool, and the adequacy of the administered test forms generated by a computer adaptive algorithm that were delivered to Utah students in the 2014-2015 school year. To evaluate the adequacy of the item pool, CRESST will rely on a team of content experts to code for cognitive complexity and the academic content standards for each of the content areas (ELA/L, Mathematics, and Science). To evaluate the adequacy of the computer adaptive tests (CATs) administered to students, the CRESST study will evaluate standards and blueprint fulfillment, as well as the informativeness, item difficulty, and reliability of the administered tests. The alignment studies are scheduled to be completed in spring 2016.

3. Evidence for Interpretation of Performance

Alignment of test content to the Utah Core Standards ensures that test scores can serve as valid indicators of the degree to which students have achieved the learning expectations detailed in the Utah Core Standards. However, the interpretation of the SAGE test scores rests fundamentally on how test scores relate to performance standards, which define the extent to which students have achieved the expectations defined in the Utah Core Standards. SAGE test scores are reported with respect to four proficiency levels, demarcating the degree to which Utah students have achieved the learning expectations defined by the Utah Core Standards. The cut score establishing the Proficient level of performance is the most critical, since it indicates that students are meeting grade level expectations for achievement of the Utah Core Standards, that they are prepared to benefit from instruction at the next grade level, and that they are on track to pursue post-secondary education or enter the workforce. Procedures used to adopt performance standards for the SAGE assessments are therefore central to the validity of test score interpretations.

Following the first operational administration of the SAGE assessments in spring 2014, a series of standard setting workshops were conducted to recommend to USOE a set of performance standards for reporting student achievement of the Utah Core Standards. Utah educators, serving as standard setting panelists, followed a standardized and rigorous procedure to recommend performance level cut scores. The workshops employed the Bookmark procedure, a widely used method in which standard setting panelists used their expert knowledge of the Utah Core Standards and student achievement to map the performance level descriptors adopted by USOE onto an ordered item book comprising an operational test form meeting all blueprint elements.

Panelists were also provided with contextual information to help inform their primarily content-driven cut score recommendations. Panelists recommending performance standards for the high school assessments were provided with information about the approximate location of the relevant ACT college ready performance standard for each of the assessments. Panelists recommending performance standard for the grades 3–8 summative assessments were provided with the approximate location of relevant National Assessment of Educational Progress (NAEP) performance standards. Panelists were asked to consider the location of these benchmark locations when making their content-based cut score recommendations. When panelists are able to use benchmark information to locate performance standards that converge across assessment systems, validity of test score interpretations is bolstered.

In addition, panelists were provided with feedback about the vertical articulation of their recommended performance standards so that they could view how the locations of their recommended cut scores for each grade level assessment sat in relation to the cut score recommendations at the other grade levels. This approach allowed panelists to view their cut score recommendations as a coherent system of performance standards, and further reinforces the interpretation of test scores as indicating not only achievement of current grade level standards, but also preparedness to benefit from instruction in the subsequent grade level.

Following recommendation of final performance standards, as well as vertical moderation sessions to ensure articulation of recommended cut scores across grade levels, the recommended cut scores were presented to a stakeholder panel for review and comment.

Based on the adopted cut scores, Table 2 shows the percentage of students meeting the SAGE Proficient standard for each assessment in spring 2014. In addition, Table 2 shows the approximate percentage of Utah students meeting the associated ACT college ready standard for high school assessments and the percentage of Utah students meeting the NAEP proficient standards at grades 4 and 8. As Table 8 indicates, the performance standards recommended and adopted for the SAGE assessments are quite consistent with relevant ACT college ready and NAEP proficient benchmarks. Moreover, because the performance standards were vertically articulated, the proficiency rates across grade levels are generally consistent.

Table 2. Percentage of Students Meeting SAGE and Benchmark Proficient Standards

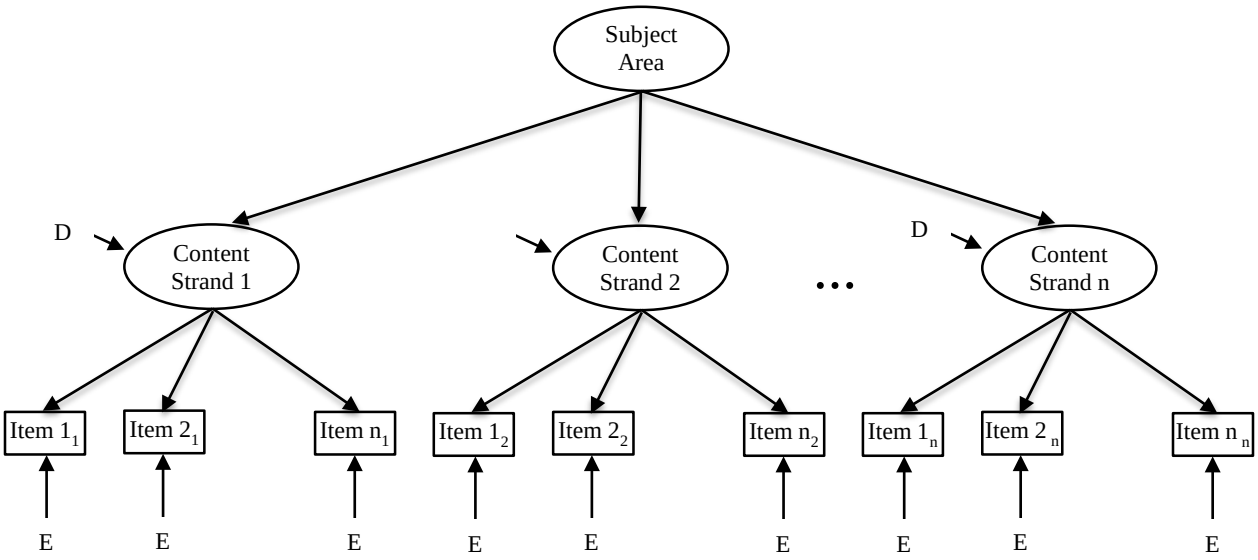
Grade	SAGE Proficient	ACT College Ready	NAEP Proficient
ELA			
3	45		
4	42		37
5	42		
6	42		
7	42		
8	41		39
9	39		
10	40		
11	38	41	
Mathematics			
3	45		
4	48		44
5	44		
6	35		
7	43		
8	38		36

Grade	SAGE Proficient	ACT College Ready	NAEP Proficient
SMI	32	31	
SMII	28	31	
SMIII	33	36	
Science			
4	42		38
5	46		
6	45		
7	42		
8	46		43
Biology	38	30	
Chem	46	39	
ESS	43	20	
Phys	45	48	

4. Evidence Based on Internal Structure

Utah's SAGE assessment represents a structural model of student achievement in grade level and course specific content areas. Within each subject area (e.g., ELA), items are designed to measure a single content strand (e.g., Reading Information, Reading Literature, Language, Writing). Content strands within each subject area are, in turn, indicators of achievement in the subject area. The form of the second-order confirmatory factor analyses is illustrated in Figure 1. As the exhibit illustrates, each item is an indicator of an academic content strand. Because items are never pure indicators of an underlying factor, each item also includes an error component. Similarly, each academic content strand serves as an indicator of achievement in a subject area. As at the item level, the content strands include an error term indicating that the content strands are not pure indicators of overall achievement in the subject area. The paths from the content strands to the items represent the first-order factor loadings, the degree to which items are correlated with the underlying academic content strand construct. Similarly, the paths from subject area achievement to the content strands represent the second-order factor loading, indicating the degree to which academic content strand constructs are correlated with the underlying construct of subject area achievement.

Figure 1: Second-Order Structural Model for SAGE Assessments



Confirmatory factor analysis was used to evaluate the fit of this structural model to student response data from the SAGE test administrations. SAGE assessments in spring 2014 were administered using only the blueprint match component of the adaptive algorithm, since there were as yet no item response theory (IRT) parameter estimates on which to adapt test information to student ability. In the absence of a common test form for all students, we constructed a single form for each grade and subject comprising frequently administered items that met content standard blueprint specifications. This approach was necessary to ensure a well-conditioned covariance matrix to support the analyses.

For each of these test forms, we examined the goodness of fit between the structural model and the operational test data. Goodness of fit is typically indexed by a χ^2 statistic, with good model fit indicated by a non-significant χ^2 statistic. The χ^2 statistic is sensitive to sample size, however, so even well-fitting models will demonstrate highly significant χ^2 statistics given a very large number of students. Therefore, fit indices such as the Comparative Fit Index (CFI; Bentler, 1990), the Tucker-Lewis Index (TLI, Tucker & Lewis, 1973), the Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Residual (SRMR) were also used to evaluate model fit. Table 3 provides a list of the goodness-of-fit statistics used to evaluate model fit, along with a guideline as to what constitutes a good fit.

In addition to testing the fit of the hypothesized SAGE second-order confirmatory factor analysis model, we examined the degree to which the second-order model improved fit over the more general one-factor model of academic achievement in each subject area. Because the second-order model is nested within the one-factor, general achievement model, a simple likelihood ratio test can be used to determine whether the added information provided by the structure of the Utah Core Standards frameworks improves model fit over a general achievement model. Results indicating improved model fit for the second-order factor model provide support for the interpretation of content standard performance above that provided by the overall subject area score. In addition to model fit, information criterion indices can be used to evaluate the gains of model fit relative to increased model complexity. Complex models often improve model fit, but do so by sacrificing

parsimony. Information indices such as Akaike's Information Criteria (AIC), the Bayesian Information Criteria (BIC), and the sample size adjusted Bayesian Information Criteria (aBIC), allow for evaluation of gains in model fit relative to model complexity.

Results of these analyses are presented in Volume 4. Across the SAGE grade level and subject area assessments, as well as the high school end-of-course assessments, the hypothesized second-order factor model fit the observed data well. Because the first-order, general achievement model also fit the data well, it was also important to verify that the fit of the hypothesized second-order model significantly improved fit over the more general achievement model. For ELA, the evidence indicated clearly that the added complexity of the second-order model significantly improved fit over the more general achievement model. For math and science, the evidence was mixed. While the increase in model fit for the second-order model was uniformly statistically significant, information criteria that take into consideration increase in model complexity in the evaluation of improved model fit were mixed, sometimes indicating preference for the second-order model, but sometimes indicating preference for the more general model.

5. Depth of Knowledge

The SAGE assessments also claim to measure subject area achievement using test items that probe student knowledge and skills across multiple depth of knowledge levels. As with the content standards, the alignment of items by depth of knowledge also represents a structural model that can be evaluated using confirmatory factor analysis. In this case, each item is an indicator of a depth of knowledge level first-order factor, and each depth of knowledge level is in turn an indicator of subject area achievement. Thus, confirmatory factor analysis was used to evaluate the fit of this depth of knowledge structural model to student response data from the SAGE test administrations. In the absence of a common test form for all students, we constructed a single form for each grade and subject comprising highly administered items that met content standard blueprint specifications. This approach was necessary to ensure a well-conditioned covariance matrix to support the analyses. We note that there were two assessments in math and one in science for which we were unable to produce an analyzable matrix.

Consistent with the hypothesized reporting model, across the SAGE grade level and subject area assessments, as well as the high school end-of-course assessments, the second-order depth of knowledge factor model fit the observed data well. Across grade level and subject area assessments, the fit of the second-order DOK model significantly improved fit over the more general achievement model. Moreover, across all grades and subject area assessments, evaluation of information criteria indicated that the improvement of fit was substantial enough to warrant increased model complexity, indicating that information about the DOK level of test items adds information beyond the general achievement factor.

6. Measurement Invariance Across Subgroups

The meaning of test scores should be the same for all students, regardless of group membership. Measurement invariance is said to hold when the likelihood of correct responding conforms to the measurement model, independent of group membership; the parameters of the measurement model are statistically equivalent across groups. The parameters of interest in measurement invariance testing include factor loadings and intercepts/thresholds. Invariance in residual variances or scale factors can also be tested, but there is consensus that it is not necessary to demonstrate invariance across groups on these parameters. Examination of measurement invariance can be conducted

using a series of multiple-group confirmatory factor analysis (CFA) models, which impose identical parameters across groups. That is, the models that investigate the invariance of factor pattern (configural invariance), factor loadings (metric or weak invariance), latent intercepts/threshold (scalar or strong invariance), and unique or residual factor variances (strict invariance) are tested across groups in that sequential order. When factor loadings and intercepts/thresholds are invariant across groups, scores on latent variables can be validly compared across the groups and the latent variables can be used in structural models hypothesizing relationships among latent variables.

Because SAGE is adaptively administered and students do not see a common set of items, to investigate measurement invariance across subgroups, we first selected from each assessment pool a set of items with high response rates from each reporting category from 2014–2015 test administration. This strategy was necessary to produce an analyzable covariance matrix containing a sample of items representing the full breadth of the content domain specified by the blueprint. The numbers of items selected varied across tests: 30–33 items across ELA assessments, 31–34 items across math assessments, and 30–37 items across the science assessments.

Results of this investigation are reported in full in Volume 4. Following the sequence of tests of measurement invariance (Millsap & Cham, 2012), we tested configural, metric, and scalar invariance models using $\chi 2$ difference test (at $\alpha \leq 0.05$) and the examination of significant differences of the Root Mean Square of Approximation (RMSEA, change in RMSEA ≤ 0.015; Chen, 2007) between the two nested invariance models. Measurement invariance was investigated across gender, ethnicity (due to small sample sizes, classified as white, Asian, or other ethnic group), special education status, limited English proficiency status, and economically disadvantage status. Invariance tests of subgroups were investigated separately for each grade and subject area test.

The null hypothesis of the $\chi 2$ difference test is that the more restricted invariance model (e.g., metric) fits the data equally well as the less restricted invariance model (e.g., configural). Given the sensitivity of the $\chi 2$ difference test to sample size, we additionally examined significant differences on this test with an examination of the RMSEA. A small change in the RMSEA between the more restricted and less restricted invariance models supports retention of the more restricted invariance model (Chen, 2007). Although the $\chi 2$ difference test should ideally be nonsignificant, all $\chi 2$ difference tests were significant or marginally significant at α=.05 due to large sample sizes. Nevertheless, we found that changes of the RMSEA between the two nested invariance models were very small (ranging from 0.000 to 0.004 for ELA; from 0.000 to 0.002 for math; and from 0.000 and 0.005 for science).

In addition, we evaluated fit indices of scalar invariance model assuming same factor pattern + identical factor loadings + identical latent intercept/threshold across subgroups. Global model fit indices included the Comparative Fit Index (CFI; Bentler, 1990) and Root Mean Square of Approximation (RMSEA). CFI values ≥ 0.90 and RMSEA values ≤ 0.08 were used to evaluate acceptable model fit. The model fit indices of the scalar invariance models for all tests suggested acceptable fit to the data. For ELA, CFI ranged from 0.893 to 0.972 and RMSEA ranged 0.007 to 0.018. For math, excluding the SM II assessment, CFI values ranged from .877 to .957 and RMSEA ranged from 0.009 to 0.019. CFI values for SM II ranged from 0.75 to 0.806 across models, indicating unacceptable fit, although RMSEA values range from .017 to .02, indicating acceptable model fit. For science, CFI values ranged from 0.860 to 0.957 and RMSEA ranged from 0.010 to 0.026.

Based on the similar magnitudes of the RMSEA (i.e., no material change across all tested models; Cheung & Rensvold, 2002) and the acceptable fit indices of the scalar invariance model to the data, SAGE test scores have the same measurement structure across gender, ethnicity (classified as White, Asian, or other ethnic groups), special education status, limited English proficiency status, and economically disadvantage status for each test.

7. Evidence for Student Growth Across Subgroups

The SAGE vertical scale provides educators, parents, and students to monitor achievement gains over time. The vertical scale also allows the possibility of examining whether there are differential patterns of growth across demographic subgroups that could point to systematic differences in instruction or development across subgroups, or limits to the generalizability of test scores for some subgroups. To examine the possibility of differential growth across demographic subgroups, a series of regression analyses were conducted predicting 2015 test scores from 2014 test scores, and controlling for demographic subgroup membership. The demographic variables were gender, ethnicity, special education status (SPEG), limited English proficiency status (LEP), and economically disadvantage status (Low Income). To evaluate effects of ethnicity, students were categorized into one of seven groups; white, African American, Hispanic, Asian, Native Haw or Pacific, American Indian or Alaskan, and Multiple. To evaluate differential growth across ethnic subgroups, we created six dummy variables contrasting white students with each of other ethnic groups (e.g., white/Hispanic, white/African American). With this dummy-variable coding approach, effects can be directly interpreted as differences in the effect of each ethnic group relative to effect of the white reference group. Gender was coded as 0 for male and 1 for female. SPEG was coded as 1 for students with special education status and 0 for non-SPEG students. LEP was coded as 1 for students with limited language learner and 0 for non-LEP students. Low Income was coded as 1 for economically disadvantaged status and 0 for non-Low Income students.

The 2014 test scores were centered on the population mean so that the initial level in the analyses represents the mean performance on the 2014 assessment. The slope represents the association between 2014 test scores and 2015 test scores, controlling for demographic subgroup membership. A positive slope indicates students with higher test scores in 2014 have higher average levels of test scores in 2015 whereas the negative slope indicates students with higher average levels of test scores in 2014 have lower average levels of test scores in 2015. The significance of the effect is less than p-value of .05.

Results of this investigation are fully reported in Volume 4. Although many individual effects attained conventional levels of statistical significance due to very large sample sizes, we focused only on highly significant effects associated with more practically significant effect sizes and effects that might point to trends across grade level and/or subject area assessments. For all grades and subject areas, the higher scores on the 2014 assessment were associated with greater growth. Also generally consistent across grade levels and subject areas, students classified as special education status showed greater rates of gain than other students during the elementary school years, but beginning in the middle school grades the pattern is reversed, with students classified as special education showing lower rates of gain relative to other students. For math especially, but this was also observed for ELA and science, females tend to demonstrate lower rates of achievement gain than males.

Generally, differential gains were not observed across ethnic groups, and observed effects were not consistent across grades or subject areas. In ELA, students classified as Hawaiian/Pacific

Islander showed lower gains than whites between the grade 10 and 11 assessments. In math, Hispanics showed lower rates of gain than whites from the grade 5 to grade 6, and SM I to SM II assessments. African Americans showed lower rates of gain than whites in math between the grade 6 and grade 7 assessments. Hawaiian/Pacific Islander students also showed lower rates of gain between the Earth Science and Biology assessments.

8. Summary

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 principal 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 achieve 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.