Validity of SAGE Test Score Interpretations

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, 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.

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 for students 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. 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 of 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, which 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):

- Inclusive assessment population
- Precisely defined constructs
- Accessible, non-biased items
- Amenable to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum readability and comprehensibility
- 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. Exhibit 1 lists the accommodations and accessibility supports currently available for the SAGE assessments.

Exhibit 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	Student can zoom in and zoom out the entire page. This capability persists throughout

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

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 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 1-2 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.

The following section details the procedures used to develop and review the items comprising the SAGE adaptive item pools.

Review Process for Items Appearing in SAGE Operational Test Administrations

In this section, we describe the item review procedures used to ensure item accuracy and alignment with the Utah Core Standards. Following a standard item review process, item reviews proceed initially through a series of internal AIR reviews before items are eligible for review by USOE content experts. Most of AIR's content staff members, who are responsible for conducting internal reviews, are former classroom teachers who hold degrees in education and/or their respective content areas. Each item passes through four internal review steps before it is eligible for review by USOE. These steps include

- Preliminary review, conducted by a group of AIR content area experts
- Content Review 1, performed by a Level 3-4 AIR content specialist
- Edit, in which a copyeditor checks the item for correct grammar/usage
- Senior Content Review, by the Level 4-5 lead content expert.

At every stage of the item review process, beginning with preliminary review, AIR's test developers analyze each item to ensure that

- The item is well-aligned with the intended content standard
- The item conforms to the item specifications for the target being assessed
- The item is based on a quality idea (i.e. it assesses something worthwhile in a reasonable way);
- The item is properly aligned to a depth of knowledge (DoK) level;
- The vocabulary used in the item is appropriate for the intended grade/age and subject matter, and takes into consideration language accessibility, bias, and sensitivity.
- The item content is accurate and straightforward
- Any accompanying graphic and stimulus materials are actually necessary to answer the question

- The item stem is clear, concise, and succinct, meaning it contains enough information to know what is being asked, is stated positively (and does not rely on negatives such as no, not, none, never, unless absolutely necessary), and it ends with a question
- For selected response items, the set of response options are succinct; parallel in structure, grammar, length, and content; sufficiently distinct from one another; and all plausible, but with only one correct option
- There is no obvious or subtle cluing within the item
- The score points for constructed-response items are clearly defined; and
- For machine-scored constructed-response (MSCR) items, the items score as intended at each score point in the rubric.

Based on their review of each item, the test developer may accept the item and classification as written, revise the item, or reject the item outright.

Items passing through the internal review process are sent to USOE for their review. At this stage, items may be further revised based on any edits or changes requested by USOE, or rejected outright. Items passing through the USOE review level then pass through three external reviews in which committees of Utah educators and stakeholders review each item's accuracy, alignment to the intended standard, and DoK level, as well as item fairness and language sensitivity. Thus, all items considered for inclusion in the SAGE item pools are initially reviewed by

- Utah content advisory committees, which check to ensure that each item is
    - aligned to the intended content standard,
    - appropriate for the grade level,
    - accurate, and
    - presented online in a way that is clear and appropriate.
- Utah fairness and sensitivity committees, which check to ensure that each item and any associated stimulus materials are free from bias, sensitive issues, controversial language, stereotyping, and statements that reflect negatively on race, ethnicity, gender, culture, region, disability, or other social and economic conditions and characteristics;
- Utah community panels also review all test items for appropriateness of test content.

Items successfully passing through this committee review process are then field tested to ensure that the items behave as intended when administered to students. Despite conscientious item development, some items perform differently than expected when administered to students. Using the item statistics gathered in field testing to review item performance is an important step in constructing valid and equivalent operational test forms.

Classical item analyses ensure that items function as intended with respect to the underlying scales. Classical item statistics are designed to evaluate the item difficulty and the relationship of each item to the overall scale (item discrimination) and to identify items that may exhibit a bias across subgroups (differential item functioning analyses).

Items flagged for review based on their statistical performance have to pass a three-stage review to be included in the final item pool from which operational forms are created. In the first stage of this review, a team of psychometricians reviews all flagged items to ensure that the data are

accurate and properly analyzed, response keys are correct and that there are no other obvious problems with the items.

USOE then reconvenes the content review and fairness and sensitivity committees to re-evaluate flagged field-test items in the context of each item's statistical performance. Based on their review of each item's performance, the content review and fairness and sensitivity committees may recommend that a flagged item be rejected or deem the item eligible for inclusion in operational test administrations.

Evidence for Validity and Consequences of Testing

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 standard 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 grade 3-8 summative assessments were provided with the approximate location of relevant 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, Exhibit 2 shows the percentage of students meeting the SAGE proficient standard for each assessment in spring 2014. In addition, Exhibit 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 Exhibit 2 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.

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

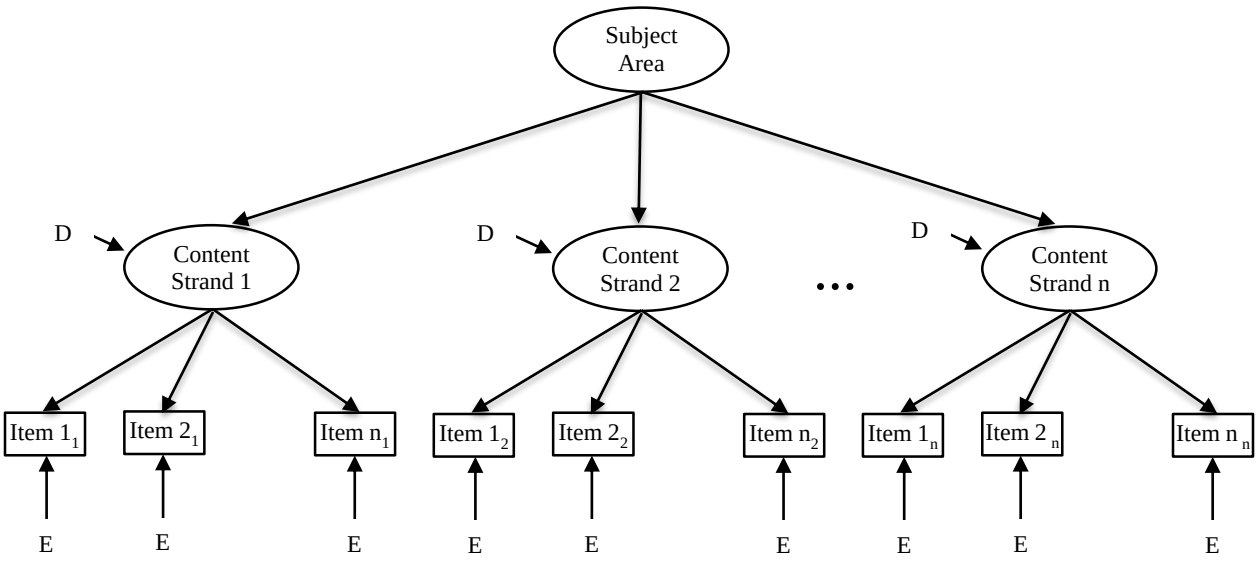
	Percent of Students Meeting Standard		
Grade	**SAGE Proficient**	**ACT College Ready**	**NAEP Proficient**
ELA			
3	45		
4	42		37
5	42		
6	42		
7	42		
8	41		39
9	39		
10	40		
11	38	41	
Mathematics			
3	45		
4	48		44
5	44		
6	35		
7	43		
8	38		36
SMI	32	31	
SMII	28	31	
SMIII	33	36	
Science			
4	42		38

5	46		
6	45		
7	42		
8	46		43
Biology	38	30	
Chemistry	46	39	
Earth and Space Science	43	20	
Physics	45	48	

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 Exhibit 3. 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.

Exhibit 3: 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 highly 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 (Tucker & Lewis, 1973), the Root Mean Square of Approximation (RMSEA), and Standardized Root Mean Residual (SRMR) were also used to evaluate model fit. Exhibit 4 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.

Exhibit 4: Guidelines for Evaluating Goodness of Fit.

Goodness-of-Fit Index	*Indication of Good Fit*
CFI	$\geq .95$
TLI	$\geq .95$
RMSEA	$\leq .05$
SRMR	$\leq .08$

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.

ELA Results

The goodness-of-fit statistics for the hypothesized SAGE second-order models in ELA are shown in Exhibit 5. All of the statistics indicate the second-order models posited by the SAGE assessments fit the data well. This pattern was true across all grades. The CFI and TLI values were all equal to or greater than .95. The RMSEA values are all .01 and SRMR values between .02 and .04, well below the values used to indicate good fit.

Exhibit 5: Goodness-of-Fit for the SAGE ELA Second-Order Model

Grade	*CFI*	*TLI*	*RMSEA*	*SRMR*
	Second-Order Models			
3	0.96	0.96	0.01	0.03
4	0.97	0.97	0.01	0.03
5	0.95	0.95	0.01	0.03
6	0.98	0.97	0.01	0.03
7	0.98	0.98	0.01	0.03
8	0.97	0.97	0.01	0.02
9	0.97	0.97	0.01	0.03
10	0.97	0.97	0.01	0.03
11	0.98	0.98	0.01	0.02

The results of the comparison between the hypothesized SAGE model and the more general achievement model are presented in Exhibit 6. The chi-square difference test shows that across grade levels, the strand-based second-order model showed significantly better fit than the general achievement first-order model. The χ^2_{Diff} *p*-values were less than .001 across all grade levels. In addition, the positive values for the information criteria indicate that the gains in fit for the second-order model justify the increased model complexity.

Exhibit 6: Difference in Fit Between Strand-Based Second-Order and General Achievement First-Order Models

Grade	x^2_{Diff}	Df_{Diff}	*p-value*	AIC_{Diff}	BIC_{Diff}	$aBIC_{Diff}$
	First-Order and Second-Order Models					
3	2850.5	5	0.000	2840.5	2796.7	2812.6
4	3228.7	5	0.000	3218.7	3174.9	3190.8
5	2568.0	5	0.000	2558.0	2514.3	2530.1
6	2846.5	5	0.000	2836.5	2792.8	2808.7
7	1250.8	5	0.000	1240.8	1197.2	1213.1
8	2485.6	5	0.000	2475.6	2432.1	2448.0
9	1325.1	5	0.000	1315.1	1271.8	1287.7
10	5540.0	5	0.000	5530.0	5487.0	5502.8
11	1413.2	5	0.000	1403.2	1360.5	1376.4

Mathematics Results

The goodness-of-fit statistics for the strand-based second-order models are shown in Exhibit 7. The models generally show good fit, although the CFI and TLI fit indices are less than the cutoff value of .95 for some of the higher grade level assessments. Even for these grades, however, the RMSEA and SRMR estimates are well below their respective .05 and .08 cut-off values. All of the statistics indicate the second-order models are a good fit for the data.

Exhibit 7: Goodness-of-Fit for the SAGE Mathematics Second-Order Model

Grade	*CFI*	*TLI*	*RMSEA*	*SRMR*
	Second-Order Models			
3	0.96	0.95	0.01	0.03
4	0.97	0.96	0.01	0.03
5	0.96	0.96	0.01	0.03
6	0.96	0.96	0.01	0.03
7	0.96	0.96	0.01	0.03
8	0.92	0.92	0.02	0.03
SMI	0.93	0.93	0.01	0.04
SMII	0.96	0.96	0.01	0.03
SMIII	0.83	0.82	0.02	0.05

The results of the comparison between the second-order, strand-based model and the first-order, general achievement model are presented in Exhibit 8. The chi-square difference test shows that across grade levels, the hypothesized second-order model provided significantly greater fit relative to the first-order model, with χ^2_{Diff} *p*-values less than .001 across grade levels. The information criteria, however, showed mixed results, indicating that the gains in model fit afforded by the second-order model may be outweighed, at least in part, by the greater complexity of that model relative to the first-order, general achievement model.

Exhibit 8: Difference in Fit Between First-Order Model and Second-Order Model Considering Strands

Grade	x^2_{Diff}	Df_{Diff}	*p-value*	AIC_{Diff}	BIC_{Diff}	$aBIC_{Diff}$
3	31.3	5	0.000	21.3	-22.6	-6.7
4	22.5	5	0.000	12.5	-31.4	-15.5
5	19.0	5	0.002	9.0	-34.7	-18.8
6	82.7	5	0.000	72.7	29.1	44.9
7	19.5	5	0.002	9.5	-33.9	-18.0
8	20.4	5	0.001	10.4	-33.0	-17.1
SMI	16.2	5	0.006	6.2	-37.3	-21.5
SMII	14.7	5	0.012	4.7	-37.9	-22.0
SMIII	34.7	5	0.000	24.7	-14.0	1.9

Science Results

The goodness-of-fit statistics for the strand-based, second-order model for the SAGE science assessments are shown in Exhibit 9. The statistics indicate good fit of the second-order models to the operational test data. The CFI and TLI values are all greater than .95, except at grade 6 which yielded fit indices below .90 in both models. The RMSEA values are less than .03 and the SRMR values less than .03 for all grade level assessments, well below the values used to indicate good model fit to the data.

Exhibit 9**: Goodness-of-Fit for the SAGE Science Second-Order Models**

Grade	*CFI*	*TLI*	*RMSEA*	*SRMR*
	Second-Order Models			
4	0.97	0.96	0.01	0.02
5	0.96	0.96	0.01	0.02
6	0.89	0.88	0.03	0.03
7	0.96	0.95	0.02	0.02
8	0.95	0.95	0.01	0.02
Bio	0.97	0.97	0.01	0.02
Chem	0.96	0.96	0.02	0.02
ESS	0.96	0.95	0.01	0.02
Phy	0.95	0.95	0.01	0.03

The results of the comparisons between the hypothesized second-order, strand-based model and first-order, general achievement model are presented in Exhibit 10. Results examining differences between the models are mixed. The second-order models generally show greater fit than general achievement, first-order model, but the difference is not statistically significant for the grade 5 and Biology assessments. The information criteria also show mixed results, with the hypothesized second-order model providing consistently preferable in some grades, but showing only mixed evidence for other grades and courses.

Exhibit 10**: Difference in Fit Between First-Order Model and Second-Order Model Considering Strands**

Grade	x^2_{Diff}	Df_{Diff}	*p-value*	AIC_{Diff}	BIC_{Diff}	$aBIC_{Diff}$
4	32.4	5	0.000	22.4	-21.4	-5.5
5	6.8	5	0.233	-3.2	-46.9	-31.0
6	161.7	6	0.000	149.7	97.3	116.4
7	94.9	5	0.000	84.9	41.8	57.7
8	112.2	4	0.000	104.2	69.4	82.1
Bio	10.4	5	0.065	0.4	-43.1	-27.3
Chem	22.2	6	0.001	10.2	-38.2	-19.2
ES	20.1	5	0.001	10.0	-30.8	-15.0
Phy	12.2	5	0.032	2.2	-36.5	-20.6

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 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.