WebbAlign

WCEPS
WISCONSIN CENTER for EDUCATION PRODUCTS & SERVICES

Alignment Analysis of Two Forms of the SAT with the Arizona Academic Standards for English Language Arts Grades 11-12, Algebra 1, and Geometry

Sara C. Christopherson and Norman L. Webb
November 25, 2020

Wisconsin Center for Educational Products and Services
Matt Messinger, Executive Director
University of Wisconsin Research Park
510 Charmany Drive, Suite 269
Madison, WI 53719

Acknowledgements

English Language Arts:

External Panelists

Cindy Jacobson	Group Leader	Wisconsin
Kimberly Hernandez		Georgia

Arizona Panelists

Kristine Cunningham	Phoenix, AZ
Melinda Escarcega	Tombstone, AZ
Marlena Sypel	Scottsdale, AZ

Mathematics:

External Panelists

Lisa Ashe	Group Leader	North Carolina
Linda Hall		Washington, D.C.

Arizona Panelists

Andrea Dales	Chandler, AZ
Anna Van Zile	Morenci, AZ
Richard Sypel	Scottsdale, AZ

This independent third-party analysis was conducted for and funded by the Arizona Department of Education. Audra Ahumada, Deputy Associate Superintendent of Assessment, was the main contact for Arizona. Margaret Bowerman and Anju Kuriakose also provided input and support for study planning and execution. Sherral Miller, Janet Swandol, Jim Patterson, and Katina Marshall of the College Board provided test forms for the study.

Table of Contents

Executive Summary

This report describes a content alignment analysis conducted during the month of November 2020 to provide information about the degree of alignment of the SAT with the Arizona Academic Standards for English Language Arts (ELA) Grade 11-12, Algebra 1, and Geometry. The analysis was conducted to provide evidence about the degree of alignment of the SAT with the corresponding Arizona standards, as pertains to fulfilling requirements as stated in Federal statute. The Every Student Succeeds Act (ESSA, 2015) provides states the flexibility to use a locally selected, nationally recognized high school academic assessment in lieu of the statewide summative assessment, provided the assessment meets certain technical criteria, including that it is aligned to and addresses the depth and breadth of the state's academic content standards.

Pursuant to Arizona statute, the State Board of Education maintains a Menu of Assessments for high school testing that includes nationally recognized high school assessments that meet policy requirements. These tests are intended to be used to measure student achievement of Arizona Academic Standards. The SAT is planned for inclusion in the Menu of Assessments, starting in 2020-2021 (Arizona State Board of Education, 2020).

Arizona students take the SAT in spring of grade 11. Alignment of the SAT ELA/literacy portions are therefore considered in relation to the grade 11-12 ELA standards. In contrast to ELA, mathematics courses are taken in different grades by different students; there is no universal grade 11 set of mathematics standards. Because mathematics is course-based, multiple factors must be taken into account to make a decision about the appropriate set of high school mathematics standards for use in a state's accountability system. According to the Arizona Department of Education (ADE), at least 90% of students in grade 11 have completed both Algebra 1 and Geometry mathematics courses. Further, prior State Board work with educator panels resulted in a consensus that Algebra 1 and Geometry standards should be prioritized in a high school summative assessment used for state accountability purposes (A. Ahumada, personal communication, October 26, 2020). Taking these key considerations into account, ADE selected the Algebra 1 and Geometry standards as the appropriate referents for alignment. The College Board identified the Arizona Algebra 1 and Geometry standards as corresponding to the SAT as well as standards from two additional courses: Algebra 2 and Quantitative Reasoning (College Board, 2020).

A two-day remote alignment institute took place on November 6-7, 2020 via Zoom video conferencing to analyze the agreement between the Arizona Academic Standards and two forms of the SAT. Three Arizona educators and two external reviewers (i.e. reviewers from other states) participated in each subject-area panel (ELA/literacy and mathematics). All panelists were selected because of their notable high school education experience and content expertise.

The study was designed to answer two main research questions:

1. What is the degree of alignment of the SAT Evidence-Based Reading and Writing section (Reading test + Writing and Language test) and Essay with the corresponding Arizona Academic Standards for grades 11-12 English Language Arts (ELA) with regards to satisfying the federal requirements within the Every Student Succeeds Act (ESSA): that a locally selected, nationally recognized high school academic assessment be aligned with state academic content standards, and address the depth and breadth of the standards?
2. What is the degree of alignment of the SAT Math test with the corresponding Arizona Academic Standards for Algebra 1 and Geometry with regards to satisfying the federal requirements within the Every Student Succeeds Act (ESSA): that a locally selected, nationally recognized high school academic assessment be aligned with state academic content standards, and address the depth and breadth of the standards?

Four alignment criteria received major attention:

- **Categorical Concurrence** between standards and assessment is met if the same or consistent categories of content appear in both documents.
- **Depth of Knowledge Consistency** between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.
- **Range of Knowledge Correspondence** is used to judge whether a comparable span of knowledge expected of students by a reporting category (domain/strand) is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.
- **Balance of Representation** is used to indicate the degree to which one content indicator (standard) is given more emphasis on the assessment than another.

The ELA/literacy portions of both test forms analyzed were considered acceptably aligned with the Arizona Academic Standards for ELA grades 11-12. Each test form would need only one item revised or replaced to fully meet the typically accepted minimum cutoffs for full alignment. If considering the full set of Algebra 1 and Geometry standards, spanning two years of coursework, then both test forms analyzed would need major adjustments to meet typically accepted alignment criteria, with approximately 24 items added, replaced, and/or revised per test form. Over half of these adjustments are required to attend to the breadth (Range of Knowledge) of the Geometry reporting category. Alignment of statewide summative assessments are typically considered in relation to one year of coursework, and not two years of coursework. If considering the alignment of the mathematics portion of the test forms with Algebra I standards only, then the test forms would be considered to need slight adjustments to meet typically accepted alignment criteria, with approximately 10 items added, replaced, and/or revised per test form.

While augmenting the SAT to attain an acceptable level of alignment is certainly possible, it should be noted that augmentation tends to be a rather expensive process and adds complexity to the administration of the tests, because items used to augment a test need to be administered separately from the college entrance test.

Introduction and Methodology

The alignment of expectations for student learning with assessments for measuring students' attainment of these expectations is an essential attribute for an effective standards-based education system. Alignment is defined as the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide an education system toward students learning what they are expected to know and do. As such, alignment is a quality of the relationship between expectations and assessments and not an attribute solely of either of these two system components. Alignment describes the match between expectations and an assessment that can be legitimately improved by changing either student expectations or the assessments. As a relationship between two or more system components, alignment is determined by using the multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997). These alignment study procedures and criteria (developed through NISE, funded by the National Science Foundation, and in cooperation with the Council of Chief State School Officers) influenced the specification of alignment criteria by the U.S. Department of Education. The Webb alignment process has been used to analyze curriculum standards and assessments in at least 30 states to satisfy or to prepare to satisfy Title I compliance as required by the United States Department of Education (USED). The corresponding methodology used to evaluate alignment has been refined and improved over the last 20 years, yielding a flexible, effective, and efficient analytical approach.

A content alignment analysis in the areas of English language arts (ELA) and mathematics was conducted during the month of November 2020 to provide information that could be used to judge the degree to which the two forms of the SAT that were used in the analysis were aligned with the Arizona Academic Standards for Grades 11-12 ELA, Algebra 1, and Geometry. As such, the study focused on the degree to which the two SAT test forms provided addressed the full depth and breadth of these state standards. The College Board provided two forms for analysis, identified as Form 07 and Form 10. These sample forms were selected from available inventory, and were typical forms, representative of SAT test forms administered in Arizona (J. Patterson, personal communication, November 11, 2020).

The alignment analysis detailed in this report was completed through the WebbAlign program, which works directly with Dr. Norman Webb, and operates out of the Wisconsin Center for Education Products and Services (WCEPS), a non-profit organization that strives to extend the reach of innovations developed at the University of Wisconsin-Madison's Wisconsin Center for Education Research (WCER). Sara Christopherson, Director of WebbAlign, led the study.

The study was designed to answer two main research questions:

1. What is the degree of alignment of the SAT Evidence-Based Reading and Writing section (Reading test + Writing and Language test) and Essay with the corresponding Arizona Academic Standards for grades 11-12 English Language Arts (ELA) with regards to satisfying the federal requirements within the Every Student Succeeds Act (ESSA): that a locally selected, nationally recognized high school academic assessment be aligned with state academic content standards, and address the depth and breadth of the standards?
2. What is the degree of alignment of the SAT Math test with the corresponding Arizona Academic Standards for Algebra 1 and Geometry with regards to satisfying the federal requirements within the Every Student Succeeds Act (ESSA): that a locally selected, nationally recognized high school academic assessment be aligned with state academic content standards, and address the depth and breadth of the standards?

The remote content alignment institute took place on November 6-7, 2020 via Zoom video conferencing. Three Arizona educators and two external reviewers participated in each subject-area panel (ELA/literacy and mathematics). One of the external reviewers served as group leader for each panel. One of the five reviewers on each panel had previously participated in one WebbAlign alignment study for Arizona. Three of the five reviewers on each panel had previously participated in multiple WebbAlign alignment studies. One of the Arizona educators on each panel was new to the WebbAlign alignment process but experienced with ADE committee work related to standards and assessments. Representation of a diversity of populations, including race/ethnicity, socioeconomic, and regional factors was considered in panelist selection. Experience with multilingual learners and with special education was also taken into account in panelist selection.

The Version 2 of the Web Alignment Tool (WATv2) was used to enter all of the content analysis codes during the institute. The WATv2 is a web-based tool connected to the server at the Wisconsin Center for Education Research (WCER). It was designed to be used with the Webb process for analyzing the alignment between assessments and standards. Prior to the institute, a group was registered on the WATv2 for each of the two panels. Each panel was assigned a group identification number and the group leader was designated. Then the reporting categories and standards were entered into the WATv2 along with the information for each assessment, including the number of items/tasks, the weight (point value) given to each item/task, and any additional comments that could help panelists find the correct item/task. A sequential account of the alignment study procedures is provided below.

Training and Coding

In the morning of the first day of the remote alignment study, all panelists attended a launch meeting via Zoom. The launch meeting included an overview of the purpose of the work and the steps of the coding processes. During the meeting, panelists participated in activities to calibrate understanding and application of the Depth of Knowledge (DOK) definitions used to describe content complexity. All reviewers had some experience with the DOK language system prior to the institute and most reviewers had extensive experience with DOK. The general training at the alignment institute reviewed the origins of DOK (to inform alignment studies of standards and assessments) and purpose (to differentiate between and among degrees of complexity), and highlighted common misinterpretations and misconceptions to help reviewers better understand and, therefore, consistently apply the DOK language system. The groups then separated into different virtual Zoom "rooms" to conduct more detailed and interactive practice with the DOK levels for each content area. A necessary outcome of training is for panelists to have a common, calibrated understanding of the DOK language system for describing categories of complexity. Definitions for each DOK level for ELA and mathematics are included in **Appendix E**.

Reviewers then continued to calibrate their use of DOK as they evaluated the complexity of a subset of the standards, first assigning DOK individually and then participating in a consensus discussion. After completing coding and discussion of the subset, the panelists reviewed previously assigned DOK levels for the standards. DOK levels had been assigned for all standards via independent coding followed by consensus discussion during a 2017 alignment study conducted for ADE. (No substantive changes to the high school Arizona Academic Standards used in this analysis occurred between 2017 and 2020.) Group leaders then facilitated discussions about any standards for which panelists wanted to clarify the intended complexity. The standards analysis is a necessary component of the alignment study but also, importantly, fosters thorough, nuanced, and calibrated understanding of the standards by panelists. Consensus DOK levels were then entered into the online data collection system, the WATv2. The consensus DOK values for all standards are given in **Appendix A** for each subject. Additional detail about the subject-area standards discussions is provided within this report.

After thoroughly discussing the standards and coming to consensus on the intended complexity of each standard, panelists then conducted individual analyses of three to five assessment items from the first SAT test form analyzed for each group. For each item, panelists worked individually to assign a DOK level to the item and then to code each item to the standard(s) that they judged the item measured, i.e. what students needed to know or do in order to successfully respond to the question. After completing individual coding of the subset of items, the full panel engaged in adjudication discussions to ensure that

all reviewers understood the processes. Reviewers then continued work, independently evaluating each assessment item. Up to three standards could be coded as corresponding to each item.

Reviewers used the following materials to conduct their analysis:

- **Coding Instructions** (a reference with all information about the coding process, provided via email);
- **WATv2** (the online data entry system);
- **Two SAT test forms + student guides** (provided via email);
- **Arizona Academic Standards** (provided via email);
- **DOK definitions for each subject area** (provided via email);
- **Slides from the launch meeting** (provided via email)

Following individual analyses of the items, reviewers participated in a debriefing discussion in which they analyzed the degree to which they had coded particular items or types of content to the standards. This overall coding process was repeated for each section on each test form to maintain calibration within each group of reviewers.

Reviewers were instructed to focus primarily on the alignment between the Arizona standards and the assessment items on the SAT test forms. However, reviewers were encouraged to offer their opinions on the standards or on the assessment tasks by writing a note about the item in the appropriate text box in the WATv2 data collection tool. Reviewers were instructed to enter a note into the WATv2 for an assessment item if the item only corresponded to a part of a standard and not the full standard. Thus, the reviewers' notes can be used to reveal if assessment items only targeted a part of the individual standards. Reviewers also could indicate whether there was a Source of Challenge issue with an item—i.e. a technical or content problem with the item that might cause the student who knows the material to give a wrong answer or enable someone who does not have the knowledge being tested to answer the item correctly. No technical issues with items were identified.

Reviewers engaged in adjudication of their results after completing the coding of each test form. After all reviewers completed coding an assessment form, the study director and group leader identified the assessment items that did not have a majority of reviewers in agreement on DOK or where the reviewers differed significantly on the DOK assigned (e.g. three different DOK values were assigned). When these substantial differences in DOK occur, it sometimes indicates a data entry error. If data are entered as intended, then it suggests that reviewers are either interpreting the DOK definitions in very different ways or are interpreting the particular assessment item in very different ways.

Reviewers also discussed items for which there were great differences in coding to a standard. The adjudication process helped panelists identify and correct any errors in coding (e.g. accidentally assigning an item to a standard that they did not intend to assign). Adjudication also helped panelists build familiarity with the standards (e.g. a reviewer might not have noticed that a particular expectation is explicit in one of the standards) as well as build common interpretation of the standards (e.g. panelists may calibrate their understanding of the meaning of certain standards that may be interpreted in different ways because of ambiguous wording or differences in the way people understand the content). Adjudication additionally helped reveal differences in interpretation of assessment items and helped reviewers to build a common understanding of exactly what knowledge, abilities, and skills particular items were assessing. Overall, adjudication is intended to foster full and appropriate interpretation of the assessment items and standards, and to ensure that panelists have coded the items as they intended. Reviewers were not required to change their results after the discussion. Reviewer agreement statistics were computed after adjudication and are included in the Findings section of this report.

Reviewers were instructed to consider the full set of grade-level/course expectations when mapping an assessment item to one (or more) standard(s). Panelists were instructed to select only the standard(s) that were necessary and sufficient for a correct response. In most cases, the expectations within a single standard were considered necessary and sufficient as related to the demands of an item. In several cases, however, panelists determined that a correct response required students to draw on the expectations from more than one standard. In other cases, reviewers thought that a single standard was sufficient, but could make reasonable arguments for coding an item to different standards. For example, ELA reviewers noted that there was some overlap between standard W.5 within the Writing domain and standards L.1, L.2, and L.3 within the Language domain because W.5 specifically expects application of these three language standards. Input from Arizona educators helped the group to decide on decision rules for assigning these standards to items. Similarly, Language standard L.4 expects students to determine the meaning of unknown words, which overlaps with expectations in Reading: RL.4 (literature) and RI.4 (informational text) to determine the meaning of words based on context. Discussion helped the panel to differentiate between and among these expectations.

If reviewers map an item to a variety of standards, it also may indicate that the assessment task is not a close fit for any standard. Reviewers may have difficulty finding where an item best fits when an assessment is coded to a set of standards that were not used in developing the assessment, as for this study. If no particular grade-level standard was targeted by a given assessment item, then the reviewers were instructed to code the item to a standard where there

was a partial, but reasonable, fit or to a conceptual category level: the domain or reporting category level. Coding to the level of a conceptual category is referred to as coding to a "generic" standard. This coding to a "generic standard" sometimes indicates that the item was inappropriate for a particular grade level or course (for example, the item might better match a standard from another grade level or course). If the item was grade-appropriate but a corresponding standard was not found, a generic coding may indicate that the item is targeting knowledge within the standards that is being interpreted differently by different parties. It is anticipated that some items on a nationally recognized college-readiness assessment test form may intentionally address assessment targets outside of the specific grade-level/course state standards.

All panelists signed non-disclosure agreements (NDAs) provided by the College Board in advance of the study and confirmed double-deletion of all test form files from both email and computer folders upon study completion.

Data Analysis

To derive the results from the analysis, the reviewers' responses were averaged. First, the value for each of the four alignment criteria (described in the next section) was computed for each individual reviewer. Then the final reported value for each criterion was found by averaging the values across all reviewers. Any variance among reviewers was considered legitimate, for example, with the reported DOK level for an item falling somewhere between the two or more assigned values. Such variation could signify differences in interpretation of an item or of the assessed content and/or a DOK that falls in between two of the four defined levels. Any large variations among reviewers in the final results represented true differences in opinion among the reviewers and were not because of coding error. These differences could be because of different standards targeting the same content knowledge or may be because an item did not explicitly correspond to any standard, but it could be inferred to relate to more than one standard. Standard deviations are reported in the tables provided in **Appendix B**, which give one indication of the variance among reviewers.

The results for each content area produced from the institute pertain only to the issue of alignment between the Arizona academic standards and the two test forms that were analyzed. Note that an alignment analysis of this nature does not serve as external verification of the general quality of the standards or assessments. Rather, only the degree of alignment is discussed in the results. For these results, the means of the reviewers' coding were used to determine whether the alignment criteria were met.

Alignment Criteria Used for This Analysis

This report describes the results of an alignment study of ELA/literacy and mathematics portions of two SAT test forms with the Arizona Academic Standards for ELA Grade 11-12, Algebra 1, and Geometry. The study addressed specific criteria related to the agreement between the expectations within the standards and the demands of the items within the assessments. Four criteria, summarized in **Table 1**, received major attention.

Table 1. Criteria used to evaluate content alignment of the SAT test forms with corresponding Arizona Academic Standards for ELA and mathematics

Criterion	Description of Criterion	Typically Used Cutoffs for Acceptable Alignment*
Categorical Concurrence	Test forms have the potential to yield sufficient evidence to make inferences about student proficiency as relates to each reporting category. The criterion of Categorical Concurrence between reporting categories and assessments is met if the same or consistent categories of content appear in both documents.	A test form has at least six items measuring content from a reporting category.
DOK Consistency	The assessment elicits work that is as cognitively demanding as the expectations in the corresponding assessment targets.	At least 50% of the assessment items corresponding to standards within a reporting category are at (or above, although not common) the Depth of Knowledge level of the corresponding standards.
Range of Knowledge	A comparable span of knowledge expected of students by a reporting category is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.	A least 50% of the standards for a reporting category are addressed by at least one related assessment item.
Balance of Representation	A single assessed standard should not be overrepresented on a test form. This criterion is used to indicate the degree to which one standard is given more emphasis on the assessment than another.	An index value of .7 or higher is obtained, based on the difference in the proportion of standards addressed by items and the proportion of items corresponding to a standard.

Details on the criteria and indices used for determining the degree of alignment between standards and assessments are provided below. For each alignment criterion, an acceptable level was defined by what would be required to ensure that a student had reasonably met the expectations within the reporting categories for each discipline. In the descriptions below, the words “domain” and “reporting category” are used to describe reporting levels.

Reporting Categories and Standards:
Study results are reported according to the reporting categories for each subject area as bulleted below. Consensus DOK values for all standards are given in **Appendix A** for each subject.

In this analysis, the reporting categories for **grade 11-12 ELA** were:

- Reading Standards for Literature (RL)
- Reading Standards for Informational Text (RI)
- Writing (W)
- Language (L)

Total number of standards: 29

The reporting categories for **Algebra 1** and **Geometry** were:

- Number and Quantity (N)
- Algebra (A)
- Functions (F)
- Statistics and Probability (S)
- Geometry (G)

Total number of standards: 82

Because the College Board identifies Arizona standards from two other courses as corresponding to the SAT, an additional “generic” category was included for items that addressed material typically taught in courses following Algebra 1 and Geometry. A parallel category was included for items that addressed material typically taught before students take Algebra 1 or Geometry. These two generic categories were:

- Below-Grade Mathematics
- Advanced Mathematics

In the descriptions on the following pages, the term “standards” may be used as an umbrella term, to refer to expectations in general. In addition to judging alignment between reporting categories and assessments on the basis of the four key alignment criteria, reviewers had the opportunity to provide narrative feedback on all items.

Categorical Concurrence

An important aspect of alignment between academic content standards and assessments is whether both address the same content categories. The Categorical Concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. *The criterion of Categorical Concurrence between content standards and assessments is met if the same or consistent categories of content appear in both documents.* This criterion was judged by determining whether the assessment included items measuring content, as explicated in the standards, from each reporting category. The analysis assumed that the assessment had to have at least six items (or points for polytomous items) for measuring content from a reporting category for a minimum acceptable level of Categorical Concurrence to exist between the domain and the assessment. The number of items/points, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale. Of course, many factors must be considered in determining what a reasonable number is, including the reliability of the subscale, the mean score, and cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and that the reliability of one item is 0.1, it was estimated that six items would produce an agreement coefficient of at least 0.63. This indicates that about 63% of the group would be consistently determined to be masters or non-masters if two equivalent test administrations were employed. The agreement coefficient would increase to 0.77 if the cutoff score is increased to one standard deviation from the mean and, with a cutoff score of 1.5 standard deviations from the mean, to 0.90.

Usually, states do not report student results by domains or require students to achieve a specified cutoff score on expectations related to a domain. If a state did do this, then the state would seek a higher agreement coefficient than 0.63. Six items were assumed as a minimum for an assessment measuring content knowledge related to a reporting category, and as a basis for making some decisions about students' knowledge of that content under the reporting category. If the mean for six items is 3.0 points and one standard deviation is equal to a one-point item, then a cutoff score set at 4.0 points would produce an agreement coefficient of 0.77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement on the subscale.

Depth of Knowledge Consistency

Standards and assessments can be aligned not only on the category of content covered by each, but also on the basis of the complexity of knowledge required by each. *Depth of Knowledge Consistency between standards and an assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.* For consistency to exist between the assessment and the reporting categories, as judged in this analysis, at least 50% of the items corresponding to a reporting category had to be at or above the Depth of Knowledge level of the corresponding content expectation. The 50% level, a conservative minimum cutoff point, is based on the assumption that a minimal passing score for any one reporting category of 50% or higher would require the student to successfully answer at least some items at or above the Depth of Knowledge level of the content expectations within the corresponding reporting categories. For example, assume an assessment included six items related to one domain and students were required to answer correctly four of those items to be judged proficient—i.e. 67% of the items. If three, 50%, of the six items were at or above the Depth of Knowledge level of the corresponding expectations, then for a student to achieve a proficient score would require the student to answer correctly at least one item at or above the Depth of Knowledge level of one expectation. If a domain had between 40% and 50% of items at or above the Depth of Knowledge levels of the expectations, then it was reported that the criterion was "weakly" met.

DOK Levels 1-4

Interpreting and assigning Depth of Knowledge levels to both standards and assessment items is an essential requirement of alignment analysis. The DOK descriptions help to clarify what the different levels represent for each subject area. Full descriptions for reading and mathematics are included in **Appendix E**.

Range of Knowledge Correspondence

For reporting categories and assessments to be aligned, the breadth of knowledge required on both should be comparable. *The Range of Knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a reporting category is the same as, or corresponds to, the span of knowledge that students need to correctly answer the assessment items/activities.* The criterion for correspondence between span of knowledge for a reporting category and an assessment considers the number of standards within the reporting category with one related assessment item/activity. Fifty percent of the standards for a reporting category must have at least one related assessment item for the alignment on this criterion to be judged acceptable. This level is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a reporting category. This assumes that each expectation for a reporting category should be given equal

weight. Depending on the balance in the distribution of items and the need to have a low number of items related to any one expectation, the requirement that assessment items need to be related to more than 50% of the expectations for a reporting category increases the likelihood that students will have to demonstrate knowledge on more than one expectation per reporting category to achieve a minimal passing score. As with the other criteria, a state may choose to make the acceptable level on this criterion more rigorous by requiring an assessment to include items related to a greater number of the expectations. However, any restriction on the number of items included on the test will place an upper limit on the number of expectations that can be assessed. Range of Knowledge correspondence is more difficult to attain if the content expectations are partitioned among a greater number of reporting categories and if there are a large number of expectations. If 50% or more of the objectives for a reporting category had a corresponding assessment item, then the Range of Knowledge correspondence criterion was met. If between 40% and 50% of the objectives for a reporting category had a corresponding assessment item, the criterion was "weakly" met.

Balance of Representation

In addition to comparable depth and breadth of knowledge, aligned reporting categories and assessments require that knowledge be distributed equally or proportionally in both. The Range of Knowledge criterion only considers the number of expectations with at least one assessment item within a reporting category; it does not take into consideration how the assessment items/activities are distributed among these expectations. *The Balance of Representation criterion is used to indicate the degree to which one standard is given more emphasis on the assessment than another.* An index is used to judge the distribution of assessment items. This index only considers the expectations for a reporting category that has at least one related assessment item per expectation. The index is computed by considering the difference in the proportion of expectations and the proportion of items assigned to the expectation. An index value of 1 signifies perfect balance and is obtained if the corresponding items related to a reporting category are equally distributed among the expectations for the given reporting category. Index values that approach 0.0 signify that a large proportion of the items assess only one or two of all of the expectations that were measured. Depending on the number of expectations and the number of items, a unimodal distribution (most items related to one expectation and only one item related to each of the remaining expectations) has an index value of less than 0.5. A bimodal distribution has an index value of around 0.55 or 0.6. Index values of 0.7 or higher indicate that items/activities are distributed among all of the expectations at least to some degree (e.g. nearly every expectation has at least two items) and is used as the acceptable level on this criterion. Index values between 0.6 and 0.7 indicate the Balance of Representation criterion has only been "weakly" met.

Source of Challenge

The Source of Challenge criterion is used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted language reporting category or expectation (i.e. construct irrelevance). Bias and sensitivity issues, as well as technical issues and error, could all be reasons for an item to have a Source of Challenge problem. Such item characteristics may result in some students not answering an assessment item or answering an assessment item incorrectly even though they possess the understanding and skills being assessed.

Cutoffs for Alignment Criteria

For overall alignment, an assessment form is reported as fully aligned if no items need replacement to meet the conditions for all of the criteria described above. Note that "fully aligned" refers to the condition of meeting the minimum acceptable levels of alignment and does not mean that an assessment has "100% alignment" with the corresponding standards. A test form is considered acceptably aligned if it needs between one and five items replaced or revised in order to meet the minimum acceptable conditions for all alignment criteria. A test form is reported to need slight adjustments if six to ten items need to be replaced or revised to meet the minimum levels of alignment criteria and is reported to need major adjustments if more than ten items need to be replaced or revised. These categories represent typically used cutoff levels in the context of submission to federal peer review.

Findings: ELA/Literacy

The ELA/literacy portion of the SAT college and career readiness assessment consists of a two-part Evidence-Based Reading and Writing section and an Essay direct-writing task. The Evidence-Based Reading and Writing section includes a Reading Test and a Writing and Language Test. A student's total score on the SAT is the sum of two section scores: Math and Evidence-Based Reading and Writing. To obtain the Evidence-Based Reading and Writing section score, the SAT Reading and Writing and Language Test scores (10-40) are summed and multiplied by 10 (200 to 800). The SAT Essay is scored separately using a 4-point rubric scored by two raters. Scores are reported for each of the three rubric dimensions. Section scores are also reported in terms of how they relate to grade-level college and career readiness benchmarks. Select items from the SAT Reading and Writing and Language Test contribute to the four sub-scores: 1) Command of Evidence, 2) Expression of Ideas, 3) Words in Context, and 4) Standard English Conventions.

A summary of item counts and allotted assessment times is shown in **Table 2**. The ELA/literacy portion of the SAT includes 96 multiple-choice items and one essay given over 150 minutes, total. There were no field test items on the ELA/literacy portions of the SAT test forms and no items were excluded from the analysis. On all test forms, all items except for the writing prompt were weighted as one (1) point. The SAT essay was weighted at 24 points, reflective of three scores, corresponding to the three-part rubric, each scored on a scale of 2-8 (resulting from summing the two raters' scores on each dimension).

Table 2. SAT item counts, types, and session times – ELA/literacy

ELA/literacy Portions of the SAT: Test Sections & Time	Total Number of Items	Item Type (across sections)
Reading 65min Writing/Language 35min Essay 50min Total 150min	52 items 44 items Total: 96 items One (1) essay	96 multiple-choice (four choices) One (1) writing sample

Source: The College Board, 2020

The passages used within the SAT test forms meet specific criteria for text complexity, ranging from grades 9-10 to postsecondary entry (College Board, 2015).

Standards

A summary of the levels of complexity within the Arizona Academic Standards for grade 11-12 ELA is given in **Table 3**. Only one of the standards included in the study (3%) was considered DOK 1. This expectation targeted command of conventions of Standard English. Seven standards (24%) were considered a DOK level 2, emphasizing work that involves both comprehension and subsequent processing of text, making basic inferences from text and using specific information from text to explain events and ideas, as well as purposeful application of language knowledge and skills. The majority of standards, 21 standards (72%), were DOK 3, emphasizing expectations for deep analysis of text and abstract thinking, including making holistic inferences based on text, and engaging in critical reading to consider aspects of author's purpose and use of textual features. These DOK 3 standards also included expectations for argumentative, explanatory, and narrative writing as well as a variety of aspects of the research process and subsequent communication of findings. No standards included in the study were considered DOK 4. A DOK 4 expectation is one that is at least as complex as a DOK 3 but also requires extended time—days, weeks, or months—to complete.

Standard 10 from the Reading Literature, Reading Informational Text, and Writing Domains were not included in the study because these standards describe expectations for ongoing reading and writing over the course of the year, and are not appropriately assessed on a single on-demand assessment. Writing standard 6 was not included because it expects collaborative use of technology and is not appropriate for assessment on an on-demand assessment intended to be completed individually.

Table 3. Expectations by Depth of Knowledge (DOK) levels for grade 11-12 ELA Arizona Academic Standards, November 2020

AZ Grade 11-12 ELA Reporting Categories	Total Number of Expectations	DOK Level	Number of Standards by Level	Percentage within RC by Level
Reading Standards for Literature (RL)	8	2 3	1 7	12.5 87.5
Reading Standards for Informational Text (RI)	9	2 3	1 8	11.1 88.9
Writing (W)	6	3	6	100.0
Language (L)	6	1 2	1 5	16.7 83.3
Total	29	1 2 3	1 7 21	3 24 72

Mapping of Items by Standards
Panelists reviewed two test forms, identified as Form 07 and Form 10. For both test forms, reviewers found that all of the test items reasonably addressed specific grade 11-12 ELA standards. There were no multiple-choice items or essay components on either SAT test form that any reviewers coded to a generic standard.

The two SAT test forms targeted similar percentages of the total Arizona Academic Standards for grade 11-12 ELA (see **Table 4**). Averaging across the two SAT test forms, the forms were found to include items that addressed around 67% of the grade 11-12 ELA standards.

Table 4. Number and percentage of Arizona Academic Standards for grade 11-12 ELA with at least one corresponding item found by a majority of reviewers

Assessment	Number of Items (including writing prompt)	Number of Standards Targeted	Percentage of Standards with at least One Corresponding Assessment Item
SAT Form 07	97	19	66%
SAT Form 10	97	20	69%

Alignment Statistics and Findings for Two SAT Test Forms and Arizona Standards for Grade 11-12 ELA
Overall alignment results are summarized in **Table 5** (below) and then detailed for each test form in the pages that follow. Based on typically accepted cutoffs for the four main alignment criteria considered in this study, both SAT test forms would be considered acceptably aligned with Arizona Academic Standards for grade 11-12 ELA. For full alignment, one multiple choice item would need to be revised or replaced on each form to resolve the weak Range of Knowledge (Form 07) or DOK Consistency (Form 10) for the Reading Standards for Literature reporting category.

Table 5. Overall alignment findings for two forms of the SAT ELA/literacy sections with Arizona Academic Standards for grade 11-12 ELA

Test Form	Alignment Findings	Approx. Number of Items that Need Revision/Replacement for Full Alignment
SAT Form 07	Acceptably Aligned	1 multiple choice item
SAT Form 10	Acceptably Aligned	1 multiple choice item

Results by Test Form

The results of the analysis for each of the four alignment criteria are provided in **Tables 6** and **7** (on the following page) for the ELA/literacy portion of each test form for the reporting categories of Reading Literature, Reading Informational Text, Writing, and Language. The approximate numbers of replaced or revised items necessary to fully meet the minimum cutoffs for alignment are provided for each test form. More detailed data on each of the criteria are given in **Appendix B**, in the first three tables for each test form.

In **Tables 6** and **7**, "YES," indicates that an acceptable level was attained between the assessment and the reporting category on the criterion. "WEAK" indicates that the criterion was nearly met, within a margin that could simply be due to error or reasonable variation in reviewer coding. "NO" indicates that the criterion was not met by a noticeable margin—10% under an acceptable level for Depth of Knowledge Consistency, 10% under an acceptable level for Range of Knowledge Correspondence, and 0.1 under an index value of 0.7 for Balance of Representation. Categorical Concurrence is reported in average number of items. Depth of Knowledge Consistency is reported by the percentage of items that were at or above the DOK of the corresponding standard. Range of Knowledge is reported as the percentage of standards within each reporting category that were targeted by one or more items. Balance of Representation is an index value, ranging from 0-1.

SAT Test Forms 07 and 10

SAT Test Form 07 fully met all alignment criteria for all reporting categories with the exception of weakly meeting the criterion of Range of Knowledge for the Reading Standards for Literature domain. This weak Range could be resolved with the adjustment of just one item. SAT Test Form 10 fully met all alignment criteria for all reporting categories with the exception of weakly meeting the criterion of Depth of Knowledge for the Reading Standards for Literature domain. This weak Depth could be resolved with the adjustment of just one item. Results were very similar across the two forms. On both forms, there was greater emphasis on informational text than on literature.

Table 6. Results for SAT Form 07 and Arizona Academic Standards for grade 11-12 ELA

AZ Grade 11-12 ELA Reporting Categories	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
Reading Literature (RL)	10	66%	47%	0.77	YES	YES	WEAK	YES
Reading Informational Text (RI)	82.6	61%	80%	0.82	YES	YES	YES	YES
Writing (W)	29.6	61%	66%	0.75	YES	YES	YES	YES
Language (L)	38.4	90%	69%	0.79	YES	YES	YES	YES

*Number of items/points

Table 7. Results for SAT Form 10 and Arizona Academic Standards for grade 11-12 ELA

AZ Grade 11-12 ELA Reporting Categories	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
Reading Literature (RL)	10.2	47%	57%	0.73	YES	WEAK	YES	YES
Reading Informational Text (RI)	81.8	74%	7%	0.82	YES	YES	YES	YES
Writing (W)	29.6	63%	66%	0.77	YES	YES	YES	YES
Language (L)	38.4	83%	84%	0.75	YES	YES	YES	YES

*Number of items/points

Writing Prompt The SAT included a single weighted writing prompt that was evaluated according to a three-part rubric that details aspects of reading, analysis, and writing. The 50-minute essay expects students to prepare a written analysis of source text and was coded to standards from three of the four reporting categories: RI.1, RI.2, and RI.3, (related to analysis of text meaning and structure), RI.5, and RI.6, (related to evaluation of author's purpose and craft), W.1-3, W.4 (related to clear composition of texts), and L.1, L.2, and L.3 (related to use of Standard American English, grammar, mechanics, language function in context, etc.).

Source of Challenge Issues and Reviewers' Comments
Reviewers were instructed to document any Source of Challenge issue and to provide any other comments they may have about an item. A Source of Challenge is a technical issue with an item that can result in a student answering the item correctly or incorrectly for the wrong reason. There were no items for which any reviewer noted a technical issue.

Reviewers wrote notes about a number of items. Some notes indicate when only part of a particular standard was targeted by an assessment task. After coding each assessment form, reviewers were asked to respond to four debriefing questions. In these comments, reviewers noted that the text passages used in the assessments were challenging but appropriate for grade 11 students. Reviewers commented that in some cases they needed to look back at standards from prior grades to trace the vertical progression of expectations in order to differentiate between and among certain grade 11-12 standards, particularly for the Language domain, because some items included aspects of standards from prior years. Reviewers expressed some concern that the SAT emphasis on informational texts and explanatory/argumentative writing would have the unintended consequence of promoting a similar emphasis within the classroom setting, despite the fact that the Arizona Academic Standards equally represent the Standards for Reading Literature and for Informational text, as well as equally represent the Writing standards for argumentative, informative/explanatory, and narrative texts. The full text of reviewers' notes and debriefing comments was provided to ADE and the College Board but has been redacted for public release.

Reliability among Reviewers

Reviewers engaged in some adjudication of their data after all reviewers finished their coding for an assessment. These discussions were used to identify any mistakes in coding. Reviewers were not required to change their coding after discussion unless they found a compelling reason. The agreement statistics shown in **Table 8**, on the following page, were computed after adjudication. The overall intraclass correlation among the ELA reviewers' assignment of DOK levels to items was high for both analyses: 0.95 and 0.99 (**Table 8**). An intraclass correlation value greater than 0.8 generally indicates a high level of agreement among the reviewers.

A pairwise comparison was used to determine the degree of reliability of reviewers coding at the reporting category and the standard level. The pairwise comparison was computed by considering for every item the coding assigned by each reviewer compared to the coding by each of the other reviewers. For example, for five reviewers a total of 10 comparisons are computed for each item. For most alignment studies, the objective pairwise agreement is higher than 0.60. The pairwise agreement for assigning objectives to items was high for both test forms, +/- 0.80 for each. For coding to the level of reporting category, a pairwise agreement of 0.90 is desired. For both test forms, the pairwise agreement for reporting category was very high. The ELA panel was comprised of five expert panelists, four of whom had previously participated between one and many time(s) in the full training and alignment processes with the WebbAlign team. The high level of experience likely influenced the high agreement.

Table 8. Intraclass and pairwise comparisons, SAT with Arizona Academic Standards for ELA Grade 11-12

Test Form	Intraclass Correlation (DOK)	Pairwise Comparison (DOK)	Pairwise Comparison (Reporting Category)	Pairwise Comparison (Objectives)
SAT Form 07	0.95	0.86	0.98	0.79
SAT Form 10	0.99	0.91	0.98	0.83

Findings: Mathematics

The SAT math test had 58 items, including 20 items where calculators were not permitted and 38 items where students were permitted to use a calculator. The College Board SAT website provides a list of brands and models of calculators that are acceptable for use on the mathematics test. Permitted calculators include most graphing calculators and all scientific calculators. More basic four-function calculators are permitted but not recommended.

The College Board defined four mathematical domains for the Arizona SAT (College Board, 2020). Most items are intended to address three of the domains: Heart of Algebra, Problem Solving and Data Analysis, and Passport to Advanced Math. Only six items per test form are intended to target the fourth domain: Additional Topics in Math (College Board, 2015). A summary of item counts and allotted assessment times is shown in **Table 9**. All items were equally weighted at one (1) point. Students were allotted 80 minutes to complete the mathematics portions of the assessment. The mathematics portions of the assessments had two types of items: multiple-choice (78%) and student-produced response items (22%), in which students fill in a grid to enter a positive whole number, decimal, or fraction.

Table 9. SAT item counts, types, and session times - mathematics

Math Portions of the SAT: Test Sections & Time		Total Number of Items	Item Type (across sections)
No Calculator	25 min	20 items (No Calculator)	Multiple-choice (78%)
Calculator	55 min	38 items (Calculator)	Student-produced response (22%)
Total	80 min	58 items total	

Source: The College Board, 2015

Standards

A summary of the levels of complexity within the Arizona Academic Standards for Algebra 1 and Geometry is given in **Table 10**. Thirteen of the mathematics standards included in the study (16%) were considered DOK 1. These expectations emphasized use of standard algorithms to conduct calculations, recognition of particular mathematics concepts, and reproduction of set procedures. Sixty standards (73%) were considered a DOK level 2, targeting work involving conceptual understanding of mathematics concepts, decision making, and/or making sense of mathematics in context. Eight standards (10%), were considered to be DOK 3, emphasizing non-routine problem solving such as developing and evaluating mathematical arguments and strategies as well as generating proofs. One standard was considered DOK 4. This standard expected application of mathematics to solve real-world design problems. A DOK 4 expectation is one that is both at least as complex as a DOK 3 but also requires extended time—days, weeks, or months—to complete. Although some components of DOK 4 standards may be reasonably assessed by on-demand

assessments, DOK 4 standards should not be expected to be fully assessed by an on-demand assessment and are more appropriate for classroom assessment contexts. Because the Standards for Mathematical Practices (MPs) are intended to be habits of mind that are developed throughout K-12 education, they are not expected to be authentically assessed on a single on-demand test such as the SAT, but rather addressed primarily in the classroom.

Table 10. Percentage of expectations by Depth of Knowledge (DOK) levels for Arizona Academic Standards for Algebra 1 and Geometry

Arizona Algebra 1 and Geometry Reporting Categories	Total Number of Expectations	DOK Level	Number of Standards by Level	Percentage within RC by Level
Number and Quantity (N)	4	2	4	100.0
Algebra (A)	17	1 2	5 12	29.4 70.6
Functions (F)	15	1 2	3 12	20.0 80.0
Statistics and Probability (S)	10	1 2	1 9	10.0 90.0
Geometry (G)	36	1 2 3 4	4 23 8 1	11.1 63.9 22.2 2.8
Total	82	1 2 3 4	13 60 8 1	16 73 10 1

Mapping of Items by Standards

Panelists reviewed two test forms, identified as Form 07 and Form 10. With a single exception, reviewers were able to find standard correlations for all items on both forms that they thought addressed topics within Algebra I and Geometry, although they noted instances for which they found only a partial correlation. Averaged across forms, 23 of the 82 Arizona Academic Standards for Algebra 1 and Geometry standards were targeted by items on each test form (**Table 11**). Of these targeted standards, 18 (averaged across both test forms) were Algebra I standards. Although only a small proportion (28%) of the total standards were assessed by the items on these test forms, it is important to recognize that the high school mathematics standards used in this study span two courses (Algebra I and Geometry). The SAT math test includes 58 items, which technically allows the capacity to meet typically accepted cutoffs for alignment criteria, even if considering the 82 standards from Algebra 1 and Geometry. However, only a low percentage of items on each test form (<10%) were found to correspond to a high school geometry standard.

Table 11. Number and percentage of Arizona Academic Standards for Algebra 1 and Geometry with at least one corresponding item found by a majority of reviewers

Assessment	Number of Items	Number of Standards Targeted	Percentage of Standards with at least One Corresponding Assessment Item
SAT Form 07	58	26	32%
SAT Form 10	58	20	24%

If no particular standard was targeted by a given assessment item on a form, reviewers were instructed to code the item at a more inclusive level, such as the domain or reporting category level. This coding would reflect that the item addressed knowledge, skills, and abilities that fit within Algebra 1 or Geometry but that weren't explicit in the standards. This situation may indicate that there is a part of the content on the assessment that is not expressly or precisely described in the standards, or that there is a part of the content within the standards that is being interpreted differently by different parties. When panelists cannot find a particular Algebra 1 or Geometry standard to which an item corresponds, it may also indicate that the item addresses content that is typically included in coursework before or after the Algebra 1 and Geometry sequence. In this case, it was known that the SAT test forms could include items that corresponded to standards from other courses in addition to Algebra 1 and Geometry. For example, in addition to Algebra 1 and Geometry, the College Board identified standards from Arizona's Algebra 2 and Quantitative Reasoning courses that corresponded to the SAT (College Board, 2020). If panelists found an item that addressed mathematics included in courses taken after Geometry, they could code it to a generic "Advanced Mathematics" domain. Similarly, if panelists found an item that included math content addressed prior to students taking Algebra 1, they could code it to a generic "Below-Grade Mathematics" domain. Two of the Arizona mathematics panelists were directly involved in the development of the state academic standards for mathematics and were very familiar with the Arizona math standards across grades and courses.

There were 12 items on SAT test form 07 that all or all-but-one of the five panelists coded to a generic domain. Most of these items (75%) targeted below-grade mathematics. There were 25 items on SAT test form 10 that all or all-but-one of the five panelists coded to a generic domain. Most of these items (68%) targeted below-grade mathematics. Items for which reviewers noted domain-level, below-grade, or mathematics expectations addressed in courses outside of Algebra 1 and Geometry are identified in **Table 12**. Item-level information subject to non-disclosure agreements has been provided to ADE and to the College Board but has been omitted for public release.

Table 12. SAT math items assigned to generic content expectations by all or all-but-one of five reviewers for each test form

Test	Generic Content Expectation	Item Number
SAT Form 07	Below-Grade Mathematics	Section 3 - #15
		Section 4 - #5
		Section 4 - #7
		Section 4 - #9
		Section 4 - #13
		Section 4 - #32
		Section 4 - #35
		Section 4 - #37
		Section 4 - #38
	Advanced Mathematics	Section 3 - #20
		Section 4 - #18
	Statistics and Probability	Section 4 - #26
SAT Form 10	Below-Grade Mathematics	Section 3 - #1
		Section 3 - #6
		Section 3 - #7
		Section 3 - #8
		Section 3 - #9
		Section 3 - #10
		Section 4 - #1
		Section 4 - #2
		Section 4 - #3
		Section 4 - #5
		Section 4 - #8
		Section 4 - #14
		Section 4 - #17
		Section 4 - #19
		Section 4 - #27
		Section 4 - #35
		Section 4 - #37
	Advanced Mathematics	Section 3 - #15
		Section 3 - #16
		Section 3 - #20
		Section 4 - #10
		Section 4 - #23
		Section 4 - #25
		Section 4 - #30
		Section 4 - #32

Alignment Statistics and Findings for Two SAT Test Forms and Arizona Standards for Algebra I and Geometry

Overall alignment results are summarized in **Table 13** below and then detailed for each test form in the pages that follow. Overall results are provided for each test form as relates to the full set of Algebra 1 and Geometry standards. Overall results are also provided for each test form as relates to the Algebra 1 standards only.

Both SAT test forms were found to need major adjustments to meet minimum cutoffs for alignment with the full set of high school Algebra 1 and Geometry standards (**Table 13**). Averaged across test forms, approximately 24 items per test form would need to be added, revised, and/or replaced to meet all alignment criteria if using the full set of high school Algebra 1 and Geometry standards. If considering Algebra 1 standards only, some alignment issues remain, but with a more limited scope. Averaged across test forms, approximately 10 items per test form would need to be added, revised, and/or replaced to meet all alignment criteria if considering Algebra 1 standards only (**Table 13**). When compared against Algebra 1 standards, and averaging results across test forms, the test forms would be considered to need slight adjustments. When compared against Algebra 1 and Geometry standards, the test forms would be considered to need major adjustments.

Table 13. Overall alignment findings for two forms of the SAT math sections with Arizona Academic Standards for Algebra 1 and Geometry and for Algebra 1 only

Test Form	**Alignment Findings**	**Approx. Number of Items that Need Revision/Replacement for Full Alignment with <u>Algebra 1 and Geometry Standards</u>**
SAT Form 07	Needs Major Adjustments	22 items
SAT Form 10	Needs Major Adjustments	26 items
Test Form	**Alignment Findings**	**Approx. Number of Items that Need Revision/Replacement for Full Alignment with <u>Algebra 1 Standards Only</u>**
SAT Form 07	Needs Slight Adjustments	9 items
SAT Form 10	Needs Major Adjustments	11 items

Results by Test Form

The results of the alignment analysis for each of the four alignment criteria are provided in **Tables 14** and **15** for the mathematics portions of the two test forms. In these tables, results are shown for the five reporting categories of Number and Quantity (N), Algebra (A), Functions (F), Statistics and Probability (S), and Geometry (G). These reporting categories include all standards from the courses of Algebra 1 and Geometry. In general, assessments used for accountability typically assess a single grade level of standards or the standards associated with a single course (e.g. Algebra 2). In this case, the College Board identified standards from four different Arizona high school math courses that have some representation on the SAT. For the purposes of this study, ADE selected standards from two of these courses (Algebra 1 and Geometry) as the referents for alignment. This is because at least 90% of AZ students have completed these two math courses by grade 11. The approximate numbers of added, replaced, or revised items necessary to meet minimum levels of alignment are provided for each test form/set of standards. More detailed data on each of the criteria are given in **Appendix B**, in the first three tables for each test form.

In **Tables 14** and **15**, "YES," indicates that an acceptable level was attained between the assessment and the reporting category on the criterion. "WEAK" indicates that the criterion was nearly met, within a margin that could simply be because of error or reasonable variation in reviewer coding. "NO" indicates that the criterion was not met by a noticeable margin—10% under an acceptable level for Depth of Knowledge Consistency, 10% under an acceptable level for Range of Knowledge Correspondence, and 0.1 under an index value of 0.7 for Balance of Representation. Categorical Concurrence is reported in average number of items. Depth of Knowledge Consistency is reported by the percentage of items that were at or above the DOK of the corresponding standard. Range of Knowledge is reported as the percentage of standards within each reporting category that were targeted by one or more items. Balance of Representation is an index value, ranging from 0-1. Alignment statistics for DOK Consistency, Range of Knowledge, and Balance are not reported for the generic below-grade and advanced math categories, because these categories do not include actual standards.

SAT Test Forms 07 and 10

For SAT test form 07, the Algebra reporting category met all alignment criteria. The Functions reporting category nearly met all alignment criteria, with a weak Range of Knowledge as the only alignment issue. This weakness could be resolved with the revision or addition of two items that target at least two standards within the Functions reporting category that are not currently targeted (and at the appropriate level of DOK). The Number and Quantity reporting category would need four items added to meet the criterion of Categorical Concurrence. If these items targeted the standards at the appropriate DOK, then

the weak DOK consistency could also be resolved. The Statistics and Probability reporting category had unmet DOK Consistency and unmet Range of Knowledge. To resolve these two alignment gaps, at least one item would need to be revised or replaced to target the corresponding standard at the appropriate DOK and at least two items would need to be added, each of which targeted an additional (currently unassessed) standard within the reporting category at the appropriate level of DOK. Reviewers identified only five items corresponding to the high school Geometry reporting category, which included 36 standards. For the test form to meet typically accepted alignment criteria for the full set of Geometry standards, approximately 13 items would need to be added, each of which would need to target a different and currently unassessed Geometry standard. Overall, for SAT Form 07, a total of approximately 22 items would need to be added, revised, and/or replaced to meet the minimum levels of acceptable alignment with the full set of Algebra I and Geometry standards. Panelists identified at least one item that addressed math content that is typically addressed in courses following high school Geometry and at least eight items that addressed mathematics included in Arizona mathematics standards for middle school.

For SAT test form 10, the Algebra reporting category met all alignment criteria. The Statistics and Probability reporting category nearly met all alignment criteria, with a weak DOK Consistency and Range of Knowledge that could be resolved with the addition of just one item that addresses a currently unassessed Statistics and Probability standard at the appropriate DOK. Panelists identified only one item corresponding to the Number and Quantity reporting category. Five items would need to be added, that targeted at least one additional standard and at the appropriate DOK level, to meet all alignment criteria. The Functions reporting category had unmet Categorical Concurrency and unmet Range of Knowledge. To resolve these two alignment gaps, at least five items would need to be added, each of which targeted an additional standard within the reporting category. Reviewers identified only three items corresponding to the high school Geometry reporting category, which includes 36 standards. For the test form to meet typically accepted alignment criteria for the full set of Geometry standards, approximately 15 items would need to be added, each of which would need to target a different and currently unassessed Geometry standard. Overall, for SAT Form 10, a total of approximately 26 items would need to be added, revised, and/or replaced to meet the minimum levels of acceptable alignment with the full set of Algebra I and Geometry standards. Panelists identified at least seven items that addressed math content that is typically addressed in courses following high school Geometry and at least 16 items that addressed mathematics included in Arizona mathematics standards for middle school.

Table 14. Results for SAT Form 07 and Arizona Academic Standards for Algebra 1 and Geometry

Arizona Algebra 1 and Geometry Reporting Categories	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
Number and Quantity (N)	2.4	40%	60%	1.00	NO	WEAK	YES	YES
Algebra (A)	20.6	61%	58%	0.76	YES	YES	YES	YES
Functions (F)	12.6	51%	40%	0.74	YES	YES	WEAK	YES
Statistics and Probability (S)	7.0	33%	38%	0.79	YES	NO	NO	YES
Geometry (G)	5.0	28%	13%	1.00	NO	NO	NO	YES
Advanced Math	1.8	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Below-grade Math	8.6	N/A	N/A	N/A	N/A	N/A	N/A	N/A

*Number of items/points

Table 15. Results for SAT Form 10 and Arizona Academic Standards for Algebra 1 and Geometry

Arizona Algebra 1 and Geometry Reporting Categories	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
Number and Quantity (N)	1.0	60%	24%	N/A	NO	WEAK	NO	N/A
Algebra (A)	17.8	69%	60%	0.76	YES	YES	YES	YES
Functions (F)	5.0	52%	20%	0.81	NO	YES	NO	YES
Statistics and Probability (S)	7.0	49%	40%	0.78	YES	WEAK	WEAK	YES
Geometry (G)	3.0	47%	7%	0.93	NO	WEAK	NO	YES
Advanced Math	7.8	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Below-grade Math	16.6	N/A	N/A	N/A	N/A	N/A	N/A	N/A

*Number of items/points

Source of Challenge Issues and Reviewers' Comments

Reviewers were instructed to document any Source of Challenge issue and to provide any other comments they may have about an item. A Source of Challenge is a technical issue with an item that can result in a student answering the item correctly or incorrectly for the wrong reason. No technical issues were identified on either of the forms reviewed.

Reviewers also wrote notes about many items on each form. For all items that were coded as below-grade, reviewers made note of the general topic targeted by the item. If an item addressed only a portion of a standard, reviewers noted the component of the standard that was targeted. Panelists commented that some items may be addressed in multiple ways, and the expectations within different standards would be invoked depending on approach. Panelists discussed and commented on their different perspectives related to which approach would be most likely for students to take. The full text of reviewers' notes and debriefing comments was provided to the ADE and to the College Board but has been redacted for public release.

Reliability among Reviewers

Reviewers engaged in some adjudication of their data after all reviewers finished their coding for an assessment. These discussions were used to identify any mistakes in coding. Reviewers were not required to change their coding after discussion unless they found a compelling reason. The agreement statistics shown in **Table 16**, on the following page, were computed after adjudication. The overall intraclass correlation among the mathematics reviewers' assignment of DOK levels to items was very high (0.91 and 0.94) for both analyses (**Table 16**). An intraclass correlation value greater than 0.80 generally indicates a high level of agreement among the reviewers.

A pairwise comparison was used to determine the degree of reliability of reviewers coding at the reporting category level and the standard level. The pairwise comparison was computed by considering for each item the coding assigned by each reviewer compared to the coding by each of the other reviewers. For example, for five reviewers a total of 10 comparisons were computed for each item. For most alignment studies, the standards pairwise agreement is higher than 0.60. The pairwise agreement for assigning standards to items was very high (0.89 and 0.87) for both test forms analyzed. For coding to the level of reporting category, a pairwise agreement of 0.90 is desired. For both test forms, the pairwise agreement for reporting category was high (0.94 and 0.97). The mathematics panel was comprised of five expert panelists, four of whom had previously participated between one and many time(s) in the full training and alignment processes with the WebbAlign team. Two of the Arizona panelists had been deeply involved in the development of the Arizona math standards. The high level of experience likely influenced the high agreement.

Table 16. Intraclass and Pairwise Comparisons, SAT with Arizona Academic Standards for Algebra 1 and Geometry

Test Form	Intraclass Correlation (DOK)	Pairwise Comparison (DOK)	Pairwise Comparison (Reporting Category)	Pairwise Comparison (Standards)
SAT Form 07	0.91	0.82	0.94	0.89
SAT Form 10	0.94	0.85	0.97	0.87

Conclusion

A content alignment analysis was conducted during the month of November 2020 to provide information about the degree of alignment of the SAT with the Arizona Academic Standards for English Language Arts (ELA) Grade 11-12, Algebra 1, and Geometry. The analysis was conducted to provide evidence about the degree of alignment of the SAT with the corresponding Arizona standards, as pertains to fulfilling requirements as stated in Federal statute. The Every Student Succeeds Act (ESSA, 2015) provides states the flexibility to use a locally selected, nationally recognized high school academic assessment in lieu of the statewide summative assessment, provided the assessment meets certain technical criteria, including that it is aligned to and addresses the depth and breadth of the state's academic content standards.

Pursuant to Arizona statute, the State Board of Education maintains a Menu of Assessments for high school testing that includes nationally recognized high school assessments that meet policy requirements. These tests are intended to be used to measure student achievement of Arizona Academic Standards. The SAT is planned for inclusion in the Menu of Assessments, starting in 2020-2021 (Arizona State Board of Education, 2020).

Arizona students take the SAT in spring of grade 11. Alignment of the SAT ELA/literacy portions are therefore considered in relation to the grade 11-12 ELA standards. In contrast to ELA, mathematics courses are taken in different grades by different students; there is no universal grade 11 set of mathematics standards. Because math is course-based, multiple factors must be taken into account to make a decision about the appropriate set of high school mathematics standards for use in a state's accountability system. According to the Arizona State Department of Education (ADE), at least 90% of students in grade 11 have completed both Algebra 1 and Geometry mathematics courses. Further, prior State Board work with educator panels resulted in a consensus that Algebra 1 and Geometry standards should be prioritized in a high school summative assessment used for state accountability purposes (A. Ahumada, personal communication, October 26, 2020). Taking these key considerations into account, ADE selected the Algebra 1 and Geometry standards as the appropriate referents for alignment. The College Board identified the Arizona Algebra 1 and Geometry standards as corresponding to the SAT as well as standards from two additional courses: Algebra 2 and Quantitative Reasoning.

A two-day remote alignment institute took place on November 6-7, 2020 via Zoom video conferencing to analyze the agreement between the Arizona academic standards and two forms of the SAT. Three Arizona educators and two external reviewers (i.e. reviewers from other states) participated in each subject-area panel (ELA/literacy and mathematics). All panelists were selected because of their notable high school education experience and content expertise.

The study was designed to answer two main research questions:
1. What is the degree of alignment of the SAT Evidence-Based Reading and Writing section (Reading test + Writing and Language test) and Essay with the corresponding Arizona Academic Standards for grades 11-12 English Language Arts (ELA) with regards to satisfying the federal requirements within the Every Student Succeeds Act (ESSA): that a locally selected, nationally recognized high school academic assessment be aligned with state academic content standards, and address the depth and breadth of the standards?
2. What is the degree of alignment of the SAT Math test with the corresponding Arizona Academic Standards for Algebra 1 and Geometry with regards to satisfying the federal requirements within the Every Student Succeeds Act (ESSA): that a locally selected, nationally recognized high school academic assessment be aligned with state academic content standards, and address the depth and breadth of the standards?

Four alignment criteria received major attention:

- **Categorical Concurrence** between standards and assessment is met if the same or consistent categories of content appear in both documents.
- **Depth of Knowledge Consistency** between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.
- **Range of Knowledge Correspondence** is used to judge whether a comparable span of knowledge expected of students by a reporting category (domain/strand) is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.
- **Balance of Representation** is used to indicate the degree to which one content indicator (standard) is given more emphasis on the assessment than another.

The ELA/literacy portions of both test forms analyzed were considered acceptably aligned with the Arizona Academic Standards for ELA grades 11-12. Each test form would need only one item revised or replaced to fully meet the typically accepted minimum cutoffs for full alignment.

If considering the full set of Algebra 1 and Geometry standards, spanning two years of coursework, then both test forms analyzed would need major adjustments to meet the typically accepted alignment criteria agreed upon and used in this analysis. Approximately 24 math items would need to be added, replaced, and/or revised per test form. Over half of these adjustments are required to attend to the breadth (Range of Knowledge) of the Geometry reporting category. Alignment of statewide summative assessments are typically considered in relation to one year of coursework, and not two years of coursework. If considering the alignment of the mathematics portion of the test forms with Algebra I standards only, then the test forms would be considered to need slight adjustments to meet typically accepted alignment criteria, with approximately 10 items added, replaced, and/or revised per test form.

On both test forms, panelist found many items that mapped to Arizona Academic Standards for middle school math and Algebra 2. These items addressed a range of math topics included in the Arizona standards across multiple domains. Because of their use on the SAT, these items can also be inferred to have value for making inferences about college and career readiness based on student test scores.

The directive for use of a college and career readiness exam as the Arizona statewide summative high school assessment for ELA and mathematics was initiated by Arizona legislators and ultimately became law (AZ Rev Stat § 15-741.02, 2016). Audra Ahumada, Deputy Associate Superintendent of Assessment for the Arizona Department of Education, noted the state perspective is that a college and career readiness exam is meaningful for high school students, taken seriously by high school students, and provides students with opportunities to demonstrate readiness for post-secondary endeavors (A. Ahumada, personal communication, November 20, 2020). However, if intended as a measure of the Arizona Algebra 1 and Geometry standards (or of the Algebra 1 standards only), the SAT would need to be augmented to satisfy (both the state and) the federal requirements that a locally selected, nationally recognized high school academic assessment be aligned with state academic content standards, and address the depth and breadth of the standards. While augmenting the SAT to attain an acceptable level of alignment is certainly possible, it should be noted that augmentation tends to be a rather expensive process and adds complexity to the administration of the tests, because items used to augment a test need to be administered separately from the college entrance test.

References

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55.

The College Board. (2020) SAT Suite of Assessments – Alignment to Arizona Standards

The College Board. (2015). *Test Specifications for the Redesigned SAT.* Downloaded November 2020, https://collegereadiness.collegeboard.org/pdf/test-specifications-redesigned-sat-1.pdf

AZ Rev Stat § 15-741.02 (2016) Accessed November 2020, https://www.azleg.gov/viewdocument/?docName=https://www.azleg.gov/ars/15/00741-02.htm

Valencia, S. W., & Wixson, K. K. (2000). Policy-oriented research on literary standards and assessment. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research: Vol. III.* Mahwah, NJ: Lawrence Erlbaum.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and mathematics education*. Council of Chief State School Officers and National Institute for Mathematics Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.