A Systematic Review of Studies Aligning SAT Math to State Math Content Standards

MARYLAND ASSESSMENT RESEARCH CENTER
UNIVERSITY OF MARYLAND, COLLEGE PARK
2025

COLLEGE OF EDUCATION
MARYLAND ASSESSMENT RESEARCH CENTER

Table of Contents

Executive Summary

Under the Every Student Succeeds Act (ESSA), states gained broader latitude to replace their statewide high school mathematics assessments with nationally recognized tests such as the SAT math or ACT math. Whether the SAT math items align with states' content standards and cognitive complexity remains critical. The results from different SAT math alignment studies were not consistent. This inconsistency may stem from alignment methodologies, evaluation criteria, alignment procedures, and panelist composition, which highlights the need for a comprehensive review of the alignment studies of SAT math to state Algebra I standards.

This paper reviews and synthesizes 14 alignment reports with 20 alignment studies on the SAT math section across 12 states, including Arizona, Connecticut, Delaware, Florida, Georgia, Illinois, Maine, Maryland, New Mexico, Rhode Island, Tennessee, and Texas. Two questions guide our analysis: (1) Which alignment methodologies were employed in previous studies aligning SAT math to state Algebra I standards and cognitive complexity? (2) To what degree do the SAT math items align with state Algebra I standards and cognitive complexity? The findings of this review aim to provide empirical evidence to state policymakers and other stakeholders about the validity of using the SAT math test in place of state Algebra I tests.

Major Findings

The major findings from this review study are summarized as follows.

1. Comparison of Algebra I Standards Across States
   1) The CCSS comprised of 70 Algebra I standards, with 4 domain and 11 subdomains. Most of the states in this review followed four main CCSS Algebra I domains. However, Florida uses only three domains. Only Arizona includes 11 subdomains. All other states include 10 subdomains.
   2) The number of Algebra I standards ranged from 482 to 63 across reviewed states. Some states subdivided existing CCSS standards (e.g., Georgia, Rhode Island), while others added new standards not in CCSS (e.g., Tennessee).

2. Alignment Methodologies
   1) Among the 20 reviewed alignment studies, 13 studies aligned SAT math items to state Algebra I standards (item-to-standard approaches), while 7 studies aligned the SAT math standards to the state standards (standard-to-standard approaches).
   2) The 13 item-to-standard alignment studies were from Maryland, Arizona, Delaware, Florida, Georgia, and Maine aligned SAT items to state standards. Among the 13 alignment studies, Maryland aligned the items on two (15%) digital SAT math forms, while the other 11 (85%) aligned items on the paper-and-pencil SAT math forms.
   3) For the item-to-standard alignment studies, the following evaluation frameworks were used.

a) Webb’s four criteria: Categorical Concurrence, Range of Knowledge, Balance of Representation, and Depth-of-Knowledge (DOK) Consistency.
b) HumRRO's four criteria adapted from Webb, focusing on content matching and item sufficiency for reporting.
c) CCSSO High Quality Assessment Criteria emphasizing cognitive demand, item quality, and content-practice connections.
d) Number of items to be added/replaced determined if items needed revision or replacement to cover all state standards.
e) Proportion of matched standards showing different degrees of alginemnt.

4) Seven alignment studies from Connecticut (2), Illinois, New Mexico, Rhode Island, Texas, and Tennessee aligned SAT standards to state standards, among which 6 were conducted on the paper-and-pencil SAT and 1 on digital SAT. In these studies, alignment was evaluated in terms of the proportion of state standards matched by at least one SAT standard.

3. Findings of the Item-to-Standard Alignment Studies

1) Alignment on Content Standards

a) Twelve alignment studies used Webb’s four criteria. While the *Algebra* and *Functions* domains generally met the criteria in most studies, *Number & Quantity* and *Statistics* were either partially met or not met.
b) HumRRO’s four criteria for evaluating item alignment were reported in the Delaware and Maine alignment studied in 2016 and the criteria were met.
c) The CCSSO evaluation criteria were reported in the Delaware and Maine alignment study as well. The criterion for assessing a balance of concepts, procedures, and applications (C2) and the degree of high-quality items (C5.2) were partially met, while the criterion for connecting practice to content (C3) and ensuring high-quality items and a variety of item types (C5.1) were fully met.
d) The criterion on the number of items to be added or replaced was applied in 6 alignment studies, with the number of items ranging from 4 (7%) to 35 (60%). Accordingly, one alignment indicated the SAT was acceptably aligned, two showed slight adjustments were needed, and three revealed major adjustments were required.
e) Regarding proportion of matched standards, the item-to-standard alignment studies had a relatively low proportion with a smaller variance (*M*=42%, *SD* = 9%), ranging from 26% to 58%, suggesting that two test forms showed strong alignment, and all other 10 alignments showed weak or unacceptable alignment.

2) Alignment on Cognitive Complexity

a) Thirteen item-to-standard alignment studies applied Webb’s four DOK levels. Although Algebra typically met DOK Consistency criteria, alignment to *Number & Quantity* and *Statistics* at appropriate DOK levels was frequently not met or only partially met.

b) Florida combined *Number & Quantity* and *Statistics*, yielding mixed results: 3 out 4 studies met or partially met alignment while the forth one lack of high-level items.

c) HumRRO's Four Criteria. Two states (Delaware and Maine) reported HumRRO's Item DOK Represents Test Specifications. The criterion was partially met.

d) CCSSO High Quality Assessment Criteria was used in two states (Delaware and Main). The criterion, Requiring a range of cognitive demand (C4) was used to assess cognitive complexity alignment. The criterion was only partially met.

4. Findings of the Standard-to-Standard Alignment Studies

   1) Among the 7 standard-to-standard alignment studies, 6 (86%) evaluated the paper and pencil version of the SAT, while 1 (14%) evaluated the digital version.

   2) Results on Content Standards

      a) The 7 standard-to-standard alignment used the proportion of matched standards as criterion. The proportion ranged from 45% to 82% (*M* = 62%, *SD* = 16%). The proportion of standards matched by at least one standard, suggesting that three alignment studies showed very strong alignment, one showed strong alignment, and three showed weak alignment.

5. Implications

   1. Alignment methodology: we strongly encourage the use of item-to-standard approach as the alignment is test form specific. Item-level evaluation will provide more detailed validity evidence related to form equivalence.
   2. Independence: we recommend having independent agency to conduct alignment studies instead of test vendors to evaluate the alignment of items they created and pulled on different test forms.
   3. Logistic considerations: factors including the number of panelists, the representation of the panelists who know better of different student populations, and the rounds of reviewing should be well defended in balancing the technical quality and cost efficiency in designing an alignment study.
   4. Evaluation criteria: given different evaluation frameworks have been used in different alignment studies. It is worth exploring the technical defensibility and the utility of the evidence extracted from each evaluation framework to decision making.
   5. Additional sources: given the narrower content coverage of the state Algebra I standards by the SAT math items, state should consider additional items or sources as technically defensible to supplement what is missing from SAT math items.
   6. Given the potential subjectivity and inconsistency in item alignment by human panelists, artificial intelligence (AI) is a promising tool for mitigating the panelists' subjectivity and increasing cost efficiency in item alignment.

1. Introduction

Initially signed into law in 1965, the Elementary and Secondary Education Act (ESEA) requires each state to establish a coherent and consistent state assessment system. The original purpose of ESEA was to ensure equity for all by increasing access to a basic education with a hope that all students will master the knowledge and skills to achieve success in college and career at high school graduation. The Every Student Succeeds Act (ESSA) provided states with greater flexibility in selecting assessments for accountability purposes. Compared with ESEA, the ESSA gives states and local education agencies a great amount of education authority from the federal government (Sharp, 2016). For example, states do not have to administer the same assessment to all students. The local educational agencies can choose an alternative nationally recognized high school academic assessment approved by the state in lieu of the state assessment as long as it aligns with the state academic standards. As a result, many states have explored nationwide aptitude tests such as the SAT or ACT in lieu of the state test for math (Camara et al., 2019).

The state assessments for general mathematics and reading/language arts for high school must be submitted for peer review (ESEA section 1111(a)(4); ESEA section 1111(b)(2)). The purpose is to facilitate state and local innovation while providing feedback to states, promoting effective implementation of state standards, and enhancing collaboration between the federal government and each state (ESSA section 1111(a)(4)(B)). The peer review process is evidence-based. States submit critical elements, which are evidence that shows the state's assessment system meets a set of established criteria, such as state plans, alignment, standard setting criteria, and methodology for identifying schools and children at risk.

As of 2024, at least 21 states have opted to use either the SAT or ACT as a substitute for their statewide mathematics assessments, with 10 states selecting the SAT (source: https://reports.ecs.org/comparisons/states-school-accountability-systems-2024-10, retrieved on January 27, 2025). However, locally selected assessments must align with state academic content standards and address these standards comprehensively in both depth and breadth. With the increasing adoption of college admissions tests for ESSA-designated statewide assessments, the National Council on Measurement in Education (NCME) has highlighted two primary concerns: the alignment of these assessments to state content standards and their alignment with students' competency levels. The first issue pertains to the content coverage of the SAT, while the second relates to its cognitive complexity. As a result, alignment studies have become essential tools for examining whether scores from different test forms can be considered interchangeable.

Despite the prevalence of using SAT in lieu of a dedicated high school math test, the alignment between the SAT math section and state content standards remain uncertain. For instance, previous alignment studies yielded different conclusions on the acceptance of SAT studies. This inconsistency probably comes from various methods, criteria, types and units of analysis, rater expertise, and the way summary results are computed and presented (Camera et al., 2019). Therefore, it is necessary to gather and systematically analyze existing alignment results to decide the SAT's suitability as a statewide assessment.

SAT

SAT stands for the Scholastic Aptitude Test designed by the College Board to serve as a standardized college admission test (College Board, 2015). It is a test that assesses knowledge and skills for the readiness for pursuing undergraduate education (Wao et al., 2017). SAT includes two subjects: 1) Evidence-based Reading and Writing, and 2) Mathematics (math hereafter). In 2015, the College Board redesigned the SAT. The subject of focus in this review, math, includes four sections: (1) Heart of Algebra, (2) Problem Solving and Data Analysis, (3) Passport to Advanced Math, and (4) Additional Topics in Math such as Geometry and Trigonometry.

The SAT has two types of delivery formats: the traditional paper-and-pencil format and a digital format which adopts computerized multistage adaptive testing. SAT math in the paper-and-pencil format consists of two sections, which add up to 58 items. The time allotted for the SAT math test is 80 minutes. The no-calculator section contains 20 questions, and the time allotted is 25 minutes. The calculator section consists of 38 questions, and the time allotted is 55 minutes. Across all SAT math items in the linear form, 45 are multiple-choice items, and 13 items are constructed-response questions (College Board, 2015).

The digital form of SAT math (SAT digital form, hereafter) consists of two equal-length portions, so-called stages. The first stage, the router stage, consists of a mix of easy, medium, and difficult items. The difficulty of the items in the second stage is targeted to the test taker's performance on the items in the first router stage. In total, the digital SAT math section consists of 44 items, and the time allotted for the SAT math test is 70 minutes. Each stage consists of 22 items, and the time allotted is 35 minutes. Across all SAT math test items in the digital form, 28-32 are multiple-choice items, and 8-12 items are constructed-response items (College Board, 2022).

Starting in March 2023, all students taking the SAT at international test centers take the digital test form. Starting in fall 2023, all students taking the SAT-related assessments take the digital tests. SAT School Day and SAT weekend administrations in the U.S. are still paper-and-pencil. Starting in spring 2024, all students will take the full SAT Suite of Assessments digitally (College Board, 2024, retrieved from https://satsuite.collegeboard.org/help-center/will-paper-and-pencil-sat-still-be-available-alongside-digital-version).

According to the College Board, the key features of the digital SAT Suite's Math section are a strong focus on the content that matters most for college and career readiness and success; an emphasis on applied problems in real-life settings in which the use of mathematical practices is integrated with the content; a balance of fluency, conceptual understanding, and application items within and across all content topics; and an emphasis on problem-solving and data analysis. The validity of the SAT has been controversial, with some studies finding it reliably predicts college performance such as GPA (Bridgeman et al., 2000; Cornwell et al., 2008; Coyle & Pillow, 2008; Noble, 2000; Patterson & Mattern, 2010; Wao et al., 2017) and some others did not (Ellrich, 2014; Soares, 2012).

SAT has been an important indicator in college admission. In 2023, 1.9 million students in the class of 2023 took the SAT in high school, and more students than ever are participating in SAT School Day (College Board, 2023). A survey conducted by the National Association for College Admission Counseling (NACAC) indicated that SAT and ACT scores are ranked second in the hierarchy of factors considered in admission decisions after high school grades (Hawkins & Lantz, 2005). About 90% of four-year academic programs in the USA require students who graduated from high school to take either the SAT and/or the ACT in order to be admitted to college or undergraduate programs (Zwick, 2007). This also leads to the act that states are using SAT or ACT as the alternative to the statewide summative assessment.

However, as a college admission test, the SAT was not specifically developed to measure the content identified in each state's content standards. The failure of tests to represent state standards with balance has serious consequences for the kind of teaching that will occur in the states using such tests. As research in Kentucky, Washington State, and elsewhere confirms, teachers pay more attention to what is on the test than to what is on the standards (Resnick et al., 2004; Stecher & Barron, 1999; Stecher et al., 2000). Thus, the most challenging requirement for existing college admissions tests is demonstrating alignment with state standards or, when used as a local option, demonstrating equivalent breadth and depth to a customized state assessment.

Common Core State Standards (CCSS)

Academic content standards are one part of the high level criteria that states hold for accountability. They are statements that specify what students are expected to know and be able to do, guiding teachers and students in teaching and learning. Content standards should be coherent and rigorous and encourage the teaching of advanced skills, be aligned with entrance requirements for credit-bearing coursework in the system of public higher education in the State and relevant career and technical education standards. States were not permitted to use their own accountability systems until the enactment of the No Child Left Behind (NCLB) Act. Nevertheless, it was claimed that states tended to bring down the demands of standards to let more students meet the criteria. Both conservative and progressive policymakers supported the development of national standards and assessments, arguing that the NCLB Act incentivized states to exploit the law by lowering academic standards. (Skinner & Fedder, 2014).

By 2004, conversations among policymakers and the conduct of the American Diploma Project had shifted policy-making towards accepting the notion of national standards and assessments (Watt, 2011). CCSS was then discussed and created in alignment with this policy. The National Governors Association (NGA), the Council of Chief State School Officers (CCSSO), known as the CCSS Initiative, contributed to the development of the CCSS. The primary goal of the CCSS Initiative is to define the knowledge and skills students need to acquire to succeed in college and their careers, providing clarity for all stakeholders (Kendall, 2011). CCSS was finally released in 2010, after being reviewed and verified by the Validation Committee for the first draft, receiving feedback from the experts to ensure aligning with college- and career-readiness standards, and reviewed by the Validation Committee again. CCSS is a set of academic content standards designed to regularize the standards for states across the CCSS Initiative and also bring other benefits such as helping intentional instruction, forming a manageable number of standards, serving as a greater pool of resources by integrating different states' resources together, increasing collegiality and professionalism, bringing a more consistent and equitable learning experience, and providing customized learning and multiple pathways (Kendall, 2011). Although the CCSS Initiative was meant to set a common standard for all states, there remain differences across states due to two reasons. First, not all states adopted CCSS. As of 2024, among the 50 states, 40 states followed the structure of CCSS, four states (i.e., Alaska, Nebraska, Texas, and Virginia) have never adopted CCSS , five states (i.e., Arizona, Florida, Indianna, Oklahoma, and South Carolina) adopted but later repealed CCSS, and one state (i.e., Minnesota) partially adopted CCSS which means they only adopted CCSS in English while crafting their own CCSS in math (https://www.datapandas.org/ranking/common-core-states). Second, states that adopt the CCSS are permitted to add an additional 15% of standards chosen on their own (cf. Skinner & Feder, 2014). This may add to the differences of state standards built upon CCSS and may affect the adoption of a national assessment (Camara et al., 2019).

Alignment Studies

Alignment and the methods of alignment have been discussed for about 2 decades (e.g., La Marca et al., 2000; Porter, 2002; Webb, 1997). Smith and O'Day (1990) and Webb (1997) defined an alignment study in education by how well all policy elements in a system work together to guide instruction and, ultimately, student learning (Resneck et al., 2004). In another view (Herman et al., 2002; Webb, 1997), an alignment study examines whether the specified standards of what students should be taught and what and how well they are learning are closely synchronized with what tests measure. According to the assessment peer review guidance, alignment is a piece of evidence to support the adequate overall validity evidence for its assessments consistent with nationally recognized professional and technical testing standards, to provide documentation to support the validity of states' assessment system. It includes documentation of adequate alignment between the State's assessments and the academic content standards the assessments are designed to measure (i.e., knowledge and process), balance of content, and cognitive complexity; documentation that the assessments address the depth and breadth of the alignment.

Different approaches can be used to conduct an alignment study depending on the alignment objectives, such as aligning standards to standards, items to standards, and standards to items (Bridgeman & Stone, 2017). The standard-to-standard alignment study reviews the test specifications or standards of two tests such as SAT and state standards on the same subject, and decide the extent to which the two sets of standards match. The item-to-standard approach is to review the items in a test form, and decide the extent to which the items target the standards, focusing on the connection between content standards and assessment items and test forms. The standards-to-item approach is to decide to what extent the standards are measured by the items. For alignment studies serving as evidence to substitute for the state test, the standard-to-standard alignment methods have been used by the College Board (2023) and the University of Connecticut (2016) while the item-to-standard approaches have been used by HumRRO (2016), Maryland Assessment Research Center (MARC, 2021, 2023), and Wisconsin Center for Education Products & Services (WCEPS, 2020). Among them, the item-to-standard approach is the most popular one.

The alignment study, also known as alignment workshop (HumRRO, 2016) and content alignment institute (Webb, 2007; WECPS, 2018) usually consists of three phases. First, in the pre-session training, panelists and reviewers are trained to assign the standards/items to the target standards and certain DOK levels. In the second phase, panelists rate the standards/items independently. Some alignment studies ask panelists to rate the first several items together and have a small discussion for calibration. After the independent rating, one to two reviewers check the rating and form the report. Additional rounds of discussion may be needed when there is a large discrepancy between the panelists' ratings.

Along with different definitions of alignment, multiple approaches have been used for alignment studies. Bhola et al. (2003) reviewed various alignment models and categorized them based on their complexity. Low-complexity models define alignment as the extent to which test items match relevant content standards (or test specifications). For example, experts rate items on a Likert scale from "not match" to "match exactly." Moderate-complexity models consider both content and cognitive complexity. An example is the Council for Basic Education (CBE), which assesses content, content balance, rigor, and item response type (CCSSO, 2002). High-complexity models incorporate additional factors such as the weight of standards (e.g., Achieve, 2001; La Marca et al., 2000; Porter, 2002), as well as applicability, equity, and fairness (e.g., Webb, 1997, 1999). Furthermore, nearly all methods for analyzing the content correspondence of specific test-standards document pairs involve judgmental evaluation of test items' content by groups of subject-matter experts (Traynor, 2017).

Another review (Forte, 2017) specifically examined five alignment methods of moderate to high complexity. Among the five approaches, Webb (1997), Achieve (2006), and the Links to Academic Learning (LAL) methods consider the blueprints during alignment. LAL also considers other aspects of the assessment such as policy and technical documents and documents related to instruction. Achieve (2006) developed a model that addressed six criteria: accuracy of the test blueprint, content centrality, performance centrality, challenge, balance, and range; the Surveys of Enacted Curriculum (SEC) alignment model (Blank, Porter, & Smithson, 2001; Porter, Smithson, Blank, & Zeidner, 2007). Some alignment methodologies developed for an item to standard approach have been widely used (e.g., Achieve, 2006; Porter & Smithson, 2001; Webb, 1997, 2005).

A recent approach in content alignment research leverages AI to automate the process of mapping course outcomes to program outcomes and educational taxonomy levels, reducing human bias and subjectivity. Machine learning models, including Random Forest, Support Vector Machine, and Logistic Regression, have been applied alongside natural language processing techniques such as TF-IDF, Universal Sentence Encoder, and Word2Vec to analyze textual alignment with high accuracy. This AI-driven method enables more precise classification of learning objectives, ensuring constructive alignment in curriculum design. The development of web-based tools integrating these models has demonstrated significant potential in streamlining alignment processes, enhancing consistency, and improving the accuracy of outcome assessments. By mitigating errors in human judgment and optimizing classification, AI facilitates a more reliable and scalable approach to content alignment in educational assessments.

Purposes of this Study

Previous Reviews on Alignment Studies

Alignment studies have been done for years. Reviews on general alignment methods are common (e.g., Anderson et al., 2015; Porter, 2002; Resnick et al., 2004; Webb, 2007). Among all reviews, some focused on the alignment methodology (e.g., Bhola et al., 2003; Forte, 2017; Martone & Sireci, 2009), a few on the alignment of SAT items to state content standards (e.g., Department of Education of Oregon, 2020; West Virginia, 2020).

Bhola et al. (2003) reviewed the alignment models and categorized them according to the complexity of models for all types of alignment studies. However, the focus was on methods rather than results. Second, math content domain was not emphasized. Similarly, Forte et al. (2017) provided a report which is a comprehensive review of alignment studies. The focus was on the methodology of all possible types of alignment studies and peer review criteria. Alignment models and procedures were comprehensively reviewed. They did not examine a specific subject, nor did they provide the results of alignment to indicate how well the exam is aligned with the standards. As a result, the review did not provide practical guidance to evaluate whether exams meet or do not meet their testing expectations. Another review discussed the importance of an alignment study and illustrated three most popular alignment models (Martone & Sireci, 2009).

Camara et al. (2019) reviewed the submission document sent to US Department of Education and peer review letters instead of the alignment study reports. They found that the alignment is generally weak. Furthermore, with the focus on validity, they emphasized that the alignment study was only a part of the evidence. After reviewing peer-review letters, they identified major claims and subclaims related to the use of college admissions tests to measure the achievement, growth, and status of a state's high school students in terms of readiness for academic post-secondary success. They analyzed the claims from four aspects of accountability: validity, equity, reliability, and benefits, and summarized the sources of evidence, type of evidence, and supporting research for each subclaim. Though the review provides important insight into the feasibility of using college admission test scores for high school academic accountability, the focus was on the validity evidence submitted by states for federal accountability without reference to SAT math. Second, the analysis was not based on empirical data. Third, the analysis approach was qualitative rather than quantitative to provide us with more nuanced and concrete results.

For alignment studies between nationally college admission tests and state standards in Algebra I, there were only a few review papers. The Departments of Education in Oregon (2020) and West Virginia (2020) reviewed all available studies on alignment between SAT and common core state standards and offered insight into the use of the SAT in lieu of state exams. Though these reviews provided a good summary of the existing alignments results, there was no discussion related to the methods used for alignment in these review papers. Given no study has reviewed both alignment methods and the results of alignment studies specifically for the SAT, our review aims to focus on the methods and results of the alignment of SAT math items to state content standards, specifically.

Current Review

Though many states have enacted the use of SAT as a part of the state accountability system, and many alignment studies have been conducted for validity evidence collection and to satisfy the peer-review requirement, alignment studies have yielded different results based on different methods. In this paper, we reviewed alignment studies that align SAT math items to state math content standards as well as align SAT math content standards to states' math content standards. Alignment studies from 12 states were identified—Arizona, Delaware, Florida, Georgia, Maine, Maryland, Connecticut, New Mexico, Rhode Island, Tennessee, Texas, and Illinois. The methods and results were synthsized across alignment studies to provide insights and lessons to state test administrators about designing an alignment study and properly interprete the alignment study results. In addition, it is expected that the findings from the reviewed studies can serve additional source of validity evidence to support decisions related to using SAT math in place of state Algebra I test scores to meet the peer-reivew requirements.

This review seeks to examine alignment studies from different states conducting studies to align SAT math items or content standards to state math content standards. The following questions were addressed in our study:

(1) What alignment methods were used in previous alignment studies?

(2) What are the findings related to the alignment of SAT math items to state math content standards?

The current review is of significance in two aspects. First, it helps states to decide to what degree the assessment and standards are aligned. Second, it helps identify what challenges the states are facing in using college admission tests as an alternative to state tests. In the following sections, we introduce the method for this review study, summarize the alignment methods, synthesize the results from different studies, and discuss possible implications of the results for the alignment of SAT math items/standards to state math assessment standards.

2. Method

Data Collection

Alignment Study Reports

The focus of this review is states' alignment studies on the SAT math section. Therefore, we searched the keywords "alignment study", "math", "SAT", study report, and the state's name (e.g., Illinois) on Google.com and Google Scholar. Twenty-six alignment studies were identified that met our criteria for inclusion. Then, two researchers independently screened the articles based on the full text. When there was disagreement, a consensus was finally reached after discussion. As a result, fourteen reports were excluded from the review study due to the following reasons:

1. not including SAT alignment results (n = 7),
2. not including a review report (n = 1, Oregon),
3. aligning old version of SAT (n = 2, Pennsylvania, NAEP), and
4. summaries without quantitative results (n = 2, North Carolina and West Virginia).

As a result, 14 reports for alignment with the math standards of 12 states were retained. The 12 states are Arizona, Connecticut, Delaware, Florida, Georgia, Illinois, Maine, Maryland, New Mexico, Rhode Island, Tennessee, and Texas. Note that two alignment study reports were found for each of Connecticut, Maryland, and Florida. Additionally, the alignment studies for Delaware and Maine were presented in one single report. The results for these 14 reports are summarized in Appendix A.

Content Standards and Blueprint of State Tests

To investigate potential differences in the state math content standards and test blueprints, we also collected the standards and the blueprint of the each included state test. If the complete blueprint was provided in the alignment study report, no extra effort was made. Otherwise, the state blueprints were searched in Google.com by using terms "Algebra I" "blueprint" along with the state's name. Given that each state may have a modified set of common core state standards, or even a different standards framework, the standards for Algebra I were also searched online. Similar to searching the blueprint of the state's test, if the alignment study report did not include the complete list of the state's standards, we searched online. For two states (i.e., Delaware and New Mexico), the documents for Algebra I were not found on the state's Department of Education website, but were found on iXL.com, an online study platform in the U.S. instead. Note that among Connecticut, Florida, and Maryland, two separate alignment studies were found. Only Florida used slightly different standards for their two alignment studies.[1] Therefore, for Connecticut and Maryland, one standard framework was identified.

Alignment Content

SAT Form. We recorded (1) the version of the SAT math, namely a paper and pencil version or a digital version. (2) The length of the SAT, i.e., the number of items in each test form

[1] There were 43 and 45 standards in the 2017 study and 2018 study, respectively.

was also recorded to ensure that the test forms in different alignment studies were the same and thus comparable to each other.

Scope of Alignment. The scope of alignment refers to the subject and grade of the targeted standards in the alignment study. For most studies, the scope of alignment was Algebra 1. For Delaware and Maine, CCSS at Grades 11-12 were aligned.

Data Analysis

Each alignment study was recorded for its identification information, methodology, and alignment results. If one test form was aligned to multiple standard frameworks (e.g., the SAT paper-pencil test specification was aligned to Texas College & Career Readiness Standards and Texas Essential Knowledge & Skills in one report), the alignment to each standard framework was coded separately. Only SAT-aligned tests were included in this review; other tests in the report were excluded due to the scope of this study. The state for which the alignment study was conducted and the ID number of the state's alignment study were recorded. In addition, the year of alignment and the agency conducting the alignment were recorded. The alignment agency was the entity responsible for conducting and authoring the report.

State Standard Frameworks

We compared and contrasted the content standards for different states. The widely accepted standards are Common Core State Standards (CCSSI, 2010) with revisions in each state. We collected information on (1) whether the standard uses the CCSS framework or not, (2) the number of domains, (3) the total number of standards, and (4) the number of standards in each domain.

Blueprint of State Tests

Blueprint was used by some alignment studies that map the items to the blueprint. Therefore, we recorded the number or proportion of items in each domain. This was only done for studies that used the items-to-standards alignment approach.

Alignment Methodology

Alignment Procedure. Alignment methods were recorded in terms of alignment agency, the number of panelists, workshop venue, and rounds of alignment discussion. (1) The number of panelists (also known as reviewers in some reports) was defined as the number of experts involved in rating. People who coordinated or facilitated the alignment study while not rating the items were not counted as panelists. (2) The workshop venue was coded as in-person or online. (3) The rounds of alignment discussions counts the rounds of discussions after the initial training session and the independent rating by panelists.

Coding Scheme. The following information was recorded for the coding scheme. (1) The degree of the alignment of an item was recorded. For instance, some alignment studies used the binary approach: aligned vs. not aligned; while some others used the trierary approach: an item can be rated as fully aligned, partially aligned, or not aligned with a standard. (2) Whether multiple-coding was allowed was recorded. Double-coding means that for each item, only one to two standards were allowed for alignment. Multiple coding means the number of standards aligned to each item could be more than two standards. (3) Alignment objects are either standards (strictly speaking, test specifications) or test items. As a result, there are two types of alignment studies, standard-to-standard alignment and item-to-standard alignment. Given the large discrepancy of methods and evaluation between these two types of alignment, we reported the results of these two methods separately in the Results section.

Cognitive Complexity Taxonomies

Webb's Depth of Knowledge. Depth of Knowledge (Webb, 1997, 1999) refers to the cognitive difficulty required in finishing a task. Four levels are included from the least cognitively complex to the most complex: (1) Recall; (2) Skill/Concept; (3) Strategic Thinking; and (4) Extended Thinking.

Cognitive Rigor. Cognitive Rigor is defined as pursuing "conceptual understanding, procedural fluency, and application with equal intensity" (NGA Center & CCSSO, 2010, Key Shifts in Mathematics section, subsection 3). Therefore, in this CCSSO rigor criteria, three levels of cognitive rigor were used.

Alignment Evaluation Framework

Alignment studies used different frameworks to evaluate whether the test items were aligned to the content standards and the DOK levels. Therefore, the criteria of alignment and cutoff values were coded and included in the data analysis.

Webb's Four Criteria. Webb (1997, 2007) developed an alignment framework to evaluate how well test items aligns to each domain or subdomain (e.g., Number and Quantity, Algebra, Functions, Statistics for Algebra I content standards) of the target standards. This framework includes four evaluation criteria: i.e., categorical concurrence, DOK consistency, range of knowledge, and balance of representation. Categorical Concurrence ensures items align with academic standards by covering identical content categories, requiring at least six items per category for reliable measurement. DOK Consistency evaluates whether items reflect the cognitive complexity of standards, with at least 50% of items meeting or exceeding the expected the DOK level. Range of Knowledge Correspondence examines whether items cover a sufficient breadth of content, requiring at least 50% of standards within a reporting category to be assessed by items on the test. Balance of Representation evaluates the equitable distribution of items across standards, using an index where values closer to 1 indicate a well-balanced assessment. Together, these indices provide a structured method for assessing item alignment with content standards at the test level.

HumRRO's Four Criteria. HumRRO adapted Webb's alignment methodology and developed four criteria for evaluating the content coverage and cognitive complexity of test items, addressing limitations in traditional methods such as lenient cutoff values (HumRRO, 2016). The first three criteria rely on alignment ratings from expert panelists, while the fourth is based on student assessment data. The criterion of 'Items Represent Intended Content' assesses the alignment of test items with content standards, with three levels as no link, partial link, or full link. The criterion of 'Items Represent Intended Categories' evaluates whether items align with designated subscore categories by comparing expected and actual distributions. The criterion of

‘Item DOK Represents Test Specifications’ examines the cognitive rigor of items against expected difficulty distributions. Finally, the criterion of ‘Item Sufficiency for Category Reporting’ determines whether reporting categories are adequately measured using psychometric modeling, including confirmatory factor analysis and reliability estimates.

CCSSO High Quality Assessment Criteria. CCSSO (2014) developed assessment alignment criteria, later refined into rubrics and scoring procedures by the Center for Assessment (NCIEA, 2016). The purpose is to evaluate the content and characteristics of college and career readiness assessments, particularly those aligned with CCSS. For mathematics, these criteria include focusing strongly on the content most needed for success in later mathematics (C1): the assessments help educators keep students on track to readiness by focusing strongly on the content most needed in each grade or course for later mathematics; assessing a balance of concepts, procedures, and applications (C2) by ensuring score points are distributed evenly across these areas. Connecting practice to content (C3) evaluates whether assessment items align with both mathematical content and practices, requiring at least 90% alignment to fully meet the standard. Requiring a range of cognitive demand (C4) assesses DOK distribution through a DOK index, which must be at least 80% and maintain appropriate representation of higher-level DOK items. Ensuring high-quality items and a variety of item types (C5) requires at least two item formats, including one that demands student-generated responses, and mandates high editorial and technical accuracy, with at least 95% of items free from errors. These criteria establish a structured framework for evaluating the alignment, rigor, and quality of mathematics assessments.

Number of Items to Be Added/Replaced. Based on the Webb’s four criteria framework, another evaluation criterion that concerns the overall alignment of the whole assessment items counts the number of items that need to be replaced or added to make the overall test items meets the alignment requirement. In this framework, an assessment is reported as ‘fully aligned’ if there is no need of item replacement to meet Webb’s four criteria. ‘Acceptable alignment’ is when a test requires minor changes, with one to five items needing replacement. When six to ten items need to be added or replaced to meet the evaluation criteria, the alignment is labeled as ‘slight adjustment. When over ten items need to be added or replaced, the alignment is labeled as ‘major adjustment. Both slight and major modification indicate progressively greater deviations from the minimum alignment criteria, based on standard federal peer review benchmarks (WCEPS, 2017).

Proportion of Matched Standards. This alignment criterion concerns the overall alignment of the standards, which was modified by College Board (2023) based on the Range of Knowledge Correspondence in Webb’s (2007) criteria. It calculates the proportion of standards that are matched by at least one item/standard in the test out of the total number of standards. In studies that align standards to standards, the proportion of standards that are matched by at least one expectation in the test specification is counted. According to Webb (2007), a test form is considered acceptably aligned if there are equal to or more than 50% of standards hit by at least one item. A test form is weakly aligned if 40-50% of standards are hit by at least one item. In

College Board's (2023) alignment report, they modified the criteria to that a test shows very strong alignment if more than 75% of standards are aligned to the test standards, 50-75% for strong alignment, and less than 50% for weak or no alignment.

3. Results

This section summarize the results synthsized across the reviewed alignment studies. The alignment content, state standards framework and blueprint, alignment methods, and alignment results are summarized. The alignment results of each study are summarized in terms of (1) the quantitative evaluation results of alignment based on the specific evaluation framework that the study used, (2) the findings of alignment results, and (3) the alignment agency's recommendations on substituting the test.

Fourteen reports were summarized over 20 alignment studies when some study reports may include more than one alignment study. Within the 20 alignment studies, 13 aligned items to standards and 7 aligned standards to standards. The agencies conducting the alignment studies included the College Board (n = 5), University of Connecticut (n = 1), Maryland Assessment Research Center (n = 4), HumRRO (n = 1), WCEPS (n = 6), and Buros Center for Testing (n = 2). No information was reported about who conducted the alignment study in the Tennessee's report (n = 1). The alignment studies were conducted from 2015 to 2023. The earliest alignment study for the redesigned paper-and-pencil SAT was in 2015 by Rhode Island. The first alignment study for the digital version of SAT was conducted in 2023 by MARC.

Comparison among the State Standard Frameworks

The review study included 14 reports from 12 states which presented 13 standard frameworks. Ten states followed CCSS: Arizona, Connecticut, Delaware, Georgia, Illinois, Maine, Maryland[2], New Mexico, Rhode Island, and Tennessee. The number of content domains and standards of Algebra I for each state are summarized in Table 1. The number of Algebra I standards in each state varies, M = 56.73, SD = 5.80, range = [43, 64]. All states that followed the CCSS framework except Georgia have four domains: (1) Number and Quantity, (2) Algebra, (3) Functions, and (4) Statistics or Statistics and Probability. Georgia used the subdomains in CCSS as domains. As a result, there were 10 domains in Georgia's Algebra I standards. Among the 12 states, Florida and Texas used a different standard framework other than CCSS. The standards in Florida are high school Mathematics Florida Standards (MAFS), which were adapted from CCSS. Texas used Texas College and Career Readiness Standards (CCRS of 2012) and Texas Essential Knowledge and Skills (TEKS). MAFS covers 3 domains: (1) Algebra, (2) Functions, and (3) Statistics and Numeracy; TEKS has 4 domains: (1) Linear Functions, (2) Quadratic Functions and Equations, (3) Exponential Functions and Equations, and (4) Number and Algebraic Methods; and Texas CCRS has 8 domains: (1) Numeric Reasoning, (2) Algebraic Reasoning, (3)Probabilistic Reasoning, (4) Statistical Reasoning, (5) Functions, (6) Problem

[2] Maryland has different Algebra I standards in 2019 and 2023.

Solving and Reasoning, and (7) Communication and Representation, (8) Connections. For a detailed list of standards in each state, see Supplementary Document 1.

Table 1

Number of Domains and Standards of Each State

States	Number of Domains	Number of Subdomains	Number of Standards
CCSS	4	11	70
Arizona	4	11	52
Connecticut	4	10	57
Delaware	4	10	63
Georgia	4	10	54
Illinois	4	10	60
Maine	4	10	63
Maryland (2019)	4	10	53
Maryland (2023)	4	10	49
New Mexico	4	10	63
Rhode Island	4	10	62
Tennessee	4	10	48
Mean	-	-	56.73

Notes. This table only lists the information of states that following the CCSS structure.

Comparison among the Blueprints of State Tests

Only six blueprints were found for 5 states' tests, including Maryland PARCC (2021), Maryland MCAP (2023), Arizona (2018), Maine (2016), Florida MFAS (2017), and Georgia (2018). No blueprint following the CCSS framework was available for Delaware. All blueprints show a similar pattern of item allocation to different domains. In terms of reporting categories, though each of the states follows the original version or a modified version of the CCSS framework in terms of their Algebra I standard, the test blueprint did not always follow the same framework. Only the blueprints of PARCC and MCAP of Maryland used a framework similar to the CCSS framework: *Number & Quantity*, *Algebra*, *Functions*, and *Statistics*. Both Arizona and Florida reported *Algebra* and *Functions* separately, and combined *Number* and *Statistics* as the third reporting category *Statistics and Quantitative Reasoning*; *Statistics and the Number System*. Maine did not report the CCSS domain of *Functions*. Georgia did not report the domain *Number & Quantity* and separated *Algebra* into *Equations* and *Expressions*. In general, the *Algebra* and *Functions* domains have the largest proportion of items allocated, with 20%-63.6% of items allocated to *Algebra*, and 16%-42.9% of items measuring *Functions*. In comparison, the *Number & Quantity* and *Statistics* domains have the fewest items allocated, with 4.2%-18.2% of items measuring *Number & Quantity* and 5.7%-20% of items measuring *Statistics*.

For all alignments except Delaware and Maine, the content standards aligned was Algebra I. For Delaware and Maine, CCSS at Grades 11-12 were aligned, where both Algebra I and Algebra II were aggregated. Given the purpose of the review, we only extracted the information for Algebra I. However, we were not able to separate the alignment results for Algebra I and Algebra II for Delaware and Maine.

Results for the Item-to-Standard Alignment Studies

Out of the 13 studies aligning SAT math items to state content standards, 11 (85%) evaluated the paper and pencil version of the SAT, while 2 (15%) evaluated the digital version. Each paper and pencil version contained 58 items in each test form, while each digital version of the SAT had 40 items. Different states used 2 to 8 panelists ($M = 6.58$, $SD = 1.93$), with Florida used the least. Only Arizona and Maryland invited national experts to rate the alignment of test items. Six alignment studies were conducted online and seven were in person. Nine alignments had only one round of discussion among panelists. One study had one to two rounds of discussion depending on the discrepancy of item ratings. Four alignments for Maryland had two or more rounds of ratings.

Coding Scheme

In three reports (Connecticut, 2016; Delaware & Maine, 2016; Florida, 2017), the three-category alignment was used: an item could be rated as fully aligned, partially aligned, or not aligned with a standard. All other studies used the two-category alignment: an item could be coded as aligned or not aligned with the state standard. All alignment studies allowed multiple coding, namely, an item can be coded or aligned to multiple content standards. Only 2 alignment studies for Florida conducted by the Buros Center for Testing (2017) constrained the number of codes up to two content standards for each item.

Item-to-Standard Alignment Results based on Webb's Four Criteria

Webb's Four Criteria were applied to alignment studies conducted for Maryland (2023 and 2021), Arizona (2021), Florida (2017 and 2018), and Georgia (2018). In total, 12 alignment studies were reviewed under this framework. For the Maryland 2021 report, the values were manually calculated based on reported data, while for the Georgia 2018 report, reporting categories were consolidated into four domains to facilitate analysis. In the Florida 2017 and 2018 reports, *Number & Quantity* was combined with *Statistics* into a reporting category called *Statistics & the Number System.* The results for each of Webb's four criteria—Categorical Concurrence, Range of Knowledge, Balance of Representation, and Depth of Knowledge Consistency—are summarized below and in Table 2.

Categorical Concurrence. Categorical Concurrence results varied across domains and states. In the *Number & Quantity* domain, this criterion was met only in the Maryland 2023 paper-and-pencil SAT alignment (1/12), while seven alignments did not meet the criterion (7/12). Florida did not assess this domain separately (4/12). For *Algebra*, all alignments across all states met this criterion (12/12). The *Functions* domain showed strong alignment, with 11 out of 12 alignments meeting the criterion; the exception was the one for Arizona (1/12). In the *Statistics* domain, alignment was met in four alignments (4/12), partially met in three Maryland 2023 alignments (3/12), and not met in one alignment in Georgia (1/12). In Florida's *Statistics &*

Table 2

Webb's Four Criteria Evaluation Results

State	Year	Number & Quantities (4 Webb's evaluation criteria)				Algebra (4 Webb's evaluation criteria)				Functions (4 Webb's evaluation criteria)				Statistics (4 Webb's evaluation criteria)			
		CC	DOK	ROK	BOR	CC	DOK	ROK	BOR	CC	DOK	ROK	BOR	CC	DOK	ROK	BOR
Maryland	2021	No	Yes	N/A	N/A	Yes	Yes	N/A	N/A	Yes	Yes	N/A	N/A	Yes	Yes	N/A	N/A
	2023	Yes	Yes	Yes	No	Yes	Yes	Yes	Weak	Yes	Yes	Yes	No	Partial	Yes	Yes	Yes
		No	No	No	N/A	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Partial	Yes	Yes	Yes
		No	No	No	N/A	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Weak	Partial	Yes	Yes	Yes
Arizona	2020	No	Weak	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Weak	Yes	Yes	No	No	Yes
		No	Weak	No	N/A	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Weak	Weak	Yes
Florida	2017	No	Yes	N/A	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No	Yes	N/A	Yes
		No	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No
	2018	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Weak	Yes	Yes	Yes	No	Yes
		Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes
Georgia	2018	No	Yes	No	N/A	Yes	Yes	Yes	N/A	Yes	Weak	No	N/A	Yes	Yes	Yes	Yes
		No	No	N/A	N/A	Yes	Yes	No	N/A	Yes	No	No	N/A	No	Yes	No	Yes

Notes. CC = Categorical Concurrence, DOK = Depth of Knowledge, ROK = Range of Knowledge, BOR = Balance of Representation.

the Number System domain, alignment was met in two Florida 2018 alignments (2/4) but not met in two Florida 2017 alignments (2/4).

Range of Knowledge. The Range of Knowledge criterion was reviewed in 11 alignments, as no data were available for Maryland alignment study in 2021. In the *Number & Quantity* domain, the criterion was met only in the Maryland 2023 paper-and-pencil SAT alignment and one Arizona alignment (2/11), while it was not met in six alignments across Maryland, Arizona, and Georgia (6/11). 1 alignment from Georgia was N/A (1/11) and 2 from Florida do not apply (2/11). For *Algebra*, the criterion was not met in one Georgia alignment (1/11) and was marked as not applicable (N/A) in another Georgia alignment due to missing data (1/11); it was met in the remaining alignments (9/11). For *Functions*, this criterion was met in three Maryland 2023 alignments (3/11), weakly met in two alignments in Arizona and Florida (2/11), and not met in six alignments across Arizona, Florida, and Georgia (6/11). In the *Statistics* domain, the criterion was met in four alignments (4/11), weakly met in one Arizona alignment (1/11), and not met in the remaining alignments (6/11). Florida's combined *Statistics & the Number System* domain did not meet this criterion in any of its alignments (4/4).

Balance of Representation. The Balance of Representation criterion was also reviewed in 11 alignments, as no data were available for Maryland 2021 alignment study. In the *Number & Quantity* domain, the criterion was met in one Arizona and two Florida alignment studies (3/11) and not met in the Maryland 2023 paper-and-pencil SAT alignment (1/11). Seven alignments, including two Maryland 2023 digital SAT alignments, were marked as N/A due to the absence of items aligned to this domain (7/11). For *Algebra*, this criterion was met in eight alignments (8/11), weakly met in the Maryland 2023 paper-and-pencil SAT alignment (1/11), and marked as N/A in two Georgia alignments due to missing data (2/11). The *Functions* domain showed alignment in seven cases, including Maryland 2023 digital SAT form 1 (7/11), weak alignment in Maryland 2023 digital SAT form 2 (1/11), and no alignment in the Maryland 2023 paper-and-pencil SAT alignment (1/11). Two Georgia alignments were marked as N/A due to missing data (2/11). In the *Statistics* domain, alignment was met in seven alignment studies (7/11), while Florida listed as a separate domain (4/11). In Florida's combined *Statistics & the Number System* domain, alignment was met in three alignments (3/4), with one alignment marked as N/A due to the absence of items (1/4).

Item-to-Standard Alignment Results based on HumRRO's Four Criteria

HumRRO's four criteria for evaluating item alignment were reported in the Delaware and Maine alignment studied in 2016. Note that this criterion showed that the items represented the intended content and categories effectively, meeting expectations in these areas, which were described in SAT specification. Additionally, item sufficiency for category reporting was achieved at the section, test, or subscore levels. However, the criterion related to items representing the DOK specified was only partially met.

Item-to-Standard Alignment Results based on CCSSO High Quality Assessment Criteria

The CCSSO evaluation criteria were also reported in the Delaware and Maine alignment studies in 2016. Specifically, only criteria C2 to C5 were used in their study. Based on the tentative cutoff values, the criterion for assessing a balance of concepts, procedures, and applications (C2) was partially met, while the criterion for connecting practice to content (C3) was fully met. Similarly, the criterion for ensuring high-quality items and a variety of item types (C5.1) was achieved at the section, test, or subscore level, but the degree of high-quality items (C5.2) was only partially met.

Number of Items to Be Added/Replaced

The criterion on the number of items to be added or replaced was applied in the alignment studies conducted for Arizona (2021), Florida (2018), and Georgia (2016) across six studies. The number of items needing revision or replacement ranged from 4 (7%) to 35 (60%). Based on this evaluation criterion, one alignment indicated the SAT was acceptably aligned (1/6), two showed slight adjustments were needed (2/6), and three revealed major adjustments were required (3/6).

Proportion of Matched Standards by Items

Among the seven reviewed study reports, two reports with four alignment studies (2 for Arizona and 2 for Florida) used this evaluation criterion in their original reports. The other 4 reports presented the frequency of matched standards or displayed the original alignment table between items and state standards except for Maryland's 2021 alignment study, we were able to calculate the proportion of the macthed standards agaist this evaluation criterion for these 4 reports related to 8 alignment studies. The item-to-standard alignment studies (12) had a relatively low proportion of matched standards with a smaller variance–on average 42% (*SD* = 9%) of state standards were aligned by at least one item, ranging from 26% to 58%. Based on the cut-off values (very strong alignment if more than 75% of items aligned to the test standards, 50-75% for strong alignment, and less than 50% for weak or no alignment) sugested for this criterion by the College Board (2023), two test forms were considered strong alignment (Florida, 2017: 53% and 58% for two SAT forms). All other 10 alignments[3] showed weak or unacceptable alignment with proportions ranging from 26% to 45%.

Cognitive Taxonomies

All of item-to-standard alignment studies utilized the Webb's DOK framework. Only the alignment studies for Delaware and Maine (HumRRO, 2016) used cognitive rigor.

Webb's Four Criteria. Four states—Maryland, Arizona, Florida, and Georgia—reported results based on Webb's four criteria, DOK Consistency, among the 12 state for the reviewed alignment studies. In the domain of *Number & Quantity*, the criterion was met only in the Maryland 2021 and 2023 paper-and-pencil SAT alignments and one Georgia alignment study (3/12). It was weakly met in one Arizona alignment (1/12), not met in two Maryland 2023 digital SAT alignments (2/12), and marked as not applicable (N/A) in one Georgia alignment due to missing data (1/12). Florida did not include this domain in its analysis (4/12). For *Algebra*, the criterion was met in 11 alignments, including both paper-and-pencil and digital SAT forms from

[3] The result for Arizona was reported across two test forms which were considered two alignments.

Maryland 2023 (11/12), but it was not met in one Florida alignment study (1/12). In the *Functions* domain, alignment was met in nine alignment studies, including Maryland 2023 alignments across both test forms (9/12), weakly met in one Georgia alignment (1/12), and not met in two alignments from Florida and Georgia (2/12). The *Statistics* domain showed alignment in six alignment studies, including Maryland 2023 alignments on both test forms (6/12), weak alignment in one Arizona alignment (1/12), and no alignment in another Arizona alignment (1/12). Florida excluded Statistics as a separate domain (4/12). Finally, in the combined "*Statistics & the Number System*" category reported by Florida, alignment was met in three Florida alignments (3/4) and not met in one alignment (1/4).

HumRRO's Four Criteria. Two states (Delaware and Maine) reported Item DOK Represents Test Specifications from HumRRO's 4 criteria. The criterion was partially met.

CCSSO High Quality Assessment Criteria was used in two states (Delaware and Main). In this criteria, Requiring a range of cognitive demand (C4) was the criterion used to assess cognitive complexity alignment. The criterion was only partially met.

Results for the Standard-to-Standard Alignment Studies

Among the 20 alignment studies, 7 studies (35%) aligned standards to standards. Among the 7 studies, 6 (86%) evaluated the paper and pencil version of the SAT, while 1 (14%) evaluated the digital version. No information was reported on alignment procedure or methodology (e.g., number of panelists, the workshop location, or the number of alignment discussions) by the College Board's reports for Connecticut, Illinois, New Mexico, Rhode Island, and Texas. Nor was it found in the Tennessee's report. Only the Connecticut alignment study report in 2016 provided the information about the alignment methodology. Nineteen panelists participated in one round of discussion to complete the alignment study. Multiple state standards were allowed to be chosen for one standard in SAT. For each SAT standard, the alignment degree was either a strong match, a weak match, or a no match. Cognitive complexity was not reported in any standards-to-standards alignment studies.

The 7 standard-to-standard alignment used the proportion of matched standards as criterion. The proportion ranged from 45% to 82% (M = 62%, SD = 16%). The proportion of standards matched by at least one item was originally reported by Florida 2018 and Georgia (2016) in 4 alignments. According to the College Board (2023), a proportion larger than 75% is considered a very strong alignment, and a proportion larger than 50% is considered a strong alignment. A proportion lower than 50% is considered a weak alignment. Based on this requirement, three alignment studies showed very strong alignment, one showed strong alignment, and three showed weak alignment.

4. Summary and Discussion

In this review, we synthesized the findings from 20 alignment studies across 12 states. Based on the accumulated evidence, the SAT is potentially an eligible substitute for Algebra I assessment in high school, with the need to supplement with additional materials or revise/replace certain items. In the following sections, we summarize the findings and discuss the implications for the SAT alignment study methodology, the validity of using SAT assessing Algebra I, and recommendations for state stakeholders to consider in designing an alignment study and using the results from an alignment study.

Differences in State Standards and Test Blueprints

The divergence in the reviewed alignment results could partly be due to the different sets of state standards. Despite the total ratification of the Common Core initiative, implementation or adoptation of the whole set of CCSS differed quite a lot across states. Even though many states have adopted CCSS, some states started to repeal the standards after adopting. Furthremore, some other states have never adopted CCSS.

The CCSS stipulates the levels of the math content and knowledge that high school students are expected to gain. However, the CCSS only specifies the standards for high school and leaves space for each state to decide what students are expected to gain for Algebra I, Algebra II, Statistics, and Geometry respectively. It was observed that one state may set one standard in Algebra I while another state may put the same standard in Algebra II. This might increase the inconsistency of findings in the alignment results across different states.

Given that state standards may be different, when referencing to the alignment results from other states, caution needs to be excercised. Among the 11 states reviewed in this study, 10 of them used the CCSS or a modified version. This is consistent with expectation when 42 out of 50 states and DC have adaopted the CCSS, suggesting that CCSS is used widely as a framework for high school learning, instruction, and assessments. However, not all blueprints of state Algebra I tests followed the CCSS framework closely, when different subscores or reporting categories were used.

Differences in Alignment Methodology

The methodology used in the reviewed alignment studies varied significantly. First, differences were observed in the number of panelists involved and the number of discussion rounds completed. Among the states, Arizona employed the most panelists, while Maryland conducted the most rounds of discussion. Webb (2007) recommends using 5-8 panelists in alignment studies, as a larger number of panelists increases reliability. However, some studies involved fewer panelists, with as few as three, potentially compromising the reliability and validity of the alignment results. While it is reasonable to consider personnel and time costs when determining the number of panelists, a balance must be struck between ensuring reliable results and cost efficiency considerations. Similarly, increasing the number of discussion rounds can enhance alignment reliability among raters, but excessive discussions may compromise the independence of individual raters. On the other hand, some alignment reports lack sufficient details about data collection procedures, making it difficult for readers to evaluate the quality of the study. For example, reports for the Connecticut SAT digital test, New Mexico, Rhode Island,

and Texas only presented alignment results without describing how the alignment was conducted. Without transparency, it becomes challenging to assess the reliability and rigor of these studies.

Second, a significant limitation was observed in the standards-to-standards alignment approach. The primary purpose of alignment studies is to determine whether test items align with state standards, rather than whether test standards from another test align with state standards. Even if the SAT standards may align well with state standards, the individual SAT items may not perfectly reflect the distribution of the state standards. As a result, the method of aligning standards to standards (standard-to-standard alignment) is insufficient to determine whether SAT items are suitable for assessing state standards. In other words, alignment at the standard level does not guarantee alignment at the item level, which undermines the validity of using the SAT math items for state Algebra I tests. Therefore, the method of standard-to-standard alignment alone is not sufficient as a validation method. To ensure a more reliable and valid investigation, we strongly recommend states adopt an item-to-standard alignment approach.

Another big difference lies in the vendors who conducted the alignment studies. According to NCME (2019), alignment evidence should be collected by an independent party and made public. Alignments conducted by the College Board violated this principle, as their studies lacked the necessary independence to ensure unbiased results. Additionally, differences in the alginment levels used for item alignment affect the precision of alignment findings. Dichotomous approaches (aligned/not aligned) provide less detailed information compared to ternary categorization, which retains more nuanced data about alignment levels.

In summary, alignment methodology varies among studies in several key areas, including the approach to alignment (standard-to-standard vs. item-to-standard), transparency of procedures, number of panelists, rounds of discussion, independence of the vendor, and the levels used to quantify the degree of alignment. To improve the quality and reliability of alignment study results, we recommend increased transparency, adherence to independence principles that test vendors should not evaluate their own products for the alignment study purposes, and the use of methodologies that capture more precise alignment data. Balancing these factors with the practical constraints of cost and time remains critical.

Differences in Evaluation Criteria

The difference of evaluation frameworks also leads to variation in alignment results and it is a barrier for comparing alignment results across states directly. In our review, nevertheless, Webb's four criteria has been used in some item-to-standard alignment studies, a quite comprehensive evaluation framework considering the content and DOK. However, as Webb (2007) highlighted, the cutoff values could be a potential issue in synthesizing the findings when random cutoff values are applied. Though some agencies aimed to solve these limitations by using multiple evaluation criteria (e.g., HumRRO, 2016), no concensus has been reached. The current decision rule of 50% of the objectives with at least one hit clearly is a minimal requirement for alignment (Webb, 2007). CCSSO high-quality criteria also mentioned that the cutoff values are very tentative. We thus call for a more standardized procedure to carry out the alignment with the limitations resolved. This is particularly a concern to count the percentage of standards matched by at least one item when the number of items is fewer items than the number of standards.

Major Findings from the Reviewed Alignment Studies

The SAT was designed for the purpose of college and career readiness. Results indicate that the SAT aligns well with states' high school math standards in core conceptual areas, particularly *Number & Quantity*, *Algebra*, and *Statistics (& Probability)*. Furthermore, while *Functions* is covered, it often has a weaker alignment compared to the other domains. However, despite the SAT's domain sampling strategy that represents most broad categories, some alignment studies have revealed gaps in breadth of knowledge. For example, nearly two-fifths of a state's specific math standards might not be measured. Indeed, educators in states like Connecticut have warned that the breadth and depth of their Core Standards may surpass what the SAT covered. This situation can lead to concerns among educators who worry that instruction may be narrowed to what is tested (Connecticut, 2016). Moreover, while not every state's standards fit perfectly, such as those of Texas, New Mexico, or Rhode Island, these states typically claim that the SAT's omission of advanced or specialized topics does not undermine its overall alignment with essential college and career readiness domains.

In conclusion, the majority of these alignment studies converge on a common ground: the SAT is a strong measure of central high school math concepts, particularly in Algebra and related quantitative reasoning skills, while occasionally leaving out some topics or not fully addressing the depth of certain domains. Consequently, many states consider augmenting the SAT with supplemental assessments or instructional measures to ensure complete coverage of their standards.

Alignment of Content Standards

Based on this review, the alignment results differed based on different alignment approaches (item-to-standard and standard-to-standard) and different forms (digital and paper-and-pencil forms). Among the reviewed studies, two states (i.e., Maryland and Connecticut) included SAT digital forms while others aligned items on the paper-and-pencil forms. The traditional paper-and-pencil version of SAT showed a clear pattern of alignment based on the 11 alignment studies from 6 states. In general, three out of four domains in Algebra I are largely covered by SAT items, such as *Algebra*, *Functions*, and *Statistics & Probability*. However, based on Webb's four criteria of content, the biggest concern was on *Number & Quantity* and *Statistics* regarding Categorical Concurrence; more concerns on *Number & Quantity*, *Functions*, and *Statistics*, regarding Range of Knowledge and *Number & Quantity* and *Functions* regarding Balance of Representation. However, it is obvious that the range of knowledge and balance of representation are not perfectly met in many alignments, indicating the distribution of SAT items has its own weights assigned to the same content standards when it was designed as a national college admission test rather than mirroring state standards. For those standards not covered by

SAT, it would be necessary for states to consider adapting the SAT by adding items or using complementary assessment results. In summary, the results from the item-to-standard alignment studies were not very strong when the proportion of CCSS measured in SAT was less than 50%; one of Webb's alignment criteria, Balance of Representation, was not met in most domains. Among the four domains, Algebra aligned the best, with most criteria satisfied. This indicates that Algebra tested in SAT has a large overlap with Algebra required in CCSS.

As noted, the SAT math includes a considerable number of items in more advanced math topics (e.g., Geometry, Trigonometry) and some lower-level math (e.g., middle school Algebra and Probability) which is particularly true for the digital SAT math for lower ability students. The coverage of additional topics in Algebra II or Geometry may be considered as more challenging topics when evaluating the overal rigor of SAT math in place of state Algebra I tests.

Since only one criterion–the proportion of matched standards–was used in both approaches, our comparison between the two approaches was based on the results under this criterion. In comparison, the results from the standard-to-standard alignment studies are much more positive than the item-to-standard alignment. Aligning SAT math standards to state Algebra I test standards is a more lenient process than aligning individual SAT test items to the state standards, making it unsurprising that the results appear more favorable. In summary, the decision to substitute a state Algebra I test with the SAT math largely depends on the alignment approach used. The synthsized findings support what was noted in Connecticut's alignment study (2016) that the SAT math covered the narrower range of the state standards.

Aligment of Cognitive Complexity

Only the item-to-standard approach align the cognitive complexity of items. All alignment studies reviewed in this study implemented Webb's DOK levels to measure the cognitive complexity of items. DOK Consistency criterion from Webb was utilized most frequently in these studies, with only one report (Delaware & Maine, 2016) referencing to other cognitive complexity evaluation frameworks from HumRRO and CCSSO.

Among the reviewed item-to-standard alignment studies, the findings varied significantly across domains. In *Number & Quantity*, DOK Consistency was met in only 3 out of 12 alignments, weakly met in 2 studies, and not met in 3 studies. Four studies for Florida did not report this domain separately. These results suggest that SAT items may not consistently capture the DOK required for this domain. In *Algebra*, alignment of DOK was the strongest, with 11 of 12 studies meeting the criterion, indicating that SAT items align well with expected cognitive complexity in this area. For *Functions*, alignment was met in 9 studies, weakly met in one, and not met in two, reflecting good match between SAT items and the expected DOK for items examining the knowledge of functions. For *Statistics*, alignment was achieved in six studies, weakly met in one case, and not met in another. In the *Statistics & the Number System* domain reported by Florida, alignment was met in three out of four alignments and not met in one, indicating alignment in cognitive complexity in this combined category. These results suggest that while some domains, particularly *Algebra*, exhibit strong alignment, others, such as *Number & Quantity*, show unmatched requirements of cognitive complexity between SAT and state standards, raising concerns about the SAT's ability to fully assess cognitive rigor across all mathematical domains.

Implications

Alginment studies often intend to collect validity evidence related to using test scores from SAT math to meet the requirements on state Algebra I tests. Content alignment and cognitive complexity alignment are two important facets of the validity evidence intended to be collect in such studies. Based on the reviews of the studies and the findings synthesized across alignment studies reviewed in this study, we would like to highlight the following considerations to be taken into account in designing an alignment study and interpreting the results from an alignment study.

1. Item-to-standard or standard-tot-standard alignment methods: we strongly encourage the use of item-to-standard approach as the alignment is test form specific. Items on one test form may not assure the same level of alignment of items on another forms. The reason is that test form construction targets at the domain or subdomains levels, not the standard level. Item level evaluation will provide more detailed validity evidence related to form equivalence.
2. Independece: we recommend having independent agency to conduct alignment studies instead of test vendors to evaluate the alignment of items they created and pulled on different test forms.
3. Logistic considerations: factors including the number of panelists, the representation of the panelists who know better of different student populations, and the rounds of reviewing should be well defended in balancing the technical quality and cost efficiency in designing an alignment study.
4. Evaluation criteria: given different evaluation frameworks have been used in different alignment studies. It is worthy exploring the technical defensibility and the utility of the evidence extracted from each evaluation framework to decision making.
5. Additional assessment sources: given the narrower content coverage of the state Algebra I standards by the SAT math items, state should consider how additional items or sources would be technically defensible to supplement what is missing from SAT math.

In conclusion, consistent with Camera and colleagues' findings, the most challenging requirement for using existing college admissions tests in place of state Algebra I tests is to demonstrate alignment of SAT math tiems to the state content standards, or to demonstrate equivalent breadth and depth to a customized state assessment (Camara et al., 2019; Landl, 2020). There is not a rule-of-thumb that guarantees a satisfactory alignment (Camara et al., 2019; Webb, 2007).

Limitations and Future Explorations

One limitation of the current study is the inclusion of the alignment studies only from some states. Due to each state's disclosure regulations, not all alignment reports were open to the public. Future research should seek to find all available alignment reports across the U.S., and do a more comprehensive review to examine whether the findings are consistent and replicable across methods and standards, especially for the digital SAT math.

To sum up, all reviewed alignment studies showed a weak to strong alignment depending on the aligned objects, i.e., items or standards. Within each alignment approach, the alignment result was consistent. For the standard-to-standard alignment, results showed a strong alignment with proportion of covered standards up to 96%. Whereas for the item-to-standard alignment, results showed a weak to moderate alignment, with only 46% of standards covered in the SAT. However, if the SAT math is deemed as a substitution for state Algebra I test to meet the policy requirements, it is highly recommended that policymakers consider the use of additional assessments such as classroom assessments on the missing topics as supplementary evaluation to avoid narrowing instruction by focusing the content standards that are only covered by the SAT math.

Given the potential subjectivity and variability introduced by human panelists, artificial intelligence (AI) offers a promising tool for increasing consistency in item alignment. For instance, Gweon and Schonlau (2023) used a pre-trained language model, Bidirectional Encoder Representations from Transformers (BERT; Delvin et al., 2019) to automatically classify open-ended survey questions into predefined content domains, showing improved performance with larger training samples. Meanwhile, studies by Qiao and Hu (2019) and Meissner et al. (2020) leveraged machine learning and natural language processing techniques to assign items to Bloom's Taxonomy levels, a common indicator of cognitive complexity. These approaches suggest that AI can streamline item alignment on content domain as well as cognitive complexity. Furthermore, in recent years, there has been a clear shift from feature-based text mining (Qiao & Hu, 2019) and classical machine learning algorithms (e.g., support vector machines; Liu et al., 2018) to more advanced transformer-based models, which often achieve higher accuracy in text classification tasks (e.g., Gweon & Schonlau, 2023; Wang et al., 2023). Yet, the direct application of the state-of-the-art AI techniques to conduct alignment between assessment items and state standards need further exploration before such technique can be fully applied to large-scale, high-stakes testing programs. Thus, future research should explore AI-based approaches in conducting item alignment, and focus on evaluating its accuracy for large-scale assessment programs given its known cost efficiency. Such efforts would not only strengthen the methodological rigor of alignment studies but also help mitigate the challenges of subjectivity and resource intensity in manual review processes.

References

Achieve, Inc. (2006). *Closing the expectations gap, 2006: An annual 50-state progress report on the alignment of high school policies with the demands of college and work*. Washington, DC: Achieve, Inc.

AlFallay, I. S. (2017). Test specifications and blueprints: Reality and expectations. *International journal of instruction*, *11*(1), 195-210.

Behuniak, P., Goldstein, J., & DiBlasio, M. (2016). *Alignment of the Connecticut core standards to the Connecticut SAT school day*. Storrs, CT: University of Connecticut. Retrieved from https://portal.ct.gov/-/media/SDE/Student Assessment/SAT/Connecticut_SAT_Alignment_Report_Final_June_2016.pdf

Bridgeman, B., & Stone, E. (2017). Selection links: Alignment of operational and external test forms [Technical report]. BRT Projects. https://brtprojects.org/wp-content/uploads/2022/08/Selection-Links-AlignmentORExt-Spring2017.pdf

Christopherson, S. C., & Webb, N. L. (2018). Alignment analysis of the ACT and SAT with the Georgia standards of excellence for American literature and composition, Algebra I, Geometry, and Biology. *Wisconsin Center for Education Products & Services. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/General%20Presentations/Independent_Alignment_Study_ACT_SAT.pdf.*

Christopherson, S. C., Webb, N. L., & Messinger, M. (2020). Alignment analysis of two forms of the SAT with the Arizona academic standards for English language arts Grades 11-12, Algebra 1, and Geometry. *Wisconsin Center for Education Products & Services.* https://www. azed. gov/sites/default/files/media/WebbAlign_WCEPS_AZ% 20SAT% 20Ali gnment% 20Report, 2011252020.

College Board. (2015). *Test specifications for the redesigned SAT*. https://collegereadiness.collegeboard.org/pdf/test-specifications-redesigned-sat-1.pdf

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171-4186.

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Gweon, H., & Schonlau, M. (2024). Automated classification for open-ended questions with BERT. *Journal of Survey Statistics and Methodology*, *12*(2), 493-504.

Herman, J., & Linn, R. (2013). *On the road to assessing deeper learning: The status of smarter balanced and PARCC assessment consortia*. CRESST Report 823. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Khan Academy. Accessed April 2, 2021. https://www.khanacademy.org.

Maryland Classroom. (2013). *Common Core and PARCC Transitioning to the New Maryland Common Core State Curriculum and Next Generation Assessments*. https://www.acpsmd.org/cms/lib/MD01907365/Centricity/Domain/47/01.%20Maryland%20Classroom%20Common%20Core%20and%20PARCC.pdf

McCormick, C. & Geisinger, K. F. (2017). Buros Center for Testing Alignment Study Full Report. *Buros Center for Testing.*

Meissner, R., Jenatschke, D., & Thor, A. (2020, October). Evaluation of approaches for automatic e-assessment item annotation with levels of Bloom's taxonomy. In *International Symposium on Emerging Technologies for Education* (pp. 57-69). Cham: Springer International Publishing.

National Center for Education Statistics. (2018). *Table 8.3. Mathematics statewide high school assessments, by state: 2017–18*. https://nces.ed.gov/programs/statereform/tab8_3.asp

Nemeth, Y., Michaels, H., Wiley, C., & Chen, J. (2016). Delaware system of student assessment and Maine comprehensive assessment system: SAT alignment to the common core state standards. *HumRRO*. Retrieved from https://www.maine.gov/doe/sites/maine.gov.doe/files/inline-files/SAT%20Alignment%20Final%20Report_revised%2008092017.pdf

Porter, A. C., & Smithson, J. L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. Yearbook-National Society for the Study of Education, 2, 60-80.

Qiao, C., & Hu, X. (2019). Text classification for cognitive domains: A case using lexical, syntactic and semantic features. *Journal of Information Science*, *45*(4), 516-528.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2004). Benchmarking and alignment of standards and testing. *Educational assessment, 9*(1-2), 1-27.

Roeber, E., Olson, J., Topol, B., Webb, N., Christopherson, S., Perie, M., Pace, J., Lazarus, S., & Thurlow, M. (2018). *Feasibility of the use of the ACT and SAT in lieu of Florida statewide assessments: Volume 1: Final Report.* Assessment Solutions Group. Retrieved from http://www.fldoe.org/core/fileparse.php/5663/urlt/ACTSATFSA.pdf

Skinner, R. R., & Feder, J. (2014). Common core state standards and assessments: Background and issues.

Wang, T., Stelter, K., Floyd, J., O'Neill, T., Hendrix, N., Bazemore, A., ... & Newton, W. (2023). Blueprinting the Future: Automatic Item Categorization using Hierarchical Zero-Shot and Few-Shot Classifiers. *arXiv preprint arXiv:2312.03561*.

Webb, N. L. (1995, April). Issues Related to Judging the Alignment of Curriculum Standards and Assessments. Paper presented at the annual meeting of the American Educational Research Association. Montreal, CA.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and mathematics education* (Research Monograph No. 6). Washington, DC: Council of Chief State Schools Officers.

Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. *Language Arts*, *28*(March), 1-9.

Webb, N. L. (2005). *Webb alignment tool: Training manual.* Madison, WI: Wisconsin Center for Education Research. Available: http://www.wcer.wisc.edu/WAT/index.aspx

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*(1), 7–25.

Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment of mathematics state-level standards and assessments: The role of reviewer agreement. *Educational Measurement: Issues and Practice*, *26*(2), 17–29.

Appendix A

Appendix A.1

Summary of the Alignment Studies Using the Item-to-Standard Approach

State	Report No.	Year	Authors/ Agency	Paper/ Digital	Rounds of Discussion	Number of Panelists	Workshop Venue	Alignment Degree	Multiple Coding	Evaluation Framework	Proportion of Matched Standards	Overall Alignment
Maryland	1	2021	MARC	paper	3+cross-validation	5	virtual	binary	allow	-	-	-
	2	2023	MARC	paper	up to 3 + cross-validation	8	virtual	binary	allow multiple coding	Webb's four criteria	43%	Weak/No Alignment
				digital	up to 3 + cross-validation	8	virtual	binary	allow multiple coding	Webb's four criteria	43%	Weak/No Alignment
				digital	up to 3 + cross-validation	8	virtual	binary	allow multiple coding	Webb's four criteria	43%	Weak/No Alignment
Arizona	3	2020	WCEPS	paper	1 + review	5	virtual	binary	allow multiple coding	1. Webb's four criteria 2. Number to be revised/ replaced	35%	Weak/No Alignment
				paper	1 + review	5	virtual	binary	allow multiple coding	1. Webb's four criteria 2. Number to be revised/ replaced		
Delaware & Maine	4	2016	HumRRO	paper	1-2	7	in-person	three categories (ternary)	allow multiple coding	1. HumRRO's four criteria 2. CCSSO criteria	44%	Weak/No alignment

Florida***	5	2017	Buros Center for Testing	paper	1	7	in-person	three categories (ternary)	allow up to 2 coding for each item	Webb's four criteria	**53%**	Strong alignment
				paper	1	8	in-person	three categories (ternary)	allow up to 2 coding for each item	Webb's four criteria	**58%**	Strong alignment
	6	2018	WCEPS	paper	1 (with review)	2	in-person	binary	allow multiple coding	1. Webb's four criteria 2. Number to be added/ replaced 3. Proportion of matched standards	44%	Weak/No alignment
				paper	1 (with review)	no info provided	in-person			1. Webb's four criteria 2. Number to be added/ replaced 3. Proportion of matched standards	40%	Weak/No alignment
Georgia§	7	2018	WCEPS	paper	1	8	in-person	binary	allow multiple coding	1. Webb's four criteria 2. Number to be added/ replaced 3. Proportion of matched standards	36%	Weak/No alignment
				paper	1	8	in-person			1. Webb's four criteria	26%	Weak/No alignment

										2. Number to be added/ replaced 3. Proportion of matched standards		

Notes. Overall alignment is based on the proportion of matched standards, where the cutoff values followed the version adapted by the College Board as weak/no alignment (proportion <50%), strong alignment (50%=< proportion <75%), very strong alignment (proportion>=75%)

Appendix A.2

Summary of the Alignment Study Using the Standard-to-Standard Approach

State	Report No.	Year	Authors/ Agency	Paper/ Digital	Rounds of Discussion	Number of Panelists	Workshop Venue	Alignment Degree	Multiple Coding	Alignment Framework	Proportion of Matched Standards	Overall Alignment
Connecticut	1	2016	University of Connecticut	paper	1 (with review)	19	in-person	three categories	allow multiple coding	Mapping list between standards	49%	Weak/No alignment
	2	2023	College Board	digital	no info provided	no info provided	no info provided	binary	allow multiple coding	Proportion of matched standards	77%	Very strong alignment
New Mexico	3	2016	College Board	paper	no info provided	no info provided	no info provided	binary	allow multiple coding	Proportion of matched standards	45%	Weak/No alignment
Rhode Island	4	2015	College Board	paper	no info provided	no info provided	no info provided	binary	allow multiple coding	Proportion of matched standards	45%	Weak/No alignment
Tennessee	5	2018	No info provided	paper	no info provided	no info provided	no info provided	binary	allow multiple coding	Mapping list between standards	58%	Strong alignment
Texas	6	2020	College Board	paper	no info provided	no info provided	no info provided	binary	allow multiple coding	Proportion of matched standards	82%	Very strong alignment
Illinois	7	no info provided	no info provided	paper	no info provided	no info provided	no info provided	binary	allow multiple coding	No info provided	75%	Very Strong alignment

Notes. Overall alignment is based on the proportion of matched standards, where the cutoff values followed the version adapted by the College Board as weak/no alignment (proportion <50%), strong alignment (50%=< proportion <75%), very strong alignment (proportion>=75%)