Smarter Balanced Assessment Consortium:

Alignment Study Report

Developed by HumRRO

12/30/2014

Alignment Study Report

EXECUTIVE SUMMARY

Background

Test validation involves marshalling evidence in support of an argument that inferences and interpretation based on test scores are warranted. A critical part of test interpretation validation is to demonstrate that the test measures what it claims to measure. For modern standards-based assessments, the contention typically is that the assessment allows claims to be made about student performance in relation to a set of content standards. The goal of this project was to gather evidence to examine the validity of Smarter Balanced summative assessments in terms of their alignment to the Common Core State Standards (CCSS).

The intent of the Smarter Balanced assessments is to make valid inferences about students and to offer valid interpretations of test scores in terms of the CCSS based on students' performance on the Smarter Balanced summative assessments. This is a challenging task because the CCSS are broad, rich, and comprehensive; students are assessed using selected and constructed responses with a variety of innovative item formats; and the assessments are administered via computer-adaptive testing (CAT).

The validity of intended inferences, interpretations, and claims is based on the connection between the CCSS and the Smarter Balanced assessments. The CCSS-Smarter Balanced connection, however, is not simple and direct but rather is supported by a sequence of component connections. The components represent increasingly focused specifications guiding the item and task development process and moving from the broad and general CCSS to specific items and tasks with their associated scoring rubrics. The strength of the validity argument for the Smarter Balanced assessments depends directly on the strength of the connections between the various components used in the development process to move from the CCSS to specific items and tasks to which students respond. A glossary of CCSS and content specification terms is at Appendix A. The connections examined in this alignment study are presented in Figure ES.1.

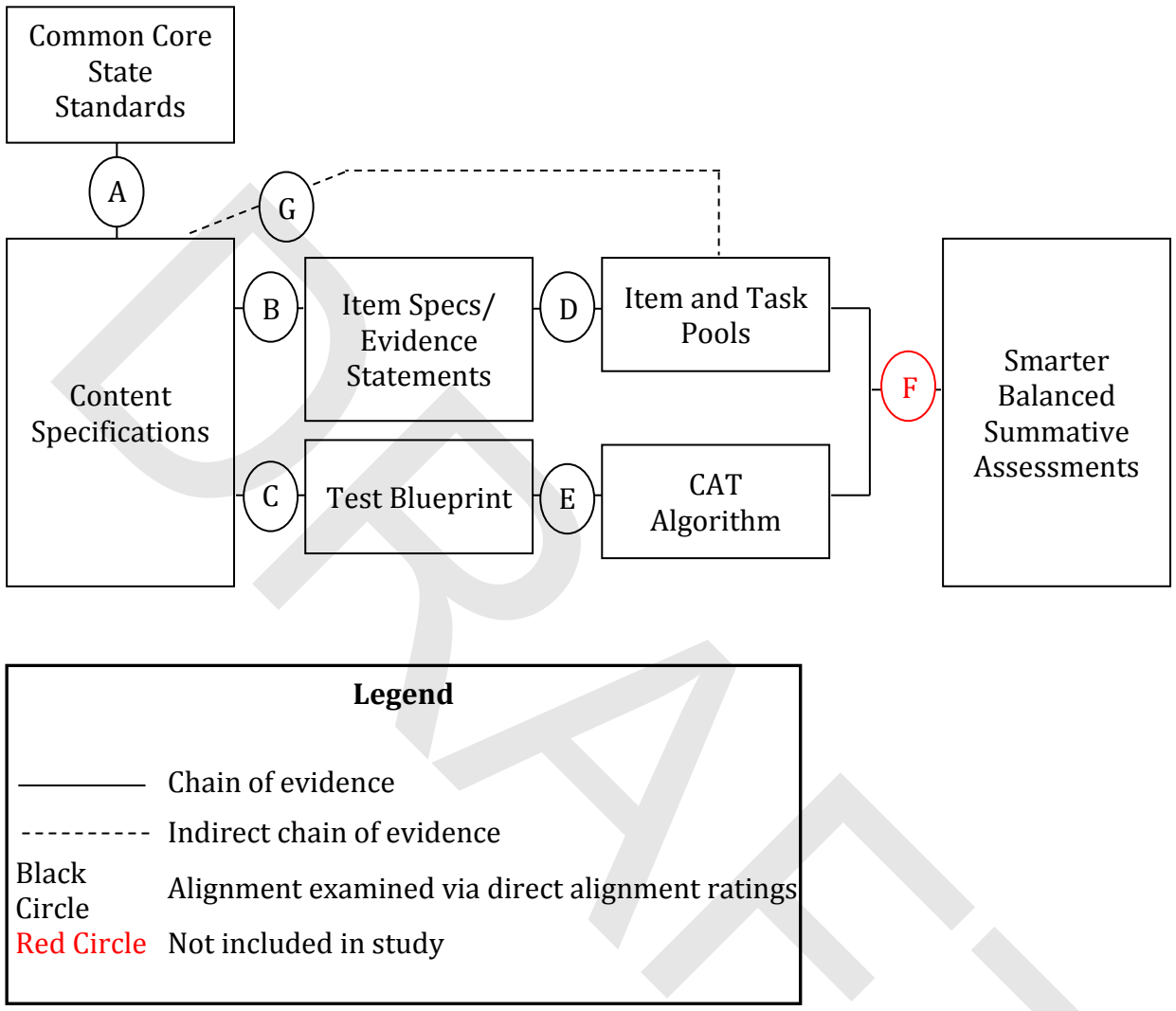


Figure ES.1. Connections examined in the alignment study.

Alignment Reviewers

A total of 223 reviewers provided data that were used to examine Connections A through D and G. The reviewers represented 19 states within the Smarter Balanced Assessment Consortium. The majority of reviewers were educators who taught in rural areas; however, there were some reviewers who also either taught or worked in suburban and urban areas. At least two-thirds of the reviewers had experience working with English language learners and more than three-fourths of them had experience working with students with disabilities.

All reviewers were trained to make specific alignment ratings. Training was specific to the type of alignment ratings the reviewers were asked to make. They were trained on the factors to consider when making their assigned ratings as well as the procedures to do so. They also were trained on how to access and navigate specific computer software to make their ratings. As reviewers completed their assigned activities, they were monitored, and remedial training and additional

guidance were provided, as needed. Based on reviewer feedback following their participation, reviewers generally agreed or strongly agreed that the presentations, training, guidance, materials and rating forms, and use of Excel and laptops were useful, and/or clear and understandable.

Alignment Procedures

Approximately half of all field test items and a limited number of performance tasks were included in this alignment study. A stratified random sampling approach was used to ensure representation across all targets and claims within each grade level and content area. The numbers of computer-adaptive test (CAT) and performance task (PT) items included in this study are provided in the body of the report. Data related to Connections A through D and G were gathered across a series of five workshops:

- Workshop 1: Examined the alignment between the Content Specifications[1] and the CCSS (Connection A) and examined the alignment between the Content Specifications and the test blueprints (Connection C).
- Workshop 2: Examined the alignment between the CCSS and the Content Specifications (Connection A).
- Workshops 3 - 5: Examined the alignment between the evidence statements and Content Specifications (Connection B), alignment between the evidence statements and items (Connection D), and alignment between items/performance tasks and Content Specifications (Connection G).

Connection E (test blueprints to the CAT algorithm) was examined by qualified HumRRO researchers familiar with computer-adaptive testing (CAT) and the development of CAT algorithms. However, because CAT documentation was incomplete at the time of the study, only an initial examination was possible.

Alignment Criteria

Connections A through D and G were evaluated and analyzed based on the following alignment criteria:

- Content Representation: The degree to which the content within an assessment component (e.g., claim, target, grade-level standard) was aligned to another assessment component (e.g., the percentage of targets that were aligned to more than one evidence statement).
- DOK Distribution: The breadth of cognitive demand associated with the elements or components included in this study that have DOK ranges assigned to them, such as claims, evidence statements or the CCSS MPs. We examined the percentage of these components at each DOK level (i.e., 1, 2, 3, and 4). Evaluating the DOK distribution included comparing ratings from reviewers in this study (for the target, evidence statement, item, etc.) to the DOK identified by the developer, which was indicated in the Content Specifications.

[1] This alignment study referred to the Smarter Balanced ELA/Literacy Content Specifications dated October 3, 2013, the updated ELA/Literacy Appendix B dated January 29, 2014, and the Mathematics Content Specifications dated June 2013. When necessary, additional information was obtained from the Smarter Balanced ELA/Literacy Item Specifications dated February 3, 2014 and the Mathematics Item Specifications dated February 5, 2014 (v 2.0).

- DOK Consistency: The extent to which the DOK of the item or evidence statement was consistent with the consensus DOK derived for the grade-level standards by the reviewers, and the Smarter Balanced claims and targets.
- Agreement between Reviewers' and Content Specifications/Developers' Ratings: The degree to which there was consistency in ratings of reviewers and Content Specifications/item developers, in terms of indication of DOK and content match.
- Agreement among Reviewers' Ratings: the degree to which the different reviewers' ratings were consistent (i.e., inter-rater reliability) in terms of DOK and content match.

Summary of Findings

A brief summary of the findings related to each connection examined in this alignment study are presented below.

Connection A: Alignment between the Content Specifications and CCSS

Alignment was examined through multiple analyses that primarily focused on the overlap of the (a) content required in both the grade-level standards and the targets, and (b) cognitive demand required in both the grade-level standards and the targets. Because the Smarter Balanced Content Specifications indicate the intended alignment, when relevant, we integrated these comparisons into the primary analyses. This was in contrast to typical alignment studies that do not take into account the intended alignment. By doing so, we were able to provide information on the relative perceptions of alignment as well as the degree of validity of the reviewers' ratings.

There were two main reviewer tasks for Connection A:

(1) Workshop 1: Four to five reviewers in each grade identified grade-level standards that they believed represented the content and knowledge required in the target, including the:
   a. Number and content of the grade-level standards identified by reviewers.
   b. Degree to which the content and knowledge required in the target was represented by the collective set of grade-level standards they identified.
   c. Cognitive demand (DOK) required by the targets (independent, blind ratings[2]).
   d. Cognitive demand (DOK) required by the grade-level standards (reviewer consensus[3], blind ratings).

(2) Workshop 2: Four to five reviewers in each grade (different reviewers from Workshop 1) identified assessment targets they believed represented the content and knowledge required in each grade-level standard.

[2] Blind ratings refer to the reviewers not having access to materials that would otherwise allow them to identify the intended cognitive demand.
[3] Reviewer consensus was achieved first by looking at the agreement among all reviewers within a group, followed by a group discussion to determine the rating identified by the majority of the group.

Results Summary

Below is a summary of the main findings related to the alignment between the Content Specifications and the CCSS. Summary results are presented by criterion for both ELA/literacy and mathematics. For mathematics, analyses were conducted using all the mathematics targets and also by disaggregating the targets by emphasis (major vs. additional and supporting). Supporting tables for the emphasis breakdown are available in the Appendix; however, no clear patterns emerged to suggest a differential alignment between the major and additional and supporting mathematics targets and the CCSS.

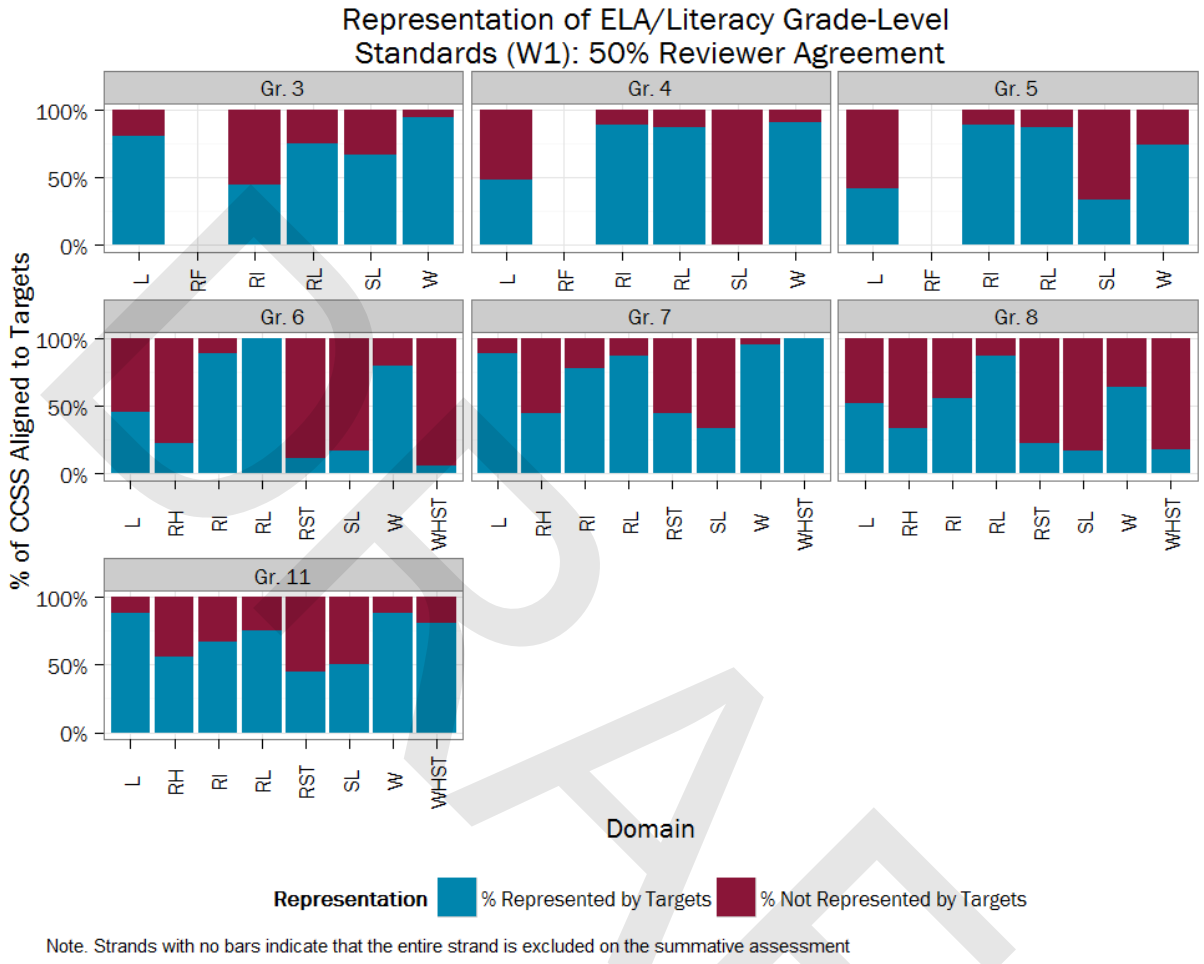
Content Representation

A pertinent piece of evidence to strengthen the validity argument involves examining the breadth of representation of the CCSS across the targets (Workshop 1) and the breadth of representation of the targets across the CCSS (Workshop 2). These results are intended to provide very granular depictions of the alignment between the content in the CCSS and the content in the targets. These results do not consider the *intended* coverage at the target level (i.e., specific mappings of grade-level standards to targets as identified in the Content Specifications). Identification of a grade-level standard (or target) is included when, at the target level (or grade-level standard level), at least 50% of the reviewers agreed on that specific mapping.

Across all grades, 64.7% of the eligible[4] ELA/literacy grade-level standards were identified by at least 50% of the reviewers. As seen in Figure ES.2, some clear patterns emerge across grades. The Speaking and Listening (SL) strand was not covered as well as other strands, specifically for the upper grades. This was expected as Smarter Balanced did not intend to measure speaking on the summative assessment. Also, the Literacy standards (RH, Reading for History/Social Studies; RST, Reading for Science & Technical Subjects; WHST, Writing for History/Social Studies and Science & Technical Subjects) for grades 6-12 were not covered as well as the grade-specific strands.

Across all grades, 76.7% of the eligible[5] mathematics grade-level standards were identified by at least 50% of the reviewers (see Figure ES.3). With the exception of high school, the domains are represented well by the targets.

[4] Grade-level standards that were solely measured by an ELA/literacy target and are not planned to be assessed on the summative assessment (e.g., Claim 3, Target 1) were excluded from the possible grade-level standards to be measured (see Appendix D).
[5] HS grade-level standards that are excluded from the summative assessment (see Appendix D).



Note. Strands with no bars indicate that the entire strand is excluded on the summative assessment

Figure ES.2. Representation of ELA/literacy grade-level standards across targets.

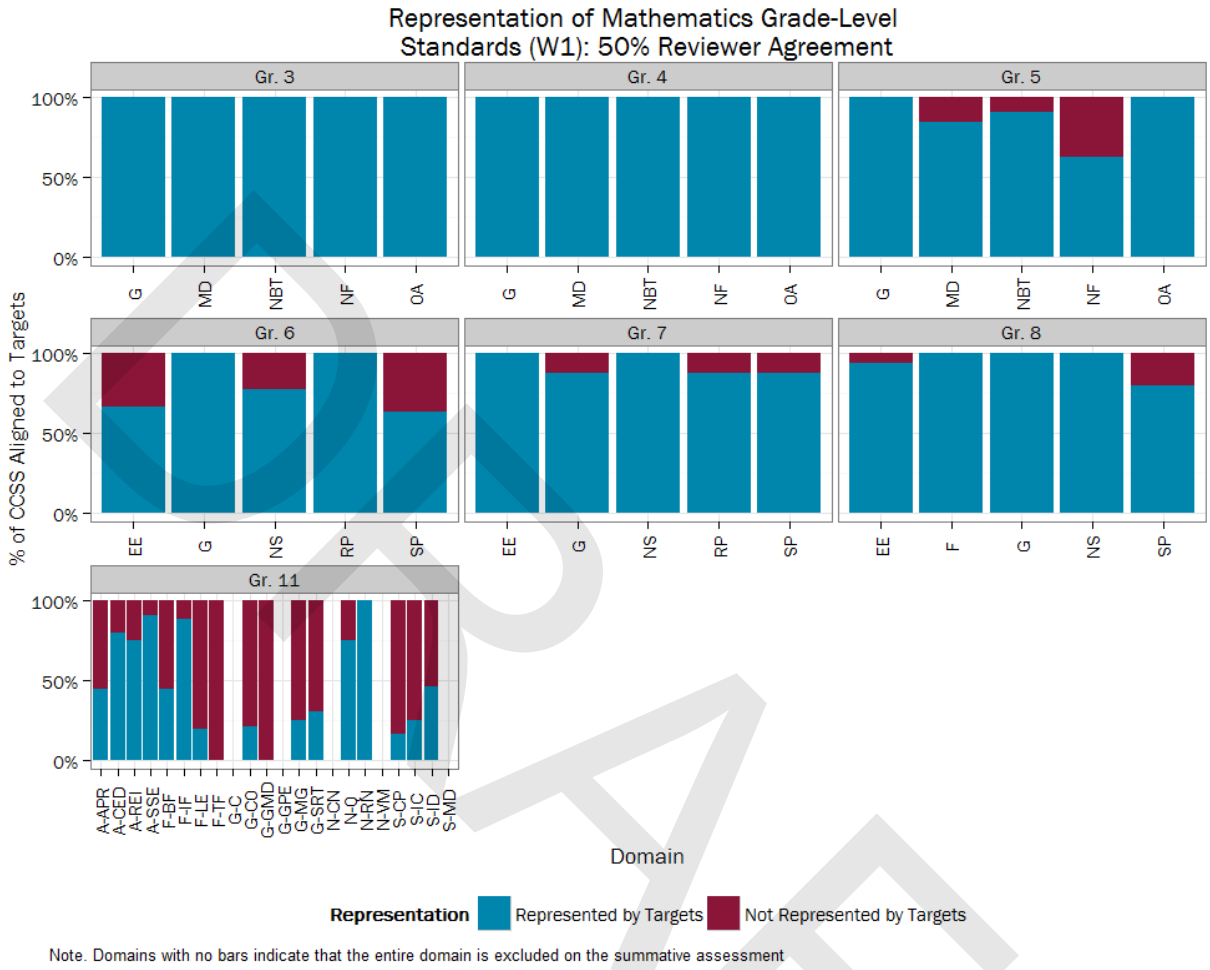
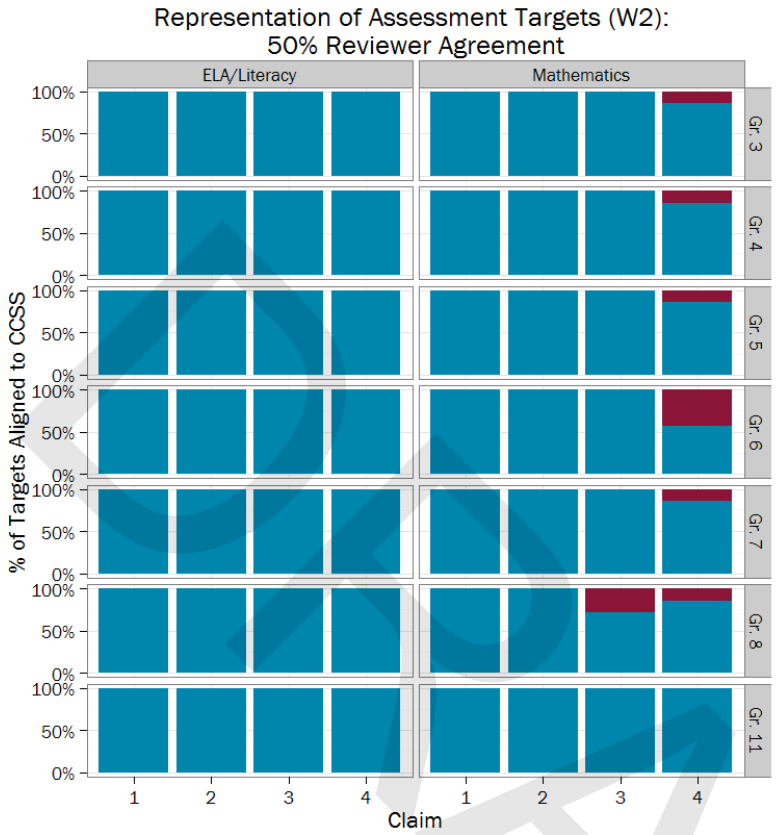


Figure ES.3. Representation of mathematics grade-level standards across targets.

Examination of aligning assessment targets to each of the grade-level standards (Workshop 2) shows that overall for ELA/literacy, 100% of the targets[6] were represented throughout the grade-level standards Figure ES.4). Across grades for mathematics, 96.4% of the targets were represented. This was not unexpected, however, because the majority of the targets that were not represented by the grade-level standards came from Claims 2-4, which were designed to be aligned to the mathematical practices rather than aligned to the grade-level standards.

[6] Please note, performance task targets were included in analyses.



Math Targets Not Represented

Grade	Claim	Target
3	4	3.4.G
4	4	4.4.G
5	4	5.4.G
6	4	6.4.E
6	4	6.4.F
6	4	6.4.G
7	4	7.4.G
8	3	8.3.D
8	3	8.3.G
8	4	8.4.G

Figure ES.4. Representation of ELA/literacy and mathematics targets across grade-level standards.

A.CR-1: Do the grade-level standards collectively reflect the content and skills required by the target?

Data to support this question were drawn from the holistic target ratings from Workshop 1. As shown in Figure ES.5, reviewers for both ELA/Literacy and mathematics rated the majority of the targets as being fully represented by the collective set of grade-level standards that they identified. Generally, Claim 3 targets, for both content areas (ELA/literacy, Speaking and Listening; mathematics, Communicating Reasoning) had the weakest alignment ratings. Because the mathematics Claims 2 – 4 targets were not intended to be aligned to the grade-level standards, this was not an unexpected finding.

Figure ES.5. Mean percentage of targets at each holistic target rating.

A.CR-3: Do the individual grade-level standards reflect the content and skills required by the intended targets?

Data to support this question were drawn from the reviewer grade-level standards mappings, which were compared to those identified in the Content Specifications.

Because reviewers could identify as many grade-level standards per target as they believed mapped, Figure ES.6 highlights that reviewers often identified more grade-level standards per target than were indicated in the Content Specifications (blue line). To determine which of the grade-level standards per target were identified by most of the reviewers, analyses were also conducted using only those grade-level standards for which at least 50% of the reviewers agreed that it mapped (green line). For ELA/literacy, there is no consistent pattern across grades; in some cases, reviewers identified a few more grade-level standards (e.g., grades 4 and 11, both green lines), a few less (e.g., grade 6, green line), quite a few more (e.g., grade 7, both blue and green lines), or the same number of grade-level standards per target (grades 3, 5, and 8, all green lines) as were indicated in the Content Specifications.

For mathematics, this analysis examined the grade-level standards identified for Claim 1 separately from those identified for targets in Claims 2 – 4. This was done because Claim 1 targets were intended to be mapped to the grade-level standards while Claims 2 – 4 targets were intended to be mapped to mathematical practices. Although reviewers identified more grade-level standards for targets in Claim 1, the number was fairly similar to the number of grade-level standards that were indicated in the Content Specifications. For mathematics Claims 2 – 4, reviewers identified substantially more grade-level standards per target than were indicated in the Content Specifications.

These results suggest that identifying which grade-level standards were most represented by the content and knowledge required by the targets was a challenging task. Many reviewers commented that the content and knowledge stated in the targets was too broadly defined (which is not necessarily to be interpreted negatively) to be able to identify accurate matches with the grade-level standards.

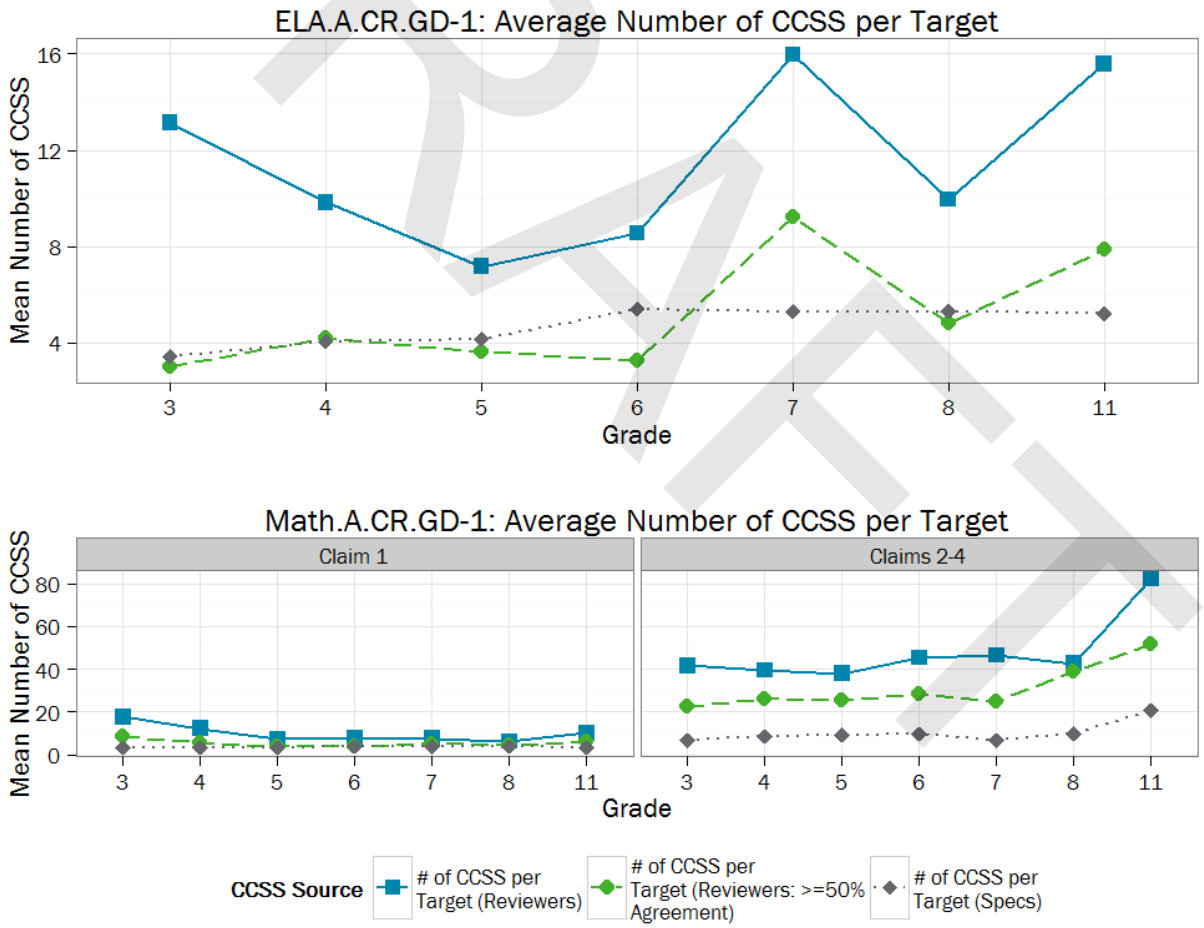


Figure ES.6. Number of grade-level standards per target comparisons for ELA/literacy and mathematics

Another analysis that provides information on how well the content in the grade-level standards overlaps with the content in the targets involves comparing the grade-level standards that at least 50% of the reviewers agreed mapped to a target. As shown in Figure ES.7, generally across grades, claims, and content areas, on average approximately 50% of the grade-level standards mapped to a target matched the intended standards. That percentages increase when the reviewers' mappings of grade-level standards are compared to the intended domains/strands, indicating that while it was more difficult for reviewers to identify all of the mapped standards identified in the specifications, they did identify the intended domains at higher rates.

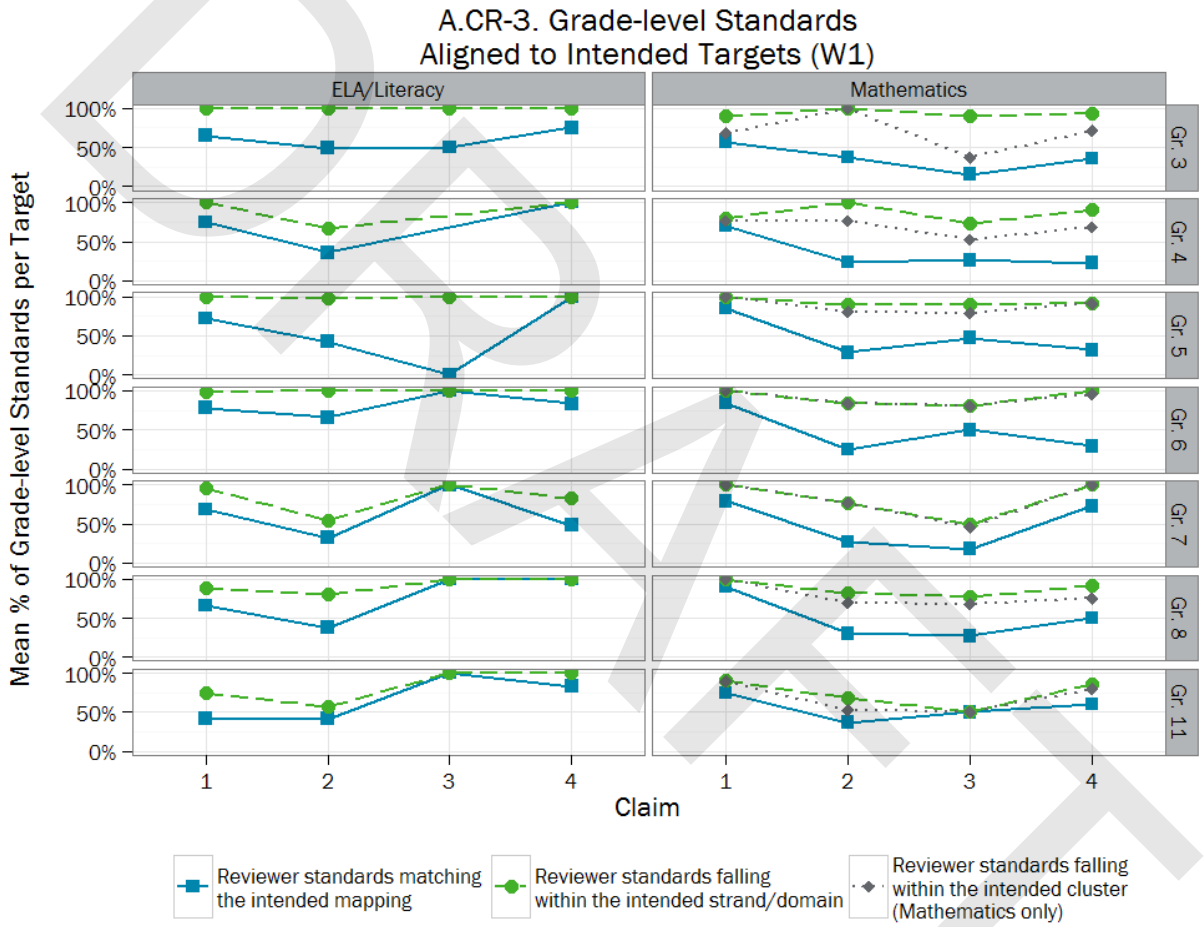


Figure ES.7. Mean percentage of reviewer grade-level standards aligned to intended target for ELA/literacy and mathematics

A.CR-5: Does each mathematical practice reflect skills required by the intended target?

Figure ES.8 shows that the typically, each mathematical practice was aligned to at least 50% of the targets in each claim. As there is no expectation that the skills in every target would represent the skills in every mathematical practice, this finding suggests that across targets the mathematical practices are generally well represented. Additionally, the larger number of mathematical practices in Claim 1 that have fewer percentages of aligned targets is not alarming as the Claim 1 targets were

developed to align to the grade-level standards and targets in Claims 2 through 4 were developed to represent the mathematical practices.

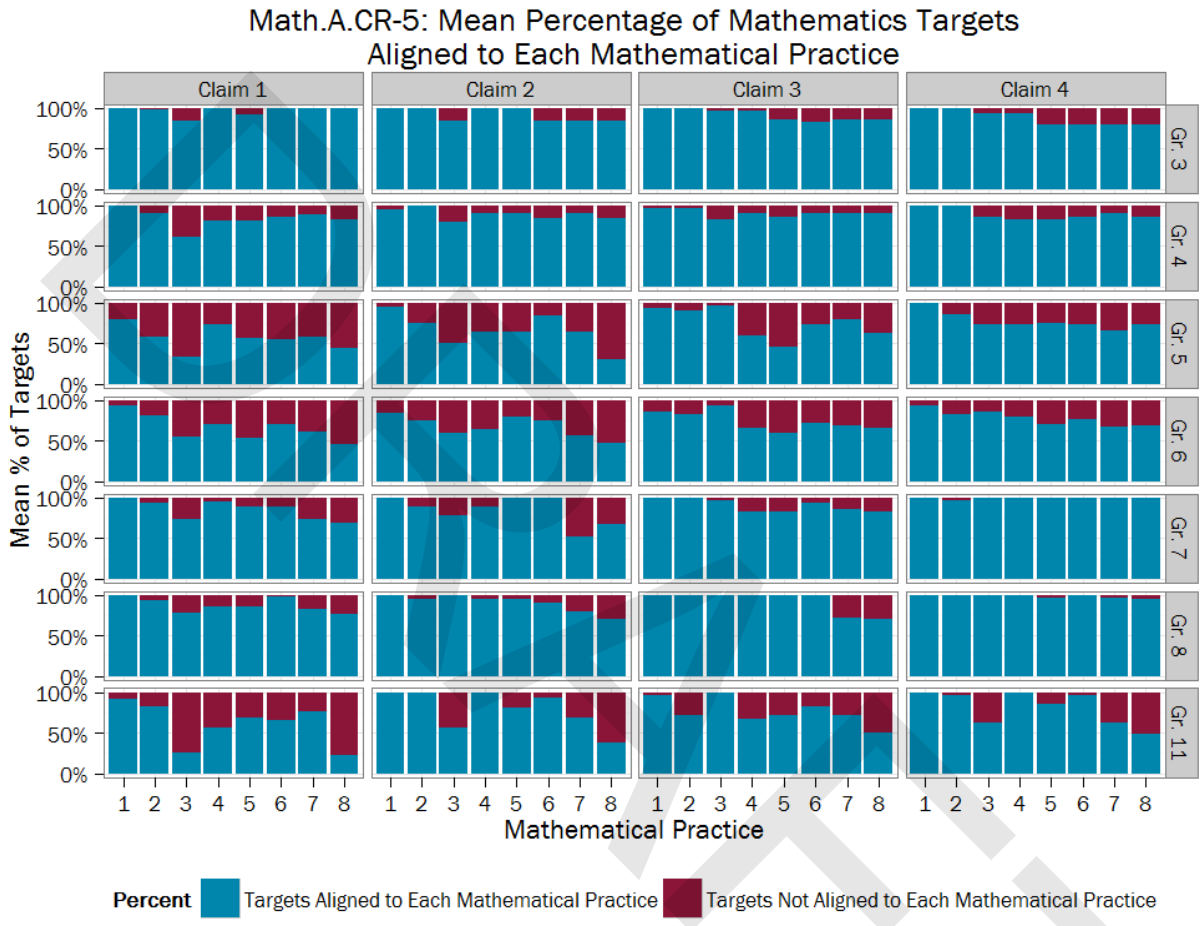


Figure ES.8. Mean percentage of targets aligned to each mathematical practice

DOK Distribution

Because the targets can require multiple levels of cognitive demand, it is useful to see how many DOK levels reviewers indicated versus that what is specified in the Content Specifications. Generally, for ELA/literacy, reviewers rated Claim 1 and 2 targets as requiring more levels of cognitive demand than what was intended and rated Claim 3[7] and 4 targets as requiring fewer levels of cognitive demand than what was intended. For mathematics, the pattern is less clear; it appears that the lower grades and high school identified more levels than what was intended, but grades 7 and 8, on average, identified fewer levels per target than did the specifications (see Figure ES.9).

[7] Of the four Claim 3 targets, three were classroom-based and therefore were excluded from our analyses.

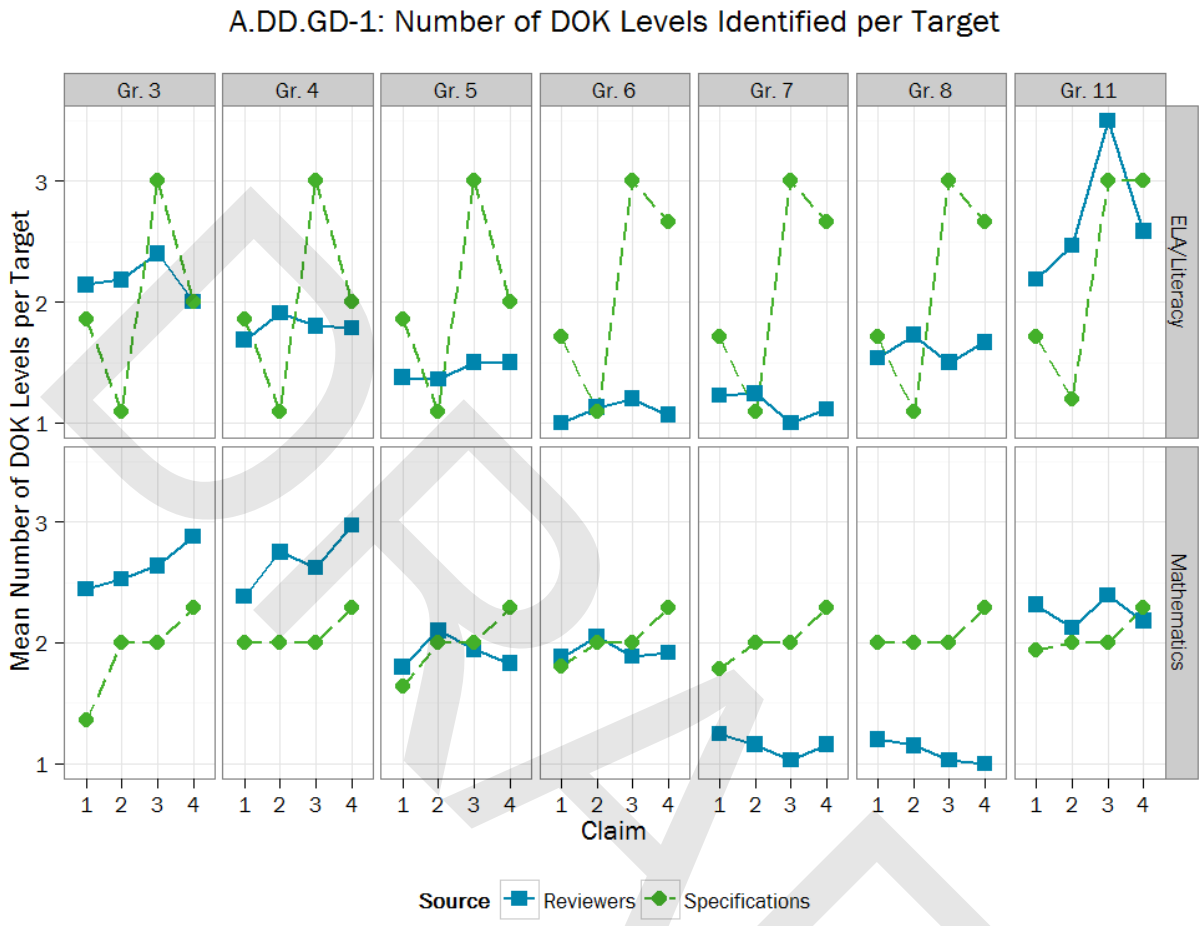


Figure ES.9. Comparison of Number of DOK levels as rated by reviewers and indicated in the Content Specifications.[8]

[8] Recall this alignment study referred to the Smarter Balanced ELA/Literacy Content Specifications dated October 3, 2013, the updated ELA/Literacy Appendix B dated January 29, 2014, and the Mathematics Content Specifications dated June 2013. When necessary, additional information was obtained from the Smarter Balanced ELA/Literacy Item Specifications dated February 3, 2014 and the Mathematics Item Specifications dated February 5, 2014 (v 2.0).

A.DD-1: Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the Content Specifications (using the max DOK level)?

Across grades, reviewers generally agreed with the maximum DOK level indicated in the Content Specifications for ELA/literacy Claim 1 and 4 targets (Figure ES.10). Reviewers generally believed the maximum DOK level was higher for Claim 2 targets as compared to the Content Specifications, which indicated they were at DOK levels 2, 3, and 4.

For mathematics, reviewers generally agreed with the maximum DOK level indicated in the Content Specifications for targets in Claims 2 – 4 (Figure ES.11). However, for Claim 1 targets, reviewers across grades believed more of these targets had a maximum DOK level that was higher than indicated in the Content Specifications.

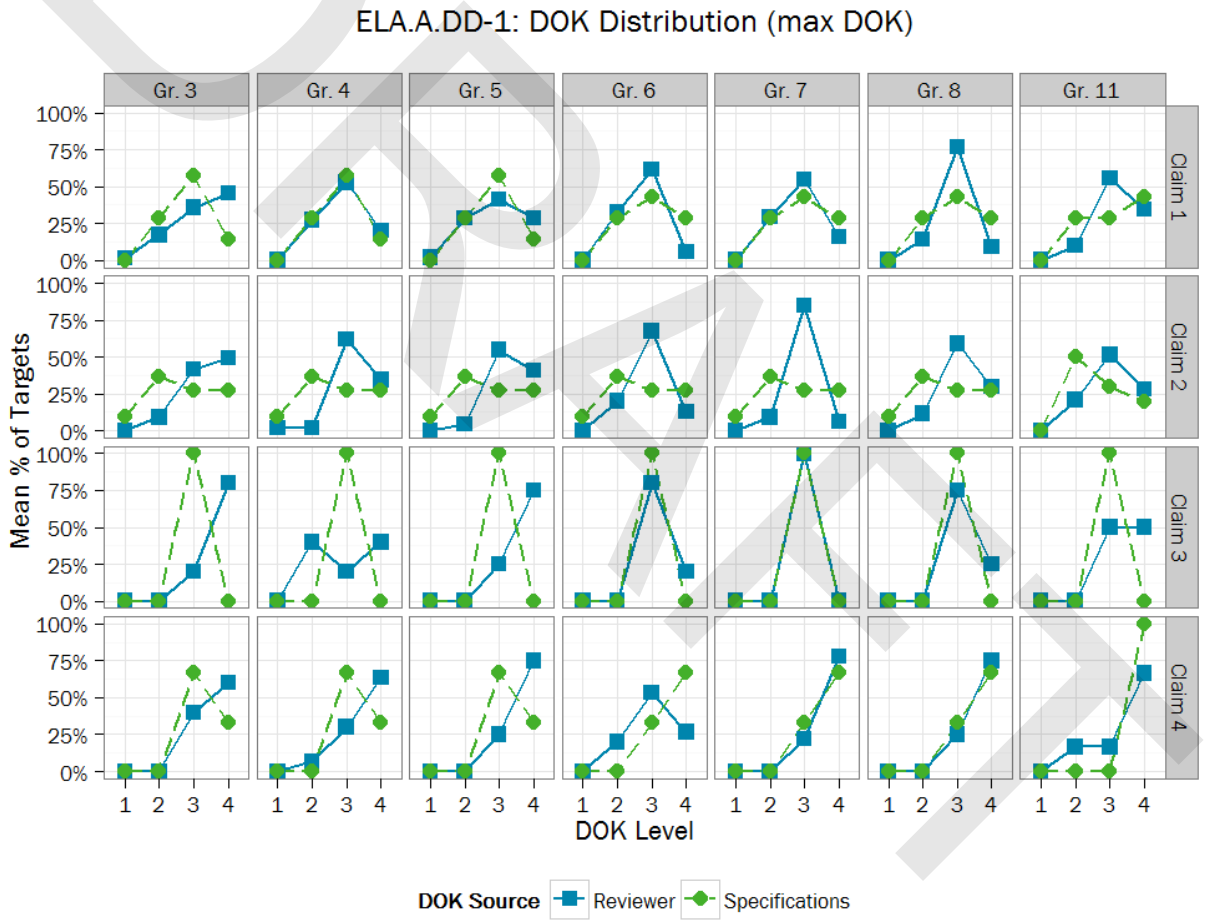


Figure ES.10. Mean percentage of ELA/Literacy targets at each (max) DOK rating.

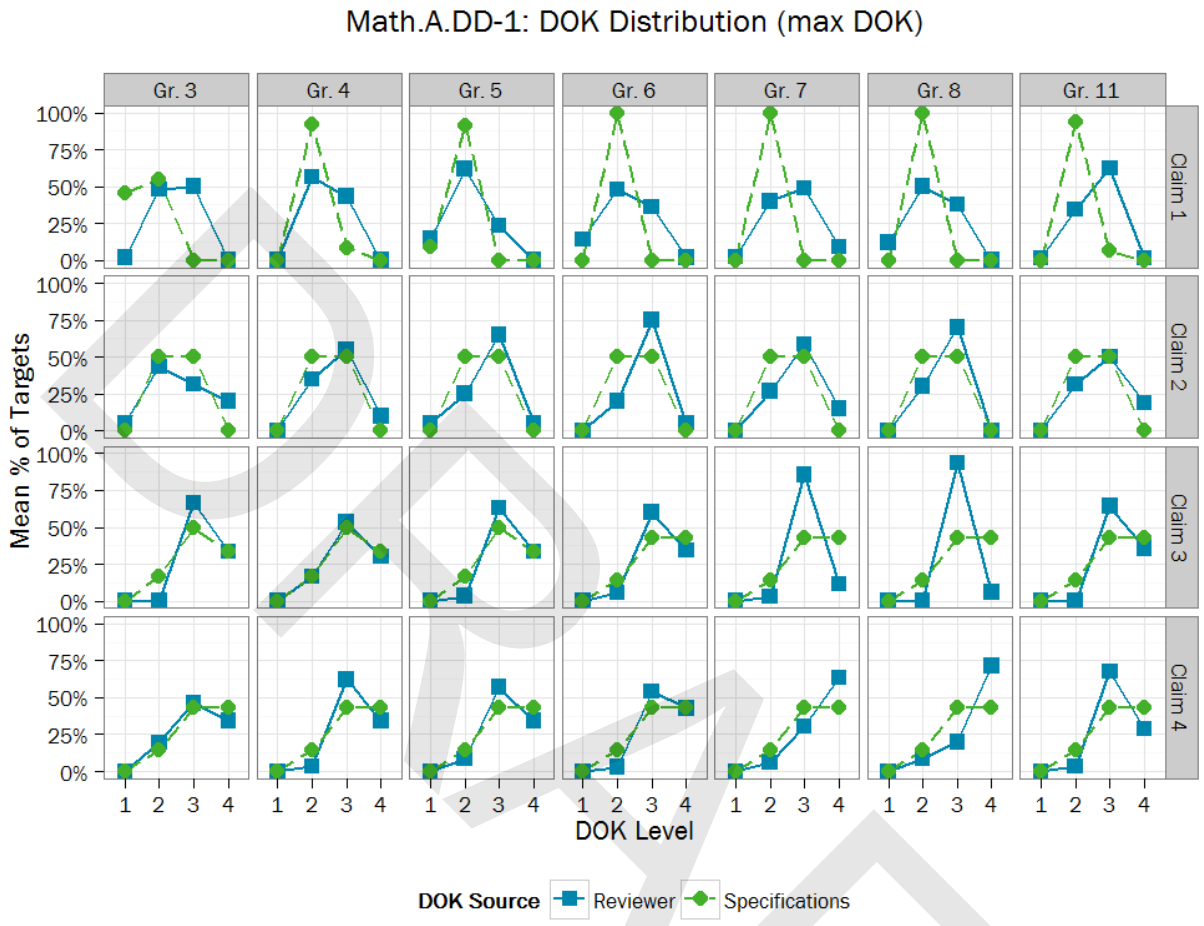


Figure ES.11. Mean percentage of mathematics targets at each (max) DOK rating.

DOK Consistency

The fact that both grade-level standards and the targets had multiple levels of cognitive demand posed a challenging obstacle for determining the DOK consistency between the two. For both ELA\literacy and mathematics targets, no real patterns emerged in identifying which targets were consistent[9] in terms of DOK levels with their mapped grade-level standards (Figures ES.12 and ES.13[10]).

[9]Consistency was defined as the DOK levels falling entirely within the range of the intended target DOK levels for all the mapped grade-level standards with 50% reviewer agreement

[10]This analysis was not conducted for mathematics Claims 2 -4 as there were no grade-level standards mapped to individual targets in Claims 2 -4. Targets from these claims were designed to be aligned to the mathematical practices.

When the DOK consistency criterion was relaxed to only require the grade-level standards mapped to a target and to have at least one DOK level fall within the range of the intended target DOK, the DOK consistency of the targets with their mapped grade-level standards increases across grades and claims for both content areas. This suggests that defining DOK consistency when multiple DOK levels are permitted makes for a challenging analysis. Requiring all of the DOK levels for a grade-level standard to fall within the range of the intended DOK of the target likely resulted in a definition of DOK consistency that was too difficult to achieve in practice. Only requiring at least one DOK level to match between the grade-level standard and the intended target, however, could result in a DOK definition that is too broadly defined. For example, assume that a target has two grade-level standards aligned to it. If the identified DOK for both grade-level standards are 2 and 3 and the intended DOK of the target are 1 and 2, the target would be labeled as being consistent in DOK with that standard. The maximum required cognitive demand for the grade-level standards, however, could reach a higher cognitive demand than what would be required by the target. Ultimately, both definitions yield useful information regarding the DOK consistency between the targets and the grade-level standards when multiple DOK levels are permitted.

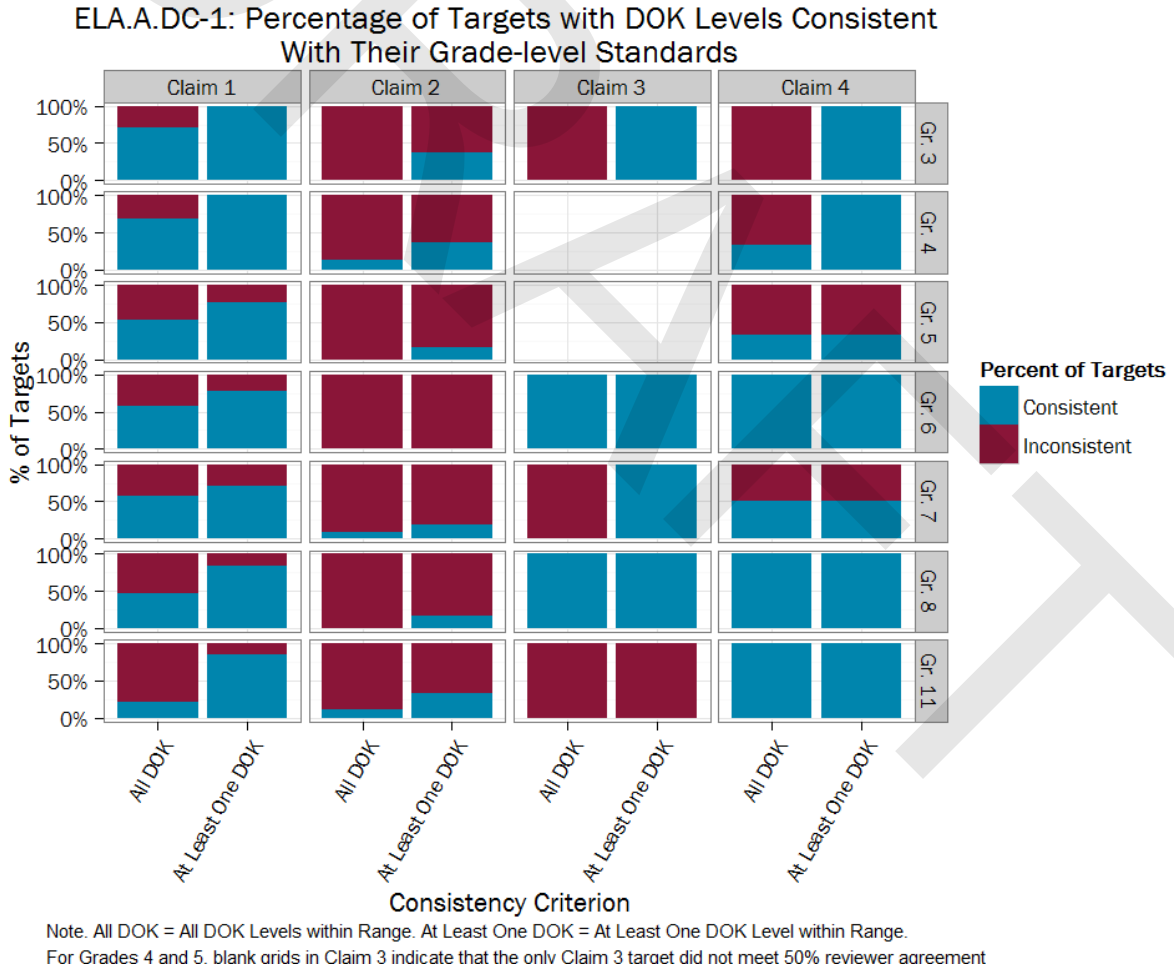


Figure ES.12. Percentage of ELA/Literacy Targets with DOK Levels Consistent with Their Grade-level Standards

Figure ES.13. Percentage of Mathematics Targets with DOK Levels Consistent with Their Grade-level Standards

To provide additional information regarding the DOK consistency, we further examined the ways that the targets and grade-level standards were inconsistent. As shown in Figures ES.14 and ES.15, which displays the mean and standard deviations of the maximum reviewer consensus DOKs for the grade-level standards mapped to the inconsistent targets, it became evident that targets were being rated as "inconsistent" because at least one of the mapped grade-level standards fell outside the

range of the target DOK. In ELA/literacy, reviewers rated the grade-level standards for Claim 2 targets as requiring *lower* levels of cognitive demand than what was intended and *higher* levels of cognitive demand for grade-level standards for the Claim 3 target. Generally, for Claim 1 targets, with the exception of grades 7 and 11, reviewers rated the grade-level standards as requiring higher levels of cognitive demand than what was intended by the targets.

In mathematics, reviewers generally rated the grade-level standards as requiring a higher upper boundary of cognitive demand than what was intended. This is particularly useful to explain why grade 7 had no targets that had DOK consistency with their mapped grade-level standards. Additionally, for mathematics, this can be partially explained because Claim 1 often measures only part of a particular standard, while Claims 2 – 4 would cover the other parts.

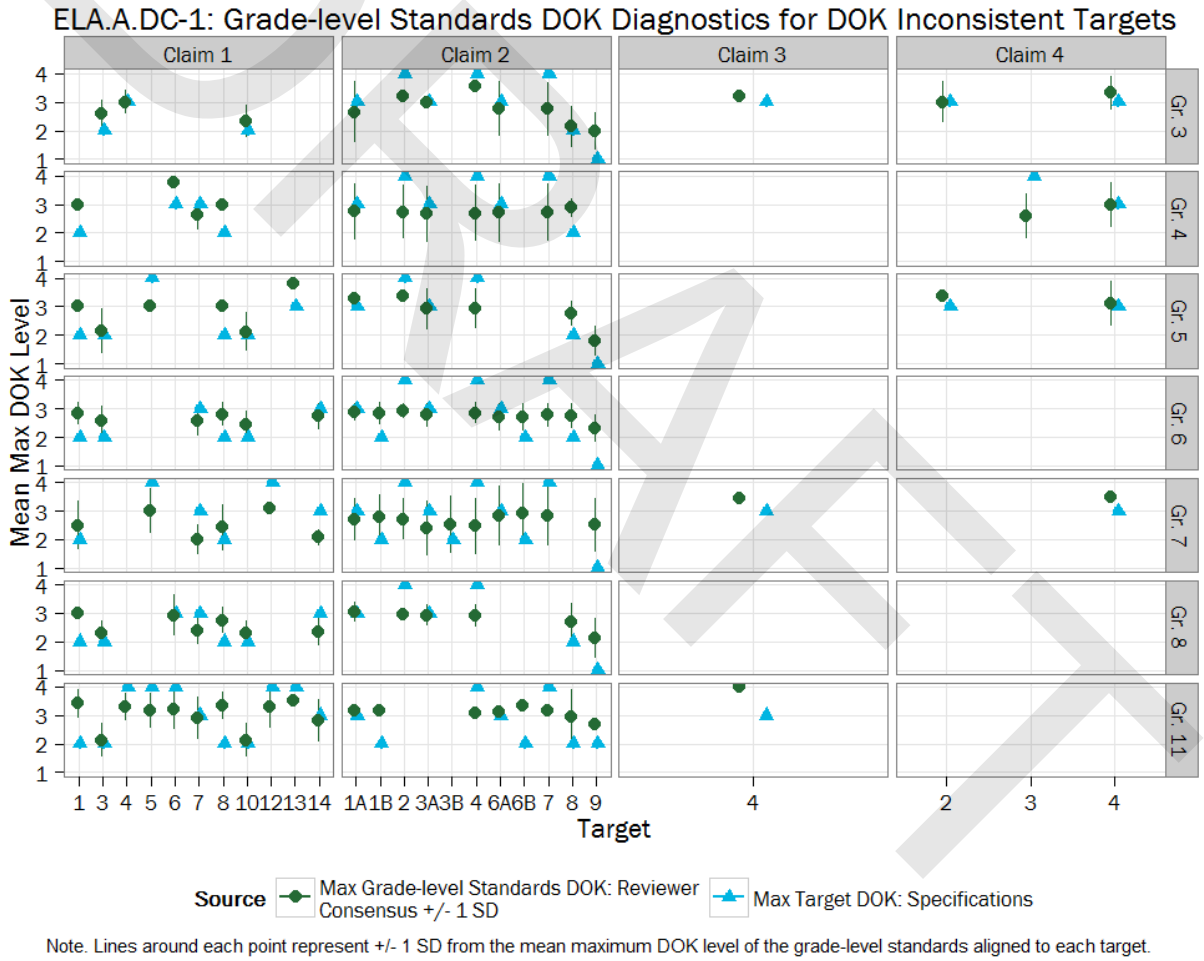


Figure ES.14. Comparison of max target DOK (specifications) and max grade-level standards DOKs (reviewers) for ELA/Literacy "DOK inconsistent" targets.

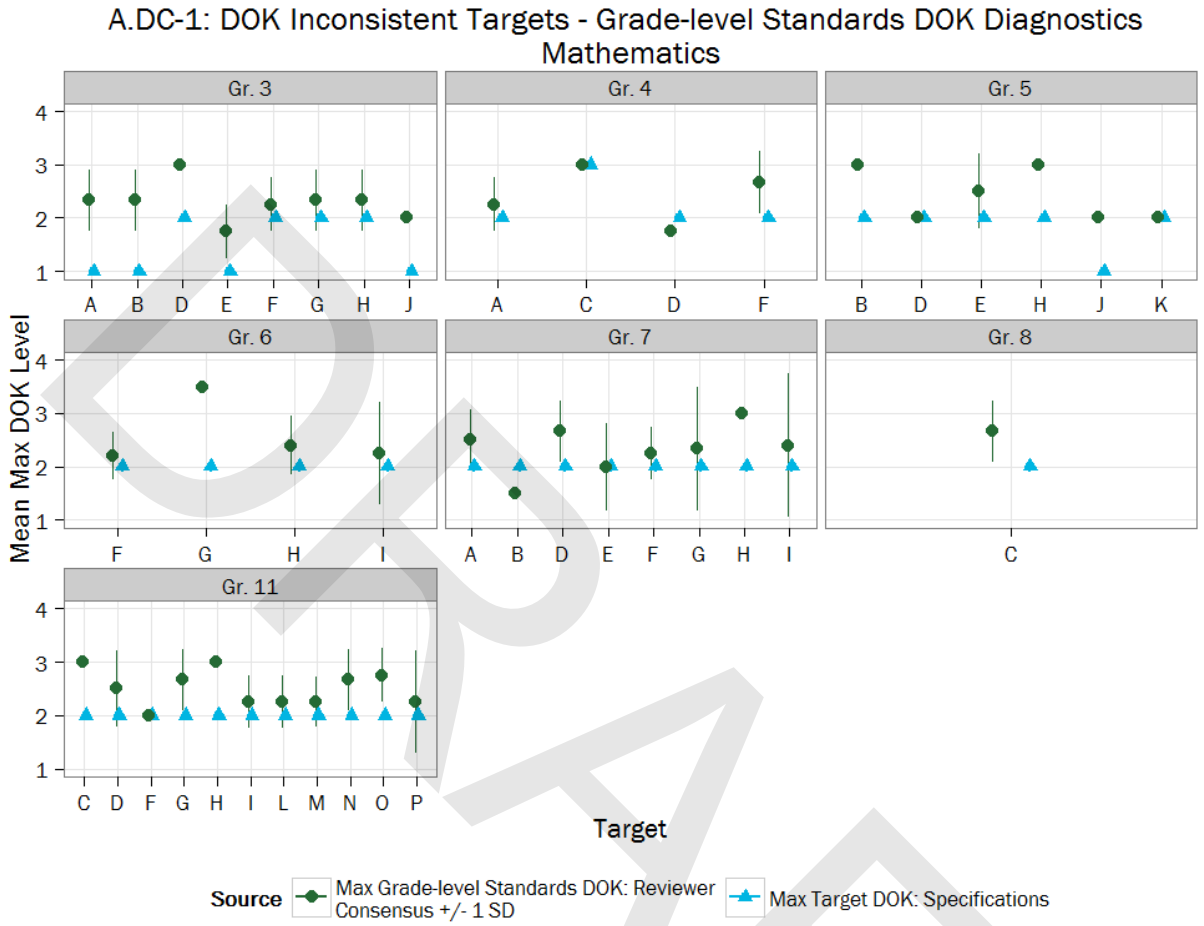


Figure ES.15. Comparison of max target DOK (specifications) and max grade-level standards DOKs (reviewers) for Mathematics Claim 1 "DOK inconsistent" targets.

Connection B: Alignment between Evidence Statements and Content Specifications

Connection B examined the alignment between the evidence statements and the Content Specifications. Alignment was examined through multiple analyses that primarily focused on the overlap between the (a) content required in both the evidence statements and the targets, and (b) cognitive demand required in both the evidence statements and the targets. The analyses provide information for the validity evidence of how well the evidence statements represent the content and the cognitive demand required by the targets. The main reviewer tasks for Connection B focused on:

(1) Workshops 3 -5: Four to five reviewers in each grade, across three workshops (for a total of 12-15 reviewers per grade) provided ratings on the evidence statements indicated for each target in the item specifications. The primary foci included:

   a. Degree to which the content and knowledge required in the target was represented by the collective set of evidence statements

b. Degree to which an individual evidence statement, as indicated in the item specifications, represents the content and knowledge required in the target (independent, verification ratings)[11]

c. Cognitive demand (DOK) required by the evidence statements (independent, blind ratings[12])

Analyses were conducted separately by content area to examine the alignment between the evidence statements and the Content Specifications. Additionally, for ELA/literacy, the evidence statements for the PT targets[13] were analyzed separately from the CAT targets. Connection B analyses focused on content representation and DOK consistency.

Results Summary

Below is a summary of the main findings related to the alignment between the evidence statements and the Content Specifications. Summary results are presented by criterion for both ELA/literacy and mathematics. For mathematics, analyses were conducted using all the mathematics targets and also by disaggregating the targets by emphasis (major vs. additional and supporting). Supporting tables for the emphasis breakdown are available in the Appendix F.

Content Representation

One of the more pertinent inferences to draw regarding the validity evidence supporting the relationship between the evidence statements and the targets is that the evidence statements collectively represent the content and knowledge required in the target. As seen in Figure ES.16, for both ELA/literacy and mathematics, across grades and claims, the targets were well-represented by the collective set of evidence statements. Additionally, although not shown here, the majority of the individual evidence statements were rated as being partially-aligned to the targets. Very few evidence statements (mathematics: M=.01 %; ELA/literacy: M=.01%) were not aligned at all to their intended targets. Evidence statements were designed by Smarter Balanced to be partially-aligned to targets; therefore, this finding confirms what was intended.

[11] By design, Smarter Balanced intended an individual evidence statement to represent only part of a target.
[12] Blind ratings refer to the reviewers not having access to materials that would otherwise allow them to identify the intended cognitive demand.
[13] Evidence statements exist for ELA/literacy Claim 4 PT targets.

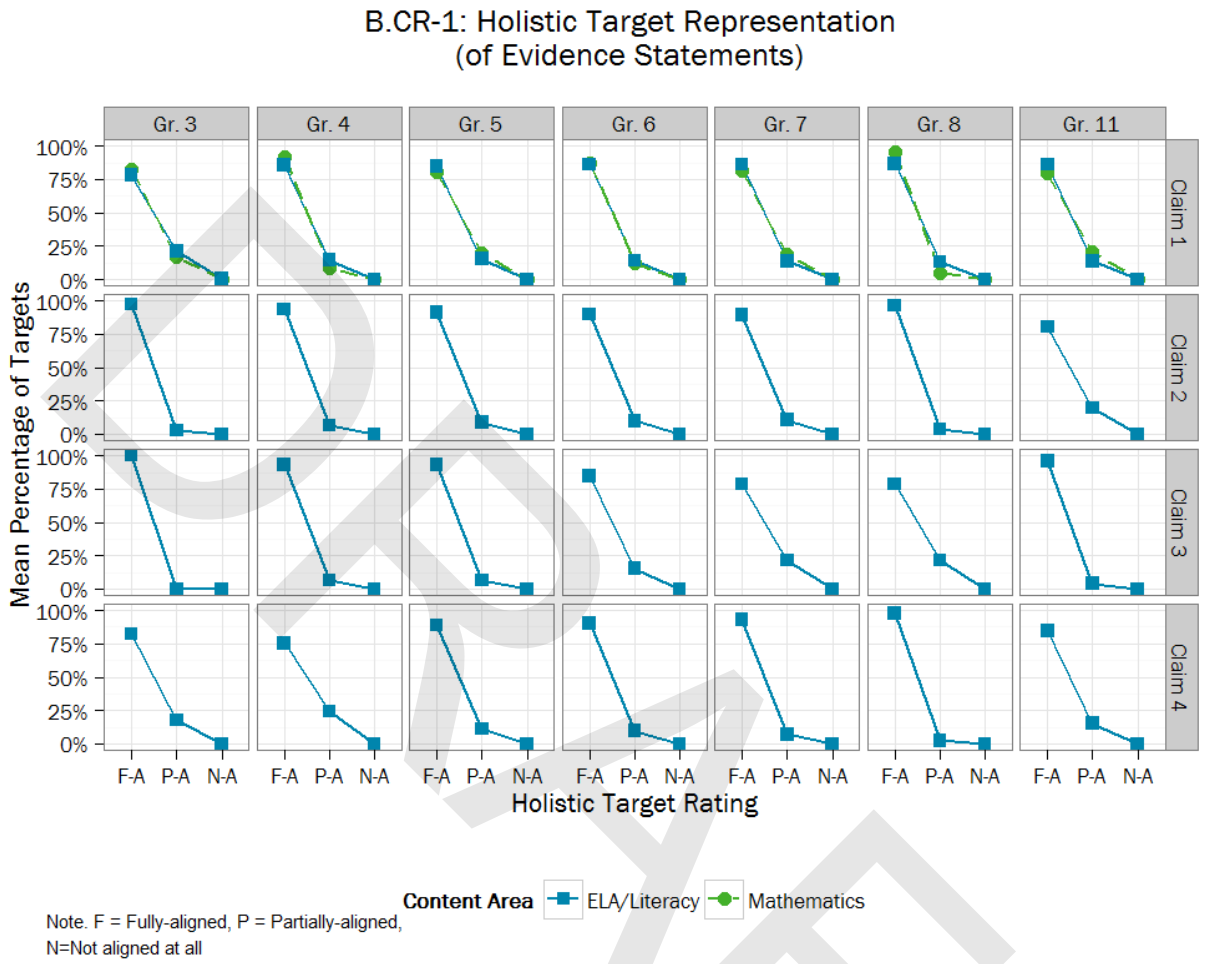


Figure ES.16. Holistic target representation of ELA/literacy and mathematics intended evidence statements.

DOK Consistency

The fact that both the evidence statements and the targets had multiple levels of cognitive demand posed a challenging obstacle for determining the DOK consistency between the two. Given the development process of the evidence statements, the expectation was that the range of the cognitive demand of the evidence statements would fall within the range of the intended target. That is, one would not want items to be developed from an evidence statement that, although overlapped with that of the intended target, had a higher upper or lower limit. For example, if a target required DOK levels 1, 2, and 3 and one evidence statement required DOK level 2 and the other required DOK level 3, then the entire range of each evidence statement's cognitive demand falls within the DOK range of the intended target. Further, if an evidence statement required DOK levels 3 and 4, although the level 3 overlaps with the intended target DOK range, the level 4 would be unexpected.

Evidence statements from ELA/Literacy Claims 1 (Comprehend Literary & Informational Text) and 3 (Speaking & Listening) generally had high percentages of evidence statements as DOK consistent with their intended targets; Claim 4 evidence statements had the fewest evidence statements that

were DOK consistent with their intended targets. For mathematics, the majority of evidence statements were rated as having DOK levels within the range of the intended targets. The outlier was grade 3, where the mean percentage of evidence statements that fell within the range of the intended target DOKs was 49.0%. Generally, it was likely that evidence statements were inconsistent with the intended targets due to the reviewers indicating that the evidence statements required a higher cognitive demand that what was intended by the target[14] (with the exception of ELA/literacy Claim 1 evidence statements, where reviewers indicated lower levels of cognitive demand that what was intended by the target).

When the DOK consistency criterion was relaxed to only require at least one evidence statement's DOK level (since evidence statements could have multiple DOK levels), the DOK consistency of the evidence statements with their intended targets increased across grades and claims for both content areas, with the exception of Claim 4 (Research & Inquiry), which remained relatively low.

Connection C: Alignment between Test Blueprint and Content Specifications

Analyses were conducted separately by content area to examine alignment between the test blueprint and the Content Specifications. After working with the CCSS and the Smarter Balanced Content Specifications, reviewers attending Workshop 1 provided holistic feedback on how representative the blueprints were to the Smarter Balanced Content Specifications and test design decisions.

Results Summary

For both ELA/literacy and mathematics, reviewers believed the blueprints were mostly to fully representative of the content and knowledge that Smarter Balanced outlined to be assessed.

Connection D: Alignment between Evidence Statements and Item/Task Pools

Item-level metadata were not available for mathematics items; therefore, we were only able to conduct evidence statement item analyses for ELA/literacy CAT items and performance tasks (PTs). Ratings for the evidence statement were made independently by reviewers; they were not made aware of the intended evidence statement(s) contained in the item metadata. Additionally reviewers were allowed to map as few or as many evidence statements to a single item as they believed was appropriate.

Results Summary

Below is a summary of the main findings related to the alignment between the evidence statements and the Smarter Balanced CAT items and performance tasks. Summary results are presented by criterion for both ELA/Literacy and mathematics.

Content Representation

Reviewers typically identified one or two evidence statements mapped to each ELA/literacy CAT item and each item making up a PT, although occasionally reviewers identified more than two items per evidence statement. Although the items were not sampled to be representative of items at the

[14] It is possible that the general nature of the targets contributed to this result.

evidence statement level, there were only 11 evidence statements across all grades that were not addressed by the items included in this alignment study. This finding suggests that the item writers included a wide range of knowledge and skills within a target when writing items so that the evidence statements were effectively represented.

We examined the match between the evidence statement(s) mapped to the first phase of CAT field test items and field test PTs as identified by reviewers compared to the evidence statement(s) intended to be mapped as indicated by item writers. We examined the match in various ways:

- Exact match: All evidence statement(s) identified by reviewers matched all the intended evidence statement(s) and reviewers identified no additional evidence statements.
- All intended evidence statements were identified: Reviewers identified all the evidence statements that the item writers identified, but the reviewers might also have identified some evidence statements that the item writers didn't intend.
- At least one identified evidence statement matched at least one intended: At least one evidence statement that a reviewer identified matched at least one of the evidence statements that the item writers identified, regardless of how many evidence statements were identified by reviewers or intended by item writers.

Depending on grade level and claim, reviewers believed one-third to almost 90% of the ELA/literacy CAT items were mapped exactly to the same evidence statement that the item writers intended. When the exact agreement was somewhat low (below 40%), it was sometimes because the reviewers tended to identify more evidence statements per item than the item writers intended. When examining whether reviewers identified all of the same evidence statements as did the item writers, but also identified some additional ones that the item writers didn't intend, the agreement increased. The agreement improved even more when reviewers identified at least one evidence statement that was intended by the item writers. The reviewers independently identified an evidence statement for each item. In some cases, there were many closely related evidence statements for a single target, which made the process potentially difficult for the reviewers to identify the "correct" evidence statement. For example, in ELA/literacy at grade 8, Claim 3, Target 9, there were 28 separate evidence statements from which the reviewers could select.

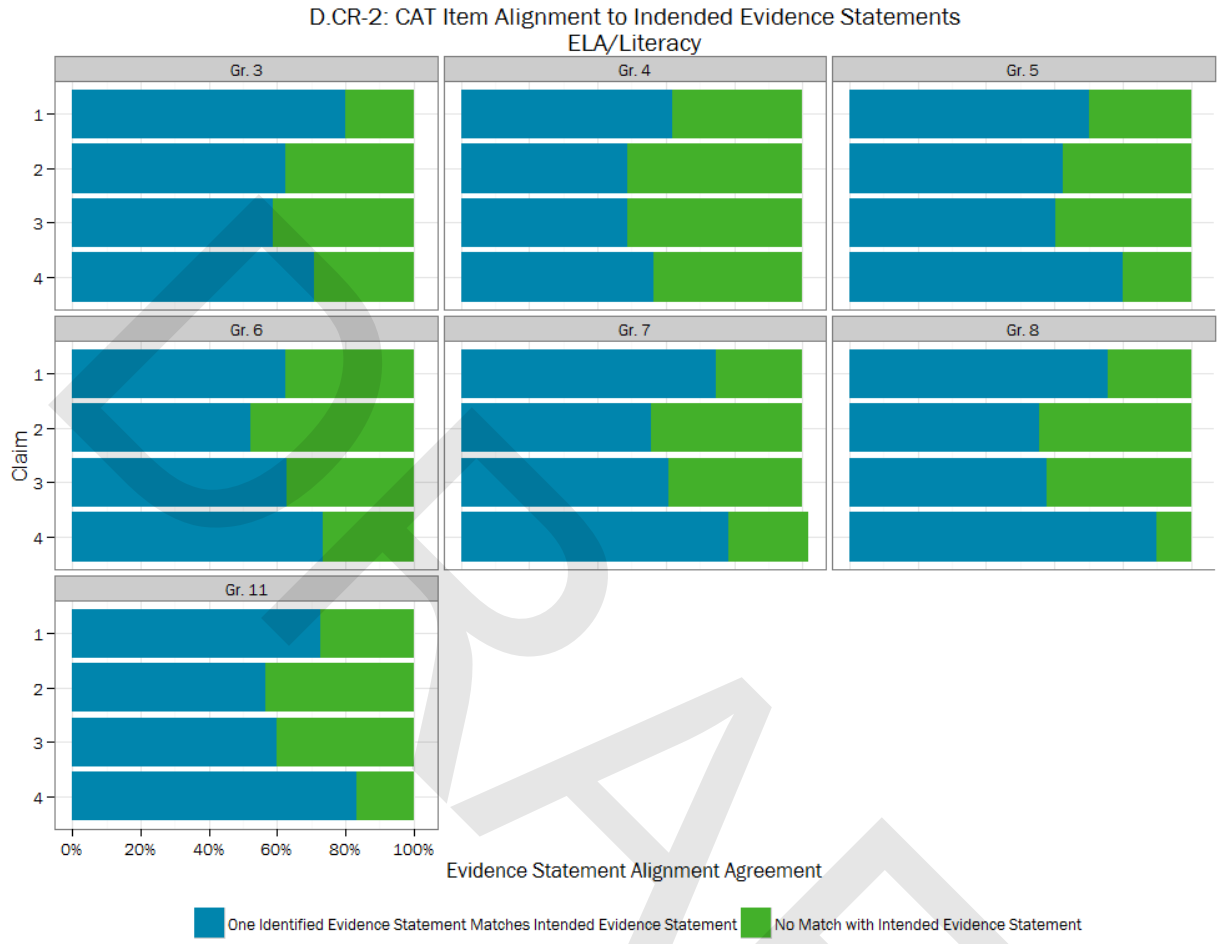


Figure ES.17. Alignment of ELA/literacy CAT items to an intended evidence statement.

Although the ELA/literacy CAT items sometimes included many mapped evidence statements, the ELA/literacy PT items selected for this study included only one evidence statement. Results for the ELA/literacy PT items were similar as those for the CAT items; reviewers identified the same evidence statement for approximately half (47.8% for grade 5) to three-fourths (77.8% for grade 6) of the PT items as did the item writers.

DOK Consistency

Because the Content Specifications did not include DOK levels for the evidence statements, DOK consistency analyses for this connection used an average of independent reviewer ratings for evidence statement DOK levels to act as the "true" DOK range for each evidence statement. We used the criterion that if at least 50% of reviewers rated an evidence statement at a level 1, a level 1 was included in the evidence statement DOK range. We took the same steps for DOK levels 2, 3, and 4 to obtain the full DOK range.

Reviewers typically rated more than half the CAT item DOK levels within a claim as being consistent with the intended DOK range of their mapped evidence statement. The exception was for Claim 4 at grades 5 – 8 and 11. . For Claim 4 items, reviewers tended to rate the items at a lower DOK level than what was identified as the lowest DOK level for their mapped evidence statements. For PTs, the item DOKs were rated as being consistent with the evidence statement DOKs more than half of the time at grades 5 – 8 and 11. At grades 3 and 4, however, more than half of the items were rated as having DOK levels below that of their mapped evidence statement. The reader is cautioned that the results for the PTs were based on a very small number of items. Combining this information with what we learned when examining Connections B and G, reviewers consistently rated the DOK level of the evidence statements at a higher level than the DOK level at which they rated items and targets.

Connection E: Alignment between Test Blueprint and CAT Algorithm

HumRRO researchers evaluated the algorithm specifications (V2. 6/17/2013) to the draft test blueprints (dated January 2014) for ELA/literacy and mathematics. At that time, no simulated test events were available for an independent review of the resulting blueprint correspondence. Moreover, the version of the specifications reviewed was incomplete because decisions had not yet been made about many of the algorithm requirements, such as defining the measurement model to use.

The American Institutes for Research (AIR) has successfully delivered computer-adaptive test events for several states. This is evidence that they are knowledgeable about the algorithm requirements that are needed. Additionally, based on the documentation, AIR was aware of the requirement to meet the Smarter Balanced blueprint in terms of DOK level and content coverage. Once decisions have been made to complete the necessary requirements for the algorithm including the final blueprints, and after item development is completed, Smarter Balanced will need to check the resulting test events against the blueprint requirements.

Connection G: Alignment between Item/Task Pools and Content Specifications

Connection G analyses examined the alignment of CAT items and PTs to the Content Specifications, which included examining item alignment to targets, grade-level standards, and mathematical practices (when applicable). Analyses were conducted for CAT and PT items for both ELA/literacy and mathematics. Approximately 50% of the available CAT item pool from each content area was included, stratified by grade, claim, and target. Because there were far fewer PTs completed at the time of the study and limited time during the workshops, our sample included three PTs each at grades 3 – 8, and six PTs at grade 11. Each ELA/literacy PT included four individual items while each mathematics PT included six individual items.

Analyses conducted on the data for Connection G were based on reviewer verification ratings. That is, reviewers were presented the intended DOK, target, and grade-level standard (as indicated in the item metadata or the Content Specifications) and asked to verify the alignment. If they disagreed with the intended DOK level or target, they were asked to provide an alternate DOK level or target that they believed was more appropriate or better aligned. To determine the reviewer-identified DOK level for each item, we used the intended DOK level if the reviewer verified it was a match and the alternate DOK level was used if it was not. Similar steps were taken to identify the reviewer-identified target.

Results Summary

Below is a summary of the main findings related to the alignment between the Smarter Balanced CAT items and performance tasks, and the Content Specifications. Summary results are presented by criterion for both ELA/literacy and mathematics.

Content Representation

Descriptive analyses were conducted to examine the distribution of CAT items and PTs across targets for both content areas. Reviewers mapped all ELA/literacy targets to at least one CAT item and the vast majority of mathematics targets, with three or fewer exceptions at any given grade. The items were fairly well distributed across targets within each claim. This was not surprising, given that the sample of items was stratified by grade, claim, and target and reviewers tended to agree with the intended mapped targets. Similar analyses were conducted for PT items; however, because only a few PTs were sampled, we were unable to provide information on potential gaps in content coverage based on the distribution of items across targets.

Descriptive analyses were also conducted to examine the distribution of items across the eight mathematical practices. The most common mathematical practice that reviewers mapped to items varied by grade and claim, with reviewers typically identifying at least one item mapped to each practice. Mathematical Practice 3 was frequently mapped to Claim 3 items across all grades. Mathematical Practices 1, 2, 4, and 6 were frequently mapped to items. Mathematical Practice 8 was found to be mapped to only a very small percentage of items across all grades and claims.

Content representation was also examined by looking at the average percentage of CAT items in ELA/literacy and mathematics that reviewers rated as being fully aligned, partially aligned, and not aligned to the intended target. Figure ES.17 presents findings for ELA/literacy CAT items and Figure ES.18 presents findings for mathematics CAT items. Reviewers typically believed the majority of items across all grades and claims were fully aligned to the intended target. This was true for both content areas, but especially so for ELA/literacy items. This is an important finding, given the CAT algorithm will ultimately use the assigned targets to pull items into test forms to ensure proper content coverage.

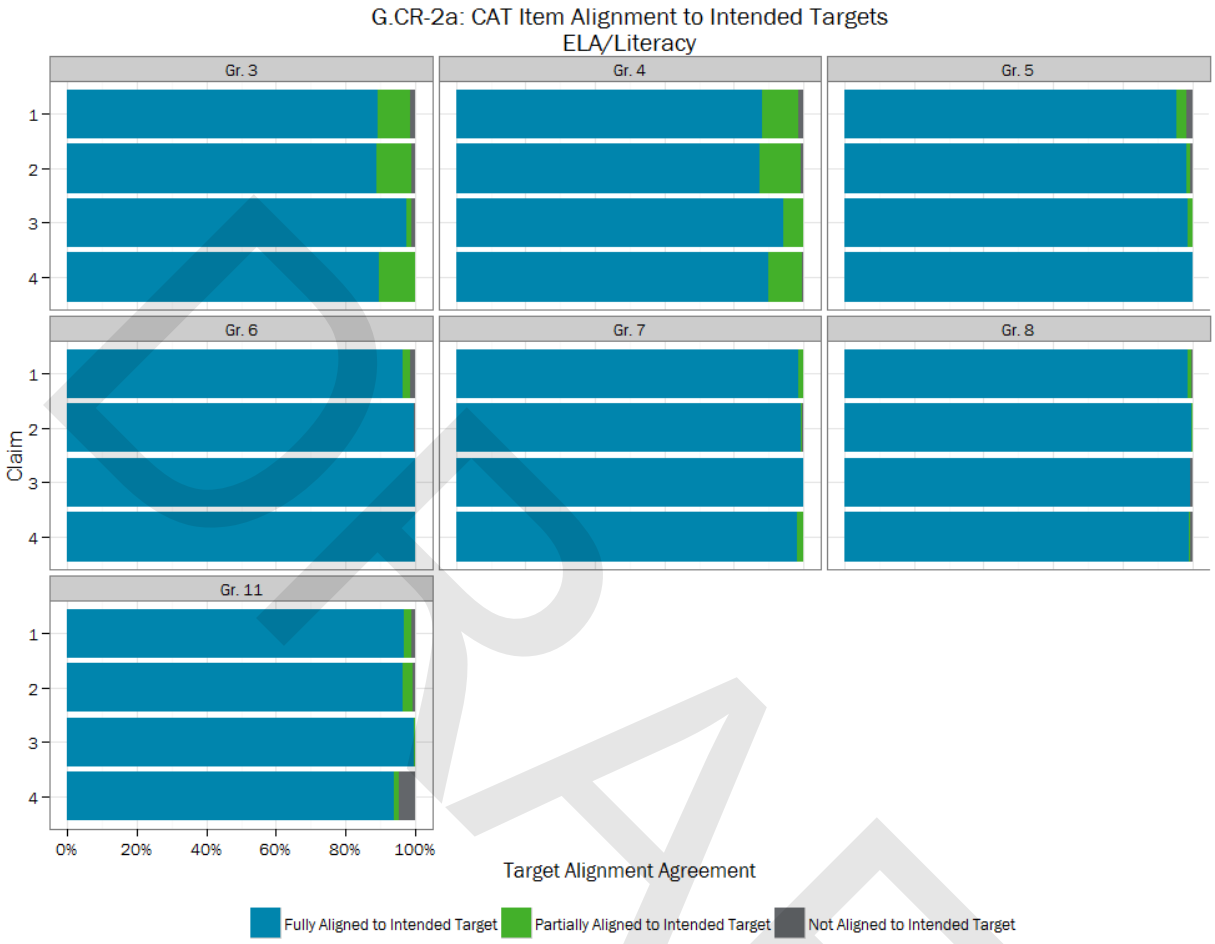


Figure ES.18. Alignment of ELA/literacy CAT items to their intended targets.

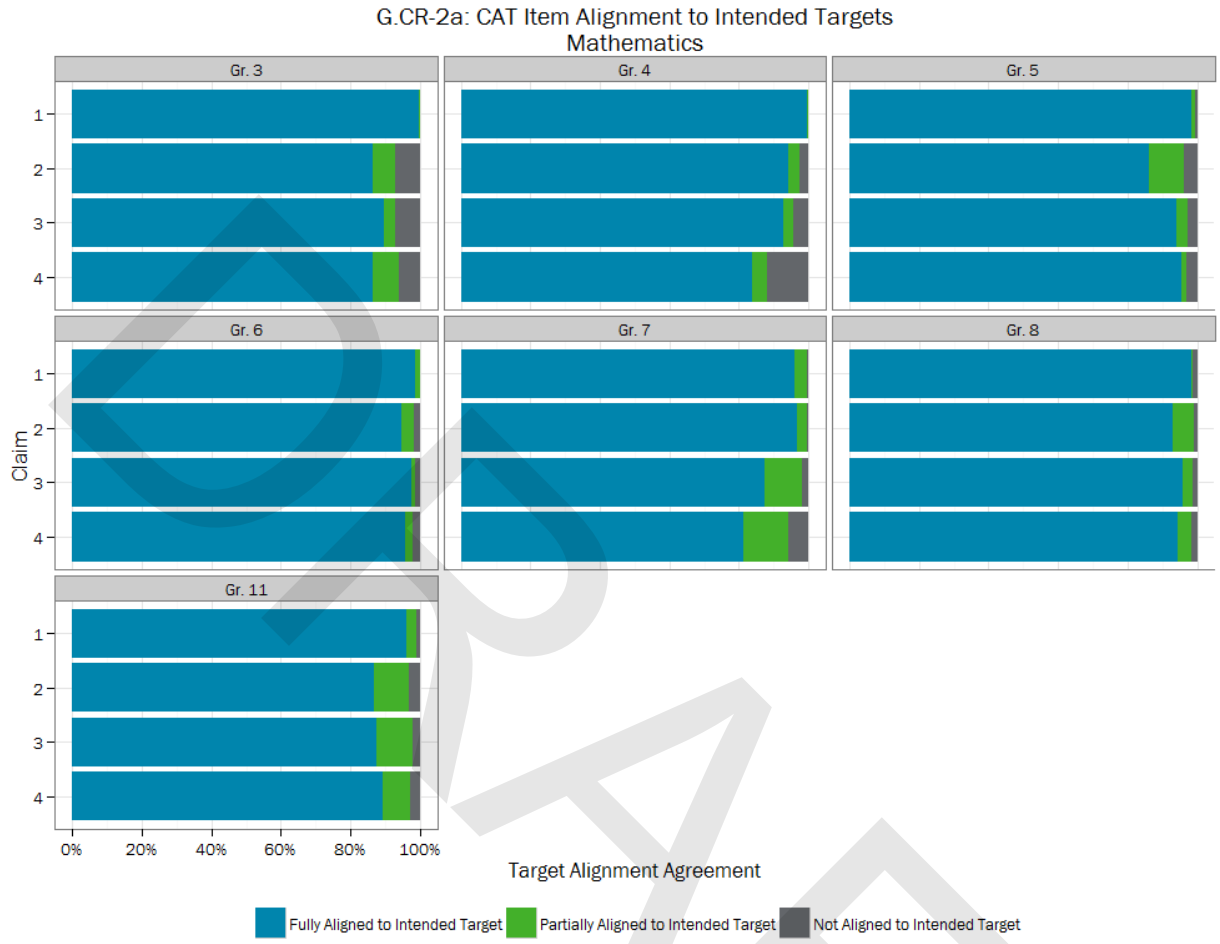


Figure ES.19. Alignment of mathematics CAT items to their intended targets.

The same analyses were conducted for PT items in both content areas to determine the extent to which items within a PT were aligned. Although not shown, alignment for items within a PT was typically very high. For most grades across content areas, more than 90% of items within a PT were believed to be fully aligned. For two grades in mathematics (grades 3 and 7), only one item (of the six within the PT), was rated as not aligned. In all other cases, there were essentially no PT items (when rounding the average percentages) that reviewers believed were not aligned.

Similar analyses were conducted to determine the alignment of CAT items and PTs to their intended grade-level standards (see Figures ES.17 and ES.18). The results were very positive across content areas, grades, and claims, with the majority of items rated by the reviewers as fully aligned. In general, alignment of the targets was slightly lower than that for the items, indicating that reviewers agreed more often with the target intended by the item writers than they did with the grade-level standard that item writers intended.

Overall, reviewers believed the majority of items within a PT across content areas and grades were fully aligned to their intended grade-level standards. The percentage of fully-aligned items was slightly higher for mathematics PT items than for ELA/literacy PT items.

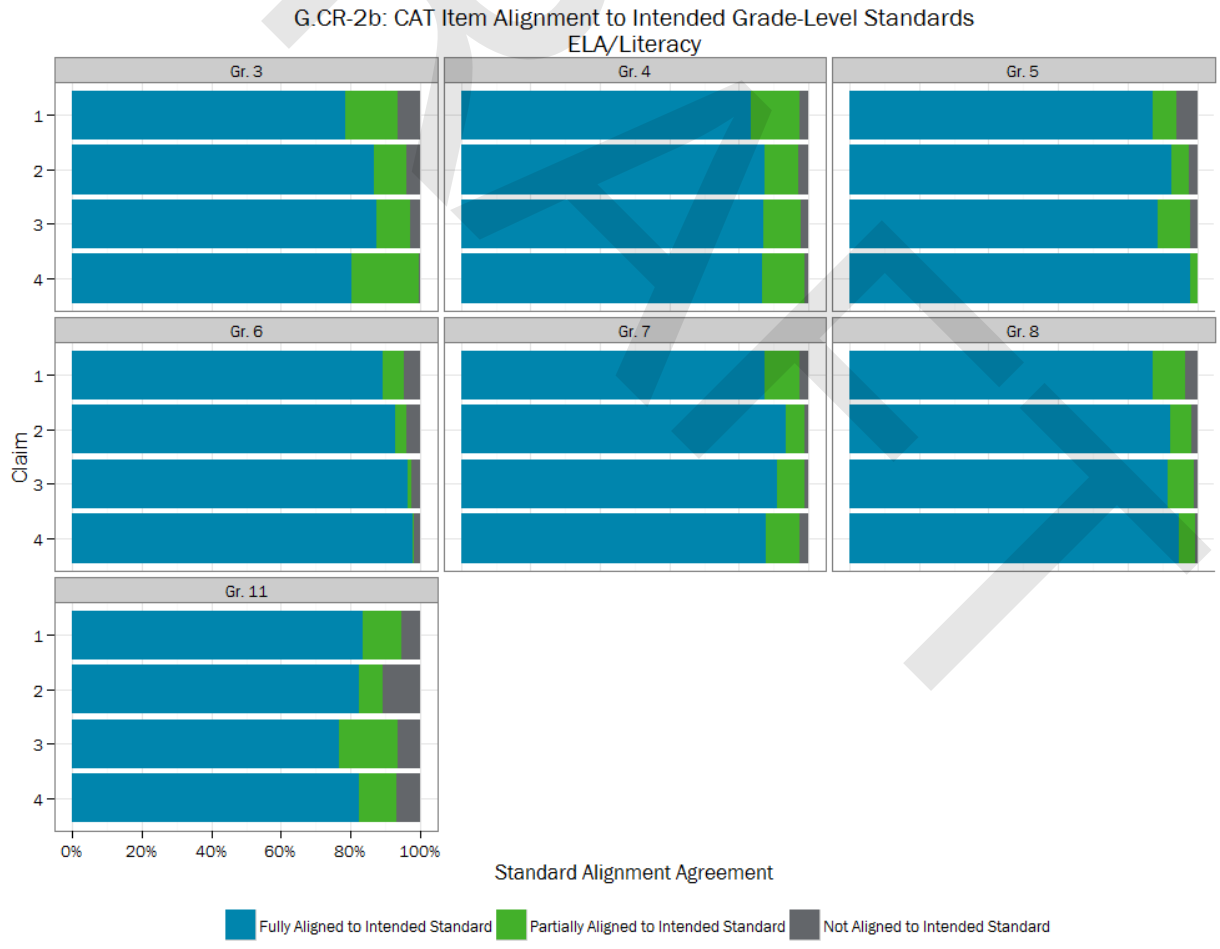


Figure ES.20. Alignment of ELA/literacy CAT items to their intended grade-level standards.

Figure ES.21. Alignment of mathematics CAT items to their intended grade-level standards.

DOK Distribution

DOK distribution of the items was examined by determining the average percentage of items that reviewers rated in each of the four DOK levels. When comparing the DOK distribution for reviewers and item writers, reviewers typically identified a similar DOK level(s) for the items as did the item writers.

DOK Consistency

DOK consistency analyses for this connection examined whether item DOK levels fell within the DOK range of the target to which they were mapped. Figure ES.21 presents the findings for ELA/literacy CAT items. Reviewers generally believed that almost all of the items within a given grade and claim were consistent with the DOK range of their mapped target, indicating the items were written at appropriate DOK levels for the content they were to cover. Figure ES.22 presents the DOK

consistency results for mathematics CAT items. Overall, reviewers believed that the DOK levels for the mathematics CAT items were within the DOK range of their mapped targets. The exception was Claim 4 (Modeling and Data Analysis) at grades 3, 4, and 5, where reviewers rated more than 20% of items with DOK levels lower than that of the DOK range of their mapped target. One possible reason for this outcome was that Claim 4 targets generally intended items to be written at DOK levels 3 and 4, and the item writers might have been challenged to do this for the lower grades. Regardless, the majority of items did fall within the DOK range of the intended target.

For all grades across both content areas, the vast majority (at least 85%) of the items within a PT were rated by the reviewers as falling within the range of the intended target. This finding provides evidence that the PT items were written to appropriate DOK levels given the content they are to assess.

Figure ES.22. DOK consistency between ELA/literacy CAT items and their mapped target.

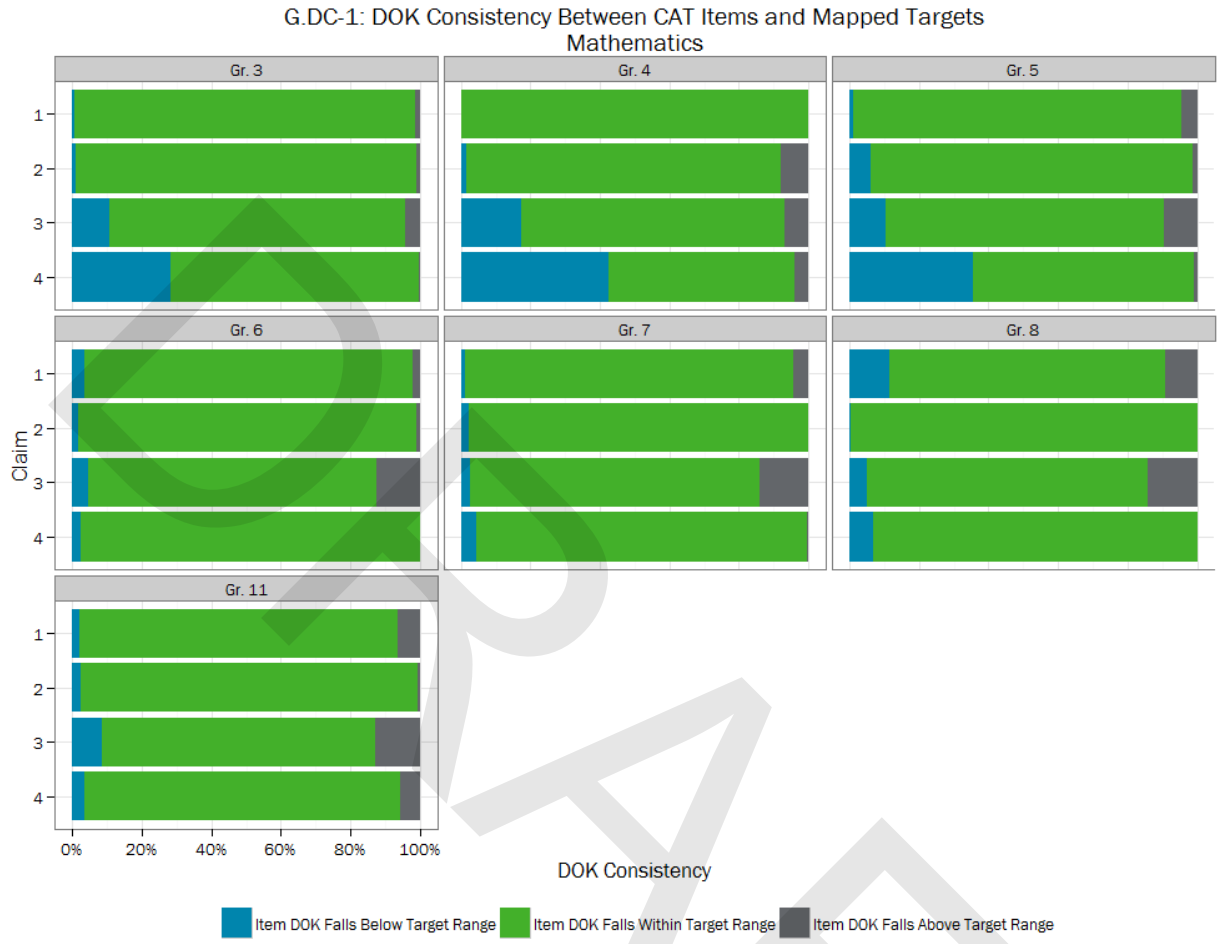


Figure ES.23. DOK consistency between mathematics CAT items and their mapped target.

Overall Alignment Summary

This alignment investigation was complex and included several elements and connections not usually included in typical alignment studies. Additionally, this study employed a two-way methodology to examine the alignment of the CCSS and Content Specifications. When all of the connections are considered, we find the overall alignment results acceptable for ELA/Literacy and mathematics. Reviewer agreement increased as they worked from standards through Content Specifications to evidence statements and then to items. Additionally, reviewer alignment agreement increased when the level of analysis was broadened. As an example, for mathematics, analyzing results at the cluster or domain level instead of the target level resulted in increased representation of the content and knowledge required.

Reviewers provided a holistic rating to indicate how well evidence statements collectively represented the content and knowledge required in the target. Reviewers generally rated the targets as being well represented by the grade-level standards they identified. Moreover, reviewers rated the majority of targets as being fully aligned to their collective set of evidence statements. Across grades,

all of the targets were at least partially represented by their collective set of evidence statements for both ELA/literacy and mathematics.

In both ELA/Literacy and mathematics, reviewers felt that the blueprints were mostly to fully representative of the content and knowledge that Smarter Balanced outlines to be assessed in the Content Specifications.

When identifying evidence statements for each of the items, reviewers typically identified at least one evidence statement that had been intended by the item writers. Reviewers' average item-level pairwise agreement was, on average, moderate to high; more than half the reviewers agreed with one another that the intended evidence statement mapped, or did not map, to its respective item. Across grades and claims, reviewers and item writers generally agreed on the average number of evidence statements that were mapped to the items. Reviewers and item writers tended to agree with the grade-level standards aligned to items across grades and claims for CAT items, and across all grades for PTs. Most importantly, the majority of reviewers agreed that the CAT and PT items fully aligned to their intended targets and the data reflect that the reviewers had high agreement with each other as well as high agreement with the item writers. Additionally, the DOK levels that reviewers verified or selected for each item tended to be consistent with the DOK levels of their mapped targets.

CONTENTS

List of Appendices

List of Tables

List of Figures

CHAPTER 1 - INTRODUCTION

Test validation involves marshalling evidence in support of an argument that inferences and interpretation based on test scores are warranted. A critical part of test interpretation validation is to demonstrate that the test measures what it claims to measure. For modern standards-based assessments, the contention typically is that the assessment allows claims to be made about student performance in relation to a set of content standards. Alignment methodologies have been developed to analyze the connections between content standards and assessments. Some widely used methodologies have been developed by Webb (1997, 2002), Porter and Smithson (2001), Achieve (2006), and others. All of these alignment methodologies—while using different definitions, procedures, and criteria--have focused on the connection between content standards and assessment items/tests.

The goal of this project was to gather evidence to examine the validity of Smarter Balanced summative assessments in terms of their alignment to the Common Core State Standards (CCSS). Evaluating the alignment between Smarter Balanced assessments and the CCSS posed the following three conceptual challenges, for which the project developed new methods:

- *How could the alignment of an assessment be evaluated when the assessment is still in development and operational test forms and events are not available?* This was the situation with Smarter Balanced when the project began. The solution was to examine test specifications.[15] There should be a clear dependent path between content standards, test specifications, test items, test forms, and the claims and interpretations about test scores.

- *Which test specifications should be examined and how should the connection between various test specifications, content standards, and test items/forms be conceived and operationally defined?* The solution was to use the array of test specifications developed by Smarter Balanced, of which there were four main components: Content Specifications, Item Specifications/Evidence Statements, Test Blueprints, and CAT Algorithm. These test specifications were compared with the CCSS content standards, the Smarter Balanced Item and Task Pools, and the Smarter Balanced tests.

- *How should the Smarter Balanced computer-adaptive test events be analyzed in terms of alignment when there are too many items, tasks, and no fixed forms to be analyzed by humans?* An alignment study typically compares a test form(s) with the content standards. That connection was of interest, although actually conducting that analysis fell outside the scope of this project since the project deadline occurred before Smarter Balanced had operational CAT test events available for analysis.

Focus on Test Specifications

Considering test specifications in an alignment study has several advantages over considering only content specifications and test items/forms. Often content specifications are too general, and needed clarification is brought to them through submitting them to the discipline of test development. Also, there often must be choices made in designing tests about what content

[15] The Smarter Balanced Content Specifications define the domain, establish claims, and provide and define the assessment targets. The Smarter Balanced Test Specifications provide the number of items of different types and their distribution across content categories and depth of knowledge levels, balance the content and depth of knowledge, and are a result of human judgment.

standards should be included on every test, which should be included on some tests, and which may not be included at all. These may be due to the limitations of large-scale, on-demand assessments (e.g., a content standard may indicate that students should be able to "Perform mental operations fluently involving addition and subtraction of whole numbers up to 100," but it is quite difficult to be sure the student is only performing mentally and not using some aids such as scratch paper or her/his fingers). More typically, content standards do not indicate specifically the level of expertise with which the student is expected to deal with the content: Should the student be able to do simple computations or apply the computational skills to solve word problems? Should the problems be somewhat similar to what the student has been instructed on, or somewhat novel? Because content standards are typically lacking these types of detailed information necessary for test development, test specifications supplement and prioritize the content standards.

Another advantage of considering test specifications in alignment studies is that alignment can be evaluated while the tests and test specifications are being developed so that potential problems are identified earlier in the process.

A third advantage of using test specifications in alignment studies is that different assessment programs' similarities and differences can be analyzed and explained in more detail. It might be, for example, that two assessment programs are somewhat different in their test specifications, but both would be judged as "overall aligned" or "not aligned" when considering only the connections between content standards and test items/forms. Consideration of test specifications in addition to content standards and test items/forms allows greater specificity in understanding and highlighting where multiple assessment programs converge or differ.

The intent of the Smarter Balanced assessments is to make valid inferences about students and to offer valid interpretations of test scores in terms of the CCSS based on students' performance on the Smarter Balanced summative assessments. This is a challenging task because the CCSS are broad, rich, and comprehensive; students are assessed using selected and constructed responses with a variety of innovative item formats; and the assessments are administered via computer-adaptive testing (CAT).

The validity of intended inferences, interpretations, and claims is based on the connection between the CCSS and the Smarter Balanced assessments. The CCSS-Smarter Balanced connection, however, is not simple and direct but rather it is supported by a sequence of component connections. The components represent increasingly focused specifications guiding the item and task development process and moving from the broad and general CCSS to specific items and tasks with their associated scoring rubrics. The strength of the validity argument for the Smarter Balanced assessments depends directly on the strength of the connections between the various components used in the development process to move from the CCSS to specific items and tasks to which students respond. These components are shown in Figure 1.1 and the connections between them are labeled by the letters A through G. A glossary of CCSS and content specification terms is at Appendix A.

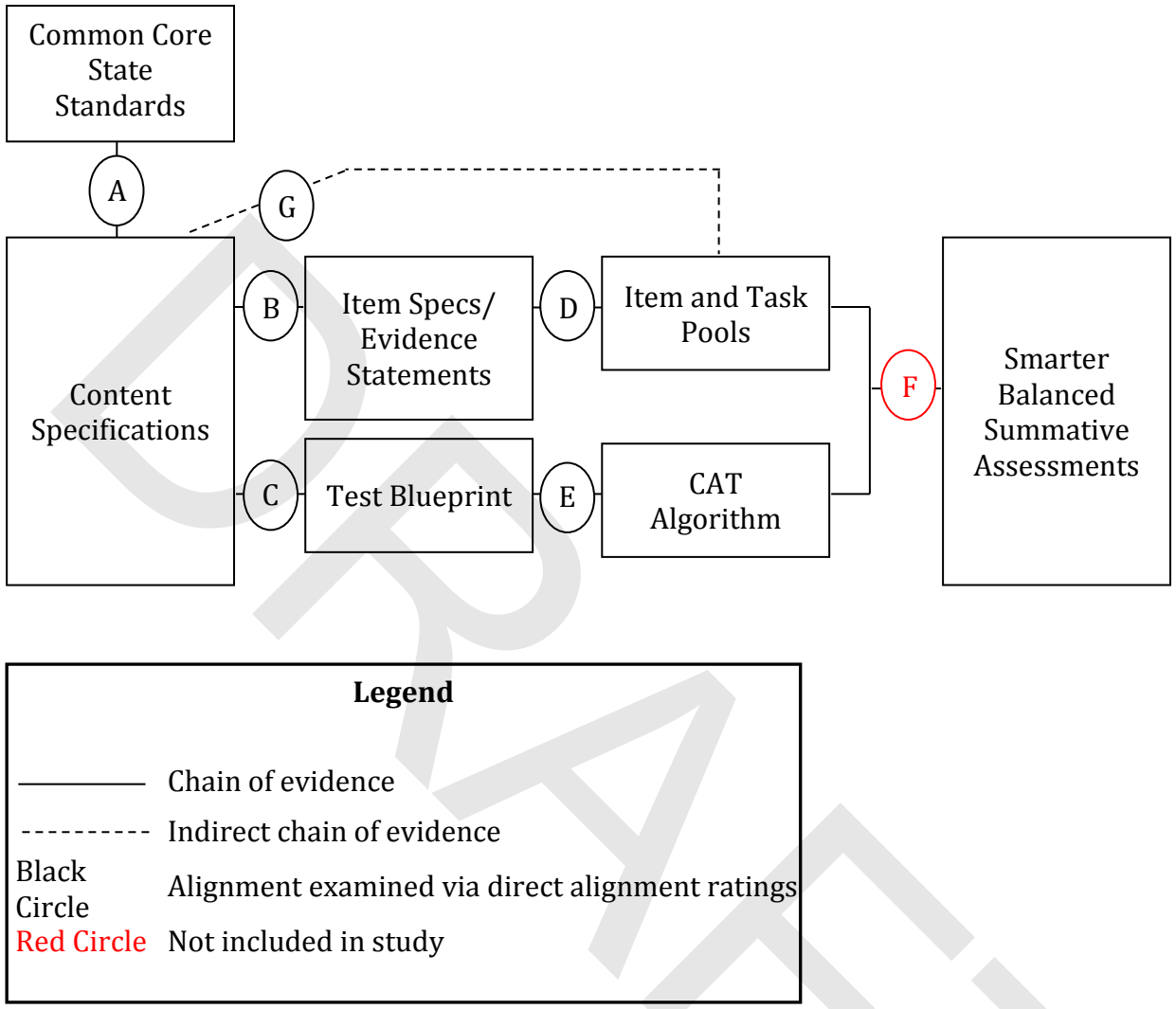


Figure 1.1. Evidence gathered to determine validity of Smarter Balanced summative assessments. (Note that Connection A examined the CCSS to the Content Specifications and the Content Specifications to the CCSS, where reasonable.)

Connection A: Alignment between Content Specifications and CCSS

The Smarter Balanced Content Specifications are a foundational document for the Consortium and, as such, the alignment of the Content Specifications (claims and targets) to the CCSS as well as the alignment of the CCSS to the Content Specifications was examined. For this alignment study, reference was made to the Smarter Balanced ELA/Literacy Content Specifications dated October 3, 2014 and the Mathematics Content Specifications dated June 2013.[16] The CCSS includes the grade-level standards in ELA/literacy and mathematics as well as the clusters, mathematical practices (MPs), and ELA/literacy anchor standards. Information was gathered about the (a) cognitive demand

[16] When necessary, additional information was obtained from the Smarter Balanced ELA/Literacy Item Specifications dated February 3, 2014 and the Mathematics Item Specifications dated February 5, 2014 (v 2.0).

of the CCSS and the Smarter Balanced claims and targets, and (b) content match between the Smarter Balanced claims/targets and the CCSS (e.g., identifying the grade-level standards that best correspond to a given target).

Connection B: Alignment between Evidence Statements and Content Specifications

The alignment of the Smarter Balanced evidence statements and the Content Specifications was examined by gathering information about the (a) cognitive demand of the evidence statements and (b) content match between the evidence statements and the claims/targets.

Connection C: Alignment between Test Blueprint and Content Specifications

Examination of the alignment between the Smarter Balanced test blueprint and the Content Specifications also informed the validity of the assessments. To do this, reviewers compared the extent to which the test blueprints covered the content as outlined in the Content Specifications. When a discrepancy was noted, the reviewers were asked to provide a description of each perceived deficiency.

Connection D: Alignment between Evidence Statements and Item/Task Pools

Additional validity evidence was gathered by examining the alignment between the pools of Smarter Balanced items to the evidence statements. Information was gathered about the (a) cognitive demand of the items and (b) the extent of agreement in content between the items and the evidence statements.

Connection E: Alignment between Test Blueprint and CAT Algorithm

A test blueprint guides the development and assembly of an assessment by specifying the content to be measured, the emphasis and balance of that content, the item types appropriate to measure the content, and the depth of knowledge for each item type. The CAT algorithm is the programmatic logic that selects the items to be administered to students based on the program's specifications. Because neither the CAT algorithm nor the Smarter Balanced test blueprints were finalized at the time of this study, HumRRO researchers reviewed the CAT Algorithm Design Report (dated June 17, 2013) and the Smarter Balanced blueprint (dated January 2014) and documented evidence of consistency regarding DOK and content requirements included in the documents.

Connection F: Alignment among Item/Task Pools, CAT Algorithm, and Summative Assessments

This connection examines how well the items selected for each test event generated from the CAT algorithm across different student ability levels are aligned with the test blueprint. Due to the lack of availability of item parameters at the time of this study, this connection was not examined. Smarter Balanced will examine this connection once field test and/or operational test data are available.

Connection G: Alignment between Item/Task Pools and the Content Specifications

For this connection, information was gathered about the (a) cognitive demand of the items and performance tasks and (b) the extent of agreement in content between the items and performance tasks and the claims/targets.

Elements Examined

The Smarter Balanced Content Specifications have several elements. For example, in mathematics there are five main content levels and depth of knowledge (DOK) indications (See Figure 1.2). The five main content levels, from most general to most specific, used in the Smarter Balanced Content Specifications are:

- Content area (e.g., mathematics)
- Claim (e.g., "Students can explain and apply mathematical concepts and carry out mathematical procedures with precision and fluency")
- Content domain (e.g., Operations and Algebraic Thinking, Number and Operations—Base Ten)
- Target (e.g., "Represent and solve problems involving multiplication and division"; "Understand properties of multiplication and the relationship between multiplication and division"; "Multiply and divide within 100")
- Standard (e.g., "Interpret products of whole numbers, e.g., interpret 5 x 7 as the total number of objects in 5 groups of 7 objects each"; "Understand division as an unknown-factor problem"; Fluently multiply and divide within 100, using strategies such as the relationship between multiplication and division [e.g., knowing that 8 x 5 = 40, one knows that 40 ÷ 5 = 8] or properties of operations")

Of these, the "target" level was developed by Smarter Balanced as a way to provide more detail about the range of content and DOK levels; the other four are also found in the CCSS. Smarter Balanced decided that much of the test blueprint would focus on the target level, rather than on the standard level (too low-level) or the content domain or claim levels (too general). Therefore, much of the analysis of the Content Specifications focused on the Smarter Balanced targets.
These levels are illustrated in Figures 1.2 and 1.3 below from the *Smarter Balanced Content Specifications* (dated October 3, 2013). Figure 1.2 displays Grade 3 Claim 1, and Figure 1.3 displays Targets A and B under the Operations and Algebraic Thinking content domain. Included in Figure 3 are references to the CCSS standards and the Smarter Balanced designated DOK levels.

Note that the Smarter Balanced Content Specifications designate targets as "major" (m) or "additional/supporting" (a/s). These designations indicate the prioritization Smarter Balanced has determined those targets will be addressed in the assessment.

A third element of the Smarter Balanced Content Specifications is the DOK indicator for each target. Some targets were assigned multiple DOK to reflect the varying levels of DOK at which content within the target could or should be assessed.

The Smarter Balanced targets were compared with the CCSS (see Connection A). Smarter Balanced Content Specifications include DOK ratings at the target level so that an individual target might have one DOK rating or it might include a range of DOK levels. Because DOK is not explicitly identified in the CCSS, the alignment study had panelists determine consensus DOK ratings. Reviewers accomplished this by using the Cognitive Rigor Matrix (CRM), a framework for thinking about cognitive demand by crossing the type of thinking that is required (based on Bloom's revised taxonomy) and the depth of content understanding required to respond (based on Webb's DOK) (Hess, 2009).[17] Since both the CCSS and the Smarter Balanced targets can be broad, it was possible to attribute more than one DOK level to them using the CRM.

[17] Note that this approach is different from that which Smarter Balanced ultimately used. Smarter Balanced used the Webb approach for identifying DOK levels.

GRADE 3 Summative Assessment Targets
Providing Evidence Supporting Claim #1

Claim

Claim #1: Students can explain and apply mathematical concepts and carry out mathematical procedures with precision and fluency.

Content for this claim may be drawn from any of the Grade 3 clusters represented below, with a much greater proportion drawn from clusters designated "m" (major) and the remainder drawn from clusters designated "a/s" (additional/supporting) – with these items fleshing out the major work of the grade. Sampling of Claim #1 assessment targets will be determined by balancing the content assessed with items and tasks for Claims #2, #3, and #4. Grade level content emphases are summarized in Appendix A and CAT sampling proportions for Claim 1 are given in Appendix B.

Target

Operations and Algebraic Thinking

Designated DOK

Target A [m]: Represent and solve problems involving multiplication and division. (DOK 1)
Items/tasks for this target require students to use multiplication and division within 100 to solve straightforward, one-step contextual word problems in situations involving equal groups, arrays, and measurement quantities such as length, liquid volume, and masses/weights of objects. These problems should be of the equal-groups and arrays-situation types, but can include more difficult measurement quantity situations. All of these items/tasks will code straightforwardly to standard 3.OA.3. Few of these tasks coding to this standard will make the method of solution a separate target of assessment. Other tasks associated with this target will probe student understanding of the meanings of multiplication and division (3.OA.1,2).

Designated content standards

Non-contextual tasks that explicitly ask the student to determine the unknown number in a multiplication or division equation relating three whole numbers (3.OA.4) will support the development of items that provide a range of difficulty necessary for populating an adaptive item bank (see section *Understanding Assessment Targets in an Adaptive Framework,* below, for further explication).

Target B [m]: Understand properties of multiplication and the relationship between multiplication and division. (DOK 1)
Whereas Target A focuses more on the practical uses of multiplication and division, Target B focuses more on the mathematical properties of these operations, including the mathematical relationship between multiplication and division.

Tasks associated with this target are not intended to be vocabulary exercises along the lines of "Which of these illustrates the distributive property?" As indicated by the CCSSM, students need not know the formal names for the properties of operations. Instead, tasks are to probe whether students are able to *use* the properties to multiply and divide.

Note: tasks that code directly to Target B will be limited to products and dividends within 100. (But see Target E under 3.NBT below.)

Figures 1.2 and 1.3: Example of Grade 3, Claim 1, Operations and Algebraic Thinking Content Domain Targets A and B from the Smarter Balanced Content Specifications for the Summative assessment of the Common Core State Standards for Mathematics, REVISED DRAFT June, 2013 (p. 31.).[18]

[18] The following definitions were used: Standards define what students should understand and be able to do; Clusters are groups of related standards. Note that standards from different clusters may sometimes be closely related; Domains are larger groups of related standards and standards from different domains may sometimes be closely related.

Reference for Analysis

Past alignment studies between two components—designated A (CCSS) and B (Smarter Balanced Content Specifications)—have underscored the importance of clearly understanding what is serving as the reference for the analysis. In cases such as this study, it is preferable to do two analyses, one with A (content standards) serving as the reference and another where B (test blueprint or specifications) serves as the reference. This is commonly referred to as a two-way alignment. The importance of two-way alignment is that one-way alignment may miss aspects of (mis)alignment. (See Figure 1.4).

Figure 1.4. Overlap and differences in the information captured in one-way and two-way alignments.

The area labelled "2" indicates that which is common between the CCSS and the Smarter Balanced Content Specifications, as expressed by the targets.

Area 1 indicates what is in the CCSS that is not in the Smarter Balanced Content Specifications. Smarter Balanced developed its Content Specifications using the CCSS, but they did not necessarily include all of the CCSS content. There are a large number of standards and not all of them could be included in a large-scale state assessment. Therefore, the Content Specifications represent Smarter Balanced's priorities. Additionally, Smarter Balanced reorganized the content found in the CCSS so that there is integration and consistency with the Smarter Balanced claims, targets, assessment design, and reporting considerations.

Area 3 indicates what is in the Smarter Balanced Content Specifications that is not in the CCSS.

If one conducted a one-way alignment study and used the CCSS as the reference (i.e., checked whether every CCSS was addressed by something in the Smarter Balanced Content Specifications), that one-way alignment study could detect areas 1 and 2, but not area 3. Similarly, a one-way alignment study using Smarter Balanced as the reference could detect areas 2 and 3, but not area 1. Thus, a two-way alignment is necessary for completeness and was employed during this study when reasonable to do so. Because Smarter Balanced reorganized and repackaged the CCSS when developing its Content Specifications, it is possible that certain content was included in the Content Specifications that is not included in the CCSS. Data collected from two-way alignment activities was intended to enable such content to be detected.

There is one additional complication with which an alignment study involving test specifications must deal. Test specifications indicate the designation made by the developers of the test specification, but there may be a question whether that designation is accurate or correct. For example, the test developer might designate a test item as measuring DOK level 3; however, expert judges might rate that item as measuring DOK level 2. Thus, an important aspect of alignment studies involving test specifications is an independent check of critical designations. Note that this independent rating may or may not result in greater alignment. For example, the test developers might have designated a CCSS as requiring a DOK level 2 and their test item as assessing DOK level 2, but the independent raters might have rated both the CCSS and the test item as being associated with DOK level 3.

CHAPTER 2 - STUDY PARTICIPANTS

Reviewer Recruitment

Educators who resided and worked within a Smarter Balanced Governing State were recruited to serve as reviewers for this alignment study. Support to recruit the requisite educators was obtained from the Teacher Involvement Coordinators (TICs). The reviewers were generally recruited from three pools of qualified educators: (a) those who were currently certified or licensed to teach or current teachers of mathematics or English language arts (ELA)/literacy in a K-12 public school or institution of higher education, (b) public school district or state education agency staff, and (c) teachers who were recently retired from a college or university. Additional qualifications required or preferred by the reviewers included:

- Had not participated in previous Smarter Balanced item development or review activities.
- Taught mathematics and/or ELA/literacy in grades 3 through 8 and/or high school within 3 years of the study; worked in a classroom content support role such as a literacy or mathematics coach, district or state content specialist, etc.; or taught in an institution of higher education in developmental and/or entry-level courses in English, composition, mathematics, statistics, or a related discipline.
- Previously reviewed part or all of the CCSS for the content area in which they reviewed items and performance tasks.
- Had previous alignment experience (preferred but not required).

Each educator was requested to participate in one on-site workshop. To ensure sufficient reviewers were available to participate in the workshops during the requisite timeframe[19], we recruited approximately 125% of the desired number. Across the then current 23 Governing States, we recruited 308 educators and ultimately selected 245 to participate in the alignment workshops; thus, 63 educators (across the two content areas) were recruited but not selected to serve as reviewers. An additional 14 educators for each content area (two per grade) were selected as alternates, for a total of 28 alternates selected across workshops. Selection of the alternates allowed for prompt replacement of any selected reviewers with last minute conflicts. Alternate reviewers were contacted only if a selected educator at a particular grade and content area could not complete the training. Table 2.1 shows the total number of qualified educators recruited for the five workshops, by grade level and content area.

Table 2.1. Number of Educators Recruited by Grade and Content Area

Content Area	Grade Level							Total
	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	High School	
Mathematics	23	23	23	23	23	23	23	161
ELA/Literacy	23	23	23	23	23	23	23	161
							Total	322

[19] Educators were needed in March 2014.

Reviewer Participation and Background

Of the 322 educators recruited to participate, a total of 226 reviewers participated across the five alignment workshops. Table 2.2 shows reviewer participation by Governing State. The content background of the reviewers is presented in Table 2.3. By design, the number of reviewers with expertise in ELA/literacy was approximately the same as that for reviewers with mathematics expertise. The reviewers' experience with English language learners (ELLs) and students with disabilities (SWDs) is shown in Table 2.4. The majority of reviewers had experience teaching both ELL and SWD students. The racial and ethnic background of the reviewers, aggregated across the alignment workshops, is presented in Table 2.5. As can be seen, the majority of reviewers were Caucasian, with some reviewers representing various racial minorities. Finally, Table 2.6 shows the various job positions of the reviewers, with the majority of them having reported to be educators.

Table 2.2. Reviewer Participation by Governing State

State	% Reviewer Representation (n)
Governing States	
California	5.8% (13)
Connecticut	7.2% (16)
Delaware	5.8% (13)
Hawaii	4.5% (10)
Idaho	0.0% (0)
Maine	2.2% (5)
Michigan	5.8% (13)
Missouri	5.8% (13)
Montana	0.0% (0)
Nevada	4.9% (11)
New Hampshire	0.5% (1)
North Carolina	9.0% (20)
North Dakota	7.2% (16)
Oregon	8.5% (19)
South Dakota	4.0% (9)
Vermont	4.5% (10)
Washington	5.4% (12)
West Virginia	5.4% (12)
Wisconsin	3.6% (8)
Wyoming	4.5% (10)
Missing	2.2% (5)
Advisory States	
Iowa	3.2% (7)
Total	100.0% (223)

Table 2.3. Reviewer Workshop Attendance by Content Area

Workshop	ELA/Literacy	Mathematics	Total
Workshop 1	36	24	60
Workshop 2	12	13	25
Workshop 3	22	22	44
Workshop 4	24	24	48
Workshop 5	23	23	46
Total	117	106	223

Table 2.4. Reviewer Experience with ELL and SWD Student Populations

Type of Student Population	% of Reviewer Experience (n)
ELL Status	
Yes	67.7% (151)
No	32.3% (72)
Disability Status	
Yes	78.9% (176)
No	21.1% (47)

Table 2.5. Racial and Ethnic Background of Workshop Reviewers

Race/Ethnic Background	% Reviewer Representation (n)
Race	
White	90.1% (201)
Black or African American	4.5% (10)
Asian	3.6% (8)
American/Indian	1.5% (3)
Hawaiian	.4 (1)
Ethnicity	
Hispanic/Latino	4.9% (11)
Not Hispanic/Latino	92.8% (207)
Missing	2.3% (5)

Table 2.6. Reviewer Job Positions

Job Position	%Reviewer Representation
Administrator/Administrative	3.6 (8)
Coach	13.9 (31)
Educator	62.8 (140)
Educator/Coach	.4 (1)
Specialist	2.7 (6)
Other	4.5% (10
Missing	12.1% (27)
Total	100.0 (223)

Reviewer Training

Educators participating in the workshops were trained to make specific alignment ratings. Training was specific to the type of alignment ratings the reviewers were asked to make. They were trained on the factors to consider when making their assigned ratings as well as the procedures to do so. They also were trained on how to access and navigate specific computer software to make their ratings. As reviewers completed their assigned activities, they were monitored, and remedial training and additional guidance were provided, as needed.

Reviewer Feedback

Participants were asked at the conclusion of the workshops to provide feedback about the training they received and their experiences completing the various workshop activities. Response rates from the reviewers ranged from 73% – 100% (refer to Table 2.7).

Table 2.7. Reviewer Response Rate to Workshop Evaluation Form

Workshop	Content Area	Response Rate (n)
1 – Charlotte, NC	ELA	92% (34)
	Math	96% (23)
2 – Charlotte, NC	ELA	92% (11)
	Math	92% (12)
3 – Charlotte, NC	ELA	100% (22)
	Math	73% (16)
4 – Los Angeles, CA	ELA	92% (22)
	Math	100% (24)
5 – Los Angeles, CA	ELA	100% (23)
	Math	78% (18)
	Missing	8% (18)
	Total	100% (223)

An overall summary of reviewer feedback across workshops is shown in Table 2.8, and Table 2.9 presents reviewer feedback by content area. Across workshops, results indicate that the reviewers generally agreed or strongly agreed that the presentations, training, guidance, materials and rating forms, and use of Excel and laptops were useful, and/or clear and understandable. These same findings were true for reviewers when looking at results by content area.

Table 2.8. Reviewer Feedback about Alignment Activities, Aggregated Across Workshops[20]

Question	Mean (n)	SD
The presentation on the CCSS and Smarter Balanced Content Specifications was useful.	4.87 (204)	1.10
The training (presentation on CCSS, Content Specifications, and reviewer's alignment tasks) was sufficient preparation to perform the alignment tasks.	4.86 (204)	1.02
The guidance provided by my group facilitator (matching targets to CCSS) was clear and understandable.	5.47 (205)	0.70
Materials and rating forms provided for the alignment tasks were clear and understandable.	5.07 (205)	0.82
Materials and rating forms provided for the alignment tasks were useful in performing the actual ratings.	5.28 (205)	0.73
The use of laptops for data entry of the alignment ratings was relatively easy.	5.54 (205)	0.68
Q8a. The training on the Excel rating forms was clear and understandable (W1 only)	5.28 (57)	0.75
Q8b. The Excel rating forms were clear and understandable. (W2-5 only)	5.28 (148)	0.73
HumRRO staff was generally courteous and helpful.	5.90 (205)	0.30

[20] Scale points and definitions: 1, Strongly disagree; 2, Disagree; 3, Somewhat disagree; 4, Somewhat Agree; 5, Agree; 6, Strongly agree.

Table 2.9. Reviewer Feedback about Alignment Activities by Content Area[21]

Question	ELA		Math	
	Mean (n)	SD	Mean (n)	SD
The presentation on the CCSS and Smarter Balanced Content Specifications was useful.	4.94 (111)	1.03	4.78 (93)	1.17
The training (presentation on CCSS, Content Specifications, and reviewer's alignment tasks) was sufficient preparation to perform the alignment tasks.	4.97 (112)	0.88	4.72 (92)	1.16
The guidance provided by my group facilitator (matching targets to CCSS) was clear and understandable.	5.47 (112)	0.70	5.47 (93)	0.70
Materials and rating forms provided for the alignment tasks were clear and understandable.	5.20 (112)	0.77	4.92 (93)	0.86
Materials and rating forms provided for the alignment tasks were useful in performing the actual ratings.	5.38 (112)	0.71	5.16 (93)	0.74
The use of laptops for data entry of the alignment ratings was relatively easy.	5.54 (112)	0.75	5.55 (93)	0.60
Q8a. The training on the Excel rating forms was clear and understandable (W1 only)	5.18 (34)	0.83	5.43 (23)	0.59
Q8b. The Excel rating forms were clear and understandable. (W2-5 only)	5.41 (78)	0.71	5.13 (70)	0.72
HumRRO staff was generally courteous and helpful.	5.90 (112)	0.30	5.90 (93)	0.30

Reviewers were encouraged to share comments about any aspect of the alignment training, especially those aspects where they had ideas for how procedures or materials could be improved. Appendix B presents a summary of the types, description, and frequency of reviewer comments.

[21] Scale points and definitions: 1, Strongly disagree; 2, Disagree; 3, Somewhat disagree; 4, Somewhat Agree; 5, Agree; 6, Strongly agree.

CHAPTER 3 - PROCEDURES

Data related to Connections A, B, C, D, and G (as described in Figure 1) were gathered across a series of five workshops and HumRRO researchers reviewed documents to address Connection E (described in subsequent sections). Table 3.1 presents the alignment data collection design, including an indication of the workshops where data were collected for each connection and the reviewer alignment methods used to collect these data. The independent reviewer method indicates that reviewers were asked to assign ratings without reference to the Smarter Balanced Content Specifications and/or Item Specifications. The reviewer verification method indicates that reviewers were asked to verify the degree to which they agreed with the developers' classifications of various attributes (e.g., target, claim). Because Connections A and B serve as the foundation for later connections, we used the independent reviewer method for ratings related to these connections so that as little bias as possible was introduced. For multiple reasons, we used the reviewer verification method to collect data related to Connections D and G. First, the ways in which items and evidence statements were categorized (through item and evidence statement metadata) represented both the content and item development expertise of item developers. Second, given the constraints of the current study (e.g., time, budget) and the large number of items and evidence statements, it was not practical to ask reviewers to independently identify evidence statements and other attributes (e.g., grade-level standards, claims, targets) to which items were aligned.

Table 3.1. Data Collection Design

Connections	Workshop #	Reviewer Method(s)
A. Content Specifications to CCSS (and CCSS to content specifications)	1, 2	Independent review & Reviewer verification
B. Evidence statements to Content Specifications	3, 4, 5	Independent review
C. Test blueprint to Content Specifications	1	Reviewer verification
D. Item/task pools to evidence statements	3, 4, 5	Reviewer verification
E. CAT algorithm to test blueprint	HumRRO researchers	Researcher review of algorithm specifications Researcher review of test blueprints
F. Items/performance tasks and CAT algorithm to the Smarter Balanced summative assessments	Not included in current study	Not included in current study
G. Items/performance tasks to content specifications	3, 4, 5	Reviewer verification

Workshops Design

As noted in Table 3.1, each workshop included tasks associated with specific connections, so that each workshop had a slightly different design. Generally, tasks in each workshop were completed by reviewers grouped by grade-level (e.g., a single grade) or grade-band (e.g., multiple grades).

Workshops 1 and 2

Workshops 1 and 2 gathered data that examined Connections A and C. For Connection A, Tables 3.2 and 3.3 present the total number of grade-level standards and targets at each grade level, and the total number of grade-level standards and targets each reviewer rated for mathematics. Workshop 1 involved tasks related to matching the targets to the grade-level standards while Workshop 2 involved tasks related to matching the grade-level standards to the targets. Workshop 1 additionally included activities regarding the alignment between the Content Specifications and test blueprint.

In Workshop 1 for mathematics, given the number of total grade-level standards and targets at grades 3–8, each group provided ratings for two grade levels. Grade 11 grade-level standards and targets were reviewed by two separate groups, to account for the larger numbers at this level (see Table 3.2). Given the numbers of grade-level standards and targets for ELA/literacy in grades 3–8, we convened one group per grade level for grades 3–8, and two groups for grade 11 (see Table 3.3). Since Workshop 2 required fewer reviewer tasks, three groups of reviewers were convened for each content area. More specifically, one group of five reviewers provided ratings for grades 3–5, one group for 6–8 and one group for high school. Thus, the design for Workshop 2 included a total of 30 reviewers.

Table 3.2. Workshop Design for Connection A – Mathematics

Grade	Individual Grade Level		Workshop 1 Grade-band Groups		
	Grade-level Standards[1]	Targets	Total Grade-level Standards	Total Targets	# Reviewers
3	35	29	70	59	5
4	35	30			
5	36	29	79	57	5
6	43	28			
7	43	27	76	55	5
8	33	28			
11	190	34	95	17	5
			95	17	5
Workshop 1 Total	415	205	415	205	25
Workshop 2 Total[2]					15

[1]There were also eight Mathematical Practices that were applicable for each grade.
[2]Workshop 2 had three groups of five reviewers (grades 3-5, 6-8, and high school), totaling 15 reviewers.

Table 3.3. Workshop Design for Connection A – ELA/Literacy

ELA/Literacy			
Workshop Grade-level Groups	**Grade-level Standards[1,2,3,4]**	**Targets**	**# of Reviewers**
3	81	29	5
4	81	29	5
5	79	29	5
6	92	29	5
7	89	29	5
8	91	29	5
HS-1	115	14	5
HS-2	113	15	5
Workshop 1 Total	741	203	40
Workshop 2 Total[5]			15[5]

[1]These counts include the ELA/Literacy grade-level standards and the Literacy in History/Social Studies, Science, & Technical Subjects.
[2]For the Literacy in History, Science, and Technology standards, they were grouped by grade-band (6-8, 9-10, 11-12). We split the 6-8 across grades 6, 7, and 8 and presented them as 6-8 standards. When rating DOK, they were given the option to note if they thought the DOK differed across those grades.
[3]These counts exclude targets that were not assessed on the summative assessments as specified in the Smarter Balanced Content Specifications.
[4]There were also 32 anchor standards that were presented to each grade-band group.
[5]Workshop 2 included three groups of five reviewers (grades 3-5, 6-8, and high school).

Workshops 3–5

Items and Evidence Statements

Approximately half of all Phase 1 items and a limited number of performance tasks were included in this alignment study. We used a stratified random sampling approach to ensure representation across all claims within each grade level and content area. Appendix C presents the plan for sampling the mathematics and ELA/literacy items and performance tasks.

All of the evidence statements associated with a particular grade were included in this alignment study. For Workshops 3–5, reviewers were asked to make ratings for Connections B, D, and G. Workshop participants reviewed the same evidence statements when making ratings related to these three connections; however, Connection B required the reviewers to make independent ratings while Connections D and G required the reviewers to verify item and evidence statement metadata.

Alignment Reviewers

Generally, workshop tasks were completed by reviewers working in grade-level (i.e., a single grade) or grade-band (i.e., multiple grades) groups. Based on the number of mathematics and ELA/literacy items and performance tasks, reviewers completed alignment activities at two grades for grades 3–8 (e.g., grades 3–4, grades 5–6), and two groups of high school reviewers (for a total of five groups). A total of 65 reviewers were required (five reviewers for each of five group for a total of 25 for mathematics and five reviewers for each of eight groups for a total of 40 for ELA/literacy) and a total of 30 reviewers were required for Workshop 2 (five reviewers for each of three groups for each content area). For Workshops 3, 4, and 5, 25 reviewers each were required for mathematics and ELA/literacy (five reviewers for each of five groups for each content area), for a total of 150 reviewers across the three workshops.

Reviewer and Researcher Tasks

Generally, reviewer tasks involved (a) identifying the DOK for each alignment component (e.g., items, grade-level standards) in each respective connection and (b) matching, or mapping, two components for each connection. More specifically, *match* was operationally defined as identifying the most appropriate component that corresponded to another component (e.g., identifying the grade-level standard(s) that best corresponded to a given item). In addition to specific instructions, reviewers were provided a graphic representation to facilitate their understanding and interpretation. Specific tasks associated with each connection were conducted by qualified educators serving as reviewers, as described above. The general activities completed by educators during each workshop included:

- Workshop 1: Examined the alignment between the Content Specifications and the CCSS (Connection A) and examined the alignment between the Content Specifications and the test blueprints (Connection C).
- Workshop 2: Examined the alignment between the CCSS and the Content Specifications (Connection A).
- Workshops 3 - 5: Examined the alignment between the evidence statements and Content Specifications (Connection B), alignment between the evidence statements and items (Connection D), and (and indirect) alignment between items/performance tasks and Content Specifications (Connection G).

Connection E (test blueprints to the CAT algorithm) was examined by qualified HumRRO researchers familiar with computer-adaptive testing (CAT) and the development of CAT algorithms. Researchers were provided the Smarter Balanced test blueprints and the CAT algorithm specifications. Because both the algorithm and blueprints were under development, researchers reviewed the documents for evidence of consistency and noted discrepancies regarding the representation of DOK and content.

Tables 3.4–3.6 outline the specific reviewer tasks associated with each connection. Because of the nature of the activities completed for Workshops 1 and 2, reviewer tasks are specified by content area (Table 3.4). The reviewer activities to be completed for Workshops 3–5 are the same for ELA/literacy and mathematics, so they are not specified by content area (Tables 3.5 and 3.6). Table 3.7 presents the specific tasks that qualified HumRRO researchers completed to provide information about the alignment of the Smarter Balanced CAT algorithm to the test blueprints (Connection E).

Table 3.4. Reviewer Tasks Associated with Connection A – Alignment between Content Specifications and CCSS

Connection	Reviewer Tasks	
	Mathematics	ELA/Literacy
A. Content Specs to CCSS	Workshop 1 1. Independently rated the DOK level for the GS[1] & MP[1] a. Reviewed and adjudicated disagreements and reached consensus 2. Independently rated the DOK level for each target 3. Matched each target to the GS[4] & MP. Reviewers were provided the targets and asked to identify each GS and MP that most closely matched or most comprehensively reflected the content described in the target a. For Claims 1–4, reviewers independently identified GSs that aligned to each target b. For all claims, reviewers indicated to what degree each MP was reflected in each target, using a 3-point Likert scale	Workshop 1 1. Independently rated the DOK for the GS[1] a. Reviewed and adjudicated disagreements and reached consensus 2. Independently rated the DOK for each target 3. Independently matched each target to the GS. Reviewers were provided the targets and asked to identify each GS that most closely matched or most comprehensively reflected the content described in the target a. Identified a primary GS and any additional GS that aligned to each target
	Workshop 2 1. Independently match each GS to targets. Reviewers will be provided the GS and identify targets that most closely matches or most comprehensively reflects the content described in them a. Identify a primary target and any additional targets that align to each GS	Workshop 2 1. Independently matched each GS to targets. Reviewers were provided the GS and identified targets that most closely matched or most comprehensively reflected the content described in them a. Identified a primary target and any additional targets that aligned to each GS

[1]GS=Grade-level standards; MP=mathematical practices
[2]For Claim 1, the Smarter Balanced targets are the actual CCSSM cluster-level headings. Since the similar language between the CCSSM cluster-level heading and target will be readily evident to reviewers, reviewers will verify the GS associated with each cluster-level heading.
[3]Reviewers will also be given the opportunity to note any additional GS that match(es) the targets.
[4]Where an item cannot be matched to a GS, reviewers will select one or more clusters as appropriate.

Table 3.5. Reviewer Tasks Associated with Connections B – Alignment between Evidence Statements and Content Specifications and C – Alignment Between Test Blueprint and Content Specifications

Connection	Reviewer Tasks
B. Evidence statements to Content Specifications	<u>Workshops 3–5</u> 1. Independently rated the DOK level for the evidence statements. Reviewers were provided the item and evidence statements and reviewers identified two DOK levels: a. The 'highest DOK' appropriate for each evidence statement. This DOK was the *highest* DOK represented in the evidence statement, regardless if it was based on the entirety of the evidence statement or if it was based on a small part of the evidence statement b. The 'most representative DOK' appropriate for each evidence statement. This DOK was the DOK that was *most representative* of the evidence statement; that is, the DOK level for which the majority of the evidence statement represented, regardless if a very small portion of the evidence statement represented a higher DOK 2. Independently matched evidence statements to claims and targets. Reviewers were provided the evidence statements and asked to identify a single claim and potentially multiple targets that aligned most closely to that evidence statement a. Reviewers identified a single claim, a primary target and any additional targets that most closely matched each evidence statement
C. Test blueprint to Content Specifications	<u>Workshops 1</u> 1. Reviewers used a Likert scale (4-point scale) to rate the extent to which the following blueprint specifications represented the content as outlined in the Content Specifications: a. Holistic rating: Reviewers provided a holistic rating that indicated the degree to which the blueprint as a whole covered the Content Specifications

Table 3.6. Reviewer Tasks Associated with Connections D – Alignment between Item/Task Pools and Evidence Statements and G- Item/Task Pools and Content Specifications

Connection	Reviewer Tasks
D. Item/task pools to evidence statements AND G. Items/ performance tasks to Content Specifications	Workshops 3–5 *Reviewer tasks for Connections D and G were the same and, thus were examined using the same set of ratings; however, the analyses conducted for these two connections were different 1. Verified the items' and performance tasks' DOK, associated claim and targets, associated GS[1], and associated MP[1], as specified in the items' and performance tasks' metadata. Reviewers used a 3-point Likert scale (1 = not aligned, 2 = partially aligned, and 3 = fully aligned) to rate the degree to which the items matched their designated metadata. Additionally, for ratings of 1 or 2, reviewers described why the item did not fully align. Reviewers provided ratings to verify the: a. Item generally represented the evidence statement to which it was written b. Item's DOK c. Item's claims d. Item's targets e. Item's GS[2] f. Item's MP for math

[1]GS=Grade-level standards; MP=mathematical practice
[2]Where an item cannot be matched to a GS, reviewers will select one or more clusters as appropriate.

Table 3.7. Researcher Tasks Associated with Connection E – Alignment between CAT Algorithm and Test Blueprint

Connection	Researcher Tasks
E. CAT algorithm to test blueprint	1. Independently evaluated the Smarter Balanced CAT algorithm specifications and compared them to the specifications outlined in the test blueprint: a. Using the documents, determined if and to what degree the test blueprint DOK requirements were included in the CAT algorithm documentation. Information was unavailable at the time of the study b. Using the documents, determined if and to what degree the test blueprint content requirements were included in the CAT algorithm documentation. Information was unavailable at the time of the study

CHAPTER 4 - ANALYSES

As stated by Webb (1997)[22], an alignment study describes "the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do."(p.3). Webb's four alignment criteria (i.e., categorical concurrence, depth of knowledge consistency, range of knowledge correspondence, and balance of representation) have been widely used in studies to evaluate the degree of alignment between state assessments and state standards on static test forms. Due to the integrated design of the CCSS, the use of evidence-based design to develop the items, and the application of a CAT design, these criteria would not provide the appropriate evidence to support the validity of the Smarter Balanced summative assessments. Therefore, each connection (A through D and G) was evaluated and analyzed based on the alignment criteria listed in Table 4.1. This table presents the general definitions of each alignment criterion that were used across each connection to inform the validity of the Smarter Balanced summative assessments. Because this study examined the degree of alignment for an assessment developed using evidence-based design and included multiple links within the broader validity chain, the alignment criteria were interpreted and defined in slightly different ways across each connection. Details about the analyses that were conducted to examine each connection are presented in Appendix D.

Table 4.1. Alignment Criteria for Analyzing Alignment Data

Alignment Criterion	Description
Content Representation (CR)	Content representation examined the degree to which the content within an assessment component was aligned to another assessment component (e.g., the percentage of targets that were aligned to more than one evidence statement).
DOK Distribution (DD)	DOK distribution examined the breadth of cognitive demand associated with the elements or components included in this study that have DOK ranges assigned to them, such as claims, evidence statements or the CCSS MPs. We examined the percentage of these components at each DOK level (i.e., 1, 2, 3, and 4). Evaluating the DOK distribution included comparing ratings from reviewers in this study to the DOK indicated by the developer, which was indicated in the Content Specifications.
DOK Consistency (DC)	DOK consistency measured the extent to which the DOK of the item or evidence statement was consistent with the consensus DOK derived for the CCSS, and the Smarter Balanced claims and targets.
Agreement Between Reviewers' and Content Specifications/Developers' Ratings (PWAS)	This measure of agreement examined the degree to which there was consistency in ratings of reviewers and Content Specifications/item developers, in terms of indication of DOK and content match.
Agreement Among Reviewers' Ratings (PWA.Reviewers)	This measure of agreement examined the degree to which the different reviewers' ratings were consistent (i.e., inter-rater reliability) in terms of DOK and content match.

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

For each of the above criteria, we conducted analyses to address the questions associated with the various connections (refer to Table 4.2).

Table 4.2. Questions Addressed by Connection and Criterion

Connection	Criterion	Question
Connection A: Alignment of Content Specifications to CCSS	Content Representation	A.CR-1: Do the grade-level standards collectively reflect the content and skills required by the target?
		A.CR-2: Do the targets collectively reflect the content and skills required by the grade-level standard?[23]
		A.CR-3. Do the individual grade-level standards reflect the content and skills required by the intended targets?
		A.CR-4: Do the individual targets reflect the content and skills required by the intended grade-level standard?
		A.CR-5. Does each mathematical practice reflect skills required by the intended target?
		A.CR-6. Do the reviewers agree with the intended mapping of targets and grade-level standards as identified in the Content Specifications?
	DOK Distribution	A.DD-1. Does the DOK distribution of the targets identified by the reviewers, match that of the distribution identified in the Content Specifications (using the max DOK level)?
		A.DD-2. Does the DOK distribution of the targets identified by the reviewers, match that of the distribution identified in the Content Specifications (using the each independent DOK level)?
		A.DD-3. Do the reviewers agree with the intended target DOK levels as identified in the content specifications?
	DOK Consistency	A.DC-1: Is the cognitive complexity required in the targets consistent with the cognitive complexity required in each targets' mapped grade-level standards/mathematical practices?
Connection B: Alignment of Evidence Statements to Content Specifications	Content Representation	B.CR-1. Do the evidence statements collectively reflect the content and skills required by the target?
		B.CR-2. Do the individual evidence statements reflect the content and skills required by the intended targets?
	DOK Consistency	B.DC-1. Do reviewers' evidence statement DOK ratings align with the DOK levels specified for the targets to which they are mapped (as indicated in the Content Specifications)?
Connection C: Alignment of Test Blueprint to Content Specifications	Content Representation	C.CR-1. To what degree are the Content Specifications represented in the draft blueprints?

[23] This question addresses part of the two-way alignment approach.

Table 4.2. (Continued)

Connection	Criterion	Question
Connection D: Alignment of Item/Task Pools to Evidence Statements	Content Representation	D.CR-1. How are the summative assessment items distributed across evidence statements?
		D.CR-2. Do the reviewers agree with the intended mapping of items to evidence statements as identified by the item developers?
	DOK Consistency	D.DC-1. Is the cognitive complexity required in the items consistent with the cognitive complexity required in each evidence statement?
Connection E: Alignment of CAT Algorithm to Test Blueprint[24]	Content Representation	E.CR-1. How well are the content requirements outlined in the test blueprint reflected in the CAT algorithm specifications?
	DOK Consistency	E.DC-1. How well are the DOK requirements outlined in the test blueprint reflected in the CAT algorithm specifications?
Connection G: Alignment of Items/ Performance Tasks to Content Specifications	Content Representation	G.CR-1. How are the summative assessment items distributed across targets, grade-level standards, and mathematical practices?
		G.CR-2. Do the reviewers agree with the intended mapping of items to targets, grade-level standards, and mathematical practices as identified by the item developers?
		G.CR-3. Do the reviewers agree with the intended mapping of items to mathematical practices as identified by the item developers?
	DOK Distribution	G.DD-1. How does the distribution of DOK of the items identified by the reviewers compare with the distribution identified in the Content Specifications?
	DOK Consistency	G.DC-1. Does the DOK of the item identified by the reviewers fall within the DOK distribution of the aligned target identified in the Content Specifications?

[24] Based on the documentation received, we were unable to address our proposed questions related to this connection.

CHAPTER 5 - RESULTS

Overall

At the conclusion of each workshop, reviewers were asked to provide an overall rating for the alignment between the Smarter Balanced targets and the CCSS or items. For Workshops 1 and 2, reviewers were asked about their opinion of the alignment between the Smarter Balanced targets and the CCSS. As can be seen in Table 5.1, the majority of reviewers believed the alignment between the Smarter Balanced targets and the CCSS was acceptable to strong. For Workshops 3, 4, and 5, reviewers were asked for their opinion of the alignment between the mathematics or ELA/literacy items and the Smarter Balanced targets. These reviewers also believed the alignment between the items and the Smarter Balanced targets was acceptable to strong.

Table 5.1. Reviewers' General Opinion of Alignment[25]

Alignment Question	Overall		ELA		Math	
	Mean (n)	SD	Mean (n)	SD	Mean (n)	SD
What is your general opinion of the alignment between the Smarter Balanced assessment targets and the CCSS? (W1-2 only)	4.46 (79)	0.63	4.53 (45)	0.59	4.37 (34)	0.69
What is your general opinion of the alignment between the Math or ELA items and the Smarter Balanced assessment targets? (W3-5 only)	4.53 (123)	0.61	4.52 (66)	0.65	4.55 (57)	0.56

Results by Connection

This section of the report describes results of the analyses conducted for each connection that was examined in the study. Within each connection, results are presented separately by content area. Table 5.2 presents the Smarter Balanced claims for ELA/literacy and mathematics. Table 5.3 provides context for interpreting the results by presenting the number of targets by content area included in the Content Specifications as well as the number of targets that were included in the analyses.

[25] Scale points and definitions: 1, Not aligned in any way; 2, Needs major improvement; 3, Needs slight improvement; 4, Acceptable alignment; 5, Strong alignment.

Table 5.2. Smarter Balanced Claims by Content Area

Smarter Balanced Claim	
ELA/Literacy	
Claim 1	Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts.
Claim 2	Students can produce effective and well-grounded writing for a range of purposes and audiences.
Claim 3	Students can employ effective speaking and listening skills for a range of purposes and audiences.
Claim 4	Students can engage in research and inquiry to investigate topics, and to analyze, integrate, and present information.
Mathematics	
Claim 1	Concepts & Procedures: Students can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.
Claim 2	Problem Solving: Students can solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem solving strategies.
Claim 3	Communicating Reasoning: Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.
Claim 4	Modeling and Data Analysis: Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.

Table 5.3. Number of Targets per Grade and Claim for ELA/Literacy and Mathematics

Grade	Claim	ELA/Literacy		Mathematics
		# Targets[26]	# Targets Included in Study[27]	# Targets
3	1	14	14	11
	2	13	11	4
	3	4	1	6
	4	4	3	7
4	1	14	14	12
	2	13	11	4
	3	4	1	6
	4	4	3	7
5	1	14	14	11
	2	13	11	4
	3	4	1	6
	4	4	3	7
6	1	14	14	10
	2	13	11	4
	3	4	1	7
	4	4	3	7
7	1	14	14	9
	2	13	11	4
	3	4	1	7
	4	4	3	7
8	1	14	14	10
	2	13	11	4
	3	4	1	7
	4	4	3	7
11	1	14	14	16
	2	13	11	4
	3	4	1	7
	4	4	3	7

[26] Some targets were excluded from this study because Smarter Balanced did not include them on the summative assessments; these skills are to be addressed in formative assessments.
[27] Three Claim 2 ELA/literacy targets are specific to PT items.

Connection A: Alignment of Content Specifications to CCSS

ELA/Literacy

Analyses were conducted separately by content area to examine the alignment between the Content Specifications and the CCSS. These analyses focused on content representation, DOK distribution, and DOK consistency. Pairwise agreement among reviewers was computed for identifying grade-level standards aligned to each target (Workshop 1) and for identifying DOK levels for each target (Workshop 1); this information is presented in the Content Representation and DOK Distribution sections below. Additionally, both CAT and performance task targets were included under Claims 2 and 4.

Content Representation

Analyses were conducted to examine the representation of ELA/literacy content between the Content Specifications and the CCSS, and to address the following questions:

- A.CR-1: Do the grade-level standards collectively reflect the content and skills required by the target?
- A.CR-2: Do the targets collectively reflect the content and skills required by the grade-level standard?[28]
- A.CR-3: Do the individual grade-level standards reflect the content and skills required by the intended targets?
- A.CR-4: Do the individual grade-level standards reflect the content and skills required by the intended targets when reviewers were asked to identify targets aligned to each grade-level standard (Workshop 2)?
- A.CR-6[29]: Do the reviewers agree with the intended mapping of targets and grade-level standards as identified in the Content Specifications?

The main reviewer tasks for examining the content representation of the targets and grade-level standards involved identifying grade-level standards that represented the targets and providing a holistic target rating to indicate how well the grade-level standards represented the target. Because Smarter Balanced does not intend to measure all grade-level standards on the summative assessment, we excluded any grade-level standard that is solely intended to be measured by a target and that is excluded from the summative assessment. See Appendix E for a list of excluded standards.

Overall, across all grades, targets, and reviewers, reviewers identified an average of 11.3 unique grade-level standards per target (SD=7.5) compared to 4.7 unique grade-level standards (SD=3.4) identified in the Smarter Balanced Content Specifications. As further shown in Table 5.A.1, the reviewers in all grades independently identified more grade-level standards per target than were indicated in the Content Specifications. Using a greater than or equal to 50% reviewer agreement rule (per target), the mean percentage of unique grade-level standards per target dropped for all grades. This suggests that using an independent identification method (where reviewers had access to the full eligible pool of grade-level standards), resulted in reviewers aligning more of the grade-level standards to the targets than what was intended.

[28] This question addresses the two-way alignment approach.
[29] Analyses related to A.CR-5 involved mathematical practices and, therefore, was not relevant to ELA/literacy.

Table 5.A.1. A.CR.GD-1 Average Number of CCSS per Target Identified by Reviewers and in Specifications

Grade	# of CCSS per Target (Reviewers)		# of CCSS per Target (Reviewers ≥ 50% Agreement)		# of CCSS per Target (Specs)	
	Mean	SD	Mean	SD	Mean	SD
3	13.1	8.9	3.0	3.4	3.5	1.9
4	9.8	7.2	4.2	3.5	4.1	2.5
5	7.2	3.7	3.6	2.1	4.2	2.8
6	8.6	3.4	3.3	2.0	5.4	4.1
7	16.0	8.2	9.2	7.7	5.3	3.9
8	10.0	4.1	4.8	2.2	5.3	3.9
11	15.6	9.3	7.9	6.0	5.2	4.0

Pairwise Agreement among Reviewers

The overall pairwise agreement among reviewers in identifying grade-level standards aligned to each target (across all grades, claims, targets, and reviewers) was 38.67%. The rather low agreement rate was likely due to a combination of the high number of grade-level standards that reviewers identified for each target and the fact that reviewers conducted a blind rating where they were permitted to choose from a lengthy list of eligible grade-level standards (Table 5.A.2).

Table 5.A.2. A.ELA.CR.PWA-1. Pairwise Percent Agreement among Reviewers' Mapping of Targets and Grade-level Standards

		Descriptives		Agreement	
Grade	Claim	Avg # of Reviewers n	# of Targets	Avg # of Reviewer Pairs	Pairwise Agreement %
3	1	5.0	14	10.0	37.3%
	2	5.0	11	10.0	34.5%
	3	5.0	1	10.0	47.5%
	4	5.0	3	10.0	12.0%
4	1	5.0	14	10.0	53.2%
	2	5.0	11	10.0	46.2%
	3	5.0	1	10.0	15.0%
	4	4.7	3	8.7	35.8%
5	1	4.0	14	6.0	55.3%
	2	4.0	11	6.0	32.1%
	3	4.0	1	6.0	61.1%
	4	4.0	3	6.0	43.9%
6	1	5.0	14	10.0	49.4%
	2	5.0	11	10.0	44.6%
	3	5.0	1	10.0	42.5%
	4	5.0	3	10.0	24.1%
7	1	3.8	14	5.4	43.9%
	2	3.0	11	3.0	56.7%
	3	3.0	1	3.0	33.3%
	4	3.0	3	3.0	27.5%
8	1	4.0	14	6.0	51.8%
	2	4.0	11	6.0	31.8%
	3	4.0	1	6.0	30.6%
	4	4.0	3	6.0	21.3%

Table 5.A.2. (Continued)

Grade	Claim	Descriptives		Agreement	
		Avg # of Reviewers n	# of Targets	Avg # of Reviewer Pairs	Pairwise Agreement %
11	1	5.0	14	10.0	48.7%
	2	3.9	11	5.7	39.7%
	3	4.0	1	6.0	30.0%
	4	4.0	3	6.0	33.0%

Table Read: For grade 3, Claim 1, there was an average of 5 reviewers who rated 14 targets. The average number of total possible pairs was 10. The pairwise agreement across all 10 pairs for all 14 targets was 37.3%. A decimal for the number of reviewers and number of pairs indicates missing data.

Findings

Findings related to each content representation (CR) question are presented below. When relevant, general descriptive statistics provide overall results for each grade across claims.

A.CR-1: Do the grade-level standards collectively reflect the content and skills required by the target?

Once reviewers identified all the grade-level standards they thought aligned to the target, they provided a holistic target representation rating to indicate how well those standards collectively represented the content and knowledge required in the target. The scale ranged from 0 to 4 ('0' = not aligned at all, '1' = small-portion aligned, '2' = somewhat aligned, '3' = mostly aligned, and '4' = fully-aligned). As seen in Table 5.A.3, reviewers generally rated the targets as being well-represented by the grade-level standards they identified. The mean alignment rating across grades and claims ranged from 3.5 to 4.0.

Table 5.A.3. A.CR-1: ELA/Literacy Target Holistic Rating (Collectively Reflected by the Grade-Level Standards)

Grade	Target Representation Rating		
	$N_{target\_ratings}$	Mean	SD
3	145	3.9	0.4
4	144	3.5	0.7
5	116	3.8	0.6
6	145	3.8	0.5
7	98	3.8	0.4
8	116	4.0	0.1
11	125	3.9	0.4

Table Read: For grade 3, across 145 target ratings (across all claims and reviewers), the average rating was 3.9, with a standard deviation of 0.4.

As shown in Table 5.A.4, across grades and claims, the ELA/literacy targets were generally represented well by their intended grade-level standards. However, reviewers for grades 4, 5, 6, 7, and 11 consistently rated the targets for Claim 3 (Speaking and Listening skills) as being less fully aligned to their intended grade-level standards than did reviewers for grades 3 and 8 for these same targets. No additional information is learned by examining the reviewer comments; however, the reviewers in grades 4 and 5 indicated that had the full list of Speaking and Listening grade-level standards, been available to use, they might have rated the target as being more fully-represented. Additionally, reviewers for grade 4 rated the targets for Claim 1 (comprehend complex literary and informational texts) as less fully aligned to their intended grade-level standards than did reviewers for the other grades.

Table 5.A.4. A.CR-1: Mean Percentage of ELA/Literacy Targets at Each Holistic Rating (Collectively Reflected by the Grade-Level Standards)

Grade	Claim	# of Targets in Claim	Holistic Target Rating				
			Fully-aligned % (n)	Mostly-aligned % (n)	Somewhat-aligned % (n)	Small-portion aligned % (n)	Not-aligned at all % (n)
3	1	14	88.6% (12.4)	8.6% (1.2)	2.9% (0.4)	0.0% (0.0)	0.0% (0.0)
	2	11	96.4% (10.6)	0.0% (0.0)	3.6% (0.4)	0.0% (0.0)	0.0% (0.0)
	3	1	100.0% (1.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	4	3	80.0% (2.4)	6.7% (0.2)	13.3% (0.4)	0.0% (0.0)	0.0% (0.0)
4	1	14	45.7% (6.4)	41.4% (5.8)	12.9% (1.8)	0.0% (0.0)	0.0% (0.0)
	2	11	85.5% (9.4)	9.1% (1.0)	5.5% (0.6)	0.0% (0.0)	0.0% (0.0)
	3	1	40.0% (0.4)	40.0% (0.4)	20.0% (0.2)	0.0% (0.0)	0.0% (0.0)
	4	3	60.0% (1.6)	33.3% (1.0)	6.7% (0.2)	0.0% (0.0)	0.0% (0.0)
5	1	14	91.1% (12.8)	5.4% (0.8)	3.6% (0.5)	0.0% (0.0)	0.0% (0.0)
	2	11	77.3% (8.5)	18.2% (2.0)	4.5% (0.5)	0.0% (0.0)	0.0% (0.0)
	3[1]	1	25.0% (0.3)	25.0% (0.3)	25.0% (0.3)	25.0% (0.3)	0.0% (0.0)
	4	3	75.0% (2.3)	16.7% (0.5)	8.3% (0.3)	0.0% (0.0)	0.0% (0.0)
6	1	14	84.3% (11.8)	12.9% (1.8)	2.9% (0.4)	0.0% (0.0)	0.0% (0.0)
	2	11	87.3% (9.6)	10.9% (1.2)	1.8% (0.2)	0.0% (0.0)	0.0% (0.0)
	3	1	40.0% (0.4)	0.0% (0.0)	40.0% (0.4)	20.0% (0.2)	0.0% (0.0)
	4	3	86.7% (2.6)	13.3% (0.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
7	1	14	80.7% (10.8)	17.5% (2.3)	1.8% (0.3)	0.0% (0.0)	0.0% (0.0)
	2	11	97.0% (10.7)	3.0% (0.3)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	3	1	66.7% (0.7)	0.0% (0.0)	33.3% (0.3)	0.0% (0.0)	0.0% (0.0)
	4	3	88.9% (2.7)	11.1% (0.3)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
8	1	14	100.0% (14.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	2	11	97.7% (10.8)	2.3% (0.3)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	3	1	100.0% (1.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	4	3	91.7% (2.8)	8.3% (0.3)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
11	1	14	92.9% (13.0)	7.1% (1.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	2	10	90.9% (9.5)	4.5% (0.5)	0.0% (0.0)	0.0% (0.0)	4.5% (0.5)
	3	1	50.0% (0.5)	25.0% (0.3)	0.0% (0.0)	25.0% (0.3)	0.0% (0.0)
	4	3	83.3% (2.5)	0.0% (0.0)	0.0% (0.0)	8.3% (0.3)	8.3% (0.3)

*Note: Claims 3 and 4, across grades, are based on small numbers of targets; percentages should not be overinterpreted.
[1]For grade 5, Claim 3, reviewers inadvertently were not allowed to map the two grade-level standards to the only target on the summative assessment (Target 4). As a result, reviewers rated Target 4 as not represented well by the standards. The descriptive comments for those reviewers, however, indicate that had they been able to consider those two standards (SL.5.2 and SL.5.3); they would have stated that the target was fully-aligned.

Table Read: For grade 3, Claim 1, reviewers rated an average of 88.6% of the ELA/literacy targets (12.4 targets) as being fully aligned to the grade-level standards identified by the reviewers, 8.6% of the targets (1.2 targets) as being mostly aligned to the grade-level standards, 2.9% of the targets (0.4 targets) as being somewhat aligned to the grade-level standards, and none of the targets as being aligned a small portion or not aligned at all to the grade-level standards.

A.CR-2: Do the targets collectively reflect the content and skills required by the grade-level standard?

Data related to this question were collected during Workshop 2. In contrast to the tasks in Workshop 1, reviewers identified all the targets they believed aligned to each grade-level standard. They provided a holistic grade-level standard representation rating to indicate how well those targets collectively represented the content and knowledge required in the grade-level standard. The scale ranged from 0 to 4 ('0' = not aligned at all, '1' = small-portion aligned, '2' = somewhat aligned, '3' = mostly aligned, and '4' = fully-aligned). This analysis provides a general indication of how well the content and knowledge required in each individual grade-level standard was measured by the targets. Based on the development and structure of the Content Specifications, there was little expectation that each grade-level standard would be collectively represented by the targets[30]. Thus, low percentages of 'fully-aligned' and 'mostly-aligned' ratings do not necessarily reflect poor alignment.

As seen in Table 5.A.5, across grades and ELA/literacy strands, reviewers generally believed that the targets represented the content and knowledge required in the grade-level standards. The exceptions were the grade-level standards in the Language and Speaking and Listening strands for grades 3 through 5. For these strands, reviewers rated less than half of the grade-level standards as being fully represented by the targets. The majority of the comments provided by the reviewers regarding this alignment were related to the lack of focus of Speaking in the targets, for which Smarter Balanced does not assess on the summative assessment.

Table 5.A.5. A.CR-2: Mean Percentage of ELA/Literacy Grade-level Standards at Each Holistic Rating

		Holistic Target Rating				
Grade	Strand	Fully-aligned % (n)	Mostly-aligned % (n)	Somewhat-aligned % (n)	Small-portion aligned % (n)	Not-aligned at all % (n)
3	L	25.3% (3.0)	16.2% (2.0)	44.5% (5.8)	14.0% (1.8)	0.0% (0.0)
	RI	64.9% (5.8)	28.8% (2.5)	3.1% (0.3)	3.1% (0.3)	0.0% (0.0)
	RL	75.0% (6.0)	15.6% (1.3)	9.4% (0.8)	0.0% (0.0)	0.0% (0.0)
	SL	25.0% (0.5)	50.0% (1.0)	0.0% (0.0)	25.0% (0.5)	0.0% (0.0)
	W	56.7% (8.5)	33.3% (5.0)	6.7% (1.0)	3.3% (0.5)	0.0% (0.0)
4	L	47.6% (6.0)	27.4% (3.5)	9.6% (1.3)	15.4% (2.0)	0.0% (0.0)
	RI	72.2% (6.5)	25.0% (2.3)	2.8% (0.3)	0.0% (0.0)	0.0% (0.0)
	RL	53.6% (3.8)	21.4% (1.5)	3.6% (0.3)	21.4% (1.5)	0.0% (0.0)
	SL	12.5% (0.3)	62.5% (1.3)	25.0% (0.5)	0.0% (0.0)	0.0% (0.0)
	W	43.8% (7.8)	53.4% (9.5)	2.8% (0.5)	0.0% (0.0)	0.0% (0.0)
5	L	47.9% (5.5)	24.2% (2.8)	13.3% (1.5)	14.6% (1.8)	0.0% (0.0)
	RI	75.4% (6.5)	21.8% (1.8)	2.8% (0.3)	0.0% (0.0)	0.0% (0.0)
	RL	73.2% (4.8)	26.8% (1.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	W	66.7% (11.8)	30.6% (5.5)	2.8% (0.5)	0.0% (0.0)	0.0% (0.0)

[30] Additional standards were also excluded for this analysis. There are some grade-level standards that are not listed at all in the Content Specifications (under either assessed or not-assessed targets). These standards were left in for Workshop 1, but reviewers were not required to use them. In Workshop 2, reviewers were required to provide ratings for all non-excluded standards. Thus, we excluded the standards that were *not present* in the Content Specifications, in addition to excluding the standards that were only being measured by targets not assessed on the summative assessment.

Table 5.A.5. (Continued)

		Holistic Target Rating				
Grade	Strand	Fully-aligned % (n)	Mostly-aligned % (n)	Somewhat-aligned % (n)	Small-portion aligned % (n)	Not-aligned at all % (n)
6	L	88.7% (12.0)	9.3% (1.3)	1.9% (0.3)	0.0% (0.0)	0.0% (0.0)
	RH	69.4% (6.3)	16.7% (1.5)	11.1% (1.0)	2.8% (0.3)	0.0% (0.0)
	RI	75.0% (6.8)	16.7% (1.5)	8.3% (0.8)	0.0% (0.0)	0.0% (0.0)
	RL	55.4% (3.8)	37.5% (2.5)	7.1% (0.5)	0.0% (0.0)	0.0% (0.0)
	RST	65.6% (5.3)	25.0% (2.0)	9.4% (0.8)	0.0% (0.0)	0.0% (0.0)
	SL	75.0% (1.5)	0.0% (0.0)	25.0% (0.5)	0.0% (0.0)	0.0% (0.0)
	W	85.0% (17.0)	12.5% (2.5)	2.5% (0.5)	0.0% (0.0)	0.0% (0.0)
	WHST	75.0% (1.5)	25.0% (0.5)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
7	L	96.4% (13.5)	1.8% (0.3)	0.0% (0.0)	1.8% (0.3)	0.0% (0.0)
	RH	83.3% (7.5)	11.1% (1.0)	5.6% (0.5)	0.0% (0.0)	0.0% (0.0)
	RI	88.9% (8.0)	8.3% (0.8)	2.8% (0.3)	0.0% (0.0)	0.0% (0.0)
	RL	89.3% (6.0)	10.7% (0.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	RST	75.0% (6.0)	18.8% (1.5)	3.1% (0.3)	3.1% (0.3)	0.0% (0.0)
	SL	100.0% (2.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	W	85.0% (17.0)	8.8% (1.8)	5.0% (1.0)	1.3% (0.3)	0.0% (0.0)
	WHST	75.0% (1.5)	25.0% (0.5)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
8	L	67.9% (9.5)	5.4% (0.8)	0.0% (0.0)	26.8% (3.8)	0.0% (0.0)
	RH	94.4% (8.5)	5.6% (0.5)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	RI	77.8% (7.0)	22.2% (2.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	RL	89.3% (6.3)	10.7% (0.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	RST	100.0% (8.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	SL	50.0% (1.0)	0.0% (0.0)	25.0% (0.5)	25.0% (0.5)	0.0% (0.0)
	W	76.2% (15.0)	20.1% (4.0)	3.8% (0.8)	0.0% (0.0)	0.0% (0.0)
	WHST	100.0% (2.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
11	L	81.8% (9.0)	9.1% (1.0)	2.3% (0.3)	6.8% (0.8)	0.0% (0.0)
	RH	94.4% (8.5)	5.6% (0.5)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	RI	96.9% (8.3)	3.1% (0.3)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	RL	89.3% (6.3)	10.7% (0.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	RST	91.7% (8.3)	5.6% (0.5)	2.8% (0.3)	0.0% (0.0)	0.0% (0.0)
	SL	50.0% (1.0)	25.0% (0.5)	0.0% (0.0)	25.0% (0.5)	0.0% (0.0)
	W	80.0% (15.8)	20.0% (4.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	WHST	100.0% (2.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)

[1]For grade 5, Claim 3, reviewers inadvertently were not allowed to map the two grade-level standards to the only target on the summative assessment (Target 4). This resulted in no targets identified for grade-level standards from the Speaking and Listening strand.

Table Read: For grade 3 Language grade-level standards, reviewers rated an average of 25.0% of them (average of 3.0 standards) as fully represented by the targets, an average of 16.2%% of them (average of 2.0 standards) as mostly aligned, an average of 44.5% of them (average of 5.8 standards), an average of 14.0% of them (average of 1.8 standards) as being a small portion aligned, and 0% of the standards as being not aligned.

A.CR-3: Do the individual grade-level standards reflect the content and skills required by the intended targets?

As noted earlier, on average, reviewers identified more ELA/literacy grade-level standards per target than what was intended in the Content Specifications. With the exception of Claim 4 (Research and Inquiry), that pattern holds true at the claim level as well (Table 5.A.6).

Table 5.A.6 A.ELA.CR-3.GD-1 Comparison of ELA/Literacy Grade-level Standards per Target as Identified by Reviewers and the Content Specifications

			Reviewer Descriptives			Specifications Descriptives		
Grade	Claim	Total number of targets in claim	Avg # of grade-level standards per target	Min # of grade-level standards per target	Max # of grade-level standards per target	Avg # of grade-level standards per target	Min # of grade-level standards per target	Max # of grade-level standards per target
		n	n	n	n	n	n	N
3	1	14	2.3	1	9	2.2	1	5
	2	11	8.1	1	26	4.9	2	8
	3	1	3.6	2	6	2	2	2
	4	3	3.3	1	7	4.3	3	6
4	1	14	2.2	1	5	2.2	1	5
	2	11	9.4	1	19	6.3	3	9
	3	1	1.4	1	2	2	2	2
	4	3	2.4	1	6	5.3	3	7
5	1	14	2.5	1	6	2.1	1	5
	2	11	4.9	1	13	6.6	2	10
	3	1	2	1	3	2	2	2
	4	3	3.4	1	6	5.3	3	7
6	1	14	3.1	1	9	3.4	1	9
	2	11	5.2	2	9	6.7	3	10
	3	1	2.2	1	4	2	2	2
	4	3	2.9	1	5	11.3	2	18
7	1	14	4.8	1	18	3.4	1	9
	2	11	16.1	4	29	6.5	3	10
	3	1	3	1	4	2	2	2
	4	3	9.2	4	12	11	2	18
8	1	14	4	1	12	3.4	1	9
	2	11	5.8	1	12	6.5	3	10
	3	1	3.5	1	6	2	2	2
	4	3	3.8	2	6	11	2	18
11	1	14	5.5	2	11	3.3	1	9
	2	10	11.7	1	29	6.1	2	9
	3	1	2.8	1	5	2	2	2
	4	3	5.8	1	11	12.3	5	19

Table Read: For grade 3, Claim 1, there are 14 targets. The reviewers, on average, identified 2.3 grade-level standards per target, with a minimum of 1 grade-level standard and a maximum of 9 grade-level standards. The average number of grade-level standards identified in the Content Specifications was 2.2, with a minimum of 1 grade-level standard and a maximum of 5 grade-level standards.

Due to the rather high number of ELA/literacy grade-level standards identified by the reviewers compared to what was intended by the Content Specifications, we imposed a ≥50% reviewer agreement rule for this analysis as a method for removing outliers. More specifically, when examining the degree to which reviewers identified the same grade-level standards as was intended by the Content Specifications, only those standards for each target that had at least 50% reviewer agreement were retained.

As shown in Table 5.A.7, overall, there were only a few ELA/literacy targets across claims and grades that did not have grade-level standards with at least 50% reviewer agreement and thus, were not included in the analysis. A fairly large average percentage of the grade-level standards per target rated as matching the intended mapping. Where there wasn't 100%, most of those grade-level standards per target were believed by the reviewers to fall within the intended strand, as specified in the Content Specifications.

Table 5.A.7. A.CR-3: Mean Percentage of ELA/Literacy Grade-level Standards Aligned to Intended Targets (Workshop 1)

Grade	Claim		≥ 50% Reviewer Agreement Descriptives		Content Representation		
		Total number of targets in claim	Number of targets included in analysis[1]	Avg # of grade-level standards per target with 50% reviewer agreement	Avg % of grade-level standards per target that matched the intended mapping	Avg % of grade-level standards per target that fell within the intended strands	Avg % of grade-level standards per target that fell outside the intended strands
		N	n	n	% (n)	% (n)	% (n)
3	1	14	13	1.4	65.4% (0.9)	100.0% (1.4)	0.0% (0.0)
	2	11	11	5.0	49.2% (2.5)	100.0% (5.0)	0.0% (0.0)
	3	1	1	4.0	50.0% (2.0)	100.0% (4.0)	0.0% (0.0)
	4	3	2	2.0	75.0% (1.5)	100.0% (2.0)	0.0% (0.0)
4	1	14	14	1.9	75.0% (1.3)	100.0% (1.9)	0.0% (0.0)
	2	11	11	7.9	35.8% (3.0)	66.1% (5.1)	33.9% (0.3)
	3	1	.	.	.	.	.
	4	3	3	1.3	100.0% (1.3)	100.0% (1.3)	0.0% (0.0)
5	1	14	14	1.8	71.9% (1.1)	100.0% (1.8)	0.0% (0.0)
	2	11	11	2.8	42.4% (1.7)	97.0% (2.7)	3.0% (0.0)
	3[2]	1	1	2.0	0.0% (0.0)	100.0% (2.0)	0.0% (0.0)
	4	3	3	2.0	100.0% (2.0)	100.0% (2.0)	0.0% (0.0)
6	1	14	14	2.4	78.1% (1.8)	97.6% (2.4)	2.4% (0.0)
	2	11	11	4.6	65.9% (3.4)	100.0% (4.6)	0.0% (0.0)
	3	1	1	1.0	100.0% (1.0)	100.0% (1.0)	0.0% (0.0)
	4	3	2	2.5	83.3% (2.0)	100.0% (2.5)	0.0% (0.0)
7	1	14	14	2.9	68.5% (1.7)	96.4% (2.8)	3.6% (0.0)
	2	11	11	17.2	32.5% (5.4)	53.6% (8.4)	46.4% (0.5)
	3	1	1	2.0	100.0% (2.0)	100.0% (2.0)	0.0% (0.0)
	4	3	3	6.3	48.3% (3.3)	83.3% (5.3)	16.7% (0.2)
8	1	14	13	3.2	66.0% (1.7)	87.8% (2.7)	12.2% (0.1)
	2	11	11	3.5	37.9% (2.1)	79.7% (3.0)	20.3% (0.2)
	3	1	1	1.0	100.0% (1.0)	100.0% (1.0)	0.0% (0.0)
	4	3	1	1.0	100.0% (1.0)	100.0% (1.0)	0.0% (0.0)

Table 5.A.7. (Continued)

Grade	Claim		≥ 50% Reviewer Agreement Descriptives		Content Representation		
		Total number of targets in claim N	Number of targets included in analysis[1] n	Avg # of grade-level standards per target with 50% reviewer agreement n	Avg % of grade-level standards per target that matched the intended mapping % (n)	Avg % of grade-level standards per target that fell within the intended strands % (n)	Avg % of grade-level standards per target that fell outside the intended strands % (n)
11	1	14	14	4.3	41.5% (1.6)	75.4% (3.1)	24.6% (0.2)
	2	10	10	7.8	41.8% (3.2)	57.6% (4.3)	42.4% (0.4)
	3	1	1	2.0	100.0% (2.0)	100.0% (2.0)	0.0% (0.0)
	4	3	3	2.7	83.3% (2.3)	100.0% (2.7)	0.0% (0.0)

[1]Number of targets with at least one standard with 50% reviewer agreement
[2]For grade 5, Claim 3, reviewers inadvertently were not allowed to map the two grade-level standards to the only target on the summative assessment. As a result, the average percentage of grade-level standards that matched the intended mapping is 0%.

Table Read: For grade 3, Claim 1, there were a total of 14 targets, while only 13 targets were included in the analysis (because one target didn't have any grade-level standards with at least 50% reviewer agreement). Of the targets included in the analysis, there was an average of 1.4 grade-level standards per target. Of the grade-level standards for which at least 50% of reviewers agreed, reviewers rated an average of 65.4% of grade-level standards as matching the intended mapping. Reviewers rated 100% of the grade-level standards per target as falling within the intended strands. Reviewers rated none of the grade-level standards per target as falling outside the intended strands.

A.CR-4: Do the individual grade-level standards reflect the content and skills required by the intended targets when reviewers are asked to identify targets aligned to each grade-level standard (Workshop 2)?

Workshop 2 data, where reviewers identified ELA/literacy targets aligned to each grade-level standard, were restructured to match that of the format of Workshop 1 data (where reviewers provided grade-level standards aligned to each target). This was done to allow for examining the two-way alignment of the targets and grade-level standards in comparison to what was intended by the Content Specifications. If the alignment was reciprocal, then the results from both analyses (A.CR-3 and A.CR-4) would be similar. If the results differed, then the alignment could be impacted by methodological (e.g., the number of grade-level standards made it difficult to perform a blind review) or content-related factors (e.g., the broad content in the target or the grade-level standards might have made the rating task very difficult).

As shown in Table 5.A.8, the task of identifying targets aligned to each grade-level standard was more difficult than identifying grade-level standards that represented the content and knowledge required in the target. The average percentage of grade-level standards per target that matched the intended mapping was approximately 46% across grades. With the exceptions of grades 8 and 11, Claim 3 (Speaking and Listening) targets were rated as being least represented by the grade-level standards.

Table 5.A.9 shows the difference in content representation of the targets by the grade-level standards using the two-way alignment method. Targets were well represented by the grade-level standards when the reviewers' task was to identify grade-level standards aligned to each target (Workshop 1). Reversing the task resulted in weaker content representation (Workshop 2). These results likely suggest that the broad nature of the targets made it more difficult to align the targets to specific standards. Workshop 2 activities permitted reviewers to rate a grade-level standard as not

represented by any targets; however, most reviewers found targets that represented at least a small amount of the content and knowledge required in the grade-level standards. This resulted in a higher number of grade-level standards being aligned to each target, thus the average percentage of grade-level standards that matched the intended mapping inherently decreased for Workshop 2.

Table 5.A.8. A.CR-4: Mean Percentage of ELA/Literacy Grade-level Standards Aligned to Intended Targets Based on Reviewers Identifying Targets Aligned to Each Grade-level Standard (Workshop 2)

Grade	Claim	≥ 50% Reviewer Agreement Descriptives			Content Representation		
		Total number of targets in claim	Number of targets included in analysis[1]	Avg number of grade-level standards per target with 50% reviewer agreement	Avg % of grade-level standards per target that matched the intended mapping	Avg % of grade-level standards per target that fell within the intended strands	Avg % of grade-level standards per target that fell outside the intended strands
		N	n	n	% (n)	% (n)	% (n)
3	1	14	14	5.5	30.8% (1.6)	91.3% (4.9)	8.7% (0.1)
	2	11	11	5.5	64.8% (3.5)	90.5% (4.8)	9.5% (0.1)
	3	1	1	19.0	10.5% (2.0)	10.5% (2.0)	89.5% (0.9)
	4	3	3	12.7	30.0% (3.7)	55.7% (7.0)	44.3% (0.4)
4	1	14	14	5.1	36.3% (1.8)	87.5% (4.4)	12.5% (0.1)
	2	11	11	6.2	76.4% (4.3)	94.6% (5.8)	5.4% (0.1)
	3	1	1	16.0	12.5% (2.0)	12.5% (2.0)	87.5% (0.9)
	4	3	3	14.3	38.5% (5.3)	97.9% (14.0)	2.1% (0.0)
5	1	14	14	3.8	50.8% (1.7)	92.7% (3.4)	7.3% (0.1)
	2	11	11	6.8	70.7% (4.6)	94.8% (6.5)	5.2% (0.1)
	3[2]	1	1	5.0	0.0% (0.0)	0.0% (0.0)	100.0% (1.0)
	4	3	3	15.3	32.1% (4.7)	100.0% (15.3)	0.0% (0.0)
6	1	14	14	5.4	50.8% (2.6)	80.9% (4.4)	19.1% (0.2)
	2	11	11	4.2	64.9% (2.9)	96.6% (4.0)	3.4% (0.0)
	3	1	1	13.0	15.4% (2.0)	15.4% (2.0)	84.6% (0.8)
	4	3	3	6.0	58.3% (3.7)	81.9% (5.0)	18.1% (0.2)
7	1	14	14	4.7	54.3% (2.4)	78.4% (3.9)	21.6% (0.2)
	2	11	10	4.2	66.5% (3.3)	97.5% (4.1)	2.5% (0.0)
	3	1	1	9.0	22.2% (2.0)	22.2% (2.0)	77.8% (0.8)
	4	3	3	5.0	73.3% (3.7)	86.7% (4.3)	13.3% (0.1)
8	1	14	14	6.4	36.4% (2.3)	77.5% (5.1)	22.5% (0.2)
	2	11	11	8.2	59.9% (4.3)	82.5% (6.4)	17.5% (0.2)
	3	1	1	4.0	50.0% (2.0)	50.0% (2.0)	50.0% (0.5)
	4	3	3	9.3	47.6% (4.3)	64.7% (6.0)	35.3% (0.4)
11	1	14	14	7.6	35.5% (2.8)	79.7% (6.0)	20.3% (0.2)
	2	10	10	10.1	62.0% (5.3)	91.3% (8.8)	8.7% (0.1)
	3	1	1	3.0	66.7% (2.0)	66.7% (2.0)	33.3% (0.3)
	4	3	3	7.0	76.0% (5.3)	89.7% (6.3)	10.3% (0.1)

These data are from Workshop 2 (A.CR-3 was from Workshop 1). Reviewers identified targets aligned to each standard

[1]Number of targets with at least one standard with 50% reviewer agreement

[2]For grade 5, Claim 3, reviewers inadvertently were not allowed to map the two grade-level standards to the only target on the summative assessment. As a result, the grade-level standards reviewers identified all fell outside the intended strands.

Table Read: For grade 3, there were 14 targets in Claim 1 and all 14 targets were included in the analysis (because they all had at least one grade-level standard with 50% reviewer agreement) and there was an average of 5.5 grade-level standards per target that had 50% reviewer agreement. Reviewers rated an average of 30.8% of the grade-level standards (average of 1.6 standards) per target that matched the intended mapping of standards and target. Reviewers rated an average of 91.3% of the grade-level standards (average of 4.9 standards) per target as falling within the intended ELA strands. Reviewers rated an average of 8.7% of the grade-level standards (average of 0.1 standards) per target as falling outside the intended ELA strands.

Table 5.A.9 A.CR-4.Supp-1 Comparison of Mean Percentage of ELA/Literacy Grade-level Standards Aligned to Intended Targets (Workshops 1. vs 2)

		CR-3 (Workshop 1)		CR-4 (Workshop 2)		Difference (CR4-CR3)		
Grade	Claim	Avg number grade-level standards per target with 50% reviewer agreement	Avg % grade-level standards per target that matched the intended mapping	Avg number grade-level standards per target with 50% reviewer agreement	Avg % grade-level standards per target that matched the intended mapping	Avg number grade-level standards per target with 50% reviewer agreement	Avg % grade-level standards per target that matched the intended mapping	
		n	% (n)	n	% (n)	n	%	n
3	1	1.4	65.4% (0.9)	5.5	30.8% (1.6)	4.1	-34.6%	0.7
	2	5.0	49.2% (2.5)	5.5	64.8% (3.5)	0.5	15.6%	1
	3	4.0	50.0% (2.0)	19.0	10.5% (2.0)	15.0	-39.5%	0
	4	2.0	75.0% (1.5)	12.7	30.0% (3.7)	10.7	-45.0%	2.2
4	1	1.9	75.0% (1.3)	5.1	36.3% (1.8)	3.2	-38.7%	0.5
	2	7.9	35.8% (3.0)	6.2	76.4% (4.3)	-1.7	40.6%	1.3
	3	0%	.	16.0	12.5% (2.0)	16.0	.	.
	4	1.3	100.0% (1.3)	14.3	38.5% (5.3)	13.0	-61.5%	4
5	1	1.8	71.9% (1.1)	3.8	50.8% (1.7)	2.0	-21.1%	0.6
	2	2.8	42.4% (1.7)	6.8	70.7% (4.6)	4.0	28.3%	2.9
	3	2.0	0.0% (0.0)	5.0	0.0% (0.0)	3.0	0.0%	0
	4	2.0	100.0% (2.0)	15.3	32.1% (4.7)	13.3	-67.9%	2.7
6	1	2.4	78.1% (1.8)	5.4	50.8% (2.6)	3.0	-27.3%	0.8
	2	4.6	65.9% (3.4)	4.2	64.9% (2.9)	-0.4	-1.0%	-0.5
	3	1.0	100.0% (1.0)	13.0	15.4% (2.0)	12.0	-84.6%	1
	4	2.5	83.3% (2.0)	6.0	58.3% (3.7)	3.5	-25.0%	1.7
7	1	2.9	68.5% (1.7)	4.7	54.3% (2.4)	1.8	-14.2%	0.7
	2	17.2	32.5% (5.4)	4.2	66.5% (3.3)	-13.0	34.0%	-2.1
	3	2.0	100.0% (2.0)	9.0	22.2% (2.0)	7.0	-77.8%	0
	4	6.3	48.3% (3.3)	5.0	73.3% (3.7)	-1.3	25.0%	0.4
8	1	3.2	66.0% (1.7)	6.4	36.4% (2.3)	3.2	-29.6%	0.6
	2	3.5	37.9% (2.1)	8.2	59.9% (4.3)	4.7	22.0%	2.2
	3	1.0	100.0% (1.0)	4.0	50.0% (2.0)	3.0	-50.0%	1
	4	1.0	100.0% (1.0)	9.3	47.6% (4.3)	8.3	-52.4%	3.3
11	1	4.3	41.5% (1.6)	7.6	35.5% (2.8)	3.3	-6.0%	1.2
	2	7.8	41.8% (3.2)	10.1	62.0% (5.3)	2.3	20.2%	2.1
	3	2.0	100.0% (2.0)	3.0	66.7% (2.0)	1.0	-33.3%	0
	4	2.7	83.3% (2.3)	7.0	76.0% (5.3)	4.3	-7.3%	3

A.CR-6: Do the reviewers agree with the intended mapping of targets and grade-level standards as identified in the Content Specifications?

The overall pairwise agreement in identifying grade-level standards aligned to each target between reviewers and the intended mapping as identified in the Content Specifications (across all grades, claims, targets, and reviewers) was 36.4%. The rather low agreement was likely a result of the high number of grade-level standards that reviewers identified for each target compared to the number of grade-level standards identified in the Content Specifications. As shown in Table 5.A.10, however, reviewer agreement with the intended mapping increased when computing the average percent of reviewers per target that agreed with at least 50% of the intended standards. This suggests that while there was low overall agreement in identifying exactly what was intended, reviewers generally agreed with at least 50% of the intended standards.

Table 5.A.10. A.CR-6: Pairwise Agreement between Reviewers' and Intended Mapping of ELA/Literacy Targets and Grade-level Standards

		Descriptives			Agreement			
Grade	Claim	# of Reviewers % (n)	# of Targets	# of Ratings	Pairwise Agree-ment	Hit All Intended Standards (but noted others) Avg % (n Reviewers)	Hit At Least 50% of the Intended Standards Avg % (n Reviewers)	Hit 0% of the Intended Standards Avg % (n Reviewers)
3	1	5.0	14	70	38.9%	11.4% (0.6)	60.0% (3.0)	28.6% (1.4)
	2	5.0	11	55	33.9%	20.0% (1.0)	63.6% (3.2)	27.3% (1.4)
	3	5.0	1	5	43.3%	40.0% (2.0)	80.0% (4.0)	20.0% (1.0)
	4	5.0	3	15	29.8%	0.0% (0.0)	26.7% (1.3)	13.3% (0.7)
4	1	5.0	14	70	51.1%	22.9% (1.1)	72.9% (3.6)	10.0% (0.5)
	2	5.0	11	55	36.1%	1.8% (0.1)	74.5% (3.7)	20.0% (1.0)
	3	5.0	1	5	30.0%	0.0% (0.0)	40.0% (2.0)	60.0% (3.0)
	4	4.7	3	14	29.8%	0.0% (0.0)	15.0% (0.7)	6.7% (0.3)
5	1	4.0	14	56	50.0%	32.1% (1.3)	82.1% (3.3)	3.6% (0.1)
	2	4.0	11	44	32.6%	11.4% (0.5)	45.5% (1.8)	27.3% (1.1)
	3[1]	4.0	1	4	0.0%	0.0% (0.0)	0.0% (0.0)	100.0% (4.0)
	4	4.0	3	12	40.9%	0.0% (0.0)	16.7% (0.7)	0.0% (0.0)
6	1	5.0	14	70	49.2%	18.6% (0.9)	65.7% (3.3)	7.1% (0.4)
	2	5.0	11	55	42.3%	3.6% (0.2)	56.4% (2.8)	18.2% (0.9)
	3	5.0	1	5	56.7%	20.0% (1.0)	100.0% (5.0)	0.0% (0.0)
	4	5.0	3	15	22.6%	0.0% (0.0)	33.3% (1.7)	6.7% (0.3)
7	1	3.8	14	53	49.3%	27.4% (1.0)	78.0% (2.9)	4.2% (0.1)
	2	3.0	11	33	35.3%	24.2% (0.7)	97.0% (2.9)	0.0% (0.0)
	3	3.0	1	3	50.0%	66.7% (2.0)	100.0% (3.0)	0.0% (0.0)
	4	3.0	3	9	23.4%	11.1% (0.3)	11.1% (0.3)	22.2% (0.7)
8	1	4.0	14	56	40.8%	25.0% (1.0)	66.1% (2.6)	10.7% (0.4)
	2	4.0	11	44	32.4%	2.3% (0.1)	47.7% (1.9)	27.3% (1.1)
	3	4.0	1	4	41.7%	50.0% (2.0)	100.0% (4.0)	0.0% (0.0)
	4	4.0	3	12	21.8%	8.3% (0.3)	25.0% (1.0)	8.3% (0.3)
11	1	5.0	14	70	31.3%	32.9% (1.6)	75.7% (3.8)	5.7% (0.3)
	2	4.0	10	40	31.3%	17.5% (0.7)	77.5% (3.1)	5.0% (0.2)
	3	4.0	1	4	51.7%	50.0% (2.0)	75.0% (3.0)	25.0% (1.0)
	4	4.0	3	12	24.4%	0.0% (0.0)	8.3% (0.3)	8.3% (0.3)

[1]For grade 5, Claim 3, reviewers inadvertently were not allowed to map the two grade-level standards to the only target on the summative assessment; therefore, there were no data to map grade-level standards to Target 4. As a result, reviewers did not identify the standards that were mapped to Target 4 in the Content Specifications.

Table Read: For grade 3, there were 5 reviewers who rated 14 targets in Claim 1 for a total of 70 ratings (pairwise comparisons with the Content Specifications). Of those ratings, the pairwise agreement of the mappings of targets and grade-level standards was 38.9%. Diagnostically, an average of 11.4% of the reviewers per target identified all of the intended standards, while also indicating additional standards. An average of 60% of the reviewers per target hit at least 50% of the intended standards and an average of 28.98% of the reviewers per target hit 0% of the intended standards.

DOK Distribution

Analyses were conducted to examine the distribution of ELA/literacy target DOK levels between the Content Specifications and the reviewers, and to address the following questions:

- A.DD-1: Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the Content Specifications (using the max DOK level)?
- A.DD-2: Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the content specifications (using the each independent DOK level)?
- A.DD-3: Do the reviewers agree with the intended target DOK levels as identified in the Content Specifications?

The main reviewer tasks for examining the DOK distribution of the targets involved reviewers providing independent DOK ratings for each target. The purpose of these analyses was to describe and compare the cognitive demand required in the targets as identified by the reviewers with the cognitive demand indicated in the Content Specifications.

Because the Content Specifications often indicate more than one DOK level per target, reviewers were also allowed to identify more than one DOK level per target. Overall, no clear patterns emerged in terms of how many cognitive demand levels reviewers indicated targets required compared to the Content Specifications (Table 5.A.11).

Table 5.A.11. DD-GD. Overall Descriptive Comparison of Reviewer and Content Specifications ELA/Literacy Target DOK Ratings

Grade	Number of DOK Levels Indicated by Reviewers			Number of DOK Levels Indicated by Content Specifications		
	N	Mean	SD	N	Mean	SD
3	145	2.2	0.6	29	1.6	0.6
4	144	1.8	0.7	29	1.6	0.6
5	116	1.4	0.5	29	1.6	0.6
6	145	1.1	0.2	29	1.6	0.6
7	98	1.2	0.4	29	1.6	0.7
8	116	1.6	0.6	29	1.6	0.7
11	126	2.3	0.7	28	1.7	0.7

A more informative pattern emerges when the number of levels used per target is disaggregated by claim (see Table 5.A.12). Across grades, more DOK levels are indicated in the Content Specifications for Claims 3 and 4 than are indicated by the reviewers.

Table 5.A.12. A.ELA.DD.GD-1 Descriptive Comparison of Reviewer and Content Specifications ELA/Literacy Target DOK Ratings by Grade and Claim

		Avg # DOK Levels Indicated per Target	
Grade	Claim	Reviewers	Content Specifications
3	1	2.1	1.9
	2	2.2	1.1
	3	2.4	3.0
	4	2.0	2.0
4	1	1.7	1.9
	2	1.9	1.1
	3	1.8	3.0
	4	1.8	2.0
5	1	1.4	1.9
	2	1.4	1.1
	3	1.5	3.0
	4	1.5	2.0
6	1	1.0	1.7
	2	1.1	1.1
	3	1.2	3.0
	4	1.1	2.7
7	1	1.2	1.7
	2	1.2	1.1
	3	1.0	3.0
	4	1.1	2.7
8	1	1.5	1.7
	2	1.7	1.1
	3	1.5	3.0
	4	1.7	2.7
11	1	2.2	1.7
	2	2.5	1.2
	3	3.5	3.0
	4	2.6	3.0

Table Read: Across grade 3, reviewers reported an average of 2.1 DOK levels for the targets in Claim 1 while the Content Specifications indicated an average of 1.9 DOK levels for the targets in Claim 1.

Pairwise Agreement among Reviewers

The overall pairwise agreement among reviewers in identifying DOK levels for each ELA/literacy target (across all grades, claims, targets, and reviewers) was 64.5% (refer to Table 5.A.13).

Table 5.A.13. A.ELA.DD.PWA-1. Pairwise Percent Agreement among Reviewers' ELA/Literacy Target DOK Ratings

		Descriptives			Agreement
Grade	Claim	Avg # of Reviewers n	# of Targets	Avg # of reviewer pairs n	Pairwise Agreement
3	1	5.0	14	10.0	63.5%
	2	5.0	11	10.0	54.8%
	3	5.0	1	10.0	66.7%
	4	5.0	3	10.0	69.4%
4	1	5.0	14	10.0	72.9%
	2	5.0	11	10.0	52.8%
	3	5.0	1	10.0	45.0%
	4	4.7	3	8.7	47.8%
5	1	4.0	14	6.0	64.5%
	2	4.0	11	6.0	68.2%
	3	4.0	1	6.0	50.0%
	4	4.0	3	6.0	61.1%
6	1	5.0	14	10.0	67.9%
	2	5.0	11	10.0	62.7%
	3	5.0	1	10.0	80.0%
	4	5.0	3	10.0	36.7%
7	1	3.8	14	5.4	66.1%
	2	3.0	11	3.0	72.7%
	3	3.0	1	3.0	100.0%
	4	3.0	3	3.0	88.9%
8	1	4.0	14	6.0	73.4%
	2	4.0	11	6.0	65.9%
	3	4.0	1	6.0	33.3%
	4	4.0	3	6.0	61.1%
11	1	5.0	14	10.0	71.9%
	2	3.9	11	5.7	70.2%
	3	4.0	1	6.0	83.3%
	4	4.0	3	6.0	54.2%

Table Read: For grade 3, Claim 1, there was an average of 5 reviewers who rated 14 targets. The average number of total possible pairs was 10. The pairwise agreement across all 10 pairs for all 14 targets was 63.5%. A decimal for the number of reviewers and number of pairs indicates missing data.

Findings

Findings related to each DOK Distribution (DD) question are presented below. When relevant, general descriptive statistics provide overall results for each grade across claims.

A.DD-1: Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the Content Specifications (using the max DOK level)?

To get a sense for whether reviewers thought the targets required higher levels of cognitive demand than what was intended, the DOK distribution of the targets as identified by the reviewers and the Content Specifications using the maximum DOK level identified was examined. For example, if the Content Specifications indicated a target required DOK levels 1 and 2, the analysis used only DOK level 2. As shown in Table 5.A.14 across grades and claims, reviewers typically rated the targets at a higher DOK level than that specified by the Content Specifications. In general, reviewers had higher mean percentages of targets as having maximum DOK levels of 3 and 4 while the Content Specifications indicated the majority of targets had maximum DOK levels of 2 and 3. Reviewers believed that virtually none of the targets had a maximum DOK level of 1, while the Content Specifications indicated one target had a maximum DOK level of 1 for Claim 2 at all grades except grade 11.

A.DD-2: Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the Content Specifications (using the each independent DOK level)?

We next examined the DOK distribution of the targets using each identified DOK level (i.e., multiple DOK levels per target). As shown in Table 5.A.15, except for grade 11, across all grades and claims, reviewers tended to rate a lower percentage of the targets at the DOK level 1 and a higher percentage of the targets at DOK level 4 than was indicated in the Content Specifications. For grade 11, across all claims, reviewers generally rated a higher percentage of the targets at each DOK level than the level that was specified in the Content Specifications.

A.DD-3: Do the reviewers agree with the intended target DOK levels as identified in the Content Specifications?

The overall pairwise agreement in identifying grade-level standards aligned to each target between reviewers and the intended mapping as identified in the Content Specifications (across all grades, claims, targets, and reviewers) was 47.3%. As shown in Table 5.A.16, reviewers usually agreed less often with the intended DOK level of targets in Claim 2 (Writing). This was likely because reviewers believed these targets required higher cognitive demand than what was intended. Except for grades 8 and 11, reviewers agreed most often with the intended DOK level of targets in Claim 1 (Comprehend Literary and Informational Texts). For grade 11, reviewers agreed most often with the intended DOK level of targets in Claim 3 (Speaking and Listening) and for grade 8, reviewers agreed most often with the intended DOK level of targets in Claim 4 (Research and Inquiry).

Table 5.A.14. A.DD-1: Reviewers' Mean Percentage of ELA/Literacy Targets at Each DOK Level (Max) by Grade and Claim Compared to Content Specifications

Grade	Claim	DOK 1		DOK 2		DOK 3		DOK 4	
		Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)
3	1	1.4% (0.2)	0.0% (0.0)	17.1% (2.4)	28.6% (4.0)	35.7% (5.0)	57.1% (8.0)	45.7% (6.4)	14.3% (2.0)
	2	0.0% (0.0)	9.1% (1.0)	9.1% (1.0)	36.4% (4.0)	41.8% (4.6)	27.3% (3.0)	49.1% (5.4)	27.3% (3.0)
	3	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	20.0% (0.2)	100.0% (1.0)	80.0% (0.8)	0.0% (0.0)
	4	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	40.0% (1.2)	66.7% (2.0)	60.0% (1.8)	33.3% (1.0)
4	1	0.0% (0.0)	0.0% (0.0)	27.1% (3.8)	28.6% (4.0)	52.9% (7.4)	57.1% (8.0)	20.0% (2.8)	14.3% (2.0)
	2	1.8% (0.2)	9.1% (1.0)	1.8% (0.2)	36.4% (4.0)	61.8% (6.8)	27.3% (3.0)	34.5% (3.8)	27.3% (3.0)
	3	0.0% (0.0)	0.0% (0.0)	40.0% (0.4)	0.0% (0.0)	20.0% (0.2)	100.0% (1.0)	40.0% (0.4)	0.0% (0.0)
	4	0.0% (0.0)	0.0% (0.0)	6.7% (0.2)	0.0% (0.0)	30.0% (0.8)	66.7% (2.0)	63.3% (1.8)	33.3% (1.0)
5	1	1.8% (0.3)	0.0% (0.0)	28.6% (4.0)	28.6% (4.0)	41.1% (5.8)	57.1% (8.0)	28.6% (4.0)	14.3% (2.0)
	2	0.0% (0.0)	9.1% (1.0)	4.5% (0.5)	36.4% (4.0)	54.5% (6.0)	27.3% (3.0)	40.9% (4.5)	27.3% (3.0)
	3	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	25.0% (0.3)	100.0% (1.0)	75.0% (0.8)	0.0% (0.0)
	4	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	25.0% (0.8)	66.7% (2.0)	75.0% (2.3)	33.3% (1.0)
6	1	0.0% (0.0)	0.0% (0.0)	32.9% (4.6)	28.6% (4.0)	61.4% (8.6)	42.9% (6.0)	5.7% (0.8)	28.6% (4.0)
	2	0.0% (0.0)	9.1% (1.0)	20.0% (2.2)	36.4% (4.0)	67.3% (7.4)	27.3% (3.0)	12.7% (1.4)	27.3% (3.0)
	3	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	80.0% (0.8)	100.0% (1.0)	20.0% (0.2)	0.0% (0.0)
	4	0.0% (0.0)	0.0% (0.0)	20.0% (0.6)	0.0% (0.0)	53.3% (1.6)	33.3% (1.0)	26.7% (0.8)	66.7% (2.0)
7	1	0.0% (0.0)	0.0% (0.0)	29.2% (3.8)	28.6% (4.0)	54.7% (7.3)	42.9% (6.0)	16.1% (2.3)	28.6% (4.0)
	2	0.0% (0.0)	9.1% (1.0)	9.1% (1.0)	36.4% (4.0)	84.8% (9.3)	27.3% (3.0)	6.1% (0.7)	27.3% (3.0)
	3	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	100.0% (1.0)	100.0% (1.0)	0.0% (0.0)	0.0% (0.0)
	4	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	22.2% (0.7)	33.3% (1.0)	77.8% (2.3)	66.7% (2.0)
8	1	0.0% (0.0)	0.0% (0.0)	14.3% (2.0)	28.6% (4.0)	76.8% (10.8)	42.9% (6.0)	8.9% (1.3)	28.6% (4.0)
	2	0.0% (0.0)	9.1% (1.0)	11.4% (1.3)	36.4% (4.0)	59.1% (6.5)	27.3% (3.0)	29.5% (3.3)	27.3% (3.0)
	3	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	75.0% (0.8)	100.0% (1.0)	25.0% (0.3)	0.0% (0.0)
	4	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	25.0% (0.8)	33.3% (1.0)	75.0% (2.3)	66.7% (2.0)
11	1	0.0% (0.0)	0.0% (0.0)	10.0% (1.4)	28.6% (4.0)	55.7% (7.8)	28.6% (4.0)	34.3% (4.8)	42.9% (6.0)
	2	0.0% (0.0)	0.0% (0.0)	20.7% (2.3)	50.0% (5.0)	51.4% (5.5)	30.0% (3.0)	28.0% (3.0)	20.0% (2.0)
	3	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	50.0% (0.5)	100.0% (1.0)	50.0% (0.5)	0.0% (0.0)
	4	0.0% (0.0)	0.0% (0.0)	16.7% (0.5)	0.0% (0.0)	16.7% (0.5)	0.0% (0.0)	66.7% (2.0)	100.0% (3.0)

Note: For each group (reviewers and specifications) the percentages across DOK levels are mutually exclusive.

Table 5.A.15. A.DD-2: Reviewers' Mean Percentage of ELA/Literacy Targets at Each DOK Level (Independent) by Grade and Claim Compared to Content Specifications

Grade	Claim	DOK 1		DOK 2		DOK 3		DOK 4	
		Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)
3	1	27.1% (3.8)	28.6% (4.0)	61.4% (8.6)	71.4% (10.0)	80.0% (11.2)	71.4% (10.0)	45.7% (6.4)	14.3% (2.0)
	2	21.8% (2.4)	18.2% (2.0)	61.8% (6.8)	36.4% (4.0)	85.5% (9.4)	27.3% (3.0)	49.1% (5.4)	27.3% (3.0)
	3	20.0% (0.2)	100.0% (1.0)	40.0% (0.4)	100.0% (1.0)	100.0% (1.0)	100.0% (1.0)	80.0% (0.8)	0.0% (0.0)
	4	0.0% (0.0)	0.0% (0.0)	40.0% (1.2)	100.0% (3.0)	100.0% (3.0)	66.7% (2.0)	60.0% (1.8)	33.3% (1.0)
4	1	14.3% (2.0)	28.6% (4.0)	61.4% (8.6)	71.4% (10.0)	72.9% (10.2)	71.4% (10.0)	20.0% (2.8)	14.3% (2.0)
	2	20.0% (2.2)	18.2% (2.0)	40.0% (4.4)	36.4% (4.0)	96.4% (10.6)	27.3% (3.0)	34.5% (3.8)	27.3% (3.0)
	3	0.0% (0.0)	100.0% (1.0)	80.0% (0.8)	100.0% (1.0)	60.0% (0.6)	100.0% (1.0)	40.0% (0.4)	0.0% (0.0)
	4	6.7% (0.2)	0.0% (0.0)	30.0% (0.8)	100.0% (3.0)	80.0% (2.2)	66.7% (2.0)	63.3% (1.8)	33.3% (1.0)
5	1	10.7% (1.5)	28.6% (4.0)	35.7% (5.0)	71.4% (10.0)	62.5% (8.8)	71.4% (10.0)	28.6% (4.0)	14.3% (2.0)
	2	4.5% (0.5)	18.2% (2.0)	15.9% (1.8)	36.4% (4.0)	75.0% (8.3)	27.3% (3.0)	40.9% (4.5)	27.3% (3.0)
	3	0.0% (0.0)	100.0% (1.0)	0.0% (0.0)	100.0% (1.0)	75.0% (0.8)	100.0% (1.0)	75.0% (0.8)	0.0% (0.0)
	4	0.0% (0.0)	0.0% (0.0)	8.3% (0.3)	100.0% (3.0)	66.7% (2.0)	66.7% (2.0)	75.0% (2.3)	33.3% (1.0)
6	1	0.0% (0.0)	14.3% (2.0)	32.9% (4.6)	57.1% (8.0)	61.4% (8.6)	71.4% (10.0)	5.7% (0.8)	28.6% (4.0)
	2	0.0% (0.0)	18.2% (2.0)	20.0% (2.2)	36.4% (4.0)	80.0% (8.8)	27.3% (3.0)	12.7% (1.4)	27.3% (3.0)
	3	0.0% (0.0)	100.0% (1.0)	0.0% (0.0)	100.0% (1.0)	100.0% (1.0)	100.0% (1.0)	20.0% (0.2)	0.0% (0.0)
	4	0.0% (0.0)	0.0% (0.0)	20.0% (0.6)	100.0% (3.0)	60.0% (1.8)	100.0% (3.0)	26.7% (0.8)	66.7% (2.0)
7	1	2.3% (0.3)	14.3% (2.0)	45.0% (5.8)	57.1% (8.0)	60.1% (8.0)	71.4% (10.0)	16.1% (2.3)	28.6% (4.0)
	2	0.0% (0.0)	18.2% (2.0)	27.3% (3.0)	36.4% (4.0)	90.9% (10.0)	27.3% (3.0)	6.1% (0.7)	27.3% (3.0)
	3	0.0% (0.0)	100.0% (1.0)	0.0% (0.0)	100.0% (1.0)	100.0% (1.0)	100.0% (1.0)	0.0% (0.0)	0.0% (0.0)
	4	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	100.0% (3.0)	33.3% (1.0)	100.0% (3.0)	77.8% (2.3)	66.7% (2.0)
8	1	5.4% (0.8)	14.3% (2.0)	53.6% (7.5)	57.1% (8.0)	85.7% (12.0)	71.4% (10.0)	8.9% (1.3)	28.6% (4.0)
	2	2.3% (0.3)	18.2% (2.0)	52.3% (5.8)	36.4% (4.0)	88.6% (9.8)	27.3% (3.0)	29.5% (3.3)	27.3% (3.0)
	3	0.0% (0.0)	100.0% (1.0)	50.0% (0.5)	100.0% (1.0)	75.0% (0.8)	100.0% (1.0)	25.0% (0.3)	0.0% (0.0)
	4	0.0% (0.0)	0.0% (0.0)	8.3% (0.3)	100.0% (3.0)	83.3% (2.5)	100.0% (3.0)	75.0% (2.3)	66.7% (2.0)
11	1	20.0% (2.8)	14.3% (2.0)	74.3% (10.4)	42.9% (6.0)	90.0% (12.6)	71.4% (10.0)	34.3% (4.8)	42.9% (6.0)
	2	50.5% (5.5)	20.0% (2.0)	88.6% (9.5)	50.0% (5.0)	79.3% (8.5)	30.0% (3.0)	28.0% (3.0)	20.0% (2.0)
	3	100.0% (1.0)	100.0% (1.0)	100.0% (1.0)	100.0% (1.0)	100.0% (1.0)	100.0% (1.0)	50.0% (0.5)	0.0% (0.0)
	4	50.0% (1.5)	0.0% (0.0)	58.3% (1.8)	100.0% (3.0)	83.3% (2.5)	100.0% (3.0)	66.7% (2.0)	100.0% (3.0)

Note: For each group (reviewers and specifications) the percentages across DOK levels are not mutually exclusive since a target could have multiple DOK levels.

Table 5.A.16. A.DD-3: Pairwise Percent Agreement between Reviewers' and Intended ELA/Literacy Target DOK Ratings

Grade	Claim	Descriptives			Agreement
		# of Reviewers % (n)	# of Targets	# of Ratings	Pairwise Agreement
3	1	5.0	14	70	67.4%
	2	5.0	11	55	38.6%
	3	5.0	1	5	53.3%
	4	5.0	3	15	64.4%
4	1	5.0	14	70	71.0%
	2	5.0	11	55	28.0%
	3	5.0	1	5	46.7%
	4	4.7	3	14	49.2%
5	1	4.0	14	56	56.0%
	2	4.0	11	44	46.6%
	3	4.0	1	4	25.0%
	4	4.0	3	12	37.5%
6	1	5.0	14	70	45.7%
	2	5.0	11	55	28.2%
	3	5.0	1	5	33.3%
	4	5.0	3	15	35.6%
7	1	3.8	14	53	51.2%
	2	3.0	11	33	31.8%
	3	3.0	1	3	33.3%
	4	3.0	3	9	38.9%
8	1	4.0	14	56	52.1%
	2	4.0	11	44	33.0%
	3	4.0	1	4	41.7%
	4	4.0	3	12	56.9%
11	1	5.0	14	70	64.3%
	2	4.0	10	40	44.6%
	3	4.0	1	4	87.5%
	4	4.0	3	12	63.2%

Table Read: For grade 3, there were 5 ELA/literacy reviewers who provided target DOK ratings for 14 targets in Claim 1, for a total of 70 ratings. Across these ratings, the pairwise agreement was 67.4%.

DOK Consistency

Analyses were conducted to examine the consistency of ELA/literacy DOK levels between the Content Specifications and the CCSS, and to address the following question:

- A.DC-1: Is the cognitive complexity required in the targets consistent with the cognitive complexity required in each targets' mapped grade-level standards?

Findings

A.DC-1: Is the cognitive complexity required in the targets consistent with the cognitive complexity required in each targets' mapped grade-level standards?

The DOK consistency analysis examined the degree to which the cognitive demand required by each of the grade-level standards aligned to a target fell within the range of cognitive demand required by the intended target. The assumptions for interpreting the results in Tables 5.A.17-5.A.18 are:

- Only the reviewers' grade-level standards that matched the intended mapping indicated in the Content Specifications were retained for this analysis.
- Of those standards that matched the intended mapping, only those grade-level standards with ≥ 50% reviewer agreement were retained for this analysis.
- *Consistency* was defined in two ways:
    a. The cognitive demand of *all* of the grade-level standards mapped to a target by the reviewers needed to fall within the range of the intended target DOK (refer to Table 5.A.17).
    b. Where the grade-level standards that were mapped to a target by reviewers had multiple DOK levels, only one of those levels had to fall within the range of the intended target DOK (refer to Table 5.A.17).
- Results here should be interpreted in relation to the reviewer agreement with the intended grade-level standard-to-target mapping. Because the DOK consistency analysis was applied only to those grade-level standards with 50% agreement *and* that matched the intended mapping, it is possible that each target had a differing percentage of mapped grade-level standards that were included.

As shown in Table 5.A.17, there was no real pattern in the percentage of targets that had DOK consistency with all of the mapped grade-level standards. It appears that Claim 2 (Writing) targets had very low consistency across grades, with grades 3, 5, 6, and 8 having had 0 targets with DOK levels consistent with the grade-level standards. Upon further investigation, the reason why so many Claim 2 targets had DOK inconsistency was because the reviewers rated the grade-level standards as requiring *lower* levels of cognitive demand than what was intended by the targets. Additionally, three of the Claim 2 targets are intended to be measured by performance tasks and were identified in the Content Specifications as having a single DOK level at level 4. This, combined with the difficulty for at least 50% of the reviewers to agree with the intended mappings resulted in fewer Claim 2 targets as being classified as having DOK consistency with the grade-level standards. In contrast to Claim 2, inconsistent target DOK levels in other claims is likely due to reviewers identifying the grade-level standards as requiring *higher* levels of cognitive demand than what was intended by the targets.

As shown in Table 5.A.18, when the DOK consistency definition was relaxed to requiring only one DOK level for each grade-level standard mapped to a target to fall within the range of the intended target DOK, the percentage of targets with DOK consistency increased. This suggests that of the grade-level standards with multiple DOK levels, the reviewers believed that part of the cognitive demand required in the grade-level standard matched that of the intended target, yet they believed there were some portions of the grade-level standards that fell outside that DOK range. For those targets that had inconsistent DOK levels, the general pattern remained that the grade-level standards required higher cognitive demand than what was intended by the target for Claims 1, 3, and 4, and lower levels of cognitive demand for Claim 2 (see the two 'Avg % of CCSS per Inconsistent Target…" columns in Table 5.A.18 to see whether reviewers rated the grade-level standards higher or lower than that of the target).

Table 5.A.17 ELA.DC-1a. Percentage of ELA/Literacy Targets with DOKs Consistent with Intended Grade-Level Standards that Matched Intended Mapping for All Targets – All CCSS within Range

Grade	Claim	Descriptives			DOK Consistency				
					Consistent	Inconsistent			
		Total number of targets in claim	Number of targets included in analysis[1]	Avg % of grade-level standards per target with ≥ 50% reviewer agreement[2]	% of Targets that Have All Mapped CCSS Consistent[3]	% of Targets With All Mapped CCSS Inconsistent	Avg % of CCSS per Inconsistent Target with Max DOK Consensus > Specs	Avg % of CCSS per Target with Min DOK Consensus < Specs	Number of Targets With < 50% agreement
		n	n	% (n)	% (n)	% (n)	% (n)	% (n)	n
3	1	14	10	47.6% (1.20)	70.0% (7)	30.0% (3)	50.0% (0.67)	33.3% (0.33)	4
	2	11	8	54.9% (3.50)	0.0% (0)	100.0% (8)	46.9% (1.38)	43.8% (1.75)	3
	3	1	1	100.0% (2.00)	0.0% (0)	100.0% (1)	50.0% (1.00)	0.0% (0.00)	0
	4	3	2	19.4% (1.50)	0.0% (0)	100.0% (2)	75.0% (1.00)	0.0% (0.00)	1
4	1	14	13	66.9% (1.38)	69.2% (9)	30.8% (4)	75.0% (0.75)	12.5% (0.25)	1
	2	11	8	46.5% (4.13)	12.5% (1)	87.5% (7)	14.3% (0.14)	67.9% (3.43)	3
	3	1	0	0.0% (.)					1
	4	3	3	26.2% (1.33)	33.3% (1)	66.7% (2)	25.0% (0.50)	75.0% (1.00)	0
5	1	14	13	63.3% (1.15)	53.8% (7)	46.2% (6)	66.7% (0.83)	16.7% (0.17)	1
	2	11	6	32.2% (3.17)	0.0% (0)	100.0% (6)	43.3% (1.00)	46.7% (1.67)	5
	3	1	0	0.0% (.)					1
	4	3	3	37.3% (2.00)	33.3% (1)	66.7% (2)	75.0% (2.00)	0.0% (0.00)	0
6	1	14	14	66.6% (1.79)	57.1% (8)	42.9% (6)	54.4% (1.17)	16.7% (0.33)	0
	2	11	10	51.4% (3.70)	0.0% (0)	100.0% (10)	35.0% (0.60)	43.5% (2.10)	1
	3	1	1	50.0% (1.00)	100.0% (1)	0.0% (0)			0
	4	3	2	37.0% (2.00)	100.0% (2)	0.0% (0)			1
7	1	14	14	64.8% (1.71)	57.1% (8)	42.9% (6)	33.3% (0.67)	66.7% (0.83)	0
	2	11	11	80.9% (5.36)	9.1% (1)	90.9% (10)	34.3% (1.50)	54.7% (3.60)	0
	3	1	1	100.0% (2.00)	0.0% (0)	100.0% (1)	100.0% (2.00)	0.0% (0.00)	0
	4	3	2	23.5% (5.00)	50.0% (1)	50.0% (1)	71.4% (5.00)	0.0% (0.00)	1

Table 5.A.17. (Continued)

Grade	Claim	Descriptives			DOK Consistency				
					Consistent	Inconsistent			
		Total number of targets in claim	Number of targets included in analysis[1]	Avg % of grade-level standards per target with ≥ 50% reviewer agreement[2]	% of Targets that Have All Mapped CCSS Consistent[3]	% of Targets With All Mapped CCSS Inconsistent	Avg % of CCSS per Inconsistent Target with Max DOK Consensus > Specs	Avg % of CCSS per Target with Min DOK Consensus < Specs	Number of Targets With < 50% agreement
		N	n	% (n)	% (n)	% (n)	% (n)	% (n)	n
8	1	14	13	60.1% (1.69)	46.2% (6)	53.8% (7)	64.3% (1.00)	28.6% (0.43)	1
	2	11	6	31.4% (3.83)	0.0% (0)	100.0% (6)	25.0% (0.33)	45.0% (2.17)	5
	3	1	1	50.0% (1.00)	100.0% (1)	0.0% (0)			0
	4	3	1	1.9% (1.00)	100.0% (1)	0.0% (0)			2
11	1	14	14	63.9% (1.64)	21.4% (3)	78.6% (11)	27.3% (0.55)	57.6% (0.73)	0
	2	10	9	57.2% (3.56)	11.1% (1)	88.9% (8)	49.6% (1.38)	27.5% (1.38)	1
	3	1	1	100.0% (2.00)	0.0% (0)	100.0% (1)	100.0% (2.00)	0.0% (0.00)	0
	4	3	3	18.0% (2.33)	100.0% (3)	0.0% (0)			0

[1]Number of targets with at least one standard with 50% reviewer agreement
[2]Standards that matched the intended mapping with greater than or equal to 50% reviewer agreement
[3]Consistent was defined as the grade-level standard DOK levels falling entirely within the range of the intended target DOK levels

Table Read: For grade 3, there were 11 targets in Claim 1. Of those targets, 10 were included in this analysis because they had at least one standard with 50% reviewer agreement. Across all Claim 1 targets, an average of 47.6% of the grade-level standards that mapped to the intended target had at least 50% reviewer agreement (an average 1.2 standards per target). DOK Consistency in the next five columns is analyzed using the standards that mapped to the intended target with 50% agreement. 70% of the 10 targets included in the analysis had DOK consistency with all of the grade-level standards included in the analysis. 30% of the targets had DOK inconsistency. Of the 3 inconsistent targets, an average of 50% of the grade-level standards per target had a maximum grade-level standard DOK higher than that of the maximum intended target DOK. Four targets were excluded from the DOK consistency analysis due to having no grade-level standards with at least 50% reviewer agreement.

Table 5.A.18 ELA.DC-1b. Percentage of ELA/Literacy Targets with DOKs Consistent with Intended Grade-Level Standards that Matched Intended Mapping for All Targets – All CCSS At Least One

Grade	Claim	Descriptives			DOK Consistency				
					Consistent	Inconsistent			
		Total number of targets in claim	Number of targets included in analysis[1]	Avg % of grade-level standards per target with ≥50% reviewer agreement[2]	% of Targets that Have All Mapped CCSS Consistent[3]	% of Targets With All Mapped CCSS Inconsistent	Avg % of CCSS per Inconsistent Target with Max DOK Consensus > Specs	Avg % of CCSS per Target with Min DOK Consensus < Specs	Number of Targets With < 50% agreement
		n	n	% (n)	% (n)	% (n)	% (n)	% (n)	n
3	1	14	10	47.6% (1.20)	100.0% (10)	0.0% (0)			4
	2	11	8	54.9% (3.50)	37.5% (3)	62.5% (5)	30.0% (0.80)	65.0% (2.60)	3
	3	1	1	100.0% (2.00)	100.0% (1)	0.0% (0)			0
	4	3	2	19.4% (1.50)	100.0% (2)	0.0% (0)			1
4	1	14	13	66.9% (1.38)	100.0% (13)	0.0% (0)			1
	2	11	8	46.5% (4.13)	37.5% (3)	62.5% (5)	20.0% (0.20)	76.0% (4.00)	3
	3	1	0	0.0% (.)					1
	4	3	3	26.2% (1.33)	100.0% (3)	0.0% (0)			0
5	1	14	13	63.3% (1.15)	76.9% (10)	23.1% (3)	66.7% (1.00)	0.0% (0.00)	1
	2	11	6	32.2% (3.17)	16.7% (1)	83.3% (5)	32.0% (1.00)	56.0% (2.00)	5
	3	1	0	0.0% (.)					1
	4	3	3	37.3% (2.00)	33.3% (1)	66.7% (2)	75.0% (2.00)	0.0% (0.00)	0
6	1	14	14	66.6% (1.79)	78.6% (11)	21.4% (3)	42.2% (1.67)	16.7% (0.33)	0
	2	11	10	51.4% (3.70)	0.0% (0)	100.0% (10)	35.0% (0.60)	43.5% (2.10)	1
	3	1	1	50.0% (1.00)	100.0% (1)	0.0% (0)			0
	4	3	2	37.0% (2.00)	100.0% (2)	0.0% (0)			1
7	1	14	14	64.8% (1.71)	71.4% (10)	28.6% (4)	50.0% (1.00)	50.0% (0.75)	0
	2	11	11	80.9% (5.36)	18.2% (2)	81.8% (9)	38.1% (1.67)	56.3% (3.78)	0
	3	1	1	100.0% (2.00)	100.0% (1)	0.0% (0)			0
	4	3	2	23.5% (5.00)	50.0% (1)	50.0% (1)	71.4% (5.00)	0.0% (0.00)	1

Table 5.A.18 (Continued)

Grade	Claim	Descriptives			DOK Consistency				
					Consistent	Inconsistent			
		Total number of targets in claim	Number of targets included in analysis[1]	Avg % of grade-level standards per target with ≥50% reviewer agreement[2]	% of Targets that Have All Mapped CCSS Consistent[3]	% of Targets With All Mapped CCSS Inconsistent	Avg % of CCSS per Inconsistent Target who's Max DOK Consensus > Specs	Avg % of CCSS per Target who's Min DOK Consensus < Specs	Number of Targets With < 50% agreement
		N	n	% (n)	% (n)	% (n)	% (n)	% (n)	n
8	1	14	13	60.1% (1.69)	84.6% (11)	15.4% (2)	0.0% (0.00)	100.0% (1.50)	1
	2	11	6	31.4% (3.83)	16.7% (1)	83.3% (5)	30.0% (0.40)	50.0% (2.40)	5
	3	1	1	50.0% (1.00)	100.0% (1)	0.0% (0)			0
	4	3	1	1.9% (1.00)	100.0% (1)	0.0% (0)			2
11	1	14	14	63.9% (1.64)	85.7% (12)	14.3% (2)	0.0% (0.00)	100.0% (1.50)	0
	2	10	9	57.2% (3.56)	33.3% (3)	66.7% (6)	54.4% (1.50)	36.7% (1.83)	1
	3	1	1	100.0% (2.00)	0.0% (0)	100.0% (1)	100.0% (2.00)	0.0% (0.00)	0
	4	3	3	18.0% (2.33)	100.0% (3)	0.0% (0)			0

[1]Number of targets with at least one standard with 50% reviewer agreement
[2]Standards that matched the intended mapping with greater than or equal to 50% reviewer agreement
[3]Consistent was defined as at least one of the grade-level standard DOK levels matched at least one DOK level of the intended target

Table Read: Table 5.A.18 columns can be interpreted the same as in Table 5.A.15. The difference is the way in which DOK consistency was defined: Here it was defined as each grade-level standard had to have at least one DOK level that matched that of the intended target.

Mathematics

Analyses were conducted separately by content area to examine the alignment between the Content Specifications and the CCSS. These analyses focused on content representation, DOK distribution, and DOK consistency. Pairwise agreement among reviewers was computed for identifying grade-level standards aligned to each target (Workshop 1) as well as for identifying DOK levels for each target (Workshop 1); this information is presented in the Content Representation and DOK Distribution sections below.

Each analysis was conducted by examining the overall targets within a claim, as well as disaggregating the targets by Claim 1 (Concepts and Procedures) emphasis. For any substantive differences that were found between the major and additional and supporting targets, findings are presented below. When substantive differences were not found, parallel tables for the emphasis breakout for each analysis are presented Appendix F.

Content Representation

Analyses were conducted to examine the representation of mathematics content between the Content Specifications and the CCSS, and to address the following questions:

- A.CR-1: Do the grade-level standards collectively reflect the content and skills required by the target?
- A.CR-2: Do the targets collectively reflect the content and skills required by the grade-level standard?[31]
- A.CR-3: Do the individual grade-level standards reflect the content and skills required by the intended targets?
- A.CR-4: Do the individual grade-level standards reflect the content and skills required by the intended targets when reviewers were asked to identify targets aligned to each grade-level standard (Workshop 2)?
- A.CR-5: Does each mathematical practice reflect skills required by the intended target?
- A.CR-6: Do the reviewers agree with the intended mapping of targets and grade-level standards as identified in the Content Specifications?

The main reviewer tasks for examining the content representation of the targets and grade-level standards involved identifying grade-level standards that represented the targets and providing a holistic target rating to indicate how well the grade-level standards represented the target. Because Smarter Balanced does not intend to measure all grade-level standards on the summative assessment, we excluded any grade-level standard that was solely intended to be measured by a target and that was excluded from the summative assessment. The list of excluded grade-level standards is presented in Appendix E.

[31] This question addresses the two-way alignment approach.

When examining how well the grade-level standards mapped to the targets, Claim 1 (Concepts and Procedures) alignment was analyzed differently than were Claims 2 – 4 (Claim 2, Problem Solving; Claim 3, Communicating Reasoning; Claim 4, Modeling and Data Analysis). Because grade-level standards were mapped to individual targets in Claim 1 within the Content Specifications, we analyzed the mappings at the target level. However, for Claims 2 – 4, we analyzed the mappings at the claim level to more accurately reflect the structure of the Content Specifications. Recall that Claims 2 – 4 were not intended to align to grade-level standards but rather they were intended to reflect the mathematical practices.

Overall, across all grades, targets, and reviewers, reviewers identified an average of 10.1 unique grade-level standards per Claim 1 target (SD=6.0) and 48.0 unique grade level standards per claim for Claims 2 – 4 (SD=15.8) compared to 3.5 unique grade-level standards (SD=1.3) indicated in the Content Specifications for Claim 1 targets and 49 unique standards (SD=29.37) per claim for Claims 2 – 4. As further shown in Table 5.A.19, reviewers in all grades independently identified more grade-level standards per target than were indicated in the Content Specifications. Using a greater than or equal to 50% reviewer agreement rule (per target), the mean percentage of unique grade-level standards per target dropped for all grades. This suggests that using an independent identification method (where reviewers had access to the full eligible pool of grade-level standards), resulted in reviewers aligning more of the grade-level standards to the targets than what was intended.

Table 5.A.19. A.CR.GD-1 Average Number of CCSS per Mathematics Target Identified by Reviewers and the Content Specifications

Grade	# of CCSS per Target (Reviewers)		# of CCSS per Target (Reviewers ≥ 50% Agreement)		# of CCSS per Target (Specs)	
	Mean	SD	Mean	SD	Mean	SD
Claim 1 (per target)						
3	17.7	5.7	8.5	5.9	3.5	1.6
4	12.3	5.8	5.6	2.2	3.5	1.2
5	7.5	5.6	3.6	1.7	3.3	1.3
6	7.8	3.7	4.0	1.9	3.9	1.2
7	7.7	3.0	5.3	3.1	3.7	0.7
8	6.2	3.0	4.2	1.9	3.8	1.2
11	10.1	5.6	6.4	3.1	3.3	1.5
Claims 2 – 4 (per claim)						
3	41.3	5.1	22.3	4.5	6.7	1.5
4	39.3	5.5	25.7	5.6	8.7	4.7
5	38.0	2.6	25.3	3.5	9.3	4.9
6	45.3	4.0	28.7	2.1	9.7	3.8
7	46.7	3.5	24.7	7.6	6.7	1.2
8	42.7	7.5	38.7	13.6	10.0	4.4
11	82.0	16.1	51.7	19.5	20.7	7.4

Pairwise Agreement among Reviewers

The overall pairwise agreement among reviewers in identifying mathematics grade-level standards aligned to each target (across all grades, claims, targets, and reviewers) was 51.3% (refer to Table 5.A.20). The moderate agreement was likely due to a combination of the relatively high number of

grade-level standards that reviewers identified for each target and the fact that reviewers conducted a blind rating where they were permitted to choose from a lengthy list of eligible grade-level standards. Across grades, reviewers generally agreed more when identifying grade-level standards for Claim 1 targets than they did for other claims; exceptions included Claim 4 for grade 4 (which was slightly more than that for Claims 1 and 2) and Claim 2 for grade 7 (which was essentially the same as that for Claim 1).

Table 5.A.20. A.Math.CR.PWA-1. Pairwise Percent Agreement among Reviewers' Mapping of Mathematics Targets and Grade-level Standards

		Descriptives			Agreement
Grade	Claim	Avg # of Reviewers n	# of Targets	Avg # of Reviewer Pairs	Pairwise Agreement %
3	1	5	11	10	54.5%
	2	5	4	10	49.1%
	3	5	6	10	50.0%
	4	5	7	10	40.5%
4	1	5	12	10	62.5%
	2	5	4	10	62.3%
	3	5	6	10	43.0%
	4	5	7	10	64.1%
5	1	5	11	10	72.4%
	2	5	4	10	50.7%
	3	5	6	10	41.1%
	4	5	7	10	62.2%
6	1	5	10	10	59.5%
	2	5	4	10	51.6%
	3	5	7	10	40.7%
	4	5	7	10	49.8%
7	1	5	9	10	62.4%
	2	5	4	10	62.6%
	3	5	7	10	44.1%
	4	5	7	10	37.1%
8	1	5	10	10	73.6%
	2	5	4	10	53.7%
	3	5	7	10	39.2%
	4	5	7	10	38.2%
11	1	4	16	6	62.2%
	2	4	4	6	46.9%
	3	4	7	6	23.3%
	4	4	7	6	36.7%

Note: Due to the structure of the Content Specifications for Claims 2 - 4, analyses for these claims occurred at the claim level (rather than the target level)

Table Read: For Grade 3, Claim 1, there was an average of 5 reviewers who rated 11 targets. The average number of total possible pairs was 10. The pairwise agreement across all 10 pairs for all 11 targets was 54.5%.

Findings

Findings related to each content representation question are presented below. When relevant, general descriptive statistics provide overall results for each grade across claims.

A.CR-1: Do the grade-level standards collectively reflect the content and skills required by the target?

Once reviewers identified all the mathematics grade-level standards they believed aligned to the target, they provided a holistic target representation rating to indicate how well those standards collectively represented the content and knowledge required in the target. The scale ranged from 0 to 4 ('0' = not aligned at all, '1' = small-portion aligned, '2' = somewhat aligned, '3' = mostly aligned, and '4' = fully-aligned). As seen in Table 5.A.21, reviewers generally rated the targets as being well-represented by the grade-level standards they identified. The mean alignment rating across grades and claims ranged from 3.3 to 3.8.

Table 5.A.21. A.CR-1: Mean Percentage of Mathematics Targets at Each Holistic Rating (Collectively Reflected by the Grade-level Standards)

Grade	Target Representation Rating		
	$N_{target\_ratings}$	Mean	SD
3	134	3.8	0.7
4	144	3.9	0.6
5	132	3.4	0.9
6	126	3.3	1.0
7	127	3.5	1.0
8	132	3.6	0.9
11	136	3.4	0.7

Table Read: For grade 3, across 139 target ratings (across all claims and reviewers), the average rating was 3.8, with a standard deviation of 0.9.

Across grades and claims, the mathematics targets were generally represented well by their intended grade-level standards, especially for Claim 1. The lower percentages of 'fully aligned' targets in Claims 2 – 4 were not necessarily unexpected. The reviewers were instructed to identify specific grade-level standards they believed aligned to each target and reviewers' comments reflected the underlying assumption that these claims do not focus on a specific type of knowledge, but rather focus on processes and mathematical practices.

Table 5.A.22. A.CR-1: Mean Percentage of Mathematics Targets at Each Holistic Rating (Collectively Reflected by the Grade-level Standards)

Grade	Claim	# of Targets in Claim	Holistic Target Rating				
			Fully-aligned	Mostly-aligned	Somewhat-aligned	Small-portion aligned	Not-aligned at all
			% (n)	% (n)	% (n)	% (n)	% (n)
3	1	11	98.2% (10.6)	1.8% (0.2)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	2	4	68.3% (2.6)	25.0% (0.8)	6.7% (0.2)	0.0% (0.0)	0.0% (0.0)
	3	6	100.0% (5.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	4	7	84.6% (5.6)	0.0% (0.0)	2.9% (0.2)	0.0% (0.0)	12.6% (0.8)
4	1	12	100.0% (11.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	2	4	90.0% (3.6)	5.0% (0.2)	0.0% (0.0)	0.0% (0.0)	5.0% (0.2)
	3	6	96.7% (5.8)	3.3% (0.2)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	4	7	82.9% (5.8)	2.9% (0.2)	8.6% (0.6)	0.0% (0.0)	5.7% (0.4)
5	1	11	79.2% (8.4)	20.8% (2.2)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	2	4	60.0% (2.2)	30.0% (0.8)	0.0% (0.0)	10.0% (0.4)	0.0% (0.0)
	3	6	38.3% (2.0)	26.7% (1.6)	21.7% (1.2)	13.3% (0.8)	0.0% (0.0)
	4	7	60.0% (4.0)	11.4% (0.8)	22.9% (1.6)	5.7% (0.4)	0.0% (0.0)
6	1	10	73.6% (7.2)	20.4% (2.0)	2.0% (0.2)	4.0% (0.4)	0.0% (0.0)
	2	4	58.3% (2.0)	15.0% (0.6)	15.0% (0.6)	11.7% (0.4)	0.0% (0.0)
	3	7	44.1% (2.4)	27.3% (1.8)	17.9% (1.0)	6.7% (0.4)	4.0% (0.2)
	4	7	55.0% (3.0)	27.9% (1.8)	11.4% (0.8)	5.7% (0.4)	0.0% (0.0)
7	1	9	95.6% (8.6)	2.2% (0.2)	0.0% (0.0)	0.0% (0.0)	2.2% (0.2)
	2	4	65.0% (2.6)	30.0% (1.0)	0.0% (0.0)	0.0% (0.0)	5.0% (0.2)
	3	7	52.1% (3.2)	42.9% (3.0)	0.0% (0.0)	0.0% (0.0)	5.0% (0.2)
	4	7	45.0% (3.0)	34.3% (2.2)	0.0% (0.0)	0.0% (0.0)	20.7% (1.0)
8	1	10	82.0% (8.0)	16.0% (1.6)	0.0% (0.0)	0.0% (0.0)	2.0% (0.2)
	2	4	70.0% (2.8)	25.0% (1.0)	5.0% (0.2)	0.0% (0.0)	0.0% (0.0)
	3	7	48.6% (3.4)	37.1% (2.6)	0.0% (0.0)	0.0% (0.0)	14.3% (1.0)
	4	7	77.1% (4.2)	22.9% (1.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
11	1	16	70.3% (11.3)	28.1% (4.5)	1.6% (0.3)	0.0% (0.0)	0.0% (0.0)
	2	4	50.0% (2.0)	43.8% (1.8)	6.3% (0.3)	0.0% (0.0)	0.0% (0.0)
	3	7	28.6% (2.0)	53.6% (3.8)	17.9% (1.3)	0.0% (0.0)	0.0% (0.0)
	4	7	14.3% (1.0)	67.9% (4.8)	14.3% (1.0)	3.6% (0.3)	0.0% (0.0)

Table Read: For grade 3, Claim 1, reviewers rated an average of 98.2% of the mathematics targets (10.6 targets) as being fully aligned to the grade-level standards that they identified, 1.8% of the targets (0.2 targets) as being mostly aligned to the grade-level standards, and none of the targets as being aligned somewhat, a small portion, or not aligned at all to the grade-level standards.

A.CR-2: Do the targets collectively reflect the content and skills required by the grade-level standard?

Data related to this question were collected during Workshop 2. In contrast to the tasks in Workshop 1, reviewers identified all the targets they believed aligned to each grade-level standard and then provided a holistic grade-level standard representation rating to indicate how well those targets collectively represented the content and knowledge required in the grade-level standard. The scale ranged from 0 to 4 ('0' = not aligned at all, '1' = small-portion aligned, '2' = somewhat aligned, '3' = mostly aligned, and '4' = fully-aligned). This analysis provided a general indication of how well the content and knowledge required in each individual grade-level standard was measured by the targets. Based on the development and structure of the Content Specifications, there was little expectation that each grade-level standard would be collectively represented by the targets. Thus, low percentages of 'fully-aligned' and 'mostly-aligned' ratings do not necessarily reflect poor alignment.

As seen in Table 5.A.23, across grades and mathematics domains, reviewers strongly believed that most of the grade-level standards were fully aligned and were well represented by the content and knowledge required in the targets. The exception was the High School Functions; for the Trigonometric Functions (F-TF) domain, reviewers rated an average of approximately 10% of the grade-level standards for this domain as being fully aligned; however, the reviewers rated the remaining approximately 90% of the grade-level standards as being mostly or somewhat aligned to that domain.

Table 5.A.23. A.CR-2: Mean Percentage of Mathematics Grade-level Standards at Each Holistic Rating

Grade	Domain	Holistic Target Rating				
		Fully-aligned % (n)	Mostly-aligned % (n)	Somewhat-aligned % (n)	Small-portion aligned % (n)	Not-aligned at all % (n)
3	G	100.0% (2.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	MD	98.6% (14.0)	0.0% (0.0)	1.4% (0.2)	0.0% (0.0)	0.0% (0.0)
	NBT	100.0% (3.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	NF	100.0% (9.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	OA	98.0% (9.0)	2.0% (0.2)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
4	G	100.0% (3.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	MD	95.6% (8.6)	4.4% (0.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	NBT	96.7% (5.8)	3.3% (0.2)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	NF	72.9% (10.2)	22.9% (3.2)	1.4% (0.2)	2.9% (0.4)	0.0% (0.0)
	OA	100.0% (5.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
5	G	90.0% (3.6)	10.0% (0.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	MD	100.0% (10.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	NBT	95.6% (8.6)	4.4% (0.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	NF	100.0% (13.2)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	OA	100.0% (3.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)

Table 5.A.23. (Continued)

Grade	Domain	Holistic Target Rating				
		Fully-aligned % (n)	Mostly-aligned % (n)	Somewhat-aligned % (n)	Small-portion aligned % (n)	Not-aligned at all % (n)
6	EE	91.9% (9.3)	5.3% (0.5)	2.8% (0.3)	0.0% (0.0)	0.0% (0.0)
	G	100.0% (5.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	NS	91.3% (15.5)	2.8% (0.5)	5.9% (1.0)	0.0% (0.0)	0.0% (0.0)
	RP	86.6% (6.8)	10.3% (0.8)	3.1% (0.3)	0.0% (0.0)	0.0% (0.0)
	SP	91.3% (7.5)	8.8% (0.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
7	EE	83.3% (6.3)	13.5% (1.0)	3.1% (0.3)	0.0% (0.0)	0.0% (0.0)
	G	84.4% (6.0)	9.4% (0.8)	3.1% (0.3)	3.1% (0.3)	0.0% (0.0)
	NS	90.8% (9.5)	4.2% (0.5)	5.0% (0.3)	0.0% (0.0)	0.0% (0.0)
	RP	80.4% (6.0)	16.1% (1.0)	3.6% (0.3)	0.0% (0.0)	0.0% (0.0)
	SP	75.4% (8.3)	20.2% (1.3)	4.4% (0.5)	0.0% (0.0)	0.0% (0.0)
8	EE	85.7% (10.8)	6.4% (0.8)	1.7% (0.3)	6.3% (0.5)	0.0% (0.0)
	F	91.4% (5.5)	3.6% (0.3)	5.0% (0.3)	0.0% (0.0)	0.0% (0.0)
	G	88.1% (7.8)	5.6% (0.5)	3.6% (0.3)	2.8% (0.3)	0.0% (0.0)
	NS	100.0% (2.7)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	SP	80.0% (3.0)	20.0% (1.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
11	A-APR	75.0% (5.3)	11.1% (1.0)	13.9% (1.3)	0.0% (0.0)	0.0% (0.0)
	A-CED	75.0% (3.5)	25.0% (1.3)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	A-REI	87.5% (12.0)	4.7% (0.8)	7.8% (1.3)	0.0% (0.0)	0.0% (0.0)
	A-SSE	75.0% (7.3)	0.0% (0.0)	25.0% (2.8)	0.0% (0.0)	0.0% (0.0)
	F-BF	74.6% (6.3)	25.4% (2.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	F-IF	97.2% (13.3)	2.8% (0.5)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	F-LE	90.0% (9.0)	10.0% (1.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	F-TF	11.1% (0.7)	33.3% (1.3)	55.6% (3.3)	0.0% (0.0)	0.0% (0.0)
	G-CO	42.4% (4.7)	15.2% (1.7)	42.4% (5.3)	0.0% (0.0)	0.0% (0.0)
	G-GMD	50.0% (1.5)	0.0% (0.0)	50.0% (1.0)	0.0% (0.0)	0.0% (0.0)
	G-MG	66.7% (2.3)	0.0% (0.0)	33.3% (1.0)	0.0% (0.0)	0.0% (0.0)
	G-SRT	69.4% (8.3)	8.3% (1.0)	22.2% (2.7)	0.0% (0.0)	0.0% (0.0)
	N-Q	66.7% (2.7)	33.3% (1.3)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	N-RN	100.0% (4.3)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	S-CP	100.0% (5.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	S-IC	100.0% (5.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	S-ID	86.7% (9.3)	13.3% (2.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)

Table Read: For grade 3 Geometry grade-level standards, reviewers rated an average of100% of them (average of 2.0 standards) as fully represented by the targets.

A.CR-3: Do the individual grade-level standards reflect the content and skills required by the intended targets?

As noted earlier, overall, reviewers identified more grade-level standards per target than what was intended in the Content Specifications. With the exception of grade 11 Claim 3, that pattern holds true at the claim level as well (refer to Table 5.A.24).

Table 5.A.24. A.Math.CR-3.GD-1 Comparison of Reviewer and Content Specifications Mathematics Target and CCSS Ratings Descriptive Statistics

			Reviewer Descriptives			Specifications Descriptives		
Grade	Claim	Total number of targets in claim	Avg # of grade-level standards per target	Min # of grade-level standards per target	Max # of grade-level standards per target	Avg # of grade-level standards per target	Min # of grade-level standards per target	Max # of grade-level standards per target
		N	n	n	n	n	n	n
3	1	11	9.6	1	23	3.5	2	8
	2	4	20.8	9	35	8	8	8
	3	6	20.4	12	35	7	7	7
	4	7	19	5	28	5	5	5
4	1	12	6.5	2	19	3.5	2	6
	2	4	19	15	24	7	7	7
	3	6	21.6	10	32	14	14	14
	4	7	24.2	17	29	5	5	5
5	1	11	4.3	2	19	3.3	2	6
	2	4	21	13	29	6	6	6
	3	6	19.8	9	32	15	15	15
	4	7	21.6	19	24	7	7	7
6	1	10	4.5	2	13	3.9	2	5
	2	4	22.4	16	32	7	7	7
	3	7	21.6	12	29	14	14	14
	4	7	25	16	34	8	8	8
7	1	9	5	1	12	3.7	3	5
	2	4	23.2	19	28	6	6	6
	3	7	23.6	10	34	6	6	6
	4	7	19.6	10	33	8	8	8
8	1	10	4.2	1	11	3.8	2	6
	2	4	20.4	11	26	7	7	7
	3	7	24.6	15	46	15	15	15
	4	7	21.2	9	46	8	8	8
11	1	16	6.1	2	18	3.3	2	8
	2	4	33.8	25	40	15	15	15
	3	7	27.5	7	50	29	29	29
	4	7	43.8	35	52	18	18	18

Note: Due to the structure of the Content Specifications for Claims 2 - 4, analyses for these claims occurred at the claim level (rather than the target level).

Table Read: For grade 3, Claim 1, there are 11 targets. The reviewers, on average, identified 9.6 grade-level standards per target, with a minimum of 1 grade-level standard and a maximum of 23 grade-level standards. The average number of grade-level standards identified in the Content Specifications was 3.5, with a minimum of 2 grade-level standards and a maximum of 8 grade-level standards.

Due to the rather high number of grade-level standards identified by the reviewers compared to what was intended by the Content Specifications, we imposed a ≥*50% reviewer agreement* rule for this analysis as a method to remove outliers. More specifically, when examining the degree to which reviewers identified the same grade-level standards as was intended by the Content Specifications, only those standards for each target that had at least 50% reviewer agreement were retained.

As shown in Table 5.A.25 across Claim 1 targets, all targets had at least one grade-level standard with at least 50% reviewer agreement and thus, all Claim 1 (Concepts and Procedures) targets were retained in the analysis. A fairly large average percentage of the grade-level standards per Claim 1 target were rated as matching the intended mapping. Where there wasn't 100%, most of those grade-level standards per target were believed by the reviewers to fall within the intended domains and to a lesser extent, to fall within the intended clusters, as indicated in the Content Specifications. This pattern remains for Claims 2 – 4 as well. Across grades, Claim 3 (Communicating Reasoning) targets generally had the weakest representation by the grade-level standards at the cluster level.

Table 5.A.25. A.CR-3: Mean Percentage of Mathematics Grade-level Standards Aligned to Intended Targets (Workshop 1)

Grade	Claim		≥ 50% Reviewer Agreement Descriptives		Content Representation		
		Total number of targets in claim	Number of targets included in analysis[1]	Avg # of grade-level standards per target with 50% reviewer agreement	Avg % of grade-level standards per target that matched the intended mapping	Avg % of grade-level standards per target that fell within the intended domains	Avg % of grade-level standards per target that fell within the intended clusters
		N	n	n	% (n)	% (n)	% (n)
3	1	11	11	8.5	57.4% (3.1)	90.9% (7.3)	68.2% (4.4)
	2	4	.	16.0	37.5% (6.0)	100.0% (16.0)	100.0% (16)
	3	6	.	19.0	15.8% (3.0)	89.5% (17.0)	36.8% (7)
	4	7	.	14.0	35.7% (5.0)	92.9% (13.0)	71.4% (10)
4	1	12	12	5.6	70.3% (3.5)	80.3% (4.4)	76.1% (4.1)
	2	4	.	17.0	23.5% (4.0)	100.0% (17.0)	76.4% (13)
	3	6	.	19.0	26.3% (5.0)	73.7% (14.0)	52.6% (10)
	4	7	.	22.0	22.7% (5.0)	90.9% (20.0)	68.1% (15)
5	1	11	11	3.6	85.1% (2.8)	100.0% (3.6)	100% (3.6)
	2	4	.	21.0	28.6% (6.0)	90.5% (19.0)	80.9% (17)
	3	6	.	19.0	47.4% (9.0)	89.5% (17.0)	78.9% (15)
	4	7	.	22.0	31.8% (7.0)	90.9% (20.0)	90.9% (20)
6	1	10	10	4.0	84.6% (3.2)	100.0% (4.0)	100.0% (4)
	2	4	.	24.0	25.0% (6.0)	83.3% (20.0)	83.3% (20)
	3	7	.	20.0	50.0% (10.0)	80.0% (16.0)	80.0% (16)
	4	7	.	24.0	29.2% (7.0)	100.0% (24.0)	95.8% (23)
7	1	9	9	5.3	79.2% (3.6)	100.0% (5.3)	100% (5.3)
	2	4	.	22.0	27.3% (6.0)	77.3% (17.0)	77.2% (17)
	3	7	.	22.0	18.2% (4.0)	50.0% (11.0)	45.4% (10)
	4	7	.	11.0	72.7% (8.0)	100.0% (11.0)	100.0% (11)

Table 5.A.25. (Continued)

Grade	Claim		≥ 50% Reviewer Agreement Descriptives		Content Representation		
		Total number of targets in claim N	Number of targets included in analysis[1] n	Avg # of grade-level standards per target with 50% reviewer agreement n	Avg % of grade-level standards per target that matched the intended mapping % (n)	Avg % of grade-level standards per target that fell within the intended domains % (n)	Avg % of grade-level standards per target that fell within the intended clusters % (n)
8	1	10	10	4.2	90.5% (3.5)	100.0% (4.2)	100.0% (4.2)
	2	4	.	23.0	30.4% (7.0)	82.6% (19.0)	69.5% (16)
	3	7	.	22.0	27.3% (6.0)	77.3% (17.0)	68.1% (15)
	4	7	.	12.0	50.0% (6.0)	91.7% (11.0)	75.0% (9)
11	1	16	16	4.5	74.9% (3.0)	90.9% (3.8)	89.6% (3.7)
	2	4	.	19.0	36.8% (7.0)	68.4% (13.0)	52.6% (10)
	3	7	.	6.0	50.0% (3.0)	50.0% (3.0)	50.0% (3)
	4	7	.	20.0	60.0% (12.0)	85.0% (17.0)	80.0% (16)

Note: Due to the structure of the specifications for Claims 2 - 4, analyses for these claims occurred at the claim level (rather than the target level)
[1]Number of targets with at least one standard with 50% reviewer agreement

Table Read: For grade 3, Claim 1, there were a total of 11 targets, with all 11 targets having at least one standard with ≥ 50% reviewer agreement. Of the targets included in the analysis, there was an average of 8.5 grade-level standards per target with ≥ 50% reviewer agreement. Of the grade-level standards for which at least 50% of reviewers agreed, reviewers rated an average of 57.4% of grade-level standards as matching the intended mapping. Reviewers rated 90.9% of the grade-level standards per target as falling within the intended domains and rated 68.2% of the grade-level standards as falling within the intended clusters.

A.CR-4: Do the individual grade-level standards reflect the content and skills required by the intended targets when reviewers are asked to identify targets aligned to each grade-level standard (Workshop 2)?

Workshop 2 data, where reviewers identified targets aligned to each grade-level standard, were restructured to match that of the format of Workshop 1 data (where reviewers provided grade-level standards aligned to each target). This was done to allow for examining the two-way alignment of the targets and grade-level standards in comparison to what was intended by the Content Specifications. If the alignment was reciprocal, then the results from both analyses (A.CR-3 and A.CR-4) would be similar. If the results differed, then the alignment could be impacted by methodological (e.g., the number of grade-level standards made it difficult to perform a blind review) or content-related (e.g., the broad content in the target or the grade-level standards might have made the rating task very difficult) factors.

As shown in Table 5.A.26, the task of identifying targets aligned to each standard was more difficult than identifying grade-level standards that represented the content and knowledge required in the target. The average percentage of grade-level standards per target that matched the intended mapping was approximately 50% across grades for Claim 1 (Concepts and Procedures), with grade 3 having the lowest percentage of target grade-level standards per target that matched the intended mapping. As expected, Claims 2 – 4, across all grades, had low percentages of grade-level standards that matched the intended mapping. However, when the match of the grade-level standards to the intended clusters and domains was examined, the percentages substantially increased.

Table 5.A.26. A.CR-4: Mean Percentage of Mathematics Grade-level Standards Aligned to Intended Targets Based on Reviewers Identifying Targets Aligned to Each Grade-level Standard (Workshop 2)

Grade	Claim	≥ 50% Reviewer Agreement Descriptives			Content Representation		
		Total number of targets in claim N	Number of targets included in analysis[1] n	Avg number of grade-level standards per target with 50% reviewer agreement n	Avg % of grade-level standards per target that matched the intended mapping % (n)	Avg % of grade-level standards per target that fell within the intended domain % (n)	Avg % of grade-level standards per target that fell within the intended cluster % (n)
3	1	11	11	15.2	22.0% (3.0)	57.1% (8.0)	32.4% (4.2)
	2	4	.	47.0	14.9% (7.0)	93.6% (44.0)	82.9% (39.0)
	3	6	.	47.0	14.9% (7.0)	85.1% (40.0)	55.3% (26.0)
	4	7	.	47.0	8.5% (4.0)	63.8% (30.0)	46.8% (22.0)
4	1	12	12	8.5	50.6% (3.4)	68.0% (5.3)	52.2% (3.7)
	2	4	.	49.0	14.3% (7.0)	91.8% (45.0)	67.3% (33.0)
	3	6	.	49.0	28.6% (14.0)	67.3% (33.0)	59.1% (29.0)
	4	7	.	49.0	10.2% (5.0)	75.5% (37.0)	48.9% (24.0)
5	1	11	11	7.7	54.1% (3.3)	73.4% (5.0)	67.2% (4.5)
	2	4	.	50.0	12.0% (6.0)	90.0% (45.0)	66% (33.0)
	3	6	.	50.0	30.0% (15.0)	90.0% (45.0)	76% (38.0)
	4	7	.	50.0	14.0% (7.0)	90.0% (45.0)	70% (35.0)
6	1	10	10	6.5	73.1% (3.9)	90.4% (5.7)	87.9% (5.3)
	2	4	.	27.0	0.0% (0.0)	85.2% (23.0)	85.1% (23.0)
	3	7	.	7.0	14.3% (1.0)	42.9% (3.0)	42.8% (3.0)
	4	7	.	12.0	0.0% (0.0)	100.0% (12.0)	100% (12.0)
7	1	9	9	5.7	77.0% (3.6)	100.0% (5.7)	100% (5.6)
	2	4	.	29.0	0.0% (0.0)	82.8% (24.0)	82.7% (24.0)
	3	7	.	7.0	0.0% (0.0)	100.0% (7.0)	85.7% (6.0)
	4	7	.	12.0	0.0% (0.0)	100.0% (12.0)	100% (12.0)
8	1	10	10	4.3	89.6% (3.6)	100.0% (4.3)	100% (4.3)
	2	4	.	20.0	0.0% (0.0)	80.0% (16.0)	80% (16.0)
	3	7	.	16.0	37.5% (6.0)	81.3% (13.0)	62.5% (10.0)
	4	7	.	13.0	0.0% (0.0)	100.0% (13.0)	100% (13.0)
11	1	16	16	5.7	64.8% (3.3)	91.1% (4.9)	81.4% (4.3)
	2	4	.	71.0	0.0% (0.0)	63.4% (45.0)	52.1% (37.0)
	3	7	.	51.0	25.5% (13.0)	80.4% (41.0)	72.5% (37.0)
	4	7	.	36.0	2.8% (1.0)	97.2% (35.0)	66.6% (24.0)

These data are from Workshop 2 (A.CR-3 was from Workshop 1). Reviewers identified targets aligned to each standard
Note: Due to the structure of the specifications for Claims 2 - 4, analyses for these claims occurred at the claim level (rather than the target level)
[1]Number of targets with at least one standard with 50% reviewer agreement
[2]For grade 5, Claim 3, reviewers inadvertently were not allowed to map the two grade-level standards to the only target on the summative assessment. As a result, all grade-level standards that reviewers identified fell outside the intended strands.

Table Read: For grade 3, Claim 1, there were a total of 11 targets, with all 11 targets having at least one standard with ≥ 50% reviewer agreement. Of the targets included in the analysis, there was an average of 15.2 grade-level standards per target with ≥ 50% reviewer agreement. Of the grade-level standards for which at least 50% of reviewers agreed, reviewers rated an average of 22% of grade-level standards as matching the intended mapping. Reviewers rated 57.1% of the grade-level standards per target as falling within the intended domains and rated 32.4% of the grade-level standards as falling within the intended clusters.

Table 5.A.27 shows the difference in content representation of the targets by the grade-level standards using the two-way alignment method. Targets were well represented by the grade-level standards when the reviewers' task was to identify grade-level standards aligned to each target. Reversing the task resulted in weaker content representation. These results suggest that the broad nature of the targets likely made it more difficult to align them to specific grade-level standards. Workshop 2 activities permitted reviewers to rate a grade-level standard as not represented by any targets; however, most reviewers found targets that represented at least a small amount of the content and knowledge required in the grade-level standards. This resulted in a higher number of grade-level standards being aligned to each target, thus the average percentage of grade-level standards that matched the intended mapping inherently decreased for Workshop 2. The weaker content representation in Claims 2 – 4 was not unexpected as the alignment for those claims was not directly related to the grade-level standards. Additionally, although not shown in Table 5.A.27, when the content representation was compared at the cluster level between the two methods, the differences were not as stark. This suggests that the two methods resulted in similar results when analyzed at the cluster level rather than at the main and sub-grade-level standard levels.

Table 5.A.27. A.Math.CR-4.Supp-1 Comparison of Mean Percentage of Mathematics Grade-level Standards Aligned to Intended Targets (Workshops 1. vs 2)

		CR-3 (Workshop 1)		CR-4 (Workshop 2)		Difference (CR4-CR3)		
Grade	Claim	Avg number grade-level standards per target with 50% reviewer agreement n	Avg % grade-level standards per target that matched the intended mapping % (n)	Avg number grade-level standards per target with 50% reviewer agreement n	Avg % grade-level standards per target that matched the intended mapping % (n)	Avg number grade-level standards per target with 50% reviewer agreement n	Avg % grade-level standards per target that matched the intended mapping %	n
3	1	8.5	57.4% (3.1)	15.2	22.0% (3.0)	6.7	-35.4%	-0.1
	2	16.0	37.5% (6.0)	47.0	14.9% (7.0)	31.0	-22.6%	1.0
	3	19.0	15.8% (3.0)	47.0	14.9% (7.0)	28.0	-0.9%	4.0
	4	14.0	35.7% (5.0)	47.0	8.5% (4.0)	33.0	-27.2%	-1.0
4	1	5.6	70.3% (3.5)	8.5	50.6% (3.4)	2.9	-19.7%	-0.1
	2	17.0	23.5% (4.0)	49.0	14.3% (7.0)	32.0	-9.2%	3.0
	3	19.0	26.3% (5.0)	49.0	28.6% (14.0)	30.0	2.3%	9.0
	4	22.0	22.7% (5.0)	49.0	10.2% (5.0)	27.0	-12.5%	0.0
5	1	3.6	85.1% (2.8)	7.7	54.1% (3.3)	4.1	-31.0%	0.5
	2	21.0	28.6% (6.0)	50.0	12.0% (6.0)	29.0	-16.6%	0.0
	3	19.0	47.4% (9.0)	50.0	30.0% (15.0)	31.0	-17.4%	6.0
	4	22.0	31.8% (7.0)	50.0	14.0% (7.0)	28.0	-17.8%	0.0
6	1	4.0	84.6% (3.2)	6.5	73.1% (3.9)	2.5	-11.5%	0.7
	2	24.0	25.0% (6.0)	27.0	0.0% (0.0)	3.0	-25.0%	-6.0
	3	20.0	50.0% (10.0)	7.0	14.3% (1.0)	-13.0	-35.7%	-9.0
	4	24.0	29.2% (7.0)	12.0	0.0% (0.0)	-12.0	-29.2%	-7.0
7	1	5.3	79.2% (3.6)	5.7	77.0% (3.6)	0.4	-2.2%	0.0
	2	22.0	27.3% (6.0)	29.0	0.0% (0.0)	7.0	-27.3%	-6.0
	3	22.0	18.2% (4.0)	7.0	0.0% (0.0)	-15.0	-18.2%	-4.0
	4	11.0	72.7% (8.0)	12.0	0.0% (0.0)	1.0	-72.7%	-8.0
8	1	4.2	90.5% (3.5)	4.3	89.6% (3.6)	0.1	-0.9%	0.1
	2	23.0	30.4% (7.0)	20.0	0.0% (0.0)	-3.0	-30.4%	-7.0
	3	22.0	27.3% (6.0)	16.0	37.5% (6.0)	-6.0	10.2%	0.0
	4	12.0	50.0% (6.0)	13.0	0.0% (0.0)	1.0	-50.0%	-6.0

Table 5.A.27. (Continued)

		CR-3 (Workshop 1)		CR-4 (Workshop 2)		Difference (CR4-CR3)		
Grade	Claim	Avg number grade-level standards per target with 50% reviewer agreement n	Avg % grade-level standards per target that matched the intended mapping % (n)	Avg number grade-level standards per target with 50% reviewer agreement n	Avg % grade-level standards per target that matched the intended mapping % (n)	Avg number grade-level standards per target with 50% reviewer agreement n	Avg % grade-level standards per target that matched the intended mapping %	n
11	1	4.5	74.9% (3.0)	5.7	64.8% (3.3)	1.2	-10.1%	0.3
	2	19.0	36.8% (7.0)	71.0	0.0% (0.0)	52.0	-36.8%	-7.0
	3	6.0	50.0% (3.0)	51.0	25.5% (13.0)	45.0	-24.5%	10.0
	4	20.0	60.0% (12.0)	36.0	2.8% (1.0)	16.0	-57.2%	-11.0

Note: Due to the structure of the Content Specifications for Claims 2 - 4, analyses for these claims occurred at the claim level (rather than the target level).

A.CR-5: Does each mathematical practice reflect skills required by the intended target?

Across all grades and claims, the average percentages of targets aligned to each mathematical practice were generally high (see Appendix J for a list of the mathematical practices). The notable exception was for grade 5, Claim 1 (Concepts and Procedures) where six of the eight mathematical practices had average percentages of aligned targets that were less than 60%. In addition, grade 11 Mathematical Practice 8 (Look for and express regularity in repeated reasoning) also had lower mean percentages of aligned targets across all claims. The mean percentages ranged from 22.1% for Claim 1 to 50.0% for Claim 3 (Communicating Reasoning). Lower percentages of alignment in Claim 1 targets were not unexpected as Claim 1 targets were designed to align more with the grade-level standards than with the mathematical practices.

Table 5.A.28. A.CR-5: Mean Percentage of Mathematics Targets Aligned to Each Mathematical Practice

Grade	Mathematical Practice	Claim 1 % (n)	Claim 2 % (n)	Claim 3 % (n)	Claim 4 % (n)
3	1	100.0% (10.8)	100.0% (3.6)	100.0% (5.6)	100.0% (5.8)
	2	98.2% (10.6)	100.0% (3.6)	100.0% (5.6)	100.0% (5.8)
	3	84.9% (9.2)	85.0% (3.0)	96.7% (5.4)	93.3% (5.4)
	4	100.0% (10.8)	100.0% (3.6)	96.7% (5.4)	93.3% (5.4)
	5	92.2% (10.0)	100.0% (3.6)	86.7% (4.8)	80.0% (4.6)
	6	100.0% (10.8)	85.0% (3.0)	83.3% (4.6)	80.0% (4.6)
	7	100.0% (10.8)	85.0% (3.0)	86.7% (4.8)	80.0% (4.6)
	8	100.0% (10.8)	85.0% (3.0)	86.7% (4.8)	80.0% (4.6)
4	1	100.0% (11.6)	95.0% (3.6)	96.7% (5.6)	100.0% (6.0)
	2	90.9% (10.8)	100.0% (3.8)	96.7% (5.8)	100.0% (6.4)
	3	61.8% (7.4)	80.0% (3.0)	83.3% (5.0)	86.7% (5.2)
	4	81.1% (9.6)	90.0% (3.4)	90.0% (5.4)	83.3% (5.0)
	5	81.2% (9.6)	90.0% (3.4)	86.7% (5.2)	83.3% (5.2)
	6	86.7% (10.2)	85.0% (3.2)	90.0% (5.4)	86.7% (5.4)
	7	88.3% (10.4)	90.0% (3.4)	90.0% (5.4)	90.0% (5.6)
	8	83.3% (9.8)	85.0% (3.2)	90.0% (5.4)	86.7% (5.4)
5	1	80.0% (8.0)	95.0% (3.4)	93.3% (5.6)	100.0% (7.0)
	2	57.9% (5.8)	75.0% (2.6)	90.0% (5.4)	85.7% (6.0)
	3	33.3% (3.4)	50.0% (1.8)	96.7% (5.8)	74.3% (5.2)
	4	73.5% (7.6)	65.0% (2.2)	60.0% (3.6)	73.3% (5.0)
	5	56.7% (5.8)	65.0% (2.4)	46.7% (2.8)	75.7% (5.2)
	6	54.9% (5.6)	85.0% (3.0)	73.3% (4.4)	74.3% (5.2)
	7	58.5% (6.0)	65.0% (2.2)	80.0% (4.8)	66.7% (4.6)
	8	44.9% (4.6)	30.0% (1.0)	63.3% (3.8)	74.3% (4.8)
6	1	93.6% (9.2)	85.0% (3.4)	85.7% (6.0)	94.3% (6.4)
	2	82.0% (8.2)	75.0% (3.0)	82.9% (5.8)	82.9% (5.6)
	3	54.4% (5.4)	60.0% (2.4)	94.3% (6.6)	85.7% (5.8)
	4	70.0% (7.0)	65.0% (2.6)	65.7% (4.6)	79.0% (5.2)
	5	54.0% (5.4)	80.0% (3.2)	60.0% (4.2)	70.0% (4.6)
	6	70.0% (7.0)	75.0% (3.0)	71.4% (5.0)	77.1% (5.2)
	7	62.0% (6.2)	56.7% (2.2)	68.6% (4.8)	66.7% (4.4)
	8	46.0% (4.6)	46.7% (1.8)	65.7% (4.6)	69.7% (4.6)

Table 5.A.28. (Continued)

Grade	Mathematical Practice	Claim 1 % (n)	Claim 2 % (n)	Claim 3 % (n)	Claim 4 % (n)
7	1	100.0% (9.0)	100.0% (3.8)	100.0% (6.6)	100.0% (5.6)
	2	93.3% (8.4)	88.3% (3.4)	100.0% (6.6)	97.1% (5.4)
	3	73.3% (6.6)	78.3% (3.0)	97.1% (6.4)	100.0% (5.6)
	4	95.6% (8.6)	88.3% (3.4)	82.9% (5.4)	100.0% (5.6)
	5	88.9% (8.0)	100.0% (3.8)	82.9% (5.4)	100.0% (5.6)
	6	88.9% (8.0)	100.0% (3.8)	94.3% (6.2)	100.0% (5.6)
	7	73.3% (6.6)	51.7% (2.0)	85.7% (5.6)	100.0% (5.6)
	8	68.9% (6.2)	66.7% (2.6)	82.9% (5.4)	100.0% (5.6)
8	1	100.0% (9.2)	100.0% (4.0)	100.0% (6.8)	100.0% (6.5)
	2	94.0% (9.4)	95.0% (3.8)	100.0% (6.8)	100.0% (5.4)
	3	78.0% (7.8)	100.0% (4.0)	100.0% (6.8)	100.0% (6.4)
	4	86.0% (8.6)	95.0% (3.8)	100.0% (6.8)	100.0% (6.6)
	5	86.0% (8.6)	95.0% (3.8)	100.0% (6.8)	97.1% (6.4)
	6	98.0% (9.8)	90.0% (3.6)	100.0% (6.8)	100.0% (6.6)
	7	82.0% (8.2)	80.0% (3.2)	72.4% (5.0)	97.1% (6.4)
	8	76.0% (7.4)	70.0% (2.8)	70.0% (4.8)	94.3% (6.2)
11	1	92.2% (14.8)	100.0% (4.0)	96.4% (6.8)	100.0% (6.8)
	2	82.8% (13.3)	100.0% (4.0)	71.4% (5.0)	95.8% (6.5)
	3	25.5% (4.0)	56.3% (2.3)	100.0% (7.0)	63.1% (4.3)
	4	56.3% (9.0)	100.0% (4.0)	67.9% (4.8)	100.0% (6.8)
	5	68.8% (11.0)	81.3% (3.3)	71.4% (5.0)	85.7% (5.8)
	6	65.6% (10.5)	93.8% (3.8)	82.1% (5.8)	95.8% (6.5)
	7	76.6% (12.3)	68.8% (2.8)	71.4% (5.0)	62.5% (4.3)
	8	22.1% (3.5)	37.5% (1.5)	50.0% (3.5)	49.4% (3.3)

Note: Lower percentages of alignment for Claim 1 targets were not unexpected based on the design of the Content Specifications

Table Read: For grade 3, Mathematical Practice 1 (Make sense of problems and persevere in solving them), reviewers rated 100% of the mean percentages of targets aligned for all four claims. The average number of targets differed by claim: 10.8 for Claim 1, 3.6 for Claim 2, 6.2 for Claim 3, and 5.8 for Claim 4.

A.CR-6: Do the reviewers agree with the intended mapping of targets and grade-level standards as identified in the Content Specifications?

The overall pairwise agreement in identifying grade-level standards aligned to each target between reviewers and the intended mapping as identified in the Content Specifications (across all grades, claims, targets, and reviewers) was 36.6%. The pairwise agreement at the cluster-level[32], however, is substantially higher (overall agreement of 64.9%).

The rather low agreement at the main and sub-standard levels likely resulted because of the combination of (a) the higher number of cluster-level standards identified for Claims 2 – 4 in the Content Specifications, and (b) the high number of grade-level standards that reviewers identified for each target compared to the number of grade-level standards identified in the Content

[32] For mathematics, clusters and domains refer to the same content.

Specifications. As shown in Table 5.A.29, however, reviewer agreement with the intended mapping increased when computing the average percent of reviewers per target that agreed with at least 50% of the intended standards. This suggests that while there was low overall agreement in identifying exactly what was intended, reviewers generally agreed with at least 50% of the intended standards.

Table 5.A.29. A.CR-6: Pairwise Agreement between Reviewers' and Intended Mapping of Mathematics Targets and Grade-level Standards

Grade	Claim	Descriptives			Agreement			
		# of Reviewers % (n)	# of Targets	# of Ratings	Pairwise Agree-ment	Pairwise Agreement (Cluster-level)	Hit All Intended Standards (but noted others) Avg % (n Reviewers)	Hit At Least 50% of the Intended Standards Avg % (n Reviewers)
3	1	4.9	11	54	44.8%	47.9%	70.0% (3.5)	100.0% (4.9)
	2	5.0	.	5	30.1%	66.0%	20.0% (1.0)	80.0% (4.0)
	3	5.0	.	5	14.3%	33.8%	0.0% (0.0)	20.0% (1.0)
	4	5.0	.	5	21.7%	50.8%	60.0% (3.0)	80.0% (4.0)
4	1	5.0	12	60	65.0%	69.8%	53.3% (2.7)	100.0% (5.0)
	2	5.0	.	5	26.6%	65.5%	0.0% (0.0)	100.0% (5.0)
	3	5.0	.	5	28.1%	52.9%	0.0% (0.0)	40.0% (2.0)
	4	5.0	.	5	20.6%	58.3%	80.0% (4.0)	100.0% (5.0)
5	1	5.0	11	55	73.6%	88.2%	21.8% (1.1)	96.4% (4.8)
	2	5.0	.	5	22.8%	62.0%	60.0% (3.0)	100.0% (5.0)
	3	5.0	.	5	36.7%	67.7%	0.0% (0.0)	60.0% (3.0)
	4	5.0	.	5	26.8%	78.4%	60.0% (3.0)	80.0% (4.0)
6	1	5.0	10	50	66.0%	88.7%	22.0% (1.1)	86.0% (4.3)
	2	5.0	.	5	26.6%	80.0%	0.0% (0.0)	100.0% (5.0)
	3	5.0	.	5	38.7%	71.0%	0.0% (0.0)	80.0% (4.0)
	4	5.0	.	5	28.8%	80.6%	40.0% (2.0)	100.0% (5.0)
7	1	5.0	9	45	65.5%	86.9%	33.3% (1.7)	88.9% (4.4)
	2	5.0	.	5	25.7%	69.3%	80.0% (4.0)	100.0% (5.0)
	3	5.0	.	5	13.0%	33.5%	20.0% (1.0)	60.0% (3.0)
	4	5.0	.	5	43.5%	82.1%	40.0% (2.0)	100.0% (5.0)
8	1	5.0	10	50	80.5%	92.7%	14.0% (0.7)	90.0% (4.5)
	2	5.0	.	5	26.6%	63.6%	40.0% (2.0)	80.0% (4.0)
	3	5.0	.	5	28.3%	54.9%	20.0% (1.0)	20.0% (1.0)
	4	5.0	.	5	36.4%	62.7%	20.0% (1.0)	80.0% (4.0)
11	1	4.0	16	64	59.2%	72.2%	64.1% (2.6)	100.0% (4.0)
	2	4.0	.	4	24.5%	46.8%	0.0% (0.0)	75.0% (3.0)
	3	4.0	.	4	20.5%	33.3%	0.0% (0.0)	0.0% (0.0)
	4	4.0	.	4	28.7%	58.9%	0.0% (0.0)	100.0% (4.0)

Note: Due to the structure of the specifications for Claims 2 - 4, analyses for these claims occurred at the claim level (rather than the target level)
Decimals in the # of Reviewers column indicate missing data

Table Read: For grade 3, there were 4.9 reviewers who rated 11 targets in Claim 1 for a total of 54 ratings (pairwise comparisons with the Content Specifications). Of those ratings, the pairwise agreement of the mappings of targets and grade-level standards was 44.8%. The pairwise agreement, rolling up to the cluster-level, is 47.9%. Diagnostically, an average of 70.0% of the reviewers per target hit all of the intended standards, while also indicating additional standards. An average of 100% of the reviewers hit at least 50% of the intended standards per target.

DOK Distribution

Analyses were conducted to examine the distribution of mathematics target DOK levels between the Content Specifications and the reviewers, and to address the following questions:

- A.DD-1: Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the Content Specifications (using the max DOK level)?
- A.DD-2: Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the Content Specifications (using the each independent DOK level)?
- A.DD-3: Do the reviewers agree with the intended target DOK levels as identified in the Content Specifications?

The main reviewer tasks for examining the DOK distribution of the targets involved reviewers providing independent DOK ratings for each target. The purpose of these analyses was to describe and compare the cognitive demand required in the targets as identified by the reviewers with the cognitive demand indicated in the Content Specifications.

Because the Content Specifications often indicate more than one DOK level per target, reviewers were also allowed to identify more than one DOK level per target. Generally, the reviewers from grades 3 and 4 indicated targets required multiple levels of cognitive demand compared to the Content Specifications, which specified fewer levels (Table 5.A.30). The reverse was true for grades 7 and 8; reviewers for grades 5, 6, and high school indicated similar numbers of DOK levels compared to the Content Specifications.

Table 5.A.30. DD-GD. Overall Descriptive Comparison of Reviewer and Specifications Target DOK Ratings for Mathematics Targets

Grade	Number of DOK Levels Indicated by Reviewers			Number of DOK Levels Indicated by Content Specifications		
	N	Mean	SD	N	Mean	SD
3	140	2.5	0.8	28	1.8	0.7
4	145	2.6	0.6	29	2.1	0.5
5	140	1.9	0.6	28	1.9	0.6
6	140	1.9	0.5	28	2.0	0.5
7	135	1.1	0.4	27	2.0	0.6
8	140	1.1	0.4	28	2.1	0.5
11	136	2.3	0.6	34	2.0	0.5

Similar patterns exist when the number of levels used per target was disaggregated by claim (refer to Table 5.A.31).

Table 5.A.31. A.Math.DD.GD-1 Descriptive Comparison of Reviewer and Specifications Target DOK Ratings by Grade and Claim

		Avg # DOK Levels Indicated per Target	
Grade	Claim	Reviewers	Content Specifications
3	1	2.4	1.4
	2	2.5	2.0
	3	2.7	2.0
	4	2.9	2.3
4	1	2.4	2.0
	2	2.8	2.0
	3	2.6	2.0
	4	3.0	2.3
5	1	1.8	1.6
	2	2.1	2.0
	3	2.0	2.0
	4	1.8	2.3
6	1	1.9	1.8
	2	2.1	2.0
	3	1.9	2.0
	4	1.9	2.3
7	1	1.2	1.8
	2	1.2	2.0
	3	1.0	2.0
	4	1.2	2.3
8	1	1.2	2.0
	2	1.2	2.0
	3	1.0	2.0
	4	1.0	2.3
11	1	2.3	1.9
	2	2.1	2.0
	3	2.4	2.0
	4	2.2	2.3

Pairwise Agreement among Reviewers

The overall pairwise agreement among reviewers when identifying DOK ratings for each target (across all grades, claims, targets, and reviewers) was 63.7% (see Table 5.A.32).

Table 5.A.32. A.Math.DD.PWA-1. Pairwise Percent Agreement between Reviewers' Target DOK Ratings

		Descriptives			Agreement
Grade	Claim	Avg # of Reviewers n	# of Targets	Avg # of reviewer pairs n	Pairwise Agreement
3	1	4.9	11	9.6	82.4%
	2	4.8	4	9.0	60.4%
	3	5.0	6	10.0	71.3%
	4	4.7	7	8.9	67.4%
4	1	5.0	12	10.0	82.5%
	2	5.0	4	10.0	79.4%
	3	5.0	6	10.0	66.8%
	4	4.9	7	9.4	76.9%
5	1	5.0	11	10.0	64.2%
	2	5.0	4	10.0	65.0%
	3	5.0	6	10.0	61.1%
	4	5.0	7	10.0	60.5%
6	1	5.0	10	10.0	61.5%
	2	5.0	4	10.0	63.8%
	3	5.0	7	10.0	58.8%
	4	5.0	7	10.0	56.4%
7	1	5.0	9	10.0	38.1%
	2	4.8	4	9.0	50.8%
	3	4.9	7	9.4	76.4%
	4	4.6	7	8.4	44.8%
8	1	5.0	10	10.0	39.5%
	2	5.0	4	10.0	52.5%
	3	4.7	7	8.9	83.6%
	4	4.7	7	8.9	45.7%
11	1	4.0	16	6.0	77.3%
	2	4.0	4	6.0	54.9%
	3	4.0	7	6.0	76.0%
	4	4.0	7	6.0	67.1%

Table Read: For grade 3, Claim 1, there was an average of 4.9 reviewers who rated 11 targets. The average number of total possible pairs was 9.6. The pairwise agreement across all 9.6 pairs for all 11 targets was 82.4%. A decimal for the number of reviewers and number of pairs indicates missing data.

Findings

Findings related to each DOK Distribution (DD) question are presented below.

A.DD-1: Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the Content Specifications (using the max DOK level)?

To get a sense for whether reviewers thought the targets required higher levels of cognitive demand than what was intended, the DOK distribution of the targets as identified by the reviewers and the Content Specifications using the maximum DOK level identified was examined. For example, if the Content Specifications indicated a target required DOK levels 1 and 2, the analysis used only DOK level 2. As shown in Table 5.A.33 generally, across grades for Claims 2 (Problem Solving), 3 (Communicating Reasoning), and 4 (Modeling and Data Analysis), the cognitive demand indicated by the reviewers and specifications was fairly similar. For Claim 1 (Concepts and Procedures), however, reviewers across grades indicated higher mean percentages of targets as requiring a higher cognitive demand that what was intended.

A.DD-2: Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the Content Specifications (using the each independent DOK level)?

We next examined the DOK distribution of the targets using each identified DOK level (i.e., multiple DOK levels per target). As shown in Table 5.A.34, generally, across claims for grades 5 and 11, reviewers indicated similar levels of cognitive demand required by the targets as did the Content Specifications. Grades 3 and 4 reviewers, indicated a larger distribution of cognitive demand that what was intended and grades 7 and 8 reviewers generally indicated the targets as requiring the higher levels of cognitive demand.

A.DD-3: Do the reviewers agree with the intended target DOK levels as identified in the Content Specifications?

The overall pairwise agreement when identifying grade-level standards aligned to each target between reviewers and the intended mapping as identified in the Content Specifications (across all grades, claims, targets, and reviewers) was 57.1%. As shown in Table 5.A.35, with the exception of grades 3, 8, and 11, reviewers had the highest level of agreement with the specifications on Claim 1 targets. Grades 7 and 8 had the lowest levels of agreement with the specifications. Grades 7 and 8 also indicated fewer levels of cognitive demand per target than what was indicated in the specifications, which explains the lower agreement rates.

Table 5.A.33. A.DD-1: Reviewers' Mean Percentage of Mathematics at Each DOK Level (Max) by Grade and Claim Compared to Content Specifications

Grade	Claim	DOK 1		DOK 2		DOK 3		DOK 4	
		Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)
3	1	1.8% (0.2)	45.5% (5.0)	48.2% (5.2)	54.5% (6.0)	50.0% (5.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	2	5.0% (0.2)	0.0% (0.0)	43.3% (1.6)	50.0% (2.0)	31.7% (1.2)	50.0% (2.0)	20.0% (0.8)	0.0% (0.0)
	3	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	16.7% (1.0)	66.7% (4.0)	50.0% (3.0)	33.3% (2.0)	33.3% (2.0)
	4	0.0% (0.0)	0.0% (0.0)	19.4% (1.2)	14.3% (1.0)	46.3% (3.0)	42.9% (3.0)	34.3% (2.4)	42.9% (3.0)
4	1	0.0% (0.0)	0.0% (0.0)	56.7% (6.8)	91.7% (11.0)	43.3% (5.2)	8.3% (1.0)	0.0% (0.0)	0.0% (0.0)
	2	0.0% (0.0)	0.0% (0.0)	35.0% (1.4)	50.0% (2.0)	55.0% (2.2)	50.0% (2.0)	10.0% (0.4)	0.0% (0.0)
	3	0.0% (0.0)	0.0% (0.0)	16.7% (1.0)	16.7% (1.0)	53.3% (3.2)	50.0% (3.0)	30.0% (1.8)	33.3% (2.0)
	4	0.0% (0.0)	0.0% (0.0)	3.3% (0.2)	14.3% (1.0)	62.4% (4.2)	42.9% (3.0)	34.3% (2.4)	42.9% (3.0)
5	1	14.5% (1.6)	9.1% (1.0)	61.8% (6.8)	90.9% (10.0)	23.6% (2.6)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	2	5.0% (0.2)	0.0% (0.0)	25.0% (1.0)	50.0% (2.0)	65.0% (2.6)	50.0% (2.0)	5.0% (0.2)	0.0% (0.0)
	3	0.0% (0.0)	0.0% (0.0)	3.3% (0.2)	16.7% (1.0)	63.3% (3.8)	50.0% (3.0)	33.3% (2.0)	33.3% (2.0)
	4	0.0% (0.0)	0.0% (0.0)	8.6% (0.6)	14.3% (1.0)	57.1% (4.0)	42.9% (3.0)	34.3% (2.4)	42.9% (3.0)
6	1	14.0% (1.4)	0.0% (0.0)	48.0% (4.8)	100.0% (10.0)	36.0% (3.6)	0.0% (0.0)	2.0% (0.2)	0.0% (0.0)
	2	0.0% (0.0)	0.0% (0.0)	20.0% (0.8)	50.0% (2.0)	75.0% (3.0)	50.0% (2.0)	5.0% (0.2)	0.0% (0.0)
	3	0.0% (0.0)	0.0% (0.0)	5.7% (0.4)	14.3% (1.0)	60.0% (4.2)	42.9% (3.0)	34.3% (2.4)	42.9% (3.0)
	4	0.0% (0.0)	0.0% (0.0)	2.9% (0.2)	14.3% (1.0)	54.3% (3.8)	42.9% (3.0)	42.9% (3.0)	42.9% (3.0)
7	1	2.2% (0.2)	0.0% (0.0)	40.0% (3.6)	100.0% (9.0)	48.9% (4.4)	0.0% (0.0)	8.9% (0.8)	0.0% (0.0)
	2	0.0% (0.0)	0.0% (0.0)	26.7% (1.0)	50.0% (2.0)	58.3% (2.2)	50.0% (2.0)	15.0% (0.6)	0.0% (0.0)
	3	0.0% (0.0)	0.0% (0.0)	2.9% (0.2)	14.3% (1.0)	85.7% (5.8)	42.9% (3.0)	11.4% (0.8)	42.9% (3.0)
	4	0.0% (0.0)	0.0% (0.0)	5.7% (0.4)	14.3% (1.0)	30.7% (2.0)	42.9% (3.0)	63.6% (4.0)	42.9% (3.0)
8	1	12.0% (1.2)	0.0% (0.0)	50.0% (5.0)	100.0% (10.0)	38.0% (3.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	2	0.0% (0.0)	0.0% (0.0)	30.0% (1.2)	50.0% (2.0)	70.0% (2.8)	50.0% (2.0)	0.0% (0.0)	0.0% (0.0)
	3	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	14.3% (1.0)	93.8% (6.2)	42.9% (3.0)	6.2% (0.4)	42.9% (3.0)
	4	0.0% (0.0)	0.0% (0.0)	8.6% (0.6)	14.3% (1.0)	20.0% (1.4)	42.9% (3.0)	71.4% (4.6)	42.9% (3.0)
11	1	1.6% (0.3)	0.0% (0.0)	34.4% (5.5)	93.8% (15.0)	62.5% (10.0)	6.3% (1.0)	1.6% (0.3)	0.0% (0.0)
	2	0.0% (0.0)	0.0% (0.0)	31.3% (1.3)	50.0% (2.0)	50.0% (2.0)	50.0% (2.0)	18.8% (0.8)	0.0% (0.0)
	3	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	14.3% (1.0)	64.3% (4.5)	42.9% (3.0)	35.7% (2.5)	42.9% (3.0)
	4	0.0% (0.0)	0.0% (0.0)	3.6% (0.3)	14.3% (1.0)	67.9% (4.8)	42.9% (3.0)	28.6% (2.0)	42.9% (3.0)

Note: For each group (reviewers and specifications) the percentages across DOK levels are mutually exclusive.

Table 5.A.34. A.DD-2: Reviewers' Mean Percentage of Mathematics Targets at Each DOK Level (Independent) by Grade and Claim Compared to Content Specifications

Grade	Claim	DOK 1		DOK 2		DOK 3		DOK 4	
		Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)
3	1	96.4% (10.4)	81.8% (9.0)	98.2% (10.6)	54.5% (6.0)	50.0% (5.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	2	85.0% (3.2)	50.0% (2.0)	95.0% (3.6)	100.0% (4.0)	51.7% (2.0)	50.0% (2.0)	20.0% (0.8)	0.0% (0.0)
	3	33.3% (2.0)	0.0% (0.0)	100.0% (6.0)	83.3% (5.0)	100.0% (6.0)	83.3% (5.0)	33.3% (2.0)	33.3% (2.0)
	4	70.3% (4.6)	28.6% (2.0)	100.0% (6.6)	71.4% (5.0)	80.6% (5.4)	85.7% (6.0)	34.3% (2.4)	42.9% (3.0)
4	1	95.0% (11.4)	91.7% (11.0)	100.0% (12.0)	100.0% (12.0)	43.3% (5.2)	8.3% (1.0)	0.0% (0.0)	0.0% (0.0)
	2	100.0% (4.0)	50.0% (2.0)	100.0% (4.0)	100.0% (4.0)	65.0% (2.6)	50.0% (2.0)	10.0% (0.4)	0.0% (0.0)
	3	50.0% (3.0)	0.0% (0.0)	100.0% (6.0)	83.3% (5.0)	83.3% (5.0)	83.3% (5.0)	30.0% (1.8)	33.3% (2.0)
	4	63.8% (4.4)	28.6% (2.0)	100.0% (6.8)	71.4% (5.0)	96.7% (6.6)	85.7% (6.0)	34.3% (2.4)	42.9% (3.0)
5	1	70.9% (7.8)	72.7% (8.0)	85.5% (9.4)	90.9% (10.0)	23.6% (2.6)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	2	40.0% (1.6)	50.0% (2.0)	95.0% (3.8)	100.0% (4.0)	70.0% (2.8)	50.0% (2.0)	5.0% (0.2)	0.0% (0.0)
	3	6.7% (0.4)	0.0% (0.0)	60.0% (3.6)	83.3% (5.0)	96.7% (5.8)	83.3% (5.0)	33.3% (2.0)	33.3% (2.0)
	4	5.7% (0.4)	28.6% (2.0)	54.3% (3.8)	71.4% (5.0)	88.6% (6.2)	85.7% (6.0)	34.3% (2.4)	42.9% (3.0)
6	1	66.0% (6.6)	80.0% (8.0)	82.0% (8.2)	100.0% (10.0)	38.0% (3.8)	0.0% (0.0)	2.0% (0.2)	0.0% (0.0)
	2	35.0% (1.4)	50.0% (2.0)	85.0% (3.4)	100.0% (4.0)	80.0% (3.2)	50.0% (2.0)	5.0% (0.2)	0.0% (0.0)
	3	5.7% (0.4)	0.0% (0.0)	54.3% (3.8)	71.4% (5.0)	94.3% (6.6)	85.7% (6.0)	34.3% (2.4)	42.9% (3.0)
	4	11.4% (0.8)	28.6% (2.0)	48.6% (3.4)	71.4% (5.0)	88.6% (6.2)	85.7% (6.0)	42.9% (3.0)	42.9% (3.0)
7	1	8.9% (0.8)	77.8% (7.0)	48.9% (4.4)	100.0% (9.0)	57.8% (5.2)	0.0% (0.0)	8.9% (0.8)	0.0% (0.0)
	2	0.0% (0.0)	50.0% (2.0)	26.7% (1.0)	100.0% (4.0)	73.3% (2.8)	50.0% (2.0)	15.0% (0.6)	0.0% (0.0)
	3	0.0% (0.0)	0.0% (0.0)	2.9% (0.2)	71.4% (5.0)	88.6% (6.0)	85.7% (6.0)	11.4% (0.8)	42.9% (3.0)
	4	0.0% (0.0)	28.6% (2.0)	8.6% (0.6)	71.4% (5.0)	43.7% (2.8)	85.7% (6.0)	63.6% (4.0)	42.9% (3.0)
8	1	26.0% (2.6)	100.0% (10.0)	56.0% (5.6)	100.0% (10.0)	38.0% (3.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
	2	0.0% (0.0)	50.0% (2.0)	45.0% (1.8)	100.0% (4.0)	70.0% (2.8)	50.0% (2.0)	0.0% (0.0)	0.0% (0.0)
	3	0.0% (0.0)	0.0% (0.0)	3.3% (0.2)	71.4% (5.0)	93.8% (6.2)	85.7% (6.0)	6.2% (0.4)	42.9% (3.0)
	4	0.0% (0.0)	28.6% (2.0)	8.6% (0.6)	71.4% (5.0)	20.0% (1.4)	85.7% (6.0)	71.4% (4.6)	42.9% (3.0)
11	1	67.2% (10.8)	87.5% (14.0)	98.4% (15.8)	100.0% (16.0)	64.1% (10.3)	6.3% (1.0)	1.6% (0.3)	0.0% (0.0)
	2	37.5% (1.5)	50.0% (2.0)	87.5% (3.5)	100.0% (4.0)	68.8% (2.8)	50.0% (2.0)	18.8% (0.8)	0.0% (0.0)
	3	10.7% (0.8)	0.0% (0.0)	92.9% (6.5)	71.4% (5.0)	100.0% (7.0)	85.7% (6.0)	35.7% (2.5)	42.9% (3.0)
	4	14.3% (1.0)	28.6% (2.0)	78.6% (5.5)	71.4% (5.0)	96.4% (6.8)	85.7% (6.0)	28.6% (2.0)	42.9% (3.0)

Note: For each group (reviewers and specifications) the percentages across DOK levels are not mutually exclusive since a target could have multiple DOK levels.

Table 5.A.35. A.DD-3: Pairwise Percent Agreement between Reviewers' and Intended Mathematics Target DOK Ratings

Grade	Claim	Descriptives			Agreement
		Avg # of Reviewers n	# of Targets	# of Ratings	Pairwise Agreement %
3	1	4.9	11	54	58.0%
	2	4.8	4	19	61.8%
	3	5.0	6	30	66.9%
	4	4.7	7	33	63.5%
4	1	5.0	12	60	85.6%
	2	5.0	4	20	66.2%
	3	5.0	6	30	62.8%
	4	4.9	7	34	66.7%
5	1	5.0	11	55	71.8%
	2	5.0	4	20	65.8%
	3	5.0	6	30	66.1%
	4	5.0	7	35	68.1%
6	1	5.0	10	50	69.7%
	2	5.0	4	20	60.0%
	3	5.0	7	35	61.0%
	4	5.0	7	35	62.4%
7	1	5.0	9	45	29.6%
	2	4.8	4	19	28.3%
	3	4.9	7	34	39.0%
	4	4.6	7	32	33.5%
8	1	5.0	10	50	41.0%
	2	5.0	4	20	33.3%
	3	4.7	7	33	40.5%
	4	4.7	7	33	23.8%
11	1	4.0	16	64	68.2%
	2	4.0	4	16	65.6%
	3	4.0	7	28	73.5%
	4	4.0	7	28	69.6%

Table Read: For grade 3, there were 4.9 mathematics reviewers who provided target DOK ratings for 11 targets in Claim 1, for a total of 54 ratings. Across these ratings, the pairwise agreement was 58.0%.

DOK Consistency

Analyses were conducted to examine the consistency of mathematics levels between the Content Specifications and the CCSS, and to address the following question:

- A.DC-1: Is the cognitive complexity required in the targets consistent with the cognitive complexity required in each targets' mapped grade-level standards?

Findings

A.DC-1: Is the cognitive complexity required in the targets consistent with the cognitive complexity required in each targets' mapped grade-level standards?

The DOK consistency analysis examined the degree to which the cognitive demand required by each of the grade-level standards aligned to a target fell within the range of cognitive demand required by the intended target. The assumptions for interpreting the results in Tables 5.A.36 - 5.A.37 were:

- The DOKs for the grade-level standards were determined by reviewer consensus in Workshop 1.
- Only the reviewers' grade-level standards that matched the intended mapping indicated in the Content Specifications were retained for this analysis.
- Of those standards that matched the intended mapping, only those standards with ≥ 50% reviewer agreement were retained for this analysis.
- *Consistency* was defined in two ways:
    a. The cognitive demand of *all* of the grade-level standards mapped to a target by the reviewers needed to fall within the range of the intended target DOK (refer to Table 5.A.36).
    b. Where the grade-level standards that were mapped to a target by reviewers had multiple DOK levels, only one of those levels had to fall within the range of the intended target DOK (refer to Table 5.A.37).
- This analysis was not conducted for Claims 2 – 4. The grade-level standards mapping indicated in the Content Specifications for these claims occurs at the claim level, rather than the individual target level. Because this analysis focuses on the consistency between the cognitive demand required in the grade-level standards and in the targets, we felt that comparing the DOK of the grade-level standards at the claim level would not result in meaningful interpretations.
- Results here should be interpreted in relation to the reviewer agreement with the intended standard-to-target mapping. Because the DOK consistency analysis was applied only to those standards with 50% agreement *and* that matched the intended mapping, it is possible that each target had a differing percentage of mapped standards that were included.

As shown in Table 5.A.36, there was no real pattern in the percentage of targets that had DOK consistency with all of the mapped grade-level standards. The percentage of targets that had DOK levels consistent with those of their mapped grade-level standards ranged quite widely from 11.1% to 90%. Upon further investigation, the reason why so many targets had DOK levels inconsistent with their mapped grade-level standards was because the reviewers rated the grade-level standards as requiring higher levels of cognitive demand than what was intended by the targets. For example, for grade 3, 72.8% (n=8) of the targets were rated as having inconsistent DOK levels. When looking at the grade-level standards for the inconsistent targets, an average of 70.8% of the grade-level standards per inconsistent target had a maximum grade-level standard with a DOK level higher than the maximum target DOK level identified in the Content Specifications. This was generally seen across grades.

As shown in Table 5.A.37, when the DOK consistency definition was relaxed to requiring only one DOK level for each grade-level standard mapped to a target to fall within the range of the intended target DOK, the percentage of targets with DOK consistency substantially increased. This suggests that of the grade-level standards with multiple DOK levels, the reviewers believed that part of the cognitive demand required in the grade-level standard matched that of the intended target, yet they believed there were some portions of the grade-level standards that fell outside that range. For those targets that had inconsistent DOK levels, the general pattern remained that the grade-level standards required higher cognitive demand than what was intended by the target.

Table 5.A.36. Math.DC-1a. Percentage of Mathematics Targets with DOKs Consistent with Intended Grade-Level Standards that Matched Intended Mapping for All Targets – All CCSS within Range

Grade	Claim	Descriptives			DOK Consistency				
					Consistent	Inconsistent			
		Total number of targets in claim	Number of targets included in analysis[1]	Avg % of grade-level standards per target with ≥50% reviewer agreement[2]	% of Targets that Have All Mapped CCSS Consistent[3]	% of Targets With All Mapped CCSS Inconsistent	Avg % of CCSS per Inconsistent Target with Max DOK Consensus > Specs	Avg % of CCSS per Target with Min DOK Consensus < Specs	Number of Targets With < 50% agreement
		N	n	% (n)	% (n)	% (n)	% (n)	% (n)	n
All Targets									
3	1	11	11	95.5% (3.09)	27.3% (3)	72.7% (8)	70.8% (2.12)	25.0% (0.75)	0
4	1	12	12	100.0% (3.50)	66.7% (8)	33.3% (4)	29.2% (1.00)	25.0% (0.50)	0
5	1	11	11	89.1% (2.73)	45.5% (5)	54.5% (6)	66.7% (1.67)	5.6% (0.17)	0
6	1	10	10	82.7% (3.10)	60.0% (6)	40.0% (4)	55.8% (1.75)	18.8% (0.75)	0
7	1	9	9	97.2% (3.56)	11.1% (1)	88.9% (8)	53.3% (1.88)	3.1% (0.12)	0
8	1	10	10	94.3% (3.50)	90.0% (9)	10.0% (1)	66.7% (2.00)	0.0% (0.00)	0
11	1	16	16	98.4% (3.25)	31.2% (5)	68.8% (11)	54.5% (1.64)	12.1% (0.27)	0

Note: Due to the structure of the specifications for Claims 2 - 4, analyses for these claims were excluded
[1]Number of targets with at least one standard with 50% reviewer agreement
[2]Standards that matched the intended mapping with greater than or equal to 50% reviewer agreement
[3]Consistent was defined as the grade-level standard DOK levels falling entirely within the range of the intended target DOK levels

Table Read: For grade 3, there were 11 targets in Claim 1. Of those targets, 11 were included in this analysis because they all had at least one standard with 50% reviewer agreement. Across all Claim 1 targets, an average of 95.5% of the grade-level standards that mapped to the intended target had at least 50% reviewer agreement (an average 3.1 standards per target). DOK Consistency in the next five columns is analyzed using the standards that mapped to the intended target with 50% agreement. 27.3% of the 11 targets included in the analysis had DOK consistency with all of the grade-level standards included in the analysis. 72.7% of the targets had DOK inconsistency. Of the 8 inconsistent targets, an average of 70.8% of the grade-level standards per target had a maximum grade-level standard DOK higher than that of the maximum intended target DOK. Zero targets were excluded from the DOK consistency analysis due to having no grade-level standards with at least 50% reviewer agreement.

Table 5.A.37. Math.DC-1b. Percentage of Mathematics Targets with DOKs Consistent with Intended Grade-Level Standards that Matched Intended Mapping for All Targets – All CCSS At Least One

Grade	Claim	Descriptives			DOK Consistency				
					Consistent	Inconsistent			
		Total number of targets in claim	Number of targets included in analysis[1]	Avg % of grade-level standards per target with >= 50% reviewer agreement[2]	% of Targets that Have All Mapped CCSS Consistent[3]	% of Targets With All Mapped CCSS Inconsistent	Avg % of CCSS per Inconsistent Target who's Max DOK Consensus > Specs	Avg % of CCSS per Target who's Min DOK Consensus < Specs	Number of Targets With < 50% agreement
		N	n	% (n)	% (n)	% (n)	% (n)	% (n)	n
All Targets									
3	1	11	11	95.5% (3.09)	100.0% (11)	0.0% (0)			0
4	1	12	12	100.0% (3.50)	100.0% (12)	0.0% (0)			0
5	1	11	11	89.1% (2.73)	90.9% (10)	9.1% (1)	100.0% (2.00)	0.0% (0.00)	0
6	1	10	10	82.7% (3.10)	80.0% (8)	20.0% (2)	75.0% (2.00)	37.5% (1.50)	0
7	1	9	9	97.2% (3.56)	22.2% (2)	77.8% (7)	56.2% (2.00)	3.6% (0.14)	0
8	1	10	10	94.3% (3.50)	90.0% (9)	10.0% (1)	66.7% (2.00)	0.0% (0.00)	0
11	1	16	16	98.4% (3.25)	100.0% (16)	0.0% (0)			0

Note: Due to the structure of the specifications for Claims 2 - 4, analyses for these claims were excluded
[1]Number of targets with at least one standard with 50% reviewer agreement
[2]Standards that matched the intended mapping with greater than or equal to 50% reviewer agreement
[3]Consistent was defined as at least one of the grade-level standard DOK levels matched at least one DOK level of the intended target

Table Read: Table 5.A.37 columns can be interpreted the same as in Table 5.A.36. The difference is that the way in which DOK consistency was defined: Here it was defined as each grade-level standard had to have at least one DOK level that matched that of the intended target. For grade 3, all of the grade-level standards for each target had at least one DOK level that matched that of the intended target.

Connection B: Alignment of Evidence Statements to Content Specifications

Analyses were conducted separately by content area to examine the alignment between the evidence statements and the Content Specifications. Additionally, for ELA/literacy, the evidence statements were analyzed separately for the CAT targets and the performance task targets (PT)[33]. Connection B analyses focused on content representation and DOK consistency.

ELA/Literacy: CAT Evidence Statements

Content Representation

Analyses were conducted to examine the representation of content between the ELA/literacy CAT evidence statements and the Content Specifications, and to address the following questions:

- B.CR-1. Do the evidence statements collectively reflect the content and skills required by the target?
- B.CR-2. Do the individual evidence statements reflect the content and skills required by the intended targets?

The main reviewer tasks for examining the content representation of the evidence statements and the targets involved verifying the target that each evidence statement represented (as indicated in the Content Specifications) and providing a holistic target rating that indicated how well the collective set of evidence statements represented the target.[34] Reviewers also provided individual evidence statement alignment ratings to indicate how well the content and knowledge in a single evidence statement measured the content and knowledge required in the target. Together, these two analyses provide an estimation of whether the evidence statements, either collectively or independently, were thought by reviewers to measure the content and knowledge required in the target.

Findings

Findings related to each content representation question are presented below. When relevant, general descriptive statistics provide overall results for each grade across claims.

B.CR-1. Do the evidence statements collectively reflect the content and skills required by the target?

Once reviewers verified that the targets identified for each evidence statement were appropriately matched, they provided a holistic rating to indicate how well those evidence statements collectively represented the content and knowledge required in the target. The scale ranged from 0 to 2 ('0'= not aligned, '1' = partially aligned, '2' = fully aligned).

[33] Evidence statements exist for ELA/literacy Claim 4 PT targets.
[34] Smarter Balanced did not intend for each evidence statement to measure all the content and knowledge required in a given target, but rather they intended the collective set of evidence statements to represent the target well.

As shown in Table 5.B.1, across grades and claims, reviewers rated the majority of targets as being fully aligned to their collective set of evidence statements (mean percentage of 75.6 – 100%). No clear patterns of alignment emerged across grades and claims. The reviewers generally identified more Claim 2 (Writing) and Claim 3 (Speaking & Listening) targets as being fully represented by the evidence statements; however, across grades, all of the targets were at least partially represented by their collective set of evidence statements.

Table 5.B.1. B.CR-1: Mean Percentage of ELA/Literacy CAT Targets at Each Holistic Rating (Collectively Reflected by the Evidence Statements), by Grade and Claim

			Holistic Target Rating		
Grade	Claim	Total Number of Targets	Fully aligned % (n)	Partially aligned % (n)	Not aligned % (n)
3	1	14	78.2% (10.8)	21.3% (2.9)	0.5% (0.1)
	2	8	97.4% (7.6)	2.6% (0.2)	0.0% (0.0)
	3	1	100.0% (1.0)	0.0% (0.0)	0.0% (0.0)
	4	3	82.2% (2.5)	17.8% (0.5)	0.0% (0.0)
4	1	14	85.6% (11.9)	14.4% (2.0)	0.0% (0.0)
	2	8	93.5% (7.5)	6.5% (0.5)	0.0% (0.0)
	3	1	93.3% (0.9)	6.7% (0.1)	0.0% (0.0)
	4	3	75.6% (2.2)	24.4% (0.7)	0.0% (0.0)
5	1	14	84.7% (11.7)	15.3% (2.1)	0.0% (0.0)
	2	8	91.3% (7.2)	8.7% (0.7)	0.0% (0.0)
	3	1	93.3% (0.9)	6.7% (0.1)	0.0% (0.0)
	4	3	88.9% (2.3)	11.1% (0.3)	0.0% (0.0)
6	1	14	86.1% (11.4)	13.9% (1.9)	0.0% (0.0)
	2	8	90.1% (7.1)	9.9% (0.8)	0.0% (0.0)
	3	1	84.6% (0.8)	15.4% (0.2)	0.0% (0.0)
	4	3	90.5% (2.4)	9.5% (0.3)	0.0% (0.0)
7	1	14	86.2% (12.1)	13.8% (1.9)	0.0% (0.0)
	2	8	89.3% (7.0)	10.7% (0.9)	0.0% (0.0)
	3	1	78.6% (0.8)	21.4% (0.2)	0.0% (0.0)
	4	3	92.9% (2.8)	7.1% (0.2)	0.0% (0.0)
8	1	14	86.7% (12.1)	13.3% (1.9)	0.0% (0.0)
	2	8	96.3% (7.4)	3.7% (0.3)	0.0% (0.0)
	3	1	78.6% (0.8)	21.4% (0.2)	0.0% (0.0)
	4	3	97.6% (2.8)	2.4% (0.1)	0.0% (0.0)

Table 5.B.1. (Continued)

			Holistic Target Rating		
Grade	Claim	Total Number of Targets	Fully aligned % (n)	Partially aligned % (n)	Not aligned % (n)
11	1	14	86.2% (12.0)	13.8% (1.9)	0.0% (0.0)
	2	8	80.5% (6.4)	19.5% (1.6)	0.0% (0.0)
	3	1	96.0% (1.0)	4.0% (0.0)	0.0% (0.0)
	4	3	84.7% (2.5)	15.3% (0.4)	0.0% (0.0)

Table Read: For grade 3, there were 14 ELA/literacy CAT targets for Claim 1. Based on the collective set of evidence statements associated with each target, reviewers rated an average of 78.2% of the targets (10.8 targets) as being fully-aligned to their collective set of evidence statements, an average of 21.3% of the targets (2.9 targets) as partially aligned, and an average of 0.5% (.1 target) as not aligned.

B.CR-2. Do the individual evidence statements reflect the content and skills required by the intended targets?

As noted earlier, in addition to providing a holistic target rating that indicated how well the target was represented by the collective set of evidence statements, reviewers also provided alignment ratings for each individual evidence statements. Although there was no expectation that an individual evidence statement would be fully-aligned to a target, this analysis provides information on whether individual evidence statements might not be interpreted as they were intended (i.e., many 'not aligned' ratings) or whether individual evidence statements might be redundant (i.e., many 'fully-aligned' ratings). Thus, the expectation here was that the majority of evidence statements would be rated as 'partially-aligned' to the targets to which they were mapped. As seen in Table 5.B.2, this outcome was supported by the data. Reviewers rated the majority of evidence statements as partially aligned to their targets (60.0% – 100%), indicating that an individual evidence statement most often reflected only some of the content and knowledge required in the target to which it was aligned. Reviewers' ratings for evidence statements being fully aligned to their target were typically much less, ranging from 0.0% – 40.0%. Some 'fully aligned' ratings were expected as some targets have only one or two evidence statements aligned to it.

Table 5.B.2. B.CR-2: Mean Percentage of ELA/Literacy CAT Evidence Statements Aligned to Targets, by Grade and Claim

			Individual Evidence Statement Ratings		
Grade	Claim	Total Number of Evidence Statements	Fully aligned % (n)	Partially aligned % (n)	Not aligned % (n)
3	1	30	23.9% (7.0)	74.5% (22.3)	1.6% (0.5)
	2	34	14.1% (4.8)	85.9% (29.1)	0.0% (0.0)
	3	3	11.1% (0.3)	88.9% (2.7)	0.0% (0.0)
	4	5	24.0% (1.2)	74.7% (3.7)	1.3% (0.1)
4	1	34	13.8% (4.7)	85.5% (28.9)	0.8% (0.3)
	2	30	5.0% (1.5)	94.7% (27.1)	0.2% (0.1)
	3	3	8.9% (0.3)	91.1% (2.7)	0.0% (0.0)
	4	6	15.6% (0.9)	84.4% (5.1)	0.0% (0.0)
5	1	34	17.8% (6.1)	81.2% (27.6)	1.0% (0.3)
	2	31	7.5% (2.3)	92.5% (28.3)	0.0% (0.0)
	3	3	15.6% (0.5)	84.4% (2.5)	0.0% (0.0)
	4	4	40.0% (1.6)	60.0% (2.4)	0.0% (0.0)
6	1	34	12.9% (4.4)	87.1% (29.6)	0.0% (0.0)
	2	34	3.8% (1.3)	96.2% (32.6)	0.0% (0.0)
	3	4	18.3% (0.7)	81.7% (3.3)	0.0% (0.0)
	4	5	32.0% (1.6)	68.0% (3.4)	0.0% (0.0)
7	1	34	15.5% (5.3)	84.5% (28.7)	0.0% (0.0)
	2	32	7.4% (2.4)	92.6% (29.6)	0.0% (0.0)
	3	4	0.0% (0.0)	100.0% (4.0)	0.0% (0.0)
	4	5	27.1% (1.4)	72.9% (3.6)	0.0% (0.0)
8	1	34	16.2% (5.5)	83.8% (28.5)	0.0% (0.0)
	2	47	4.5% (2.1)	95.5% (43.6)	0.0% (0.0)
	3	4	0.0% (0.0)	100.0% (4.0)	0.0% (0.0)
	4	5	22.9% (1.1)	77.1% (3.9)	0.0% (0.0)
11	1	34	18.0% (6.0)	81.9% (27.6)	0.1% (0.0)
	2	31	7.0% (2.2)	93.0% (28.6)	0.0% (0.0)
	3	5	16.0% (0.8)	83.2% (4.1)	0.8% (0.0)
	4	5	25.6% (1.3)	74.4% (3.7)	0.0% (0.0)

Table Read: For grade 3, there were 30 ELA/literacy CAT evidence statements for Claim 1. Reviewers rated an average of 23.9% of the evidence statements (average of 7 evidence statements) as being fully aligned to its target, an average of 74.5% of the evidence statements (average of 22.3 evidence statements) as being partially aligned to its target, and an average of 1.6% of the evidence statements (average of .5 evidence statements) as not being aligned to its target.

DOK Consistency

Analyses were conducted to examine the consistency of DOK levels between the ELA/literacy CAT evidence statements and the Content Specifications, and to address the following question:

- B.DC-1. Do reviewers' evidence statement DOK ratings align with the DOK levels specified for the targets to which they are mapped (as indicated in the Content Specifications)?

The DOK consistency analysis examined the degree to which the cognitive demand required by each evidence statement aligned to a target fell within the range of the cognitive demand required by the intended target. As such, we defined DOK consistency as the entire range of cognitive demand required by an evidence statement that was within the range of cognitive demand required by the intended target.

Pairwise Agreement

The overall pairwise agreement between reviewers in identifying DOK ratings for each target (across all grades, claims, targets, and reviewers) was 51.8% (see Table 5.B.3).

Table 5.B.3. Pairwise Percent Agreement between Reviewers' ELA/Literacy CAT Evidence Statement DOK Ratings

Grade	Claim	Pairwise Agreement	# of ES
3	1	59.0%	30
	2	47.0%	34
	3	40.8%	3
	4	46.3%	5
4	1	57.9%	34
	2	52.3%	30
	3	44.9%	3
	4	46.3%	6
5	1	58.5%	34
	2	53.5%	31
	3	50.3%	3
	4	42.8%	4
6	1	59.5%	34
	2	53.4%	34
	3	53.5%	4
	4	54.5%	5
7	1	61.3%	34
	2	49.3%	32
	3	50.6%	4
	4	47.7%	5

Table 5.B.3. (Continued)

Grade	Claim	Pairwise Agreement	# of ES
8	1	59.3%	34
	2	48.9%	46
	3	50.6%	4
	4	48.6%	5
11	1	55.3%	34
	2	53.1%	31
	3	52.2%	5
	4	51.4%	5

Each evidence statement had a total of 12-15 reviewer ratings across Workshops 3-5.

Findings

Findings related to DOK Consistency are presented below. When relevant, general descriptive statistics provide overall results for each grade across claims.

B.DC-1. Do reviewers' evidence statement DOK ratings align with the DOK levels specified for the targets to which they are mapped (as indicated in the Content Specifications)?

To get a sense for whether reviewers thought the evidence statements required the same level of cognitive demand than what was intended, we examined the DOK distribution of the evidence statements in two ways: using the maximum DOK rating and using each DOK rating independently (since evidence statements could have more than one DOK level). As seen in Table 5.B.4. Generally, the reviewers indicated that the majority of evidence statements required maximum DOK levels of 2 and 3.

Table 5.B.4. Reviewers' Mean Percentage of ELA/Literacy CAT Evidence Statements at Each DOK Level (Max)

Grade	Claim	DOK 1 % (n)	DOK 2 % (n)	DOK 3 % (n)	DOK 4 % (n)
3	1	8.9% (2.7)	39.0% (11.7)	39.2% (11.7)	12.9% (3.9)
	2	30.4% (10.2)	31.8% (10.8)	33.6% (11.4)	4.2% (1.4)
	3	6.7% (0.2)	42.2% (1.3)	44.4% (1.3)	6.7% (0.2)
	4	4.0% (0.2)	38.7% (1.9)	48.0% (2.4)	9.3% (0.5)
4	1	6.7% (2.3)	37.9% (12.9)	43.2% (14.7)	12.2% (4.1)
	2	26.9% (7.9)	31.2% (9.3)	37.2% (11.1)	4.7% (1.4)
	3	11.1% (0.3)	51.1% (1.5)	35.6% (1.1)	2.2% (0.1)
	4	2.2% (0.1)	35.6% (2.1)	48.9% (2.9)	13.3% (0.8)
5	1	11.0% (3.7)	31.4% (10.7)	43.7% (14.9)	13.9% (4.7)
	2	31.1% (9.6)	33.9% (10.5)	31.3% (9.7)	3.7% (1.1)
	3	22.2% (0.7)	31.1% (0.9)	44.4% (1.3)	2.2% (0.1)
	4	11.7% (0.5)	23.3% (0.9)	48.3% (1.9)	16.7% (0.7)

Table 5.B.4. (Continued)

Grade	Claim	DOK 1 % (n)	DOK 2 % (n)	DOK 3 % (n)	DOK 4 % (n)
6	1	9.4% (3.2)	30.8% (10.5)	43.6% (14.8)	16.1% (5.5)
	2	32.1% (10.9)	37.6% (12.8)	23.2% (7.9)	7.1% (2.4)
	3	0.0% (0.0)	13.3% (0.5)	78.3% (3.1)	8.3% (0.3)
	4	1.3% (0.1)	4.0% (0.2)	17.3% (0.9)	77.3% (3.9)
7	1	8.4% (2.9)	33.8% (11.5)	47.7% (16.2)	10.1% (3.4)
	2	25.3% (8.1)	42.1% (13.4)	26.2% (8.4)	6.5% (2.1)
	3	0.0% (0.0)	19.6% (0.8)	69.6% (2.8)	10.7% (0.4)
	4	0.0% (0.0)	2.9% (0.1)	37.1% (1.9)	60.0% (3.0)
8	1	10.1% (3.4)	33.4% (11.4)	46.6% (15.9)	9.9% (3.4)
	2	33.0% (15.1)	46.5% (21.2)	17.3% (7.9)	3.1% (1.4)
	3	0.0% (0.0)	17.3% (0.7)	73.1% (2.9)	9.6% (0.4)
	4	0.0% (0.0)	5.7% (0.3)	28.6% (1.4)	65.7% (3.3)
11	1	4.4% (1.5)	31.0% (10.5)	45.5% (15.4)	19.1% (6.5)
	2	16.1% (5.0)	44.6% (13.8)	29.3% (9.1)	9.9% (3.1)
	3	0.0% (0.0)	13.6% (0.6)	61.6% (3.0)	24.8% (1.2)
	4	0.8% (0.0)	2.4% (0.1)	30.4% (1.5)	66.4% (3.3)

Note: The percentages across DOK levels are mutually exclusive.

As shown in Table 5.B.5, the DOK distribution of the evidence statements broadens when examining each identified level, indicating that reviewers thought evidence statements required varying levels of cognitive demand, typically ranging from levels 1 to 3.

Table 5.B.5. Reviewers' Mean Percentage of ELA/Literacy CAT Evidence Statements at Each DOK Level (Independent)

Grade	Claim	DOK 1 % (n)	DOK 2 % (n)	DOK 3 % (n)	DOK 4 % (n)
3	1	37.7% (29.9)	69.3% (29.9)	48.5% (29.9)	12.9% (29.9)
	2	53.2% (33.8)	42.8% (33.8)	35.9% (33.8)	4.2% (33.8)
	3	40.0% (3.0)	71.1% (3.0)	44.4% (3.0)	6.7% (3.0)
	4	24.0% (5.0)	70.7% (5.0)	53.3% (5.0)	9.3% (5.0)
4	1	31.6% (33.9)	69.6% (33.9)	52.3% (33.9)	12.2% (33.9)
	2	54.8% (29.7)	50.4% (29.7)	41.7% (29.7)	4.7% (29.7)
	3	48.9% (3.0)	73.3% (3.0)	35.6% (3.0)	2.2% (3.0)
	4	23.3% (6.0)	68.9% (6.0)	60.0% (6.0)	13.3% (6.0)
5	1	19.6% (34.0)	45.5% (34.0)	53.7% (34.0)	13.9% (34.0)
	2	52.4% (30.8)	43.2% (30.9)	32.8% (30.9)	3.7% (30.8)
	3	40.0% (3.0)	57.8% (3.0)	44.4% (3.0)	2.2% (3.0)
	4	16.7% (4.0)	36.7% (4.0)	61.7% (4.0)	16.7% (4.0)
6	1	18.3% (33.9)	48.7% (33.9)	56.0% (33.9)	16.1% (33.9)
	2	52.5% (33.9)	45.7% (33.9)	26.5% (33.9)	7.1% (33.9)
	3	16.7% (4.0)	46.7% (4.0)	85.0% (4.0)	8.9% (3.7)
	4	8.0% (5.0)	18.7% (5.0)	73.3% (5.0)	77.3% (5.0)

Table 5.B.5. (Continued)

Grade	Claim	DOK 1 % (n)	DOK 2 % (n)	DOK 3 % (n)	DOK 4 % (n)
7	1	12.0% (34.0)	52.9% (34.0)	53.4% (34.0)	10.1% (34.0)
	2	38.0% (31.9)	50.8% (31.9)	28.9% (31.9)	6.5% (31.9)
	3	10.7% (4.0)	41.1% (4.0)	76.8% (4.0)	10.7% (4.0)
	4	0.0% (4.9)	18.6% (5.0)	62.9% (5.0)	60.0% (5.0)
8	1	13.4% (34.0)	48.3% (34.0)	52.9% (34.0)	9.9% (34.0)
	2	41.3% (45.7)	51.8% (45.7)	18.6% (45.7)	3.1% (45.6)
	3	7.7% (4.0)	44.2% (4.0)	78.8% (4.0)	9.6% (4.0)
	4	0.0% (5.0)	12.9% (5.0)	54.3% (5.0)	65.7% (5.0)
11	1	22.4% (33.9)	52.1% (33.9)	59.6% (33.9)	19.1% (33.9)
	2	52.5% (31.0)	60.0% (31.0)	38.8% (30.9)	9.9% (30.9)
	3	25.4% (4.8)	52.6% (4.9)	77.2% (4.9)	24.8% (4.9)
	4	8.8% (5.0)	21.6% (5.0)	61.6% (5.0)	66.4% (5.0)

Note: The percentages across DOK levels are not mutually exclusive since an evidence statement could have multiple DOK levels.

As shown in Table 5.B.6, reviewers across grades indicated that the DOK level for the evidence statements in Claim 4 (Research & Inquiry) were not within the range of their intended target (2.4% – 25.6%). Generally, evidence statements from Claims 1 (Comprehend Literary & Informational Text) and 3 (Speaking & Listening) had high percentages of evidence statements as DOK consistent with their intended targets. When the DOK consistency criterion was relaxed to only require at least one evidence statement's DOK level (since evidence statements could have multiple DOK levels), the DOK consistency of the evidence statements with their intended targets increased across grades and claims. This was true except for Claim 4 (Research & Inquiry), which remained relatively low. Upon further investigation, for the evidence statements that had a range of DOKs that did not fall within the range of their intended targets, reviewers indicated that the cognitive demand required by the evidence statements was higher than that required by the target to which it was aligned. This pattern was true for Claims 2 – 4; Claim 1 DOK inconsistency was likely due to reviewers indicating that the cognitive demand required by the evidence statements was lower than what was required by their intended targets.

Table 5.B.6. B.DC-1: Mean Percentage of ELA/Literacy CAT Evidence Statements with DOK Levels Consistent with the Intended Targets

		Consistent		Inconsistent	
Grade	Claim	ES Within Range of Intended Target % (n)	ES DOK Match at Least One Intended Target DOK % (n)	ES max DOK > Max DOK of Intended Target % (n)	ES min DOK < Min DOK of Intended Target % (n)
3	1	71.0% (21.3)	94.7% (28.3)	6.4% (1.9)	23.6% (7.1)
	2	48.1% (16.2)	75.4% (25.5)	44.2% (15.1)	9.4% (3.2)
	3	93.3% (2.8)	95.6% (2.9)	6.7% (0.2)	0.0% (0.0)
	4	25.3% (1.3)	70.7% (3.5)	57.3% (2.9)	24.0% (1.2)
4	1	76.0% (25.8)	96.3% (32.7)	6.9% (2.3)	17.9% (6.1)
	2	45.6% (13.5)	82.7% (24.5)	45.2% (13.5)	14.0% (4.1)
	3	97.8% (2.9)	97.8% (2.9)	2.2% (0.1)	0.0% (0.0)
	4	25.6% (1.5)	68.9% (4.1)	62.2% (3.7)	23.3% (1.4)
5	1	83.3% (28.3)	94.9% (32.3)	10.0% (3.4)	7.8% (2.7)
	2	53.3% (16.5)	79.7% (24.6)	40.8% (12.6)	6.7% (2.1)
	3	97.8% (2.9)	97.8% (2.9)	2.2% (0.1)	0.0% (0.0)
	4	20.0% (0.8)	36.7% (1.5)	65.0% (2.6)	16.7% (0.7)
6	1	75.1% (25.5)	93.5% (31.7)	7.1% (2.4)	18.1% (6.1)
	2	57.2% (19.4)	82.5% (28.0)	40.0% (13.6)	3.7% (1.3)
	3	91.7% (3.7)	98.3% (3.9)	8.3% (0.3)	0.0% (0.0)
	4	4.0% (0.2)	18.7% (0.9)	94.7% (4.7)	8.0% (0.4)
7	1	77.5% (26.4)	92.9% (31.6)	3.3% (1.1)	19.1% (6.5)
	2	55.2% (17.6)	74.8% (23.9)	41.5% (13.3)	3.2% (1.0)
	3	89.3% (3.6)	96.4% (3.9)	10.7% (0.4)	0.0% (0.0)
	4	2.9% (0.1)	18.6% (0.9)	97.1% (4.9)	0.0% (0.0)
8	1	80.7% (27.4)	92.2% (31.4)	1.7% (0.6)	17.6% (6.0)
	2	54.0% (24.7)	64.6% (29.6)	43.2% (19.7)	3.2% (1.4)
	3	90.4% (3.6)	96.2% (3.8)	9.6% (0.4)	0.0% (0.0)
	4	5.7% (0.3)	12.9% (0.6)	94.3% (4.7)	0.0% (0.0)
11	1	71.8% (24.4)	92.9% (31.5)	5.9% (2.0)	23.3% (7.9)
	2	67.5% (20.9)	87.6% (27.2)	26.7% (8.3)	9.9% (3.1)
	3	75.2% (3.7)	91.6% (4.5)	24.8% (1.2)	0.0% (0.0)
	4	2.4% (0.1)	21.6% (1.1)	96.8% (4.8)	8.8% (0.4)

Table Read: For grade 3, Claim 1, reviewers rated an average of 71.0 % of the ELA/literacy CAT evidence statements (21.3 evidence statements) as having their range of DOK levels fall within the range for that of their intended target. Reviewers rated an average of 94.7% of the evidence statements (28.3 evidence statements) as having at least one DOK level the same as that of their intended target (or within the range of the intended target). Reviewers rated an average of 6% of the evidence statements (1.9 evidence statements) as having a higher DOK level than that of their intended target. Reviewers rated an average of 24% of the evidence statements (7.1 evidence statements) as having the minimum evidence statement DOK level lower than that of minimum DOK level of their intended target.

ELA/Literacy: Performance Task Evidence Statements

Content Representation

The main reviewer tasks for examining the content representation of the ELA/literacy performance task evidence statements and the targets were the same as the tasks for the ELA/literacy CAT evidence statements. The only difference was that reviewers were asked to make ratings for the performance task evidence statements. The same questions we examined for the CAT evidence statements were examined for the PT evidence statements.

Findings

Findings related to each content representation question are presented below. When relevant, general descriptive statistics provide overall results for each grade across claims.

B.CR-1. Do the evidence statements collectively reflect the content and skills required by the target?

As shown in Table 5.B.7, reviewers across grades and claims rated the majority of ELA/literacy PT targets as being fully aligned to their collective set of evidence statements (mean percentage of evidence statements ranged from 85.4% – 100%).

Table 5.B.7. B.CR-1: Mean Percentage of ELA/Literacy PT Targets at Each Holistic Rating (Collectively Reflected by the Evidence Statements), by Grade and Claim

			Holistic Target Rating		
Grade	Claim	Total Number of Targets	Fully aligned % (n)	Partially aligned % (n)	Not aligned % (n)
3	4	3	90.0% (2.7)	10.0% (0.3)	0.0% (0.0)
4	4	3	90.0% (2.7)	10.0% (0.3)	0.0% (0.0)
5	4	3	86.7% (2.6)	13.3% (0.4)	0.0% (0.0)
6	4	3	90.0% (2.6)	10.0% (0.3)	0.0% (0.0)
7	4	3	100.0% (3.0)	0.0% (0.0)	0.0% (0.0)
8	4	3	100.0% (3.0)	0.0% (0.0)	0.0% (0.0)
11	4	3	85.4% (2.6)	14.6% (0.4)	0.0% (0.0)

Table Read: For grade 3, there were 3ELA/literacy PT targets for Claim 4. Based on the collective set of evidence statements associated with each target, reviewers rated an average of 90.0% of the targets (2.7 targets) as being fully aligned to their collective set of evidence statements, an average of 10.0% of the targets (0.3 targets) as partially aligned, and none of the targets as not aligned.

B.CR-2. Do the individual evidence statements reflect the content and skills required by the intended targets?

As was seen with the alignment of the individual CAT evidence statements, the expected pattern of higher percentages of partially aligned evidence statements also existed for ELA/literacy PT evidence statements.

Table 5.B.8. B.CR-2: Mean Percentage of ELA/Literacy PT Evidence Statements Aligned to Targets, by Grade and Claim

		Individual Evidence Statement Ratings			
Grade	Claim	Total Number of Evidence Statements	Fully-aligned % (n)	Partially-aligned % (n)	Not-aligned % (n)
3	4	5	28.0% (1.4)	72.0% (3.6)	0.0% (0.0)
4	4	6	20.0% (1.2)	80.0% (4.8)	0.0% (0.0)
5	4	4	40.0% (1.6)	60.0% (2.4)	0.0% (0.0)
6	4	5	18.0% (0.9)	82.0% (4.1)	0.0% (0.0)
7	4	5	42.0% (2.1)	58.0% (2.9)	0.0% (0.0)
8	4	5	42.0% (2.1)	58.0% (2.9)	0.0% (0.0)
11	4	6	22.9% (1.4)	77.1% (4.6)	0.0% (0.0)

Table Read: For grade 3, there were 5 ELA/literacy PT evidence statements for Claim 4 (Research and Inquiry). Reviewers rated an average of 28.0% of the evidence statements (2.7 statements) as being fully aligned, an average of 72.0% of the evidence statements (3.6 statements) as partially aligned, and there were no evidence statements that were not aligned to their intended target.

DOK Consistency

B.DC-1. Do reviewers' evidence statement DOK ratings align with the DOK levels specified for the targets to which they are mapped (as indicated in the Content Specifications)?

Pairwise Agreement

The overall pairwise agreement between reviewers in identifying DOK ratings for each target (across all grades, claims, targets, and reviewers) was 46.0% (see Table 5.B.9).

Table 5.B.9. Pairwise Percent Agreement between Reviewers' ELA/Literacy PT Evidence Statement DOK Ratings

Grade	Claim	Pairwise Agreement	# of ES
3	4	35.6%	5
4	4	34.2%	6
5	4	43.6%	4
6	4	55.7%	5
7	4	51.3%	5
8	4	51.3%	5
11	4	50.4%	6

Findings

Findings related to the DOK consistency of the ELA/literacy PT evidence statements with the PT Claim 4 (Research & Inquiry) targets are presented below.

To get a sense for whether reviewers thought the evidence statements required the same level of cognitive demand than what was intended, we examined the DOK distribution of the evidence statements in two ways: using the maximum DOK rating and using each DOK rating independently (since evidence statements could have more than one DOK level). As seen in Table 5.B.10. Generally, the reviewers indicated that the majority of evidence statements required maximum DOK levels of 3 and 4, which are higher than what was required for the CAT evidence statements.

Table 5.B.10. Reviewers' Mean Percentage of ELA/Literacy PT Evidence Statements at Each DOK Level (Max)

Grade	Claim	DOK 1 % (n)	DOK 2 % (n)	DOK 3 % (n)	DOK 4 % (n)
3	4	4.0% (2.0)	12.0% (6.0)	50.0% (25.0)	34.0% (17.0)
4	4	1.7% (1.0)	15.0% (9.0)	45.0% (27.0)	38.3% (23.0)
5	4	0.0% (0.0)	15.0% (6.0)	37.5% (15.0)	47.5% (19.0)
6	4	2.2% (1.0)	2.2% (1.0)	22.2% (10.0)	73.3% (33.0)
7	4	0.0% (0.0)	4.0% (2.0)	40.0% (20.0)	56.0% (28.0)
8	4	0.0% (0.0)	4.0% (2.0)	34.0% (17.0)	62.0% (31.0)
11	4	0.0% (0.0)	4.2% (4.0)	31.3% (30.0)	64.6% (62.0)

Note: The percentages across DOK levels are mutually exclusive.

As shown in Table 5.B.11, the DOK distribution of the evidence statements broadens when examining each identified level, indicating that reviewers thought evidence statements required varying levels of cognitive demand, typically ranging from levels 2 to 4.

Table 5.B.11. Reviewers' Mean Percentage of ELA/Literacy PT Evidence Statements at Each DOK Level (Independent)

Grade	Claim	DOK 1 % (n)	DOK 2 % (n)	DOK 3 % (n)	DOK 4 % (n)
3	4	16.0% (8.0)	22.0% (11.0)	58.0% (29.0)	34.0% (17.0)
4	4	13.3% (8.0)	26.7% (16.0)	53.3% (32.0)	38.3% (23.0)
5	4	2.5% (1.0)	17.5% (7.0)	72.5% (29.0)	47.5% (19.0)
6	4	2.2% (1.0)	13.3% (6.0)	64.4% (29.0)	73.3% (33.0)
7	4	0.0% (0.0)	12.0% (6.0)	66.0% (33.0)	56.0% (28.0)
8	4	0.0% (0.0)	12.0% (6.0)	64.0% (32.0)	62.0% (31.0)
11	4	3.1% (3.0)	12.5% (12.0)	46.9% (45.0)	64.6% (62.0)

Note: The percentages across DOK levels are not mutually exclusive since an evidence statement could have multiple DOK levels.

As shown in Table 5.B.12, the DOK consistency of the ELA/literacy PT evidence statements with the PT targets was higher for the upper grades than it was for the lower grades (grades 3-5 range from 33.3% – 40.0%; grades 6-8 range from 82.2% - 83.3%). Similar to the pattern seen with the ELA/literacy CAT evidence statements, relaxing the consistency criterion to only require at least one of the evidence statement's DOK level (since evidence statements could have multiple DOK levels), the DOK consistency of the evidence statements with their intended targets increased across grades. Divergent to the pattern seen with the CAT evidence statements was that, of those evidence statements whose cognitive demand did not fall entirely within the range of the cognitive demand required by their intended targets, reviewers mostly indicated that cognitive demand of the evidence statements was *lower* than what was required by their intended targets.

Table 5.B.12. B.DC-1: Mean Percentage of ELA/Literacy PT Evidence Statements with DOK Levels Consistent with the Intended Targets

		Consistent		Inconsistent	
Grade	Claim	ES Within Range of Intended Target % (n)	ES DOK Match at Least One Intended Target DOK % (n)	ES max DOK > Max DOK of Intended Target % (n)	ES min DOK < Min DOK of Intended Target % (n)
3	4	40% (2.0)	56% (2.8)	16% (0.8)	46% (2.3)
4	4	33% (2.0)	48% (2.9)	25% (1.5)	45% (2.7)
5	4	35% (1.4)	73% (2.9)	33% (1.3)	33% (1.3)
6	4	82% (4.1)	96% (4.8)	2% (0.1)	16% (0.8)
7	4	86% (4.3)	96% (4.8)	2% (0.1)	12% (0.6)
8	4	84% (4.2)	96% (4.8)	4% (0.2)	12% (0.6)
11	4	83% (5.0)	94% (5.6)	4% (0.3)	13% (0.8)

Table Read: For grade 3, Claim 4, reviewers rated an average of 40.0% of the ELA/literacy PT evidence statements (2.0 statements) as having a DOK level within the DOK range of the intended target. Reviewers rated an average of 56.0% of the evidence statements (2.8 statements) as matching at least one of the DOK levels as that of the intended target. Reviewers rated an average of 16% of the evidence statements (0.8 statements) as having a DOK level greater than the maximum DOK level of the intended target. Reviewers rated an average of 46.0% of the evidence statements (2.3 statements) as having a DOK level less than the minimum DOK level of the intended target.

Mathematics: CAT Evidence Statements

Analyses were conducted separately by content area to examine the alignment between the mathematics CAT evidence statements and the Content Specifications. These analyses focused on content representation and DOK consistency.

Evidence statements only exist for Claim 1 mathematics targets. Each analysis was conducted by examining the overall targets, as well as disaggregating the targets by Claim 1 (Concepts and Procedures) emphasis. For any substantive differences that were found between the major and additional and supporting targets, findings are presented below. When substantive differences were not found, parallel tables for the emphasis breakout for each analysis are presented Appendix F.

Content Representation

Analyses were conducted to examine the representation of content between the mathematics CAT evidence statements and the Content Specifications, and to address the following questions:

- B.CR-1. Do the evidence statements collectively reflect the content and skills required by the target?
- B.CR-2. Do the individual evidence statements reflect the content and skills required by the intended targets?

The main reviewer tasks for examining the content representation of the mathematics CAT evidence statements and the targets involved reviewers verifying the target that each evidence statement represented (as indicated in the Content Specifications) and providing a holistic rating to indicate

how well the collective set of evidence statements represented their intended target.[35] Reviewers also provided individual evidence statement alignment ratings to indicate how well the content and knowledge in a single mathematics CAT evidence statement measured the content and knowledge required in its intended target. Together, these two analyses provide an estimation of whether the evidence statements, either collectively or independently, measured the content and knowledge required in the target.

Findings

Findings related to each content representation question are presented below. When relevant, general descriptive statistics provide overall results for each grade across claims.

B.CR-1. Do the evidence statements collectively reflect the content and skills required by the target?

Once reviewers verified that the targets identified for each mathematics evidence statement were appropriately matched, they provided a holistic rating to indicate how well those evidence statements collectively represented the content and knowledge required in the target. The scale ranged from 0 to 2 ('0'= not aligned, '1' = partially aligned, '2' = fully aligned).

As shown in Table 5.B.13, reviewers across grades and claims rated the majority of targets as being fully aligned to their collective set of mathematics CAT evidence statements (mean percentage of (79.5% - 95.4%). All targets across grades were at least partially represented by their collective set of evidence statements.

Table 5.B.13. B.CR-1: Mean Percentage of Mathematics Targets at Each Holistic Rating (Collectively Reflected by the Evidence Statements), by Grade and Claim

			Holistic Target Rating		
Grade	Claim	Total Number of Targets	Fully-aligned % (n)	Partially-aligned % (n)	Not-aligned % (n)
3	1	11	82.8% (9.0)	17.2% (1.9)	0.0% (0.0)
4	1	12	91.6% (10.9)	8.4% (1.0)	0.0% (0.0)
5	1	11	80.6% (8.9)	19.4% (2.1)	0.0% (0.0)
6	1	10	87.3% (8.3)	12.7% (1.2)	0.0% (0.0)
7	1	9	81.2% (7.3)	18.8% (1.7)	0.0% (0.0)
8	1	10	95.4% (9.5)	4.6% (0.5)	0.0% (0.0)
11	1	16	79.5% (12.6)	20.5% (2.7)	0.0% (0.0)

Table Read: For grade 3, there were 11 targets for mathematics Claim 1. Based on the collective set of evidence statements associated with each target, reviewers rated an average of 82.8% of the targets (9.0 targets) as being fully aligned to their collective set of evidence statements, an average of 17.2% of the targets (1.9 targets) as partially aligned, and 0% as not aligned.

[35] Smarter Balanced did not intend for each evidence statement to measure all the content and knowledge required in a given target, but rather they intended the collective set of evidence statements to represent the target well.

B.CR-2. Do the individual evidence statements reflect the content and skills required by the intended targets?

As noted earlier, in addition to providing a holistic target rating that indicated how well the target was represented by its collective set of mathematics CAT evidence statements, reviewers also provided alignment ratings for each individual mathematics evidence statement. Although there was no expectation that an individual evidence statement would be fully aligned to a target, this analysis provides information on whether individual mathematics CAT evidence statements might not be interpreted as intended (i.e., many 'not aligned' ratings) or whether individual evidence statements might be redundant (i.e., many 'fully aligned' ratings). Thus, the expectation here was that the majority of evidence statements would be rated as 'partially aligned' to the targets to which they were mapped. As seen in Table 5.B.14, this outcome was supported by the data. Reviewers rated the majority of evidence statements as partially aligned to their targets (71.4% – 98.0%), indicating that an individual evidence statement most often reflected only some of the content and knowledge required in the target to which it was aligned. Reviewers' ratings for evidence statements being fully aligned to their target were typically much less, ranging from 1.1% – 27.8%. Some 'fully aligned' ratings were expected as some targets only have one or two evidence statements aligned to it.

Table 5.B.14. B.CR-2: Mean Percentage of Mathematics CAT Evidence Statements Aligned to Targets, by Grade and Claim

			Individual Evidence Statement Ratings		
Grade	Claim	Total Number of Evidence Statements	Fully-aligned % (n)	Partially-aligned % (n)	Not-aligned % (n)
3	1	30	26.4% (7.9)	72.9% (21.4)	0.7% (0.2)
4	1	44	27.8% (12.2)	71.4% (31.3)	0.8% (0.4)
5	1	32	4.4% (1.4)	94.8% (30.3)	0.8% (0.3)
6	1	50	1.1% (0.5)	98.0% (49.0)	0.9% (0.5)
7	1	35	5.9% (2.1)	92.1% (32.2)	2.0% (0.7)
8	1	43	4.9% (2.1)	93.8% (39.3)	1.3% (0.5)
11	1	54	10.1% (5.2)	89.6% (47.6)	0.4% (0.2)

Table Read: For grade 3, there were 30 mathematics CAT evidence statements for Claim 1. Reviewers rated an average of 26.4% of the evidence statements (average of 7.9 evidence statements) as being fully aligned to its target, an average of 72.9% of the evidence statements (average of 21.4 evidence statements) as being partially aligned to its target, and an average of 0.7% of the evidence statements (average of .2 evidence statements) as not being aligned to its target.

DOK Consistency

Analyses were conducted to examine the consistency of DOK levels between the mathematics CAT evidence statements and the Content Specifications, and to address the following question:

- B.DC-1. Do reviewers' evidence statement DOK ratings align with the DOK levels specified for the targets to which they are mapped (as indicated in the Content Specifications)?

The DOK consistency analysis examined the degree to which the cognitive demand required by each mathematics CAT evidence statement that aligned to a target fell within the range of the cognitive demand required by the intended target. As such, we defined DOK consistency as the entire range of cognitive demand required by an evidence statement that was within the range of cognitive demand required by the intended target.

Pairwise Agreement

The overall pairwise agreement among reviewers in identifying DOK ratings for each target (across all grades, claims, targets, and reviewers) was 59.39% (see Table 5.B.15).

Table 5.B.15. Pairwise Percent Agreement among Reviewers' Mathematics Evidence Statement DOK Ratings

Grade	Claim	Pairwise Agreement	# of Evidence Statements
3	1	72.7%	30
4	1	69.5%	44
5	1	67.5%	32
6	1	58.1%	50
7	1	49.5%	35
8	1	44.4%	42
11	1	54.0%	54

Note: Each evidence statement had a total of 12-15 reviewer ratings across Workshops 3-5

Findings

Findings related to DOK consistency are presented below. When relevant, general descriptive statistics provide overall results for each grade across claims.

> ***B.DC-1. Do reviewers' evidence statement DOK ratings align with the DOK levels specified for the targets to which they are mapped (as indicated in the Content Specifications)?***

To get a sense for whether reviewers thought the evidence statements required the same level of cognitive demand than what was intended, we examined the DOK distribution of the evidence statements in two ways: using the maximum DOK rating and using each DOK rating independently (since evidence statements could have more than one DOK level). As seen in Table 5.B.16. Generally, the reviewers indicated that the majority of evidence statements required maximum DOK levels of 2, with very few evidence statements requiring higher levels of cognitive demand. Given that these are evidence statements for Claim Targets, this is not an unexpected finding.

Table 5.B.16. A.DD-1: Reviewers' Mean Percentage of Mathematics Evidence Statements at Each DOK Level (Max) by Grade and Claim Compared to Content Specifications

Grade	Claim	DOK 1 % (n)	DOK 2 % (n)	DOK 3 % (n)	DOK 4 % (n)
3	1	23.3% (7.0)	55.2% (16.6)	19.0% (5.7)	2.4% (0.7)
4	1	24.8% (10.9)	56.7% (24.9)	17.4% (7.6)	1.1% (0.5)
5	1	28.2% (9.0)	58.9% (18.8)	12.1% (3.9)	0.8% (0.3)
6	1	23.3% (11.7)	54.9% (27.5)	21.3% (10.7)	0.4% (0.2)
7	1	22.9% (8.0)	55.2% (19.3)	19.3% (6.8)	2.6% (0.9)
8	1	25.5% (10.7)	51.4% (21.5)	20.0% (8.4)	3.1% (1.3)
11	1	20.5% (11.1)	57.2% (30.9)	21.0% (11.3)	1.3% (0.7)

Note: The percentages across DOK levels are mutually exclusive

As shown in Table 5.B.17, the DOK distribution of the evidence statements broadens when examining each identified level, indicating that reviewers thought evidence statements required varying levels of cognitive demand, mostly levels 1 and 2.

Table 5.B.17. A.DD-2: Reviewers' Mean Percentage of Mathematics Evidence Statements at Each DOK Level (Independent)

Grade	Claim	DOK 1 % (n)	DOK 2 % (n)	DOK 3 % (n)	DOK 5 % (n)
3	1	91.2% (30.0)	76.4% (29.9)	21.4% (30.0)	2.4% (29.9)
4	1	88.1% (44.0)	74.0% (44.0)	18.5% (44.0)	1.1% (44.0)
5	1	76.8% (31.9)	70.7% (31.9)	12.9% (31.9)	0.8% (31.9)
6	1	62.4% (50.0)	74.6% (49.9)	21.8% (49.3)	0.4% (49.1)
7	1	38.5% (35.0)	61.8% (35.0)	20.4% (35.0)	2.6% (35.0)
8	1	38.4% (41.5)	59.4% (40.8)	21.1% (40.4)	3.1% (40.3)
11	1	55.4% (54.0)	71.6% (53.8)	21.9% (53.9)	1.3% (53.9)

Note: The percentages across DOK levels are not mutually exclusive since an evidence statement could have multiple DOK levels

As shown in Table 5.B.18, for all grades except grade 3, the majority of mathematics CAT evidence statements (71.0% – 79.9%) were rated as having DOK levels within the range of the intended targets. The average percentage of grade 3 Claim 1 (Concepts & Problems) evidence statements with DOK levels within the range of the intended target was 49.0%. Especially for grade 3, reviewers believed that the DOK for the evidence statement was higher than that for its intended target. When the DOK consistency criterion was relaxed to only require at least one evidence statement's DOK level (since evidence statements could have multiple DOK levels) to match that of the intended target, the DOK consistency of the evidence statements with their intended targets increased across grades. Upon further investigation, for the evidence statements whose range of DOK levels did not fall within the range of their intended targets, reviewers indicated that the cognitive demand required by the evidence statements was higher than that required by the intended target. This was particularly true for grade 3.

Table 5.B.18. B.DC-1: Mean Percentage of Mathematics AT Evidence Statements with DOK Levels Consistent with the Intended Targets

		Consistent		Inconsistent	
Grade	Claim	ES Within Range of Intended Target % (n)	ES DOK Match at Least One Intended Target DOK % (n)	ES max DOK > max DOK of Intended Target % (n)	ES min DOK < min DOK of Intended Target % (n)
3	1	49.0% (14.7)	96.4% (28.9)	45.0% (13.4)	11.0% (3.3)
4	1	79.9% (35.1)	99.5% (43.8)	16.0% (6.9)	4.0% (1.9)
5	1	74.7% (23.9)	95.0% (30.3)	17.0% (5.4)	11.0% (3.5)
6	1	73.7% (36.9)	95.9% (47.9)	22.0% (10.9)	5.0% (2.7)
7	1	71.0% (24.8)	81.3% (28.5)	22.0% (7.7)	7.0% (2.5)
8	1	76.9% (32.2)	83.3% (34.9)	23.0% (9.7)	0.0% (0.0)
11	1	74.0% (39.9)	90.5% (48.8)	21.0% (11.1)	6.0% (3.0)

Table Read: For grade 3, Claim 1, reviewers rated an average of 49.0 % of the mathematics CAT evidence statements (14.7 evidence statements) as having their range of DOK levels fall within the range for that of their intended target. Reviewers rated an average of 96.4% of the evidence statements (28.9 evidence statements) as having at least one DOK level the same as that of their intended target (or within the range of the intended target). Reviewers rated an average of 45% of the evidence statements (13.4 evidence statements) as having a higher DOK level than that of their intended target. Reviewers rated an average of 11% of the evidence statements (3.3 evidence statements) as having the minimum evidence statement DOK level lower than that of minimum DOK level of their intended target.

Connection C: Alignment of Test Blueprint to Content Specifications

ELA/Literacy

Analyses were conducted separately by content area to examine alignment between the test blueprint and the Content Specifications. This analysis focused on content representation.

After working with the CCSS and the Smarter Balanced Content Specifications, reviewers attending Workshop 1 provided holistic feedback on how representative the blueprints were to the Smarter Balanced Content Specifications and test design decisions. Examples of decisions include how reading genre is emphasized across grades and which targets are assessed or not assessed on the summative assessment.

Content Representation

Analysis was conducted to examine the representation of ELA/literacy content between the test blueprint and the Content Specifications, and to address the following question:

- C.CR-1. To what degree are the Content Specifications represented in the draft blueprints?

Findings

Findings related to the representation of ELA/literacy content between the test blueprint and Content Specifications is presented below.

C.CR-1: To what degree are the Content Specifications represented in the draft blueprints?

Reviewers felt the ELA/literacy blueprints were mostly to fully representative of the content and knowledge that Smarter Balanced outlined to be assessed.

Table 5.C.1. C.CR-1a: ELA/Literacy Blueprint Rating N-Counts, Means, Standard Deviations, Median, Number of Comments

Grade	N	Mean	Standard Deviation	Median	Number of Comments
Grade 3	5	3.4	0.90	4.0	5
Grade 4	5	3.4	0.55	3.0	3
Grade 5	3	3.3	0.58	3.0	2
Grade 6	5	3.2	0.45	3.0	4
Grade 7	4	3.5	0.58	3.5	3
Grade 8	4	4.0	0.00	4.0	0
Grade 11	8	4.0	0.00	4.0	4

Reviewers were asked to provide explanations and comments for their ratings. Each of the groups discussed the ELA/literacy content and blueprint before making their final independent ratings. However, since they had discussed the content, many of the reviewers' comments from each grade/content area group were identical. Reviewers commented on the minimum number of items students would receive and the coverage of reading across genres.

Table 5.C.2. C.CR-1b: Summary of ELA/Literacy Blueprint Representativeness Comments

Grade	Number of Comments	Summary of Comments
Grade 3	5	The blueprint is representative or mostly representative of the intended targets as defined by Smarter Balanced. Some categories do not appear to have enough items to reflect proficiency (no specific examples were provided).
Grade 4	3	Recommendation for more DOK 1 items and a better representation of Reading for Information and Reading for Literacy and fewer items for Speaking and Listening. Writing seems to reflect a high emphasis on writing conventions.
Grade 5	2	The emphasis on Speaking and Listening seems to be high. There seems to be a lot of DOK 4 items required, and maybe not enough of the lower DOKs. Recommendation for more emphasis on Reading for Information.
Grade 6	4	The blueprints are well articulated, but feel like more examples are needed to accurately rate the representation of the claims and targets to the blueprint. The stated percentages in the content specifications and those on the blueprint do not seem to match.
Grade 7	3	Expected to see more questions about reasoning and evaluation across the tests. Would like a greater percentage of the items to be at the higher DOK levels. The stated percentages in the content specifications and those on the blueprint do not seem to match.
Grade 8	0	Not applicable
Grade 11	4	The comments focused on policies such as the training of teachers in science and social studies to share accountability for the ELA standards. Reading should be assessed in the performance tasks since reading is implicitly being measured.

Mathematics

Content Representation

Analysis was conducted to examine the representation of mathematics content between the test blueprint and the Content Specifications, and to address the following question:

- C.CR-1. To what degree are the Content Specifications represented in the draft blueprints?

Findings

Findings related to the representation of mathematics content between the test blueprint and Content Specifications is presented below.

C.CR-1: To what degree are the Content Specifications represented in the draft blueprints?

Reviewers felt the mathematics blueprints were mostly to fully representative of the content and knowledge that Smarter Balanced outlined was to be assessed.

Table 5.C.3. C.CR-1a: Mathematics Blueprint Rating N-Counts, Means, Standard Deviations, Median, and Number of Comments by Grade

Grade	N	Mean	Standard Deviation	Median	Number of Comments
Grade 3	5	4.0	0.00	4.0	2
Grade 4	5	3.8	0.45	4.0	2
Grade 5	5	3.8	0.45	4.0	1
Grade 6	5	3.6	0.55	4.0	2
Grade 7	5	4.0	0.00	4.0	0
Grade 8	5	3.8	0.45	4.0	1
Grade 11	8	3.3	0.89	3.5	4

Reviewers commented on the distribution of the mathematics content and the minimum numbers of items required. There were two grade 11 groups. One group reviewed all of the CCSS and Content Specifications relating to Claim 1 (Concepts and Procedures), and the other group reviewed the materials for Claims 2 – 4 (Claim 2, Problem Solving; Claim 3, Communicating Reasoning; Claim 4, Modeling and Data Analysis).

Table 5.C.4. C.CR-1b: Summary of Mathematics Blueprint Representativeness Comments

Grade	Number of Comments	Summary of Comments
Grade 3	2	The specifications have strong alignment to the blueprint. Recommend more DOK 3 items to assess transfer of knowledge.
Grade 4	2	Claims 2-4 were fairly even in my mapping while the blueprint puts more emphasis on communicating reasoning. Recommend more DOK 3 items to assess transfer of knowledge.
Grade 5	1	Assessment Target C should have more items. Fifth graders need to be asked to synthesize.
Grade 6	2	Unsure if Modeling and Data Analysis can be accurately measured with 2 performance task items. Would like to understand what is required of students to obtain credit for DOK 4 items.
Grade 7	0	
Grade 8	1	Would like to see the concept of congruence better represented.
Grade 11	4	The number of questions in each area seemed appropriately distributed. Did not feel adequately prepared for this task since only familiar with Claim 1. Concerns there are not enough items that use real-world questions. I feel that many of the depths of knowledge required of the students by the standards fall into the category 3 range although very few in claim 1 have DOK 3. Additionally, the majority of the questions on the test are in reference to functions and equations and this does not seem equitable to me. There are many other high school standards being taught besides these. If this test is reflecting what you expect an 11th grader to know then those standards should emphasized.

Connection D: Alignment of Item/Task Pools to Evidence Statements

Analyses were conducted separately by content area to examine alignment between the item/task pools and the evidence statements. These analyses focused on content representation and DOK consistency. This section includes only ELA/literacy items because metadata, which provided the evidence statements intended by the item writer, were only available for ELA/literacy items.

ELA/Literacy Computer-Adapted Test (CAT) Items

Analyses related to this connection examined the evidence statement to item ratings that were collected during Workshops 3 – 5.

Content Representation

Data were gathered to examine the mappings of evidence statements to ELA/literacy CAT items. These data were collected for the same items and from the same reviewers as the data found later in Connection G. The reader can find additional descriptive data in the report section that describes Connection G: Alignment of Item/Task Pools and Content Specifications.

Pairwise Agreement

We examined the extent to which reviewers' ratings of the evidence statements that they identified for each CAT item agreed with the evidence statements that were indicated by the item writers. This agreement was examined in two ways. One way examined the extent to which the evidence statements identified by the reviewers' matched exactly the evidence statements indicated by the item writer. The second way examined the extent to which at least one of the evidence statements that the reviewers' identified matched at least one evidence statement that was indicated by the item writers. For both types of agreement, reviewers' average item-level pairwise agreement was between approximately 57 – 85%; on average, more than half the reviewers agreed with one another that the intended evidence statement mapped to its respective item (see Table 5.D.1).

Table 5.D.1. Pairwise Agreement of Reviewers' ELA/Literacy CAT Item Identified Evidence Statement Mappings to Evidence Statements Intended by Item Writers, by Grade and Claim

Grade	Claim	# Reviewers	# items	Avg # Items Per Reviewer	Avg Pairwise Agreement for Exact Match	Avg Pairwise Agreement For at Least One Match
3	1	15	185	62	80.9%	80.5%
	2	15	85	29	61.2%	57.2%
	3	15	93	31	62.4%	62.4%
	4	15	66	22	62.7%	62.7%
4	1	15	170	56	71.6%	71.6%
	2	15	90	30	66.0%	66.4%
	3	15	91	30	69.7%	69.7%
	4	15	56	19	66.1%	66.1%
5	1	15	158	53	70.6%	70.6%
	2	15	87	29	62.3%	60.5%
	3	15	50	17	58.7%	58.7%
	4	15	53	18	77.7%	77.7%

Table 5.D.1. (Continued)

Grade	Claim	# Reviewers	# items	Avg # Items Per Reviewer	Avg Pairwise Agreement for Exact Match	Avg Pairwise Agreement For at Least One Match
6	1	15	198	66	71.4%	71.4%
	2	15	94	31	67.0%	60.0%
	3	15	82	27	70.6%	70.6%
	4	15	61	20	64.3%	64.3%
7	1	13	181	59	78.0%	78.0%
	2	13	94	32	70.1%	66.7%
	3	13	81	28	68.6%	68.6%
	4	13	58	19	76.8%	76.8%
8	1	13	176	61	78.9%	78.9%
	2	13	82	25	68.6%	63.9%
	3	13	100	33	65.0%	65.0%
	4	13	54	18	85.1%	85.1%
11	1	25	628	105	74.7%	74.3%
	2	25	300	50	69.6%	67.6%
	3	25	337	56	67.6%	67.6%
	4	25	192	32	71.7%	71.7%

Table Read: For grade 3, Claim 1, there were 15 reviewers across the three alignment workshops who provided item-level evidence statement ratings for 185 items, for an average of 62 items per reviewer. The average reviewer item-level pairwise agreement for an exact match with all of the intended evidence statements was 80.9%. The average reviewer item-level pairwise agreement for at least one identified evidence statement matching one of the intended evidence statements was 80.5%.

Findings

Findings related to each content representation question are presented below. When relevant, general descriptive statistics are provided to describe overall results for each grade across claims.

D.CR-1 (CAT): How are the summative assessment items distributed across evidence statements?

Table 5.D.2 presents the average number of items mapped to each evidence statement by reviewers, along with the minimum and maximum number of items that reviewers, on average, mapped to evidence statements. Because reviewers could map multiple evidence statements to each item, this analysis considered all mapped evidence statements. The list of all ELA/literacy evidence statements that were not mapped to a single item is presented in Appendix H.

Table 5.D.2. D.CR-1. Average Number of ELA/Literacy Items Mapped to Each Evidence Statement and Minimum and Maximum Average Numbers of Items Assigned to each Evidence Statement

Grade	Claim	# of Reviewers	Avg # Items per Reviewer	Avg # Items Per Evidence Statement	Min Avg items Per Evidence Statement	Max Avg items Per Evidence Statement
3	1	15	62	9.2	1	15
	2	15	29	4.8	1	14
	3	15	31	15.0	14	16
	4	15	22	10.6	1	14
4	1	15	56	9.2	1	15
	2	15	30	5.1	1	12
	3	15	30	6.7	1	13
	4	15	19	8.0	1	12
5	1	15	53	9.2	1	17
	2	15	29	6.3	1	17
	3	15	17	10.0	1	23
	4	15	18	11.6	1	15
6	1	15	66	9.3	1	17
	2	15	31	5.0	1	14
	3	15	27	11.2	1	19
	4	15	20	10.5	2	14
7	1	13	59	8.8	1	15
	2	13	32	4.6	1	13
	3	13	28	9.6	1	17
	4	13	19	11.2	9	13
8	1	13	61	8.5	1	16
	2	13	25	4.1	1	13
	3	13	33	10.3	1	17
	4	13	18	12.6	9	15
11	1	25	105	17.8	1	26
	2	25	50	10.0	1	26
	3	25	56	21.3	1	27
	4	25	32	13.3	1	25

Table Read: For Grade 3, Claim 1, 15 reviewers provided evidence statement ratings across three workshops. On average, reviewers provided evidence statement ratings for 62 items. For those evidence statements they mapped to items, they found an average of 9.2 items mapped to each. The minimum number of items mapped to an individual evidence statement was 1 and the maximum was 15.

D.CR-2 (CAT): Do the reviewers agree with the intended mapping of items to evidence statements as identified by the item developers?

Across grades and claims, reviewers and item writers generally agreed on the average number of evidence statements that were mapped to items. Typically, more than half the items were found to have an exact match between the evidence statements identified by the reviewers and those identified by the item writers; the exception was at Claim 2 (Writing) for all grades. In some cases, a slightly higher percentage of items were mapped to all of the intended evidence statements, plus they included additional evidence statements. In these cases, they were not counted as an exact match, but were counted in a second category of match, where reviewers identified all intended evidence statements, but may have included more (in the second agreement column of Table 5.D.3). At Claim 2 there were sometimes targets with many associated evidence statements; this likely made the task extra challenging for reviewers, resulting in the low percentages at this claim. For about one-third of the grade and claims, the percentage of items where at least one item matched at least one of the intended evidence statements resulted in a slightly higher match percentage compared to when reviewers selected all the intended evidence statements.

DOK Consistency

Data were gathered to examine the extent to which the DOK level of each ELA/literacy evidence statement (as identified by the reviewers) was consistent with the DOK level of its associated CAT item (as indicated by the Content Specifications). As these data were collected for the same items and from the same reviewers as data gathered for Connection G: Alignment of Item/Task Pools and Content Specifications, we do not duplicate these findings here. The reader is referred to the report section on Connection G: Alignment of Item/Task Pools and Content Specifications, for additional descriptive data, such as reviewer pairwise agreement results.

Table 5.D.3. Average Percentage of ELA/Literacy CAT Item Evidence Statement(s) Aligned to Intended Evidence Statement(s), by Grade and Claim

Grade	Claim	# of Reviewers	# of Items	Avg # of Items Per Reviewer	Avg # Evidence Statements per Item Rated	Avg # Evidence Statements per Item Intended	Agreement Statistics			
							Exact Agreement %(n)	Selected All Intended Evidence Statements (but may have included more) %(n)	At Least One Rating Matched At Least One Intended Evidence Statement %(n)	No Agreement with Intended Evidence Statement %(n)
3	1	15	154	62	1.0	1.0	78.5% (48)	78.5% (48)	79.9% (49)	20.1% (13)
	2	15	58	29	1.5	1.3	41.3% (11)	49.3% (14)	62.6% (18)	37.4% (11)
	3	15	86	31	1.0	1.0	58.8% (18)	58.8% (18)	58.8% (18)	41.2% (13)
	4	15	42	22	1.0	1.0	70.9% (15)	70.9% (15)	70.9% (15)	29.1% (7)
4	1	15	121	56	1.0	1.0	60.8% (34)	60.9% (34)	62.1% (34)	37.9% (22)
	2	15	63	30	1.4	1.1	39.3% (12)	45.1% (14)	48.6% (15)	51.4% (15)
	3	15	70	30	1.0	1.0	48.7% (16)	48.7% (16)	48.7% (16)	51.3% (14)
	4	15	33	19	1.0	1.0	56.5% (10)	56.5% (10)	56.5% (10)	43.5% (9)
5	1	15	152	53	1.0	1.0	67.4% (35)	70.0% (37)	70.0% (37)	30.0% (16)
	2	15	81	29	1.6	1.2	41.7% (12)	58.6% (17)	62.5% (18)	37.5% (11)
	3	15	45	17	1.1	1.0	56.4% (9)	60.3% (10)	60.3% (10)	39.7% (7)
	4	15	53	18	1.0	1.0	76.7% (13)	79.9% (14)	79.9% (14)	20.1% (4)
6	1	15	140	66	1.0	1.0	60.1% (39)	61.8% (40)	62.4% (41)	37.6% (25)
	2	15	66	31	1.4	1.2	35.3% (11)	43.0% (14)	52.3% (16)	47.7% (15)
	3	15	65	27	1.1	1.0	59.6% (16)	63.0% (17)	63.0% (17)	37.0% (10)
	4	15	35	20	1.0	1.0	71.3% (14)	73.4% (15)	73.4% (15)	26.6% (5)
7	1	13	171	59	1.0	1.0	73.0% (42)	74.8% (44)	74.8% (44)	25.2% (15)
	2	13	85	32	1.2	1.2	44.2% (13)	49.2% (15)	55.7% (18)	44.3% (14)
	3	13	81	28	1.0	1.0	57.9% (16)	60.8% (17)	60.8% (17)	39.2% (11)
	4	13	58	19	1.0	1.0	77.8% (15)	78.2% (15)	78.2% (15)	23.6% (4)

Table 5.D.3. (Continued)

Grade	Claim	# of Reviewers	# of Items	Avg # of Items Per Reviewer	Avg # Evidence Statements per Item Rated	Avg # Evidence Statements per Item Intended	Agreement Statistics			
							Exact Agreement %(n)	Selected All Intended Evidence Statements (but may have included more) %(n)	At Least One Rating Matched At Least One Intended Evidence Statement %(n)	No Agreement with Intended Evidence Statement %(n)
8	1	13	175	61	1.0	1.0	74.7% (45)	75.6% (46)	75.6% (46)	24.4% (15)
	2	13	78	25	1.7	1.3	38.4% (9)	46.6% (12)	55.5% (14)	44.5% (11)
	3	13	97	33	1.0	1.0	54.0% (18)	57.6% (19)	57.6% (19)	42.4% (14)
	4	13	54	18	1.0	1.0	89.2% (16)	90.0% (16)	90.0% (16)	10.0% (2)
11	1	25	628	105	1.0	1.0	70.3% (73)	70.9% (74)	72.6% (76)	27.4% (29)
	2	25	300	50	1.3	1.2	48.4% (24)	51.6% (26)	56.5% (28)	43.5% (22)
	3	25	337	56	1.0	1.0	59.4% (33)	59.9% (33)	59.9% (33)	40.1% (23)
	4	25	192	32	1.0	1.0	82.7% (27)	83.4% (27)	83.4% (27)	16.6% (5)

Table Read: There were 15 reviewers across three workshops that provided ratings for grade 3, Claim 1 item evidence statement mappings for 154 items, for a total of 925 ratings. On average, reviewers identified 1 evidence statement mapped to each grade 3, Claim 1 item and there was an average of 1 evidence statement mapped to these items intended by item writers. On average, reviewers reported 78.5% of grade 3, Claim 1 items were mapped to all the intended evidence statements, which was equal to an average of 48 items mapped to the intended evidence statement per rater. For this grade and claim, the same percentage of items, on average, were mapped to the intended evidence statement, even though reviewers may have included additional evidence statements, meaning that reviewers did not select additional evidence statements beyond the intended. An average of 79.9% of items (or an average of 49 items) included at least one evidence statement that matched at least one of the intended evidence statements, and an average of 20.1% (or 13 items per rater) that did not match any of the intended evidence statements.

Findings

D.DC-1 (CAT). Is the cognitive complexity required in the items consistent with the cognitive complexity required in each evidence statement?

Table 5.D.4 presents the consistency between the DOK levels of the items, as verified by the reviewers, and the DOK range of the mapped evidence statement, as independently identified by these same reviewers. For all grades, but especially for grades 3 and 4, reviewers' ratings of DOK consistency between the CAT items and their mapped evidence statements was generally high across all claims. The exception was Claim 4 (Research and Inquiry) for grades 5 – 8 and 11; reviewers' rated most Claim 4 CAT items as requiring a lower DOK level than the DOK range identified for the evidence statement mapped to that item.

We typically conducted DOK consistency analyses using the intended DOK levels, as indicated in the Content Specifications. However, Smarter Balanced did not include DOK levels for evidence statements, so DOK ranges were generated to address this question based on average evidence statement DOK ratings made by reviewers prior to beginning the item-rating tasks. For example, if 50% or more of the reviewers agreed that an evidence statement was at DOK 2 and 3 levels, and fewer than 50% of the reviewers rated that evidence statement as a DOK level 1 or a DOK level 4, the DOK range for that evidence statement became DOK levels 2 and 3. Based on the generated DOK range for the evidence statements, the reviewers tended to rate the DOK of the evidence statement at a higher level than they rated the DOK level of the items, targets, and grade-level standards. Connection G examines DOK consistency between items and the intended DOK of their mapped target; Connection G provides DOK consistency information based on the intent by Smarter Balanced.

Table 5.D.4. Average Percentage of ELA CAT Items Rated as Having DOK Levels Consistent and Inconsistent with Range of Mapped Evidence Statements as Identified by Reviewers, by Grade and Claim

Grade	Claim	# of Reviewers	# of Items	Avg # Items Per reviewer	Consistent	Inconsistent	
					Avg items falling within the range of the identified Evidence Statement %(n)	Avg items with DOKs higher than the highest of the identified Evidence Statement range %(n)	Avg items with DOKs lower than the lowest of the identified evidence statement range %(n)
3	1	15	191	61	85.6% (52)	9.5% (6)	4.9% (3)
	2	15	99	26	78.5% (20)	10.8% (3)	10.7% (3)
	3	15	94	31	72.3% (22)	20.9% (7)	6.8% (2)
	4	15	66	22	91.1% (20)	0.3% (0)	8.6% (2)
4	1	15	175	47	83.2% (40)	11.1% (5)	5.6% (2)
	2	15	94	31	98.6% (31)	0.0% (0)	1.4% (0)
	3	15	93	33	83.8% (28)	13.8% (4)	2.5% (1)
	4	15	57	17	98.9% (17)	1.1% (0)	0.0% (0)
5	1	15	163	52	64.5% (34)	15.4% (8)	20.1% (10)
	2	15	92	29	87.0% (25)	3.2% (1)	9.8% (3)
	3	15	92	16	67.7% (11)	20.9% (3)	11.4% (2)
	4	15	53	17	18.1% (3)	2.0% (0)	79.9% (14)
6	1	15	207	66	67.0% (44)	22.7% (15)	10.3% (7)
	2	15	107	30	95.2% (29)	0.8% (0)	4.0% (1)
	3	15	86	27	69.0% (19)	0.2% (0)	30.8% (8)
	4	15	61	20	6.0% (1)	0.4% (0)	93.6% (19)
7	1	13	215	58	79.3% (46)	18.6% (11)	2.1% (1)
	2	13	103	31	96.9% (30)	0.4% (0)	2.6% (1)
	3	13	96	28	70.1% (20)	0.6% (0)	29.3% (8)
	4	13	58	19	5.3% (1)	0.0% (0)	94.7% (18)

Table 5.D.4. (Continued)

Grade	Claim	# of Reviewers	# of Items	Avg # Items Per reviewer	Consistent	Inconsistent	
					Avg items falling within the range of the identified Evidence Statement %(n)	Avg items with DOKs higher than the highest of the identified Evidence Statement range %(n)	Avg items with DOKs lower than the lowest of the identified Evidence Statement range %(n)
8	1	13	191	61	80.6% (49)	14.9% (9)	4.5% (3)
	2	13	92	25	92.7% (23)	1.9% (1)	5.4% (1)
	3	13	100	33	52.1% (17)	0.0% (0)	47.9% (16)
	4	13	54	7	4.2% (1)	0.0% (0)	95.8% (6)
11	1	25	696	100	84.9% (89)	9.7% (10)	5.4% (1)
	2	25	360	61	96.5% (47)	1.6% (0)	1.9% (14)
	3	25	342	74	74.8% (42)	0.1% (0)	25.1% (32)
	4	25	192	18	3.0% (1)	0.0% (0)	97.0% (17)

Table Read: Table Read: At grade 3, claim 1, there were 191 CAT items rated across all workshops, and 15 total reviewers for a total of 905 unique ratings used in the analysis. On average, reviewers felt that most (85.6%, or an average of 52 items per reviewer) of item DOK levels fell within the DOK range of the mapped evidence statement, as rated by reviewers. On average, 9.5% of items were rated as having a DOK higher than the identified range, and 4.9% of items as having a lower DOK than the intended range.

ELA/Literacy Performance Tasks (PT)

Analyses related to this connection examined the evidence statement to item ratings that were collected during Workshops 3 – 5.

Content Representation

Data were gathered to examine the mappings of evidence statements to ELA/literacy PTs. These data were collected for the same PTs and from the same reviewers as the data found later in Connection G. The reader can find additional descriptive data in the report section that describes Connection G: Alignment of Item/Task Pools and Content Specifications.

Pairwise Agreement

We examined the extent of agreement in reviewers' ratings of the evidence statements that they identified for each PT compared with the evidence statements that were indicated by the item writers. Across grades, reviewers' average item-level pairwise agreement ranged from 64% in grade 3 to 100% in grade 6. It should be noted that reviewers rated a very small number of PTs so that each rating had a relatively large impact on the average percentage.

Table 5.D.5. Pairwise Agreement of Reviewers' ELA/Literacy PT Identified Evidence Statement Mappings to Evidence Statements Intended by Item Writers, by Grade and Claim

Grade	# of Performance Tasks	# of Reviewers	Avg # of Items Per PT*	Avg Evidence Statement Pairwise Agreement Between Reviewers
3	3	10	3	64.0%
4	3	10	3	74.0%
5	3	9	3	76.3%
6	3	8	3	100.0%
7	3	10	3	69.3%
8	3	9	3	96.0%
11	6	17	3	78.8%

* There were 4 items per PT; however, there was an average of only 3 items per PT with evidence statement ratings made by reviewers. For an average of one item per PT, reviewers tended to leave the evidence statement information blank.

Table Read: At grade 3, 10 reviewers across two workshops provided ratings on 3 PTs. There was an average of approximately 3 items per PT with evidence statement ratings provided by reviewers. Pairwise agreement on the match between reviewers' ratings of the evidence statements mapped to the PTs and the evidence statement indicated for those same PTs in the Content Specifications was 64%.

Findings

D.CR-1 (PT): How are the summative assessment items distributed across evidence statements?

Table 5.D.6 presents the average number of ELA/literacy PTs that reviewers mapped to a single evidence statement. Because the maximum number of individual items rated to a single evidence statement within a PT was one, the analysis to address this question was performed differently than the analysis conducted to examine the CAT items. Therefore, we did not first average items within a PT and then by grade, but rather we did so at the grade level alone. Recall that only a small sample of PTs was reviewed, so that covering the full range of evidence statements was highly unlikely. For this reason, we do not include diagnostic information about the evidence statements that were not covered in this study, as we did in the previous section that described findings for the CAT items.

Table 5.D.6 Average Number of ELA/Literacy PT Items Mapped to Each Evidence Statement and Minimum and Maximum Average Numbers of Items Assigned to each Evidence Statement

Grade	# Performance Tasks	Avg # Items per Performance Task	# of Reviewers	Avg # PT Items Per Evidence Statement	Min Avg PT items Per Evidence Statement	Max Avg PT items Per Evidence Statement
3	3	3	10	3.3	1	5
4	3	3	10	2.9	1	5
5	3	3	9	2.1	1	4
6	3	3	10	3.0	1	5
7	3	3	10	2.8	1	5
8	3	3	10	4.4	2	5
11	6	3	10	2.6	1	4

Note: There were 4 items per PT; however, there was an average of approximately 3 items per PT with valid evidence statement ratings

Table Read: Items from 3 PTs were rated at grade 3, with an average of 3 items with evidence statement ratings made by reviewers per PT. A total of 10 reviewers provided ratings on PTs across two workshops. Reviewers included an average of 3.3 items (across all PTs). Reviewers rated a minimum of 1 PT item per evidence statement and a maximum of 5 PT items per evidence statement.

D.CR-2 (PT): Do the reviewers agree with the intended mapping of items to evidence statements as identified by the item developers?

For grades 3, 4, 6, 8, and 11, reviewers agreed more than 70% of the time with the intended evidence statement mapping for items within a PT. This agreement was lower, however, for grades 5 and 7. For these grades, reviewers agreed approximately 50% of the time with the intended evidence statement mapping for items within a PT.

Table 5.D.7 Average Percentage of ELA/Literacy PT Evidence Statements Aligned to Intended, by Grade (Averaged First within PT, then Grade)

Grade	# of Performance Tasks	# of Reviewers	# of Items	Avg # Items per PT	Evidence Statement Ratings	
					Agreement between Rater and Intended %(n)	Disagreement %(n)
3	3	10	10	3	73.9% (2)	26.1% (1)
4	3	10	10	3	75.0% (2)	25.0% (1)
5	3	9	9	3	47.8% (2)	52.2% (1)
6	3	8	10	3	77.8% (3)	22.2% (1)
7	3	10	10	3	58.3% (2)	41.7% (1)
8	3	9	10	3	77.6% (3)	22.4% (1)
11	6	17	20	3	73.4% (2)	26.6% (1)

Note: There were 4 items per PT; however, there was an average of approximately 3 items per PT with valid evidence statement ratings

Table Read: At grade 3, there were three PTs with a total of 10 items with evidence statement ratings provided by reviewers across the two workshops where PT items were rated. There were 10 reviewers who provided ratings across the workshops, for a total of 50 grade 3 PT item ratings. An average of 73.9% of item evidence statements within a PT were found to align to the intended evidence statement, while 26.1% of items did not.

DOK Consistency

Data related to this criterion were collected for the same items and from the same reviewers as the data gathered for Connection G: Alignment of Item/Task Pools and Content Specifications; findings for this criterion are not duplicated here, but rather can be found in the report section on Connection G: Alignment of Item/Task Pools and Content Specifications, including additional descriptive data, such as reviewer pairwise agreement results.

Findings

D.DC-1 (PT). Is the cognitive complexity required in the items consistent with the cognitive complexity required in each evidence statement?

Recall that Smarter Balanced did not include DOK levels for evidence statements, so DOK ranges were generated to address this question based on average DOK ratings. For example, if 50% or more of the reviewers agreed that an evidence statement was at DOK 2 and 3 levels, and fewer than 50% of the reviewers rated that evidence statement as a DOK level 1 or a DOK level 4, the DOK range for that evidence statement became DOK levels 2 and 3. Based on the generated DOK range for the evidence statements, the reviewers tended to rate the DOK of the evidence statement at a higher level than they rated the DOK level of the items, targets, and grade-level standards.

Table 5.D.8 presents the consistency between the DOK level(s) of the PT items, as verified by the reviewers, and the range of DOK levels of the evidence statement mapped to those PT items. Similar to the ELA/literacy CAT item findings, the items within a PT were often found to be of lower difficulty than the DOK range identified for the evidence statement, particularly in the lower grade-levels. Based on our findings, we found consistency in reviewers providing higher DOK level ratings

for evidence statements compared to those for items, assessment targets, and grade-level standards.

Table 5.D.8. Average Percentage of ELA/Literacy PT Items Rated as Having DOK Levels Consistent and Inconsistent with Identified Range of Mapped Evidence Statement, by Grade

Grade	# of Performance Tasks	# of Reviewers	# of Items	# of items per PT	Consistent	Inconsistent	
					Avg items falling within the range of the identified Evidence Statement %(n)	Avg items with DOKs higher than the highest of the identified Evidence Statement range %(n)	Avg items with DOKs lower than the lowest of the identified Evidence Statement range %(n)
3	3	10	12	4	33.3% (1)	0.0% (0)	66.7% (2)
4	3	10	12	4	42.2% (1)	0.0% (0)	57.8% (2)
5	3	10	12	4	57.8% (2)	0.0% (0)	42.2% (1)
6	3	9	12	4	76.1% (3)	0.0% (0)	23.9% (1)
7	3	10	12	4	95.6% (2)	0.0% (0)	4.4% (0)
8	3	10	12	4	82.2% (2)	0.0% (0)	17.8% (1)
11	6	17	24	4	95.8% (2)	0.0% (0)	4.2% (0)

Table Read: At grade 3 there were 3 ELA/literacy PTs with 4 items each, for a total of 12 individual items. There were 10 reviewers (across two workshops) that rated the PTs. Reviewers rated an average of 33.3% of items (average of 1 item per reviewer, per PT) within a PT as falling within the DOK range of the intended evidence statement. Reviewers did not rate any of the items as falling above the intended DOK range and 66.7% of the items as falling below the DOK range of the intended target.

Connection E: Alignment of CAT Algorithm to Test Blueprint

Analysis was conducted separately by content area to examine the alignment between the CAT algorithm and the test blueprint. Three HumRRO researchers evaluated the algorithm specifications (V2. 6/17/2013) to the draft test blueprints (dated January 2014) for ELA/literacy and mathematics. At the time of this version, no simulated test events were available for an independent review of the resulting blueprint correspondence. Moreover, the version of the specifications reviewed was incomplete because, at the time, no decision had been made on many of the algorithm requirements such as defining the measurement model to use.

We are aware that the American Institutes for Research (AIR) has successfully delivered computer-adaptive test events to several states. This is evidence that they are knowledgeable about the algorithm requirements that are needed. Additionally, based on the documentation, AIR is aware of the requirement to meet the Smarter Balanced blueprint in terms of DOK level and content coverage. Once decisions have been made to complete the necessary requirements for the algorithm including the final blueprints, and after item development is completed, we recommend Smarter Balanced check the resulting test events against the blueprint requirements.

Connection G: Alignment of Items/Tasks to Content Specifications

Analyses were conducted separately by content area to examine alignment between the items/tasks and the Content Specifications. These analyses focused on content representation, DOK distribution, and DOK consistency.

ELA/Literacy CAT Items

A sample of 50% of all ELA/literacy CAT items were included in the alignment review. This sample was stratified by grade, claim, and target. Additionally, we reviewed the targets with only a very small number of total items and hand selected additional items, when available, where there were fewer than five items.

Table 5.G.1 presents the total number of ELA/literacy CAT items rated at each workshop by grade, claim, and overall. As shown, each group rated items across all claims; a fairly similar number of items were included for grades 3 – 8; however, more items were included at grade 11 because there was a larger pool of items for this grade.

Table 5.G.1 also presents the number of reviewers who provided item-level ratings at Workshops 3, 4, and 5 for each grade. Each workshop included two high school groups, as represented in the table. Although analyzed separately, reviewers rated items for two grades so that the same reviewers rated items for grades 3 and 4, the same reviewers rated items for grades 5 and 6, and the same reviewers rated items for grades 7 and 8. For each workshop, there were typically 4 or 5 reviewers who completed the ratings for their assigned grades. The exception was the group of reviewers who rated items for grades 7 and 8 at Workshop 3, where three reviewers completed their assigned ratings.

Table 5.G.1. Numbers of ELA/Literacy Reviewers and ELA/Literacy CAT Items Rated, by Workshop and Overall

Grade/ Group	Claim	# of Workshop 3 Reviewers	# of Workshop 4 Reviewers	# of Workshop 5 Reviewers	Total Reviewers	# of Workshop 3 Items	# of Workshop 4 Items	# of Workshop 4 Items	Total Items
3	1	5	5	5	15	64	67	60	191
	2					33	30	36	99
	3					29	30	35	94
	4					23	24	19	66
4	1	5	5	5	15	48	66	61	175
	2					45	26	23	94
	3					35	22	36	93
	4					12	24	21	57
5	1	5	5	5	15	58	48	57	163
	2					28	29	35	92
	3					27	41	23	91
	4					18	16	19	53
6	1	5	5	5	15	77	59	71	207
	2					31	47	29	107
	3					21	31	34	86
	4					25	18	18	61
7	1	3	5	5	13	78	51	86	215
	2					32	48	23	103
	3					25	37	34	96
	4					22	21	15	58
8	1	3	5	5	13	49	69	73	191
	2					46	21	25	92
	3					32	37	31	100
	4					18	19	17	54

Table 5.G.1. (Continued)

Grade/ Group	Claim	# of Workshop 3 Reviewers	# of Workshop 4 Reviewers	# of Workshop 5 Reviewers	Total Reviewers	# of Workshop 3 Items	# of Workshop 4 Items	# of Workshop 4 Items	Total Items
11 - Group 1	1	4	5	4	13	115	117	112	344
	2					53	50	55	158
	3					59	53	72	184
	4					38	44	26	108
11 - Group 2	1	4	4	4	12	110	119	123	352
	2					76	62	64	202
	3					44	59	55	158
	4					35	25	24	84

Content Representation

Analyses were conducted to examine the representation of content between the ELA/literacy CAT items and the Content Specifications, and to address the following questions:

- G.CR-1: How are the summative assessment items distributed across targets and grade-level standards?
- G.CR-2: Do the reviewers agree with the intended mapping of items to targets and grade-level standards, as identified by the item developers?

Pairwise Agreement among Reviewers

Tables 5.G.2 and 5.G.3 present the pairwise agreement of reviewers' ratings of ELA/literacy CAT items and targets, and the reviewers' ratings of ELA/literacy items and grade-level standards, respectively. For each item, the pairwise agreement was calculated by determining the percent of pairs of reviewers who agreed on their ratings. These percentages were then averaged to obtain the values presented.

As can be seen in Table 5.G.2, the average pairwise agreement among reviewers' ratings for ELA/literacy CAT items within a claim was very high (i.e., 95.1% - 100%).

Table 5.G.2. Pairwise Agreement for ELA/literacy CAT Item Target Ratings among Reviewers, by Grade and Claim

Grade	Claim	# items	# of Reviewers	Ave # of Items Per Reviewer	Avg Pairwise Agreement
3	1	191	15	63	96.6%
	2	99	15	31	99.6%
	3	94	15	31	97.9%
	4	66	14	22	100.0%
4	1	175	15	49	95.8%
	2	94	12	32	97.9%
	3	93	12	33	100.0%
	4	57	12	18	100.0%
5	1	163	15	55	95.1%
	2	91	15	30	98.7%
	3	92	15	30	99.6%
	4	53	15	18	100.0%
6	1	207	15	69	96.2%
	2	107	15	35	99.3%
	3	86	15	28	100.0%
	4	61	15	20	100.0%
7	1	215	13	71	99.0%
	2	103	13	34	99.0%
	3	96	13	33	99.6%
	4	58	13	19	100.0%

Table 5.G.2. (Continued)

Grade	Claim	# items	# of Reviewers	Ave # of Items Per Reviewer	Avg Pairwise Agreement
8	1	191	13	66	98.8%
	2	92	13	28	100.0%
	3	100	13	33	99.2%
	4	54	13	18	99.3%
11	1	696	25	114	97.3%
	2	360	25	54	98.1%
	3	342	25	56	99.7%
	4	192	25	31	98.2%

Table Read: At grade 3, Claim 1, 191 ELA/literacy CAT items were rated across the three workshops. There were 15 total reviewers, with each reviewer rating an average of 63 ELA/literacy CAT items. The average item-level pairwise agreement among these reviewers for rating the ELA/literacy CAT items and targets was 96.6%

Although not as high as the agreement among reviewers' ratings for target, reviewer agreement was typically high across grades and claims. The lowest pairwise agreement among reviewers was for grade 11, Claim 3 items, with an average agreement of 63.8%.

Table 5.G.3. Pairwise Agreement for ELA/literacy CAT Item Grade-level Standard Ratings among Reviewers, by Grade and Claim

Grade	Claim	# items	# of Reviewers	Avg Items Per Rater	Avg Pairwise Agreement
3	1	189	15	62	68.7%
	2	99	15	31	80.5%
	3	93	15	31	76.8%
	4	66	14	22	72.1%
4	1	175	15	48	72.1%
	2	94	12	33	78.0%
	3	90	12	32	77.7%
	4	57	12	18	81.4%
5	1	163	15	54	79.8%
	2	91	15	30	87.2%
	3	87	15	29	76.9%
	4	53	15	18	95.5%
6	1	205	15	68	82.4%
	2	107	15	35	87.9%
	3	81	15	27	92.6%
	4	61	15	19	92.4%
7	1	215	13	71	81.7%
	2	103	13	35	86.9%
	3	94	13	32	84.2%
	4	58	13	19	74.1%

Table 5.G.3. (Continued)

Grade	Claim	# items	# of Reviewers	Avg Items Per Rater	Avg Pairwise Agreement
8	1	191	13	66	79.3%
	2	92	13	27	87.0%
	3	100	13	34	85.8%
	4	54	13	18	90.6%
11	1	695	25	115	77.3%
	2	359	25	55	77.1%
	3	342	25	56	63.8%
	4	192	25	33	69.0%

Table Read: At grade 3, Claim 1, 189 ELA/literacy CAT items were rated across the three workshops. There were 15 total reviewers, with each reviewer seeing an average of 62 ELA/literacy CAT items. The average item-level pairwise agreement among these reviewers for rating the ELA/literacy CAT items and grade-level standards was 68.7%

Findings

G.CR-1 (CAT): How are the summative assessment items distributed across assessment targets?

Table 5.G.4 presents the average number of ELA/literacy CAT items assigned to each target. The last two columns of this table present the average minimum number of items assigned to any one target within the claim and the maximum number of items assigned to a target within the claim. Targets included in this analysis were those that the reviewers verified as being aligned or, if they disagreed with the alignment, the reviewer provided an alternate target and that alternate target was used. For ELA/literacy CAT items, all targets were represented by at least one item.

Table 5.G.4. Average, Minimum, and Maximum Number of ELA/literacy CAT Items Mapped to Each Target, Averaged across Reviewers, by Grade and Claim

Grade	Claim	# of Reviewers	Avg # Items Per Reviewer	Avg # Items Per Target	Min Avg items Per Target	Max Avg items Per Target
3	1	15	63	14.6	13	15
	2	15	31	10.5	7	14
	3	15	31	15.0	15	15
	4	14	22	13.0	12	14
4	1	15	49	13.4	11	15
	2	12	32	7.9	3	12
	3	12	33	12.0	12	12
	4	12	18	11.0	10	12

Table 5.G.4. (Continued)

Grade	Claim	# of Reviewers	Avg # Items Per Reviewer	Avg # Items Per Target	Min Avg items Per Target	Max Avg items Per Target
5	1	15	55	14.3	10	15
	2	15	30	9.7	1	15
	3	15	30	8.0	1	15
	4	15	18	15.0	15	15
6	1	15	69	14.6	10	15
	2	15	35	13.0	10	15
	3	15	28	15.0	15	15
	4	15	20	15.0	15	15
7	1	13	71	12.7	10	13
	2	13	34	11.4	8	13
	3	13	33	13.0	13	13
	4	13	19	13.0	13	13
8	1	13	66	12.9	12	13
	2	13	28	10.0	3	13
	3	13	33	13.0	13	13
	4	13	18	13.0	13	13
11	1	25	114	25.0	25	25
	2	25	54	17.3	1	25
	3	25	56	25.0	25	25
	4	25	31	23.3	23	24

G.CR-2 (CAT): Do the reviewers agree with the intended mapping of items to targets and grade-level standards as identified by the item developers?

Across grades and claims, the vast majority of reviewers agreed that the ELA/literacy CAT items fully aligned to their intended targets. Grades 3 and 4 had the lowest percentage of items rated as being fully aligned to the targets; however, across grades and claims, at least 87% of the items were rated as being fully aligned to their intended targets.

Table 5.G.5. Average Percentage of ELA/literacy CAT Items Rated as Fully Aligned, Partially Aligned, or Not Aligned to Intended Target, by Grade and Claim

					Target Verification Rating		
Grade	Claim	# of Reviewers	# of Items	Ave # of Items Per Reviewer	Fully- Aligned %(n)	Partially-Aligned %(n)	Not - Aligned %(n)
3	1	15	191	63	89.5% (56)	9.0% (6)	1.5% (1)
	2	15	99	31	89.1% (27)	9.9% (4)	1.1% (0)
	3	15	94	31	97.7% (31)	1.2% (0)	1.1% (0)
	4	14	66	22	89.5% (20)	10.5% (2)	0.0% (0)
4	1	15	175	49	88.3% (42)	10.3% (6)	1.4% (1)
	2	12	94	32	87.4% (29)	12% (3)	0.6% (0)
	3	12	93	33	94.2% (31)	5.8% (2)	0.0% (0)
	4	12	57	18	90.0% (16)	9.6% (2)	0.3% (0)
5	1	15	163	55	95.3% (52)	3.1% (2)	1.5% (1)
	2	15	92	30	98.2% (30)	1.2% (0)	0.6% (0)
	3	15	91	30	98.9% (30)	1.1% (0)	0.0% (0)
	4	15	53	18	100.0% (18)	0.0% (0)	0.0% (0)
6	1	15	207	69	96.6% (67)	2.1% (1)	1.3% (1)
	2	15	107	35	99.7% (35)	0.0% (0)	0.3% (0)
	3	15	86	28	100.0% (28)	0.0% (0)	0.0% (0)
	4	15	61	20	100.0% (20)	0.0% (0)	0.0% (0)
7	1	13	215	71	98.6% (70)	1.2% (1)	0.2% (0)
	2	13	103	34	99.4% (34)	0.2% (0)	0.5% (0)
	3	13	96	33	100.0% (33)	0.0% (0)	0.0% (0)
	4	13	58	19	98.3% (19)	1.7% (0)	0.0% (0)
8	1	13	191	66	98.8% (65)	1.0% (1)	0.2% (0)
	2	13	92	28	99.6% (28)	0.4% (0)	0.0% (0)
	3	13	100	33	89.5% (33)	0.0% (0)	0.4% (0)
	4	13	54	18	99.2% (18)	0.4% (0)	0.4% (0)
11	1	25	696	114	96.9% (111)	2.1% (2)	1.0% (1)
	2	25	360	54	96.5% (52)	2.7% (2)	0.8% (0)
	3	25	342	56	99.6% (56)	0.3% (0)	0.1% (0)
	4	25	192	31	94.0% (31)	1.3% (0)	4.7% (0)

Table Read: For grade 3, Claim 1, 15 reviewers rated a total of 191 ELA/literacy CAT items, with each reviewer rating an average of 63 items. On average, the reviewers found 89.5% of the items fully aligned to their intended target (an average of 56 items per reviewer), 9% of the items partially aligned to their intended target (6 items), and an average of 1.5% of the items not aligned to their intended target (1 item per reviewer).

Across grades and claims, reviewers rated the majority of ELA/literacy CAT items as fully aligned to their intended grade-level standards. The percentages were generally higher for the full alignment of items in Claims 2 (Writing) and 3 (Speaking and Listening). The exception was grade 11, which had the lowest average percentage of items (76.6%) fully aligned to their intended grade-level standard.

Table 5.G.6. Average Percentage of ELA/literacy CAT Items Rated as Fully Aligned, Partially Aligned, or Not Aligned to Intended Grade-level Standard, by Grade and Claim

CCSS Verification Rating								
Grade	Claim	# of Reviewers	# of CAT Items	Avg CAT Items Per Rater	# Excluded CAT Items[36]	Fully-Aligned %(n)	Partially-Aligned %(n)	Not - Aligned %(n)
3	1	15	191	62	0	78.4% (48)	15.3% (9)	6.3% (4)
	2	15	99	31	0	86.8% (26)	9.1% (3)	4.1% (1)
	3	15	93	31	0	87.4% (27)	9.8% (3)	2.8% (1)
	4	14	66	22	0	80.2% (18)	19.5% (4)	0.3% (0)
4	1	15	175	48	0	83.5% (40)	14.1% (8)	2.4% (1)
	2	12	94	33	1	87.3% (28)	10.0% (3)	2.8% (1)
	3	12	90	32	0	87.1% (28)	10.6% (3)	2.3% (1)
	4	12	57	18	0	86.6% (15)	12.4% (2)	1.0% (0)
5	1	15	163	54	0	87.2% (47)	6.9% (4)	5.9% (3)
	2	15	92	30	0	92.4% (28)	5.1% (2)	2.5% (1)
	3	15	86	29	1	88.5% (25)	9.5% (3)	2.0% (0)
	4	15	53	18	0	97.9% (17)	2.1% (0)	0.0% (0)
6	1	15	207	68	2	89.3% (61)	6.1% (4)	4.6% (3)
	2	15	107	35	0	92.9% (33)	3.2% (1)	3.9% (1)
	3	15	81	27	3	96.4% (26)	1.0% (0)	2.5% (1)
	4	15	55	19	0	97.7% (19)	0.4% (0)	1.9% (0)
7	1	13	215	71	0	87.2% (62)	10.3% (7)	2.5% (2)
	2	13	103	35	0	93.5% (32)	5.2% (2)	1.3% (1)
	3	13	94	32	1	91.1% (30)	7.7% (2)	1.1% (0)
	4	13	58	19	0	87.7% (16)	9.8% (2)	2.6% (1)
8	1	13	191	66	0	87.2% (57)	9.4% (7)	3.4% (2)
	2	13	92	27	0	92.3% (25)	6.2% (1)	1.5% (0)
	3	13	100	34	0	91.7% (31)	7.5% (3)	0.8% (0)
	4	13	54	18	0	94.7% (17)	4.9% (1)	0.4% (0)
11	1	25	695	115	1	83.6% (96)	11% (12)	5.4% (6)
	2	25	360	55	0	82.3% (45)	7.0% (4)	10.7% (5)
	3	25	342	56	0	76.6% (44)	17.0% (9)	6.3% (4)
	4	25	192	33	0	82.3% (27)	11.1% (3)	6.6% (3)

Table Read: For grade 3, Claim 1, 15 reviewers rated a total of 191 ELA/literacy CAT items, with each reviewer rating an average of 62 items. No items were excluded due to errors in grade-level standard metadata. On average, reviewers rated 78.4% of the items as fully aligned to their intended grade-level standard (average of 48 items per reviewer), 15.3% of the items partially aligned to their intended grade-level standard (9 items), and 6.3% of the items not aligned to their intended grade-level standard (4 items).

[36] Some items were excluded due to errors in grade-level standard metadata.

DOK Distribution

Analyses were conducted to examine the distribution of DOK levels between the ELA/literacy CAT items and the Content Specifications, and to address the following question:

- G.DD-1: How does the distribution of DOK of the items identified by the reviewers compare with the distribution identified in the Content Specifications?

Pairwise Agreement

The average pairwise agreement among reviewer ratings of ELA/literacy CAT item DOK level is presented in Table 5.G.7. The average pairwise agreement among reviewers ranged from 72.8% and 97.9%, indicating that reviewers tended to agree with the DOK levels as indicated in the Content Specifications.

Table 5.G.7. Reviewer Pairwise Agreement for ELA/literacy CAT Item DOK Ratings, by Grade and Claim

Grade	Claim	# items	# Reviewers	Avg # Items per Reviewer	Avg Pairwise Agreement
3	1	191	15	63	83.8%
	2	99	15	31	88.4%
	3	94	15	31	90.1%
	4	66	14	22	97.9%
4	1	175	15	49	79.1%
	2	94	12	32	87.0%
	3	93	12	33	83.2%
	4	57	12	18	92.3%
5	1	163	15	55	81.7%
	2	91	15	30	86.8%
	3	92	15	30	72.8%
	4	53	15	18	88.3%
6	1	207	15	69	73.8%
	2	107	15	35	90.0%
	3	86	15	28	80.6%
	4	61	15	20	87.0%
7	1	215	13	71	79.0%
	2	103	13	34	89.5%
	3	96	13	33	85.6%
	4	58	13	19	86.1%
8	1	191	13	66	80.5%
	2	92	13	28	85.2%
	3	100	13	33	88.3%
	4	54	13	18	92.8%
11	1	696	25	114	76.4%
	2	360	25	54	90.4%
	3	342	25	56	75.3%
	4	192	25	31	91.6%

Table Read: At grade 3, Claim 1, 191 items were rated by 15 reviewers across three workshops, with each reviewer rating an average of 63 ELA/literacy CAT items. The average item-level pairwise agreement among these reviewers for the DOK level of the item was 83.8%

Findings

G.DD-1 (CAT): How does the distribution of DOK of the items identified by the reviewers compare with the distribution identified in the Content Specifications?

Table 5.G.8 provides a summary of the how the ELA/literacy CAT items were distributed across the four DOK levels in terms of their DOK level, as indicated in the Content Specifications and as verified by the reviewers. For DOK levels 1 – 3, reviewers agreed with the Content Specifications about the distribution of items across DOK levels. However, for DOK level 4, reviewers believed a smaller mean percentage of items were at a DOK level 4 than was intended by the Content Specifications. This was particularly true for grades 6, 7, 8, and 11

.

Table 5.G.8. Distribution of ELA/literacy CAT Items Across DOK Levels, Average Percentage of Items Rated, and Percentage of Items with DOK Level as Indicated by Content Specifications

Grade	Claim	# of Reviewers	total # of Items	Ave # Items Per Reviewer	DOK 1		DOK 2		DOK3		DOK 4	
					Avg Items Rated %(n)	Items Intended %	Avg Items Rated %(n)	Items Intended %	Avg Items Rated %(n)	Items Intended %	Avg Items Rated %(n)	Items Intended %
3	1	15	191	63	9.2% (6)	10.0%	46.1% (29)	45.6%	39.5% (25)	37.2%	5.2% (3)	7.3%
	2	15	99	31	39.6% (13)	43.4%	44.5% (15)	47.5%	15.8% (3)	9.1%	0.0% (0)	0.0%
	3	15	94	31	17.9% (6)	20.2%	36.7% (11)	33.0%	45.3% (14)	46.8%	0.0% (0)	0.0%
	4	14	66	22	0.3% (0)	0.0%	98.8% (22)	100.0%	0.9% (0)	0.0%	0.0% (0)	0.0%
4	1	15	175	49	9.4% (4)	10.3%	38.6% (19)	38.3%	44.3% (21)	39.4%	7.7% (4)	12.0%
	2	12	94	32	38.5% (14)	42.6%	50.0% (16)	50.0%	11.5% (3)	7.5%	0.0% (0)	0.0%
	3	12	93	33	18.5% (6)	20.4%	36.8% (12)	30.1%	44.7% (15)	49.5%	0.0% (0)	0.0%
	4	12	57	18	0.0% (0)	0.0%	94.4% (17)	100.0%	4.9% (1)	0.0%	0.7% (0)	0.0%
5	1	15	163	55	7.3% (4)	8.0%	43.9% (24)	43.6%	43.5% (23)	40.5%	5.4% (3)	8.0%
	2	15	92	30	47.5% (14)	47.8%	42.1% (13)	42.4%	10.4% (3)	9.8%	0.0% (0)	0.0%
	3	15	91	30	20.5% (7)	18.7%	40.2% (12)	29.7%	39.3% (12)	51.7%	0.0% (0)	0.0%
	4	15	53	18	0.0% (0)	0.0%	93.9% (17)	100.0%	4.4% (1)	0.0%	1.7% (0)	0.0%
6	1	15	207	69	5.4% (4)	4.4%	34.4% (23)	33.8%	45.8% (32)	35.8%	14.5% (10)	26.1%
	2	15	107	35	40.0% (14)	40.2%	49.9% (17)	49.5%	9.5% (3)	10.3%	0.6% (0)	0.0%
	3	15	86	28	14.6% (4)	17.4%	40.5% (12)	33.7%	44.6% (13)	48.8%	0.2% (0)	0.0%
	4	15	61	20	0.0% (0)	0.0%	93.6% (19)	100.0%	1.7% (0)	0.0%	4.7% (1)	0.0%
7	1	13	215	71	3.0% (2)	3.3%	30.6% (21)	31.6%	47.1% (34)	31.2%	19.3% (14)	34.0%
	2	13	103	34	36.2% (14)	41.8%	53.6% (17)	48.5%	9.8% (4)	9.7%	0.4% (0)	0.0%
	3	13	96	33	14.6% (5)	16.7%	39.5% (13)	34.4%	45.4% (15)	49%	0.5% (0)	0.0%
	4	13	58	19	0.0% (0)	0.0%	94.7% (18)	100.0%	5.3% (1)	0.0%	0.0% (0)	0.0%

Table 5.G.8. (Continued)

Grade	Claim	# of Reviewers	total # of Items	Ave # Items Per Reviewer	DOK 1		DOK 2		DOK3		DOK 4	
					Avg Items Rated %(n)	Items Intended %	Avg Items Rated %(n)	Items Intended %	Avg Items Rated %(n)	Items Intended %	Avg Items Rated %(n)	Items Intended %
8	1	13	191	66	2.2% (1)	3.1%	30.2% (19)	29.8%	50.1% (33)	33.0%	17.5% (12)	34.0%
	2	13	92	28	30.3% (9)	34.8%	54.3% (15)	52.2%	14.1% (4)	13.0%	1.3% (0)	0.0%
	3	13	100	33	12.3% (4)	18.0%	39.9% (13)	36.0%	47.8% (16)	46.0%	0.0% (0)	0.0%
	4	13	54	18	0.0% (0)	0.0%	95.8% (17)	100.0%	4.2% (1)	0.0%	0.0% (0)	0.0%
11	1	25	696	114	2.9% (3)	4.0%	35.1% (41)	34.2%	51.1% (59)	37.5%	10.9% (13)	24.3%
	2	25	360	54	31.8% (19)	33.3%	54.1% (32)	56.7%	11.8% (6)	10.0%	2.2% (0)	0.0%
	3	25	342	56	11.9% (7)	18.4%	40.2% (23)	33.6%	47.3% (27)	48.0%	0.7% (0)	0.0%
	4	25	192	31	0.0% (0)	0.0%	95.6% (31)	100.0%	4.1% (1)	0.0%	0.3% (0)	0.0%

Table Read: For grade 3, claim 1, 15 reviewers rated a total of 191 ELA/literacy CAT items, with reviewers rating an average of 63 items each. For DOK level 1, reviewers rated an average of 9.2% of the items (average of 6 out of 63) as falling within the range of DOK level 1 compared to an intended 10% of the items falling within that range. For DOK level 2, reviewers rated an average of 46.1% of the items (average of 29 out of 63) as falling within the range of DOK level 2 compared to an intended 45.6% of items falling within that range. For DOK level 3, reviewers rated an average or 39.5% of the items (average of 25 out of 63) as falling within the DOK level 3 compared to an intended 37.2% of items falling within that range. For DOK level 4, reviewers rated an average of 5.2% (average of 3 out of 63) of the items as falling within the range of DOK level 4 compared to an intended 7.3% of the items falling within that range.

DOK Consistency

Analyses were conducted to examine the consistency of DOK levels between ELA/literacy CAT items and the Content Specifications, and to address the following question:

- G.DC-1: Does the DOK of the item identified by the reviewers fall within the DOK distribution of the aligned target identified in the Content Specifications?

The DOK ratings used for this analysis were the same as those used to analyze the DOK distribution of the ELA/literacy CAT items and, therefore, are not duplicated here. The reader is referred to the earlier report section that describes the ELA/literacy CAT item DOK distribution for information about reviewer pairwise agreement.

Findings

G.DC-1 (CAT): Is the cognitive complexity required in the items consistent with the cognitive complexity required in each assessment target?

Across grades and claims, reviewers believed the DOK level for the vast majority of the ELA/literacy CAT items fell within the range of DOK levels for the intended target (93.6 – 100%) (see Table 5.G.9). Reviewers rated only a few of the items as having a DOK level higher than the highest DOK level of the intended target or having a DOK level lower than the lowest DOK level of the intended target.

Table 5.G.9. Average Percentage of ELA/literacy CAT Items Rated as Having DOK Levels Consistent and Inconsistent with Intended Range of Mapped Target

Grade	Claim	# of Reviewers	# CAT Items	Ave # CAT Items Per Reviewer	Consistent	Inconsistent	
					Avg CAT items falling within the range of the intended Target %(n)	Avg CAT items with DOKs higher than the highest of the intended Target range %(n)	Avg CAT items with DOKs lower than the lowest of the intended Target range %(n)
3	1	15	191	63	96.8% (60)	0.4% (0)	2.8% (3)
	2	15	99	31	99.2% (31)	0.6% (0)	2.8% (2)
	3	15	94	31	100.0% (31)	0.0% (0)	0.2% (0)
	4	14	66	22	98.8% (22)	0.9% (0)	0.0% (0)
4	1	15	175	49	97.0% (47)	0.9% (0)	0.3% (0)
	2	12	94	32	97.2% (32)	1.5% (1)	2.1% (1)
	3	12	93	33	100.0% (33)	0.0% (0)	1.3% (0)
	4	12	57	18	94.4% (17)	5.6% (1)	0.0% (0)
5	1	15	163	55	96.9% (53)	0.9% (1)	0.0% (0)
	2	15	92	30	97.7% (30)	1.6% (1)	2.2% (1)
	3	15	91	30	100.0% (30)	0.0% (0)	0.7% (0)
	4	15	53	18	93.9% (17)	6.1% (1)	0.0% (0)
6	1	15	207	69	94.4% (65)	0.9% (1)	0.0% (0)
	2	15	107	35	97.5% (34)	2.3% (1)	4.7% (3)
	3	15	86	28	99.8% (28)	0.2% (0)	0.2% (0)
	4	15	61	20	93.6% (19)	6.4% (1)	0.0% (0)
7	1	13	215	71	98.0% (69)	1.0% (1)	0.0% (0)
	2	13	103	34	97.7% (34)	1.5% (1)	1.0% (1)
	3	13	96	33	99.3% (33)	0.7% (0)	0.8% (0)
	4	13	58	19	94.7% (18)	5.3% (1)	0.0% (0)

Table 5.G.9. (Continued)

Grade	Claim	# of Reviewers	# CAT Items	Ave # CAT Items Per Reviewer	Consistent	Inconsistent	
					Avg CAT items falling within the range of the intended Target %(n)	Avg CAT items with DOKs higher than the highest of the intended Target range %(n)	Avg CAT items with DOKs lower than the lowest of the intended Target range %(n)
8	1	13	191	66	96.3% (64)	2.0% (1)	0.0% (0)
	2	13	92	28	96.9% (27)	3.1% (1)	1.6% (1)
	3	13	100	33	100.0% (33)	0.0% (0)	0.0% (0)
	4	13	54	18	95.8% (17)	4.2% (1)	0.0% (0)
11	1	25	696	114	95.5% (110)	1.7% (2)	0.0% (0)
	2	25	360	54	97.5% (57)	2.4% (0)	2.9% (3)
	3	25	342	56	99.3% (56)	0.7% (0)	0.1% (0)
	4	25	192	31	95.6% (31)	4.4% (1)	0.0% (0)

Table Read: For grade 3, Claim 1, 15 reviewers rated 191 ELA/literacy CAT items, with reviewers rating an average of 63 items each. Reviewers rated 96.8% of the items (60 out of 63) as falling within the range of DOK levels associated with the intended target. Reviewers rated only a few of the items as outside the range of DOK levels for the intended target—0.4% of the items (0 out of 63) were rated as having a DOK higher than the highest DOK for the intended target and 2.8% of the items (3 out of 63) were rated as having a DOK lower than the lowest DOK for the intended target.

Mathematics CAT Items

Table 5.G.8 presents the number of reviewers who rated the mathematics CAT items at Workshops 3, 4, and 5 for each grade level. Similar to ELA/literacy, there were two groups of reviewers who rated grade 11 items. Although analyzed separately, reviewers rated items for two grades so that the same reviewers rated items for grades 3 and 4, the same reviewers rated items for grades 5 and 6, and the same reviewers rated items for grades 7 and 8. For each workshop, there were typically 4 or 5 reviewers who completed the ratings for their assigned grades. The exceptions were grades 7, 8, and 11, where three reviewers completed their assigned ratings.

Table 5.G.8 also presents the total number of items rated at each workshop by grade, claim, and overall. As shown, each group rated items across all claims; a fairly similar number of items were included for grades 3 – 8; however, more items were included at grade 11 because there was a larger pool of items for this grade.

Table 5.G.8. Numbers of Mathematics Reviewers and Mathematics CAT Items Rated, by Workshop and Overall

Grade/ Group	Claim	# of Workshop 3 Reviewers	# of Workshop 4 Reviewers	# of Workshop 5 Reviewers	Total Reviewers	# of Workshop 3 Items	# of Workshop 4 Items	# of Workshop 4 Items	Total Items
3	1	5	5	5	15	113	117	106	336
	2					12	15	22	49
	3					28	25	29	82
	4					18	16	17	51
4	1	5	5	5	15	105	97	110	312
	2					15	18	22	55
	3					28	33	16	77
	4					23	23	24	70
5	1	5	5	5	15	98	89	97	284
	2					11	27	15	53
	3					30	29	32	91
	4					25	20	21	66
6	1	5	5	5	15	92	112	110	314
	2					36	13	9	58
	3					35	21	24	80
	4					22	13	10	45
7	1	3	5	4	12	91	89	99	279
	2					16	19	9	44
	3					27	25	23	75
	4					10	8	13	31
8	1	3	4	4	11	95	88	93	276
	2					27	11	10	48
	3					32	29	24	85
	4					20	14	15	49

Table 5.G.8. (Continued)

Grade/ Group	Claim	# of Workshop 3 Reviewers	# of Workshop 4 Reviewers	# of Workshop 5 Reviewers	Total Reviewers	# of Workshop 3 Items	# of Workshop 4 Items	# of Workshop 4 Items	Total Items
11 - Group 1	1	4	4	5	13	168	181	175	524
	2					28	21	22	71
	3					53	50	60	163
	4					30	27	22	79
11 - Group 2	1	4	3	4	11	188	199	171	558
	2					18	22	11	51
	3					46	39	67	152
	4					25	20	31	76

Content Representation

Analyses were conducted to examine the representation of mathematics content between the mathematics CAT items and the Content Specifications, and to address the following questions:

- G.CR-1: How are the summative assessment items distributed across targets, grade-level standards, and mathematical practices?
- G.CR-2: Do the reviewers agree with the intended mapping of items to targets and grade-level standards, as identified by the item developers?
- G.CR-3: Do the reviewers agree with the intended mapping of items to mathematical practices, as identified by the item developers?

Pairwise Agreement

Tables 5.G.9, 5.G.10, and 5.G.11 present the pairwise agreement of reviewers' ratings of mathematics CAT items and targets, reviewers' ratings of mathematics CAT items and grade-level standards, and reviewers' ratings of mathematics CAT items and mathematical practices, respectively. For each item, the pairwise agreement was calculated by determining the percent of pairs of reviewers who agreed on their ratings. These percentages were then averaged to obtain the values presented.

The average pairwise agreement between reviewer ratings of mathematics CAT item mappings to targets is presented in Table 5.G.9. The average pairwise agreement among reviewers for items within a claim was generally high, with the only average pairwise agreement below 80% occurring for Claim 4 (Modeling and Data Analysis) at grades 3 and 4.

Table 5.G.9. Pairwise Agreement for Mathematics CAT Item and Target Ratings between Reviewers, by Grade and Claim

Grade	Claim	# CAT Items	# of Reviewers	Avg # CAT Items per Reviewer	Avg Pairwise Agreement
3	1	336	15	111	99.3%
	2	51	15	17	86.5%
	3	84	15	28	85.1%
	4	53	14	17	78.9%
4	1	312	14	103	99.8%
	2	58	14	20	90.2%
	3	77	14	26	90.0%
	4	72	14	24	70.7%
5	1	282	15	88	98.4%
	2	56	13	20	83.5%
	3	91	13	30	89.2%
	4	69	13	22	89.1%
6	1	315	13	105	98.0%
	2	54	13	18	89.2%
	3	73	13	23	94.7%
	4	42	13	13	92.1%

Table 5.G.9. (Continued)

Grade	Claim	# CAT Items	# of Reviewers	Avg # CAT Items per Reviewer	Avg Pairwise Agreement
7	1	279	13	92	97.5%
	2	45	13	15	99.1%
	3	79	13	26	83.7%
	4	38	13	11	78.5%
8	1	276	13	91	98.1%
	2	43	13	14	87.2%
	3	83	13	28	96.0%
	4	48	13	16	90.3%
11	1	1080	24	179	98.3%
	2	123	24	21	88.9%
	3	314	24	50	87.1%
	4	156	23	25	85.4%

Table Read: At grade 3, Claim 1, 336 items were rated across the three workshops. There were 15 reviewers across the workshops, with each reviewer rating an average of 111 mathematics CAT items. The average item-level pairwise agreement among these reviewers for rating the mathematics CAT items and targets was 99.3%

Table 5.G.10 presents the average item-level pairwise agreement among reviewer ratings of mathematics CAT item mappings to grade-level standards, by grade and claim. As shown, the average pairwise agreement was generally high. The average percentages ranged from a low of 71.5% (grade 7, Claim 2, Problem Solving) to 98.7% (grade 3, Claim 1, Concepts and Procedures).

Table 5.G.10. Pairwise Agreement for Mathematics CAT Items and Grade-Level Standard Ratings between Reviewers, by Grade and Claim

Grade	Claim	# CAT Items	# of Reviewers	Avg # CAT Items Per Reviewer	Pairwise Agreement
3	1	336	15	112	98.7%
	2	39	11	19	90.5%
	3	56	11	28	80.9%
	4	36	10	18	81.7%
4	1	312	14	104	95.4%
	2	43	10	21	85.6%
	3	49	10	25	88.2%
	4	49	10	24	86.1%
5	1	282	15	87	91.5%
	2	45	9	23	92.2%
	3	61	9	30	88.9%
	4	44	9	21	91.9%
6	1	315	14	105	94.5%
	2	30	9	15	87.6%
	3	50	9	25	93.9%
	4	29	9	14	84.6%

Table 5.G.10. (Continued)

Grade	Claim	# CAT Items	# of Reviewers	Avg # CAT Items Per Reviewer	Pairwise Agreement
7	1	279	13	92	85.6%
	2	28	9	15	71.5%
	3	51	9	25	82.7%
	4	28	9	13	76.1%
8	1	276	13	91	85.8%
	2	27	9	13	80.9%
	3	58	9	30	94.5%
	4	35	9	16	83.3%
11	1	1081	24	179	84.9%
	2	77	16	19	83.9%
	3	202	16	49	84.5%
	4	97	15	24	85.5%

Table Read: At grade 3, Claim 1, 15 reviewers rated 339 items across the three workshops, with each reviewer rating an average of 112 mathematics CAT items. The average item-level pairwise agreement among these reviewers of mathematics CAT items and grade-level standards was 98.7%

Table 5.G.11 presents the average pairwise agreement between reviewer ratings of mathematics CAT item mappings to mathematical practice. The first pairwise agreement among reviewers reflects whether the primary mathematical practice the reviewers identified matched an intended mathematical practice. The second pairwise agreement among reviewers reflects whether at least one of the mathematical practices the reviewers identified (primary or additional) matched an intended mathematical practice. As shown, the pairwise agreement was fairly similar regardless of whether a primary mathematical practice matched or at least one of the mathematical practices matched. The percentages across grades and claims ranged between 50% – 80%.

Table 5.G.11. Pairwise Agreement for Mathematics CAT Items and Mathematical Practice Mappings between Reviewers, by Grade and Claim

Grade	Claim	# Reviewers	# CAT items	Avg # CAT Items Per Reviewer	Avg Pairwise Agreement with Primary Match	Avg Pairwise Agreement With at Least One Match
3	2	15	33	16	73.1%	74.8%
	3	14	82	37	71.7%	70.4%
	4	14	40	19	72.8%	64.8%
4	2	14	34	16	76.7%	59.6%
	3	14	70	31	67.6%	63.0%
	4	14	38	18	76.2%	69.8%
5	2	12	37	19	66.8%	73.1%
	3	11	90	40	72.0%	67.4%
	4	11	54	25	74.5%	70.0%

Table 5.G.11. (Continued)

Grade	Claim	# Reviewers	# CAT items	Avg # CAT Items Per Reviewer	Avg Pairwise Agreement with Primary Match	Avg Pairwise Agreement With at Least One Match
6	2	12	44	22	69.2%	64.8%
	3	12	65	36	58.4%	69.3%
	4	12	32	16	76.8%	71.5%
7	2	13	30	15	64.1%	55.4%
	3	13	69	34	58.3%	54.3%
	4	13	30	14	60.8%	54.4%
8	2	13	37	17	79.1%	56.1%
	3	13	77	38	71.4%	52.4%
	4	13	27	13	72.6%	50.4%
11	2	24	93	19	69.6%	65.7%
	3	24	293	59	69.3%	62.7%
	4	23	150	31	63.3%	63.4%

Findings

G.CR-1a (CAT): How are the summative assessment items distributed across assessment targets?

Table 5.G.12 presents the average number of mathematics CAT items assigned to each target. The last two columns of this table present the average minimum number of items assigned to any one target within the claim and the maximum number of items assigned to a target within the claim. Targets included in this analysis were those that the reviewers verified as being aligned or, if they disagreed with the alignment, the reviewer provided an alternate target and that alternate target was used. There were 12 total mathematics targets across all grades and claims that had no items mapped to them. A list of these targets is presented in Appendix I.

G.CR-1b (CAT): How are the summative assessment items distributed across mathematical practices?

Table 5.G.13 presents the average number of items at claims 2, 3, and 4 reviewers mapped to each of the eight mathematical practices. Reviewers were informed that claim 1 items were not intended to map to mathematical practices, and therefore they were not included. The most common mathematical practice mapped to items varied by grade and claim, with reviewers typically, on average, identifying at least one item mapped to each practice. Mathematical practice 3 was frequently mapped to claim 3 items across all grades. Mathematical practices 1, 2, 4, and 6 were also frequently mapped to items. Mathematical practice 8 was found to be mapped to only a very small percentage of items across all grades and claims. Reviewers could map more than one mathematical practice to each item; therefore, totals across rows may equal more than 100%.

Table 5.G.12. Average, Minimum, and Maximum Mathematics CAT Items Mapped to Each Target, Averaged Across Reviewers, by Grade and Claim

Grade	Claim	# of Reviewers	Avg # CAT Items Per Reviewer	Avg # CAT Items Per Target	Min Avg CAT Items Per Target	Max Avg items Per CAT Target
3	1	15	111	14.5	10	15
	2	15	17	13.5	10	15
	3	15	28	14.7	13	15
	4	14	17	12.2	9	14
4	1	14	103	12.8	4	14
	2	14	20	12.5	9	14
	3	14	26	13.8	13	14
	4	14	24	13.4	12	14
5	1	15	88	13.2	9	15
	2	13	20	11.5	8	13
	3	13	30	13.0	13	13
	4	13	22	10.0	5	13
6	1	13	105	13.0	13	13
	2	13	18	12.5	11	13
	3	13	23	11.1	8	13
	4	13	13	10.4	8	13
7	1	13	92	13.0	13	13
	2	13	15	10.3	5	13
	3	13	26	12.3	8	13
	4	13	11	8.5	6	13
8	1	13	91	13.0	13	13
	2	13	14	11.8	8	13
	3	13	28	13.0	13	13
	4	13	16	10.7	4	13
11	1	24	179	24.0	24	24
	2	24	21	22.0	16	24
	3	24	50	23.1	23	24
	4	23	25	20.0	13	23

Table Read: At grade 3, Claim 1, 15 reviewers across three workshops provided target ratings for an average of 111 items. Reviewers mapped an average of 14.5 items to each of the selected targets. The minimum average number of items mapped to a target was 10 and the maximum number of items mapped to a target was 15.

Table 5.G.13. Average Number of Mathematics CAT Items Mapped to Each Mathematical Practice, Averaged Across Reviewers, by Grade and Claim

Grade	Claim	Mathematical Practice							
		MP 1	MP 2	MP 3	MP 4	MP 5	MP 6	MP 7	MP 8
3	2	51.1%(9)	35.7%(5)	2.6%(0)	38.2%(6)	7.3%(1)	6.3%(1)	1.7%(0)	0.9%(0)
	3	20.1%(6)	49.2%(14)	49.0%(13)	23.2%(7)	5.6%(2)	7.5%(2)	8.3%(2)	0.0%(0)
	4	37.1%(7)	50.6%(9)	6.0%(1)	34.7%(6)	5.4%(1)	15.0%(3)	5.1%(1)	1.1%(0)
4	2	42.6%(9)	43.9%(8)	6.4%(1)	28.1%(5)	7.6%(2)	22.2%(4)	2.5%(1)	0.6%(0)
	3	20.5%(4)	35.4%(9)	47.6%(13)	19.3%(4)	4.9%(1)	14.7%(4)	8.6%(2)	1.9%(0)
	4	31.6%(8)	43.2%(10)	3.5%(1)	39.6%(9)	4.1%(1)	10.7%(2)	6.4%(2)	4.4%(1)
5	2	70.7%(15)	47.9%(8)	1.9%(0)	16.6%(3)	10.7%(2)	49.9%(11)	9.8%(3)	0.0%(0)
	3	52.6%(16)	47.7%(14)	50.4%(15)	14.6%(4)	5.8%(2)	42.1%(12)	23.9%(7)	3.2%(1)
	4	69.1%(16)	47.8%(10)	10.1%(2)	39.3%(9)	4.9%(1)	43.1%(9)	19.5%(4)	7.3%(1)
6	2	64.2%(11)	47.0%(8)	3.6%(1)	35.5%(7)	13.7%(2)	41.0%(7)	12.7%(2)	2.3%(1)
	3	41.6%(9)	44%(10)	75.0%(18)	8.4%(2)	10.2%(3)	35.7%(8)	21.2%(5)	1.2%(0)
	4	69.3%(10)	50.4%(7)	9.7%(1)	42.1%(6)	17.9%(3)	42.7%(6)	13.0%(2)	3.1%(0)
7	2	44.2%(6)	45.3%(7)	4.1%(1)	25.4%(4)	2.0%(0)	22.7%(3)	8.5%(2)	2.0%(0)
	3	3.2%(1)	28.8%(8)	58.5%(15)	6.4%(2)	1.3%(0)	18.4%(5)	16.0%(4)	8.3%(2)
	4	34.3%(4)	32.7%(4)	4.4%(1)	30.4%(4)	5.1%(1)	14.3%(2)	16.6%(2)	4.5%(0)
8	2	28.0%(4)	34.3%(5)	4.8%(1)	31.4%(4)	7.0%(1)	23.8%(4)	16.2%(2)	5.1%(1)
	3	4.4%(1)	30.6%(8)	52.2%(15)	8.0%(2)	2.9%(1)	18.8%(5)	18.7%(5)	7.3%(2)
	4	12.9%(2)	37.1%(7)	11.0%(2)	42.6%(7)	5.9%(1)	13.2%(2)	18.2%(3)	1.8%(0)
11	2	41.7%(8)	43.1%(9)	1.7%(0)	32.2%(6)	5.7%(1)	12.3%(2)	7.5%(1)	2.4%(0)
	3	21.6%(11)	38.7%(20)	48.3%(23)	9.0%(5)	6.7%(3)	7.8%(4)	14.9%(7)	1.2%(1)
	4	35.4%(9)	34.5%(9)	5.3%(1)	63.5%(16)	4.9%(1)	12.6%(3)	5.7%(2)	2.3%(1)

Table Read: At grade 3, claim 2, reviewers mapped an average of 51.1% of items (or, approximately 9 items) to mathematical practice 1. They mapped an average of 35.7% (5 items) to mathematical practice 2, 2.6% (less than 1 item) to mathematical practice 3, 38.2% (6 items) to mathematical practice 4, 7.3% (1 item) to mathematical practice 5, 6.3% (1 item) to mathematical practice 6, 1.7% (less than 1 item) to mathematical practice 7, and 0.9% (less than 1 item) to mathematical practice 8.

G.CR-2 (CAT): Do the reviewers agree with the intended mapping of items to targets and grade-level standards as identified by the item developers?

The vast majority of reviewers rated the mathematics CAT items as being fully aligned to their intended target. This was especially true for items in Claim 1 (Concepts and Procedures). The percentages were generally lower for the alignment of items in Claims 2, 3, and 4; however, at least 81.3% of reviewers rated the items as fully aligned to their intended target (see Table 5.G.14).

Table 5.G.14. Average Percentage of Mathematics CAT Items Rated as Fully Aligned, Partially Aligned, or Not Aligned with Intended Target, by Grade and Claim

					Target Verification Rating		
Grade	Claim	# of Reviewers	# CAT Items	Ave # CAT Items Per Reviewer	Fully- Aligned %(n)	Partially-Aligned %(n)	Not – Aligned %(n)
3	1	15	336	111	99.6% (111)	0.4% (0)	0.0% (0)
	2	15	49	17	86.3% (15)	6.7% (1)	7.0% (1)
	3	15	82	28	89.5% (25)	3.3% (1)	7.2% (2)
	4	14	51	17	86.6% (15)	7.5% (1)	5.9% (1)
4	1	14	312	103	99.8% (103)	0.2% (0)	0.0% (0)
	2	14	55	20	94.2% (18)	3.3% (1)	2.5% (1)
	3	14	77	26	92.6% (24)	3.0% (1)	4.4% (1)
	4	14	70	24	84.0% (20)	4.1% (1)	11.9% (3)
5	1	15	284	88	98.4% (86)	0.9% (1)	0.7% (1)
	2	13	53	20	86.3% (17)	9.9% (2)	3.8% (1)
	3	13	91	30	94.2% (28)	3.0% (1)	2.8% (1)
	4	13	66	22	95.3% (21)	1.5% (0)	3.1% (1)
6	1	13	314	105	98.5% (104)	1.4% (1)	0.1% (0)
	2	13	58	18	94.6% (17)	3.8% (1)	1.5% (0)
	3	13	80	23	97.6% (23)	1.0% (0)	1.4% (0)
	4	13	45	13	95.6% (13)	2.4% (0)	2.0% (0)
7	1	13	279	92	95.9% (89)	3.7% (3)	0.3% (0)
	2	13	44	15	96.7% (15)	2.9% (0)	0.4% (0)
	3	13	75	26	87.5% (23)	10.8% (3)	1.8% (0)
	4	13	31	11	81.3% (9)	12.8% (1)	5.9% (1)
8	1	13	276	91	98.3% (90)	0.2% (0)	1.4% (1)
	2	13	48	14	93.1% (13)	6.0% (1)	1.0% (0)
	3	13	85	28	95.9% (27)	2.8% (1)	1.3% (0)
	4	13	49	16	94.4% (15)	3.8% (1)	1.8% (0)
11	1	24	1082	179	96.2% (172)	2.7% (5)	1.1% (2)
	2	24	122	21	86.7% (18)	10.0% (2)	3.3% (1)
	3	24	315	50	87.6% (44)	10.3% (5)	2.1% (1)
	4	23	155	25	89.4% (23)	7.6% (2)	3.0% (1)

Table Read: For grade 3, Claim 1, 15 reviewers rated a total of 336 items, with each reviewer rating an average of 111 mathematics CAT items. Reviewers reported an average of 99.6% of the items were fully aligned to their intended target (111 items), .4% of the items were partially aligned to their intended target (0 items, on average), and none of the items were not aligned to their intended target.

Across grades and claims, reviewers rated the vast majority of mathematics CAT items as being fully aligned to their intended grade-level standards. The lowest percentage of mathematics CAT items that were rated as being fully aligned to their intended grade-level standard was for grade 7, Claim 2 (79%).

Table 5.G.15. Average Percentage of Mathematics CAT Items Rated as Fully Aligned, Partially Aligned, or Not Aligned to Intended Grade-level Standard, by Grade and Claim

						CCSS Verification Rating		
Grade	Claim	# of Reviewers	# CAT Items	# CAT Excluded Items[37]	Avg # CAT Items Per Reviewer	Fully- Aligned %(n)	Partially-Aligned %(n)	Not Aligned %(n)
3	1	15	336	0	112	98.5% (110)	0.5% (1)	1.0% (1)
	2	11	37	0	19	93.6% (18)	4.5% (1)	2.0% (0)
	3	11	54	0	28	90.6% (25)	3.5% (1)	5.8% (2)
	4	10	33	0	18	91.3% (16)	4.5% (1)	4.3% (1)
4	1	14	312	0	104	97.4% (101)	1.9% (2)	0.6% (1)
	2	10	40	0	21	92.3% (20)	6.1% (1)	1.7% (0)
	3	10	49	0	25	93.8% (23)	4.0% (1)	2.2% (1)
	4	10	46	0	24	93.3% (23)	5.0% (1)	1.7% (0)
5	1	15	283	1	87	95.0% (83)	3.5% (3)	1.4% (1)
	2	9	42	0	23	95.3% (22)	3.5% (1)	1.2% (0)
	3	9	61	0	30	92.8% (28)	6.1% (2)	1.1% (0)
	4	9	41	0	21	96.2% (20)	2.9% (1)	0.9% (0)
6	1	14	314	0	105	97.3% (102)	2.5% (3)	0.1% (0)
	2	9	22	0	15	94.6% (14)	4.8% (1)	0.7% (0)
	3	9	45	0	25	97.2% (24)	2.8% (1)	0.0% (0)
	4	9	23	0	14	92.7% (13)	7.3% (1)	0.0% (0)
7	1	13	279	0	92	90.0% (83)	8.0% (7)	2.0% (2)
	2	9	27	0	15	79.0% (12)	12.0% (2)	9.0% (1)
	3	9	48	0	25	88.3% (22)	4.7% (1)	7.0% (2)
	4	9	21	0	13	82.8% (11)	14.8% (2)	2.5% (0)

[37] Some items were excluded due to errors in grade-level standard metadata.

Table 5.G.15. (Continued)

						CCSS Verification Rating		
Grade	Claim	# of Reviewers	# CAT Items	# CAT Excluded Items[38]	Avg # CAT Items Per Reviewer	Fully- Aligned %(n)	Partially-Aligned %(n)	Not Aligned %(n)
8	1	13	276	0	91	92.0% (84)	6.0% (5)	2.0% (2)
	2	9	21	0	13	88.0% (12)	10.0% (1)	2.0% (0)
	3	9	52	0	30	95.5% (28)	2.3% (1)	2.2% (1)
	4	9	29	0	16	90.7% (15)	6.9% (1)	2.4% (0)
11	1	24	1082	0	179	89.6% (161)	8.0% (14)	2.4% (4)
	2	16	75	0	19	88.8% (17)	4.8% (1)	6.4% (1)
	3	16	210	0	49	90.0% (43)	7.1% (4)	3.0% (2)
	4	15	99	0	24	88.9% (21)	6.9% (2)	4.2% (1)

Table Read: For grade 3, Claim 1, there were 15 reviewers who rated a total of 191 mathematics CAT items, with each reviewer rating an average of 61 items. There were no items removed due to errors in the metadata. On average, reviewers rated 78.4% of the items as fully aligned to their intended grade-level standard (average of 48 items per reviewer), 15.3% of the items partially aligned to their intended grade-level standard, (9 items), and 6.3% of the items were not aligned to their intended grade-level standard (4 items).

G.CR-3(CAT): Do the reviewers agree with the intended mapping of items to mathematical practices as identified by the item developers?

Reviewers were instructed to provide the primary mathematical practice that mapped to the item as well as any number of additional mathematical practices they believed mapped mathematics item in Claims 2 – 4. Table 5.G.16 shows that reviewers typically identified one-third or fewer of the mathematics CAT items as matching the primary mathematical practice as that intended by the item writers. The mapping agreement increased, however, when examining the average percentage of items where at least one of the mathematical practices identified by the reviewer matched at least one mathematical practice that was identified by the item writer.

[38] Some items were excluded due to errors in grade-level standard metadata.

Table 5.G.16. Average Percentage of Mathematics CAT Items Comparing Reviewer Identified Mathematical Practices to Intended Mathematical Practices Identified by Item Writers, by Grade and Claim

Grade	Claim	# of Reviewers	# CAT Items	Avg CAT Items per rater	Mathematical Practice Mapping				
					Avg # MPs rated per item	Avg # MPs intended per item	Rater Primary Mathematical Practice matched an Intended Mathematical Practice	At Least One Rating Matched At Least One Intended Mathematical Practices	No Agreement with Intended Mathematical Practice(s)
3	2	15	33	16	1.4	2.0	33.2% (5)	72.9% (8)	27.1% (3)
	3	14	82	37	1.7	2.0	33.5% (9)	64.5% (18)	35.5% (10)
	4	14	41	19	1.7	2.2	23.7% (4)	66.4% (10)	33.6% (5)
4	2	14	34	16	1.6	2.3	20.1% (3)	57.8% (8)	42.2% (5)
	3	14	70	31	1.5	2.1	29.1% (7)	65.2% (15)	34.8% (8)
	4	14	38	18	1.5	2.5	30.7% (4)	71.9% (10)	28.1% (4)
5	2	12	37	19	2.1	2.1	32.6% (5)	78.3% (11)	21.7% (3)
	3	11	90	40	2.4	1.8	35.9% (11)	68.8% (20)	31.2% (9)
	4	11	54	25	2.4	2.1	30.7% (6)	73.1% (14)	26.9% (5)
6	2	12	44	22	2.1	1.5	28.6% (5)	69.2% (11)	30.8% (6)
	3	12	65	36	2.2	1.3	50.3% (12)	75.7% (18)	24.3% (6)
	4	12	32	16	2.4	1.7	25.1% (3)	69.4% (9)	30.6% (4)
7	2	13	30	15	1.5	1.3	37.8% (4)	49.0% (5)	51.0% (6)
	3	13	69	34	1.4	1.5	41.1% (10)	64.8% (16)	35.2% (8)
	4	13	30	14	1.6	1.7	24.9% (3)	48.1% (6)	51.9% (5)
8	2	13	37	17	1.5	1.6	22.8% (3)	38.0% (5)	62.0% (9)
	3	13	77	38	1.4	1.3	34.4% (10)	48.9% (14)	51.1% (14)
	4	13	27	13	1.5	1.8	21.8% (2)	55.2% (6)	44.8% (5)

Table 5.G.16. (Continued)

Grade	Claim	# of Reviewers	# CAT Items	Avg CAT Items per rater	Mathematical Practice Mapping				
					Avg # MPs rated per item	Avg # MPs intended per item	Rater Primary Mathematical Practice matched an Intended Mathematical Practice	At Least One Rating Matched At Least One Intended Mathematical Practices	No Agreement with Intended Mathematical Practice(s)
11	2	24	93	19	1.4	1.8	28.0% (4)	49.2% (7)	50.8% (8)
	3	24	293	59	1.5	1.5	31.3% (14)	49.7% (22)	50.3% (23)
	4	23	150	31	1.7	2.2	22.6% (6)	72.4% (18)	27.6% (7)

Table Read: At grade 3, Claim 2, 15 reviewers across 3 workshops identified mathematical practices for 33 items, for an average of 16 items per reviewer. Reviewers mapped an average of 1.4 mathematical practices to each mathematics CAT item, compared to an average of 2.0 mathematical practices intended by item writers. Reviewers identified the same primary mathematical practice as did the item writer for an average of 33.3% of the items (or an average of 5 items per reviewer); at least one of the mathematical practices that reviewers identified matched an intended mathematical practice for 72.9% of the items (or an average of 8 items per reviewer). For approximately 27.1% of the items (an average of 8 items), reviewers identified a mathematical practice was different than the mathematical practice that was intended by the item writer.

DOK Distribution

Analyses were conducted to examine the distribution of DOK levels between the mathematics CAT items and the Content Specifications, and to address the following question:

- G.DD-1: How does the distribution of DOK of the items identified by the reviewers compare with the distribution identified in the Content Specifications?

To address this question, reviewers were asked to verify the DOK level of the mathematics CAT items that was indicated in the Content Specifications. If the reviewers disagreed with the DOK level indicated, they provided a DOK level they believed to be more appropriate to that item. In those instances, the alternate DOK level(s) provided by the reviewers were used in this analysis.

Pairwise Agreement

Table 5.G.17 presents the average item pairwise agreement between reviewers' DOK ratings for mathematics CAT items. Across grades and claims, reviewer agreement was generally high. The average agreement ranged from 71.9% (grade 5, Claim 3) to 97.4% (grade 7, Claim 2, Problem Solving).

Table 5.G.17. Pairwise Agreement for Mathematics CAT Item DOK Ratings between Reviewers, by Grade and Claim

Grade	Claim	# CAT Items	# Reviewers	Avg # CAT Items per Reviewer	Avg Pairwise Agreement
3	1	336	15	111	93.4%
	2	51	15	17	87.2%
	3	84	15	28	81.4%
	4	53	14	17	79.8%
4	1	312	14	103	84.3%
	2	58	14	20	86.7%
	3	77	14	26	91.7%
	4	72	14	24	78.8%
5	1	283	15	88	83.1%
	2	56	13	20	82.4%
	3	91	13	30	71.9%
	4	69	13	22	79.8%
6	1	315	13	105	83.1%
	2	54	13	18	87.8%
	3	73	13	23	76.0%
	4	42	13	13	72.8%
7	1	279	13	92	88.9%
	2	45	13	15	97.4%
	3	79	13	26	82.4%
	4	38	13	11	79.7%

Table 5.G.17. (Continued)

Grade	Claim	# CAT Items	# Reviewers	Avg # CAT Items per Reviewer	Avg Pairwise Agreement
8	1	276	13	91	79.9%
	2	43	13	14	90.7%
	3	83	13	28	83.9%
	4	48	13	16	73.3%
11	1	1081	24	179	83.1%
	2	123	24	21	85.3%
	3	314	24	50	79.8%
	4	156	23	25	79.3%

Table Read: At grade 3, Claim 1, 15 reviewers rated 336 items, with each reviewer rating an average of 111 mathematics CAT items. The average item-level pairwise agreement among reviewers' ratings for mathematics CAT item DOK was 93.4%

Findings

G.DD-1 (CAT): How does the distribution of DOK of the items identified by the reviewers compare with the distribution identified in the Content Specifications?

Table 5.G.18 provides a summary of how the mathematics CAT items were distributed in terms of their DOK level, as indicated in the Content Specifications and as verified by the reviewers. Across grades and claims, reviewers generally agreed with the DOK levels that were indicated in the Content Specifications. Reviewers consistently believed there was a slightly higher percentage of Claim 1 (Concepts and Procedures) items at DOK level 1 than was indicated in the Content Specifications.

Table 5.G.18. Distribution of Mathematics CAT Items across DOK Levels, Average Percentage of Items Rated, and Percentage of Item DOK Levels as Indicated in the Content Specifications

Grade	Claim	# of Reviewers	# CAT Items	Ave # CAT Items Per Reviewer	DOK 1		DOK 2		DOK3		DOK 4	
					Avg CAT Items Rated %(n)	CAT Items Intended %(n)	Avg CAT Items Rated %(n)	CAT Items Intended %(n)	Avg CAT Items Rated %(n)	CAT Items Intended %(n)	Avg CAT Items Rated %(n)	CAT Items Intended %(n)
3	1	15	336	111	66.0% (73)	65.5%	33.8% (37)	34.5%	0.2% (0)	0.0%	0.0% (0)	0.0%
	2	15	49	17	5.9% (1)	10.2%	88.8% (15)	85.7%	5.3% (1)	4.1%	0.0% (0)	0.0%
	3	15	82	28	0.7% (0)	0.0%	53.7% (15)	51.2%	45.1% (13)	48.8%	0.4% (0)	0.0%
	4	14	51	17	16.6% (3)	17.7%	53.3% (9)	47.1%	29.3% (5)	35.3%	0.7% (0)	0.0%
4	1	14	312	103	47.3% (49)	44.2%	52.1% (54)	55.1%	0.6% (1)	0.6%	0.0% (0)	0.0%
	2	14	55	20	1.5% (0)	0.0%	83.9% (16)	87.3%	14.6% (3)	12.7%	0.0% (0)	0.0%
	3	14	77	26	0.0% (0)	1.3%	45.7% (11)	45.5%	54.3% (14)	53.3%	0.0% (0)	0.0%
	4	14	70	24	10.6% (3)	12.9%	67.2% (16)	70.0%	21.5% (5)	17.1%	0.8% (0)	0.0%
5	1	15	284	88	50.3% (43)	44.4%	49.4% (44)	55.6%	0.3% (0)	0.0%	0.0% (0)	0.0%
	2	13	53	20	10.0% (2)	13.2%	81.0% (16)	77.4%	8.9% (2)	9.4%	0.0% (0)	0.0%
	3	13	91	30	2.9% (1)	0.0%	56.4% (17)	51.7%	40.7% (12)	48.4%	0.0% (0)	0.0%
	4	13	66	22	1.5% (0)	1.5%	69.8% (16)	68.2%	28.7% (6)	30.3%	0.0% (0)	0.0%
6	1	13	314	105	56.3% (59)	51.9%	43.6% (46)	48.1%	0.1% (0)	0.0%	0.0% (0)	0.0%
	2	13	58	18	1.7% (0)	0.0%	79.3% (14)	86.2%	18.9% (4)	13.8%	0.0% (0)	0.0%
	3	13	80	23	1.7% (0)	0.0%	42.6% (10)	37.5%	55.4% (13)	61.3%	0.2% (0)	1.3%
	4	13	45	13	1.6% (0)	6.7%	61.8% (9)	57.8%	34.6% (5)	33.3%	2.0% (0)	2.2%
7	1	13	279	92	44.1% (41)	41.9%	55.9% (52)	58.1%	0.0% (0)	0.0%	0.0% (0)	0.0%
	2	13	44	15	2.4% (0)	0.0%	87.8% (14)	88.6%	9.8% (1)	11.4%	0.0% (0)	0.0%
	3	13	75	26	0.0% (0)	0.0%	25.0% (7)	17.3%	75% (19)	82.7%	0.0% (0)	0.0%
	4	13	31	11	8.2% (1)	6.5%	56.3% (7)	51.6%	35.6% (4)	41.9%	0.0% (0)	0.0%

Table 5.G.18. (Continued)

Grade	Claim	# of Reviewers	# CAT Items	Ave # CAT Items Per Reviewer	DOK 1		DOK 2		DOK3		DOK 4	
					Avg CAT Items Rated %(n)	CAT Items Intended %(n)	Avg CAT Items Rated %(n)	CAT Items Intended %(n)	Avg CAT Items Rated %(n)	CAT Items Intended %(n)	Avg CAT Items Rated %(n)	CAT Items Intended %(n)
8	1	13	276	91	29.2% (27)	28.6%	70.7% (65)	71.4%	0.1% (0)	0.0%	0.0% (0)	0.0%
	2	13	48	14	0.5% (0)	0.0%	85.5% (12)	87.5%	14.0% (2)	12.5%	0.0% (0)	0.0%
	3	13	85	28	0.6% (0)	0.0%	37.7% (10)	32.9%	61.7% (17)	65.9%	0.0% (0)	1.2%
	4	13	49	16	0.4% (0)	2.0%	58.2% (9)	49.0%	41.4% (7)	49.0%	0.0% (0)	0.0%
11	1	24	1082	179	34.7% (62)	29.3%	64.6% (116)	70.5%	0.7% (1)	0.2%	0.0% (0)	0.0%
	2	24	122	21	5.1% (1)	2.5%	87.5% (18)	88.5%	7.4% (2)	9.0%	0.0% (0)	0.0%
	3	24	315	50	0.6% (0)	0.0%	39.8% (20)	31.8%	59.5% (30)	68.3%	0.1% (0)	0.0%
	4	23	155	25	0.9% (0)	0.7%	39.1% (10)	34.2%	55.0% (14)	56.8%	5.1% (2)	8.4%

Table Read: For grade 3, Claim 1, 15 reviewers rated a total of 336 mathematics CAT items, with reviewers rating an average of 111 items each. For DOK level 1, reviewers rated an average of 66% of the items (73 out of 111) as falling within the range of DOK level 1 compared to an intended 65.5% of the items falling within that range. For DOK level 2, reviewers rated an average of 33.8% of the items (37 out of 111) as falling within the range of DOK level 2 compared to an intended 34.5% of items falling within that range. For DOK level 3, reviewers rated an average or .2% of the items (0 out of 111) as falling within the DOK level 3 compared to an intended 0% of items falling within that range. For DOK level 4, reviewers rated an average of 0% of the items (0 out of 111) as falling within the range of DOK level 4 compared to an intended 0% of the items falling within that range.

DOK Consistency

Analyses were conducted to examine the consistency of DOK levels between mathematics CAT items and the Content Specifications, and to address the following question:

- G.DC-1: Does the DOK of the item identified by the reviewers fall within the DOK distribution of the aligned target identified in the Content Specifications?

Pairwise Agreement

Because DOK ratings used for DOK consistency are consistent with those used above for DOK distribution, mathematics CAT item DOK pairwise comparisons are presented in the Connection G, DOK distribution section.

Findings

D.DC-1 (CAT): Is the cognitive complexity required in the items consistent with the cognitive complexity required in each assessment target?

Across grades and claims, reviewers rated at least half of the mathematics CAT items (53.3% – 100%) as falling within the range of DOK level for the intended target (see Table 5.G.18). For grades 3, 4, and 5, reviewers consistently believed fewer mathematics Claim 4 (Modeling and Data Analysis) items fell within the range of DOK level of the intended target and consistently believed these items' DOK level was lower than the range specified for the intended target. For grades 6, 7, 8, and 11, reviewers consistently believed fewer mathematics Claim 3 (Communicating Reasoning) items fell within the range of the DOK level of the intended target and consistently believed these items' DOK level was higher than the range specified for the intended target.

Table 5.G.19. Average Percentage of Mathematics CAT Items Rated as Having DOK Levels Consistent and Inconsistent with Intended Range of Mapped Target

Grade	Claim	# of Reviewers	# CAT Items	Ave # CAT Items Per Reviewer	Consistent	Inconsistent	
					Avg CAT items falling within the range of the intended Target %(n)	Avg CAT items with DOKs higher than the highest of the intended Target range %(n)	Avg CAT items with DOKs lower than the lowest of the intended Target range %(n)
3	1	15	336	111	98.1% (109)	1.2% (1)	0.7% (1)
	2	15	49	17	98.1% (16)	1.1% (0)	0.8% (0)
	3	15	82	28	85.3% (24)	4.2% (1)	10.5% (3)
	4	14	51	17	71.5% (13)	0.4% (0)	28.2% (5)
4	1	14	312	103	99.9% (103)	0.1% (0)	0.0% (0)
	2	14	55	20	90.6% (18)	7.8% (2)	1.5% (0)
	3	14	77	26	76.0% (19)	6.8% (2)	17.3% (5)
	4	14	70	24	53.3% (13)	4.1% (1)	42.6% (10)
5	1	15	284	88	94.7% (83)	4.4% (4)	0.9% (1)
	2	13	53	20	92.8% (18)	1.1% (0)	6.1% (1)
	3	13	91	30	80.0% (24)	9.7% (3)	10.3% (3)
	4	13	66	22	63.8% (15)	0.8% (0)	35.5% (8)
6	1	13	314	105	94.3% (99)	2.3% (2)	3.4% (4)
	2	13	58	18	97.4% (18)	0.8% (0)	1.7% (0)
	3	13	80	23	82.9% (20)	12.6% (3)	4.5% (1)
	4	13	45	13	97.5% (13)	0.0% (0)	2.5% (0)

Table 5.G.19. (Continued)

Grade	Claim	# of Reviewers	# CAT Items	Ave # CAT Items Per Reviewer	Consistent	Inconsistent	
					Avg CAT items falling within the range of the intended Target %(n)	Avg CAT items with DOKs higher than the highest of the intended Target range %(n)	Avg CAT items with DOKs lower than the lowest of the intended Target range %(n)
7	1	13	279	92	94.5% (87)	4.3% (4)	1.2% (1)
	2	13	44	15	97.6% (15)	0.0% (0)	2.4% (0)
	3	13	75	26	83.4% (22)	13.9% (4)	2.7% (1)
	4	13	31	11	95.2% (11)	0.4% (0)	4.4% (1)
8	1	13	276	91	79.3% (73)	9.2% (8)	11.4% (11)
	2	13	48	14	99.5% (14)	0.0% (0)	0.5% (0)
	3	13	85	28	80.8% (22)	14.3% (4)	5.0% (1)
	4	13	49	16	93.3% (15)	0.0% (0)	6.7% (1)
11	1	24	1082	179	91.4% (164)	6.5% (12)	2.1% (4)
	2	24	122	21	96.7% (20)	0.8% (0)	2.6% (1)
	3	24	315	50	78.6% (41)	12.9% (5)	8.5% (4)
	4	23	155	25	90.8% (23)	5.7% (2)	3.5% (1)

Table Read: For grade 3, Claim 1, 15 reviewers rated 336 mathematics CAT items, with reviewers rating an average of 111 items each. Reviewers rated 98.1% of the items (109 out of 111) as falling within the range of DOK levels associated with the intended target. Reviewers rated only a couple of the items as outside the range of DOK levels for the intended target—1.2% of the items (1 out of 111) were rated as having a DOK higher than the highest DOK for the intended target and 0.7% of the items (1 out of 111) was rated as having a DOK lower than the lowest DOK for the intended target.

ELA/Literacy PT Items

To address questions for this connection as well as Connection D (Alignment of Item/Task Pools and Evidence Statements), the ELA/literacy PT analyses were based on ratings provided for three PTs for grades 3 – 8 and six PTs for grade 11, across two workshops. Each ELA/literacy PT was comprised of four individual items. Reviewers viewed the entire PT (just as a student would view the item), while providing ratings separately for each corresponding item. Table 5.G.20 provides a summary of the number of reviewers who provided PT item ratings at each workshop and the number of PTs they rated. In addition, the table provides the total number of PTs and individual PT items rated during the workshops.

For all Connection D and G analyses, averages were computed by first finding the average percentage of items by grade, PT, and rater, and then averaging the rater averages by PT to obtain a PT level statistic. Finally, the values were averaged by grade. Readers should keep in mind that only a limited number of PTs were reviewed and rated and the results are not necessarily representative of the entire item pool; therefore, generalizations beyond the PTs included might not be appropriate.

Table 5.G.20. Numbers of ELA/Literacy Reviewers and ELA/Literacy PTs Rated, by Workshop and Overall

Grade/ Group	Claim	# of Workshop 4 Reviewers	Total Reviewers	# Workshop 4 PTs	# Workshop 5 PTs	# Items Per PT	Total PT Items
3	1	5	5	1	2	4	12
4	1	5	5	1	2	4	12
5	1	5	5	1	2	4	12
6	1	5	5	1	2	4	12
7	1	5	5	1	2	4	12
8	1	5	5	1	2	4	12
11 - Group 1	1	5	5	1	2	4	12
11 - Group 2	1	4	4	1	2	4	24

Content Representation

Analyses were conducted to examine the representation of ELA/literacy content between the ELA/literacy PT items and the Content Specifications, and to address the following questions:

- G.CR-1: How are the summative assessment items distributed across targets and grade-level standards?
- G.CR-2: Do the reviewers agree with the intended mapping of items to targets and grade-level standards, as identified by the item developers?

Pairwise Agreement

Tables 5.G.21 and 5.G.22 present the pairwise agreement among reviewers' ratings of ELA/literacy PT items and targets and reviewers' ratings of ELA/literacy PT items and grade-level standards, respectively. For each item, the pairwise agreement was calculated by determining the percent of pairs of reviewers who agreed on their ratings. These percentages were then averaged to obtain the values presented.

The average pairwise agreement among reviewers for items within a claim was found to be between 89.2% (grades 7 and 8) and 100% (grade 11) (Table 5.G.21).

Table 5.G.21. Average Pairwise Comparison of ELA/Literacy PT Item and Target Ratings among Reviewers, by Grade

Grade	# Reviewers	# PTs	# items per PT	# items	Avg Pairwise Agreement
3	10	3	4	12	93.3%
4	10	3	4	12	96.7%
5	10	3	4	12	96.7%
6	8	3	4	12	90.8%
7	10	3	4	12	89.2%
8	10	3	4	12	89.2%
11	17	6	4	24	100.0%

Table Read: At grade 3, 10 reviewers across two workshops rated items for 3 PTs. Each PT included 4 items, for a total of 12 items. The average pairwise agreement among reviewers for target mappings to ELA/literacy PT items was 93.3%.

As shown in Table 5.G.22, there was an average of 70% (grade 4) to 82.5% (grade 8) pairwise agreement among reviewer ratings for the ELA/literacy PT item and grade-level standards mapping, as verified by the reviewers.

Table 5.G.22. Average Pairwise Comparison of ELA/Literacy PT Item and Grade-Level Standard Ratings among Reviewers, by Grade

Grade	# PTs	# of Reviewers	# items	Avg # of Items per PT	Avg Pairwise Agreement
3	3	10	12	4	78.3%
4	3	10	11	4	70.0%
5	3	10	12	4	68.6%
6	3	8	11	4	70.9%
7	3	10	11	4	81.8%
8	3	10	8	3	82.5%
11	6	17	23	4	74.9%

Table Read: At grade 3, 10 reviewers across two workshops rated items for 3 PTs. Each PT included an average of 4 items, for a total of 12 items. The average pairwise agreement among reviewers for grade-level standard mappings to ELA/literacy PT items was 78.3%

Findings

G.CR-1 (PT): How are the summative assessment items distributed across assessment targets?

Table 5.G.23 presents the average number of items mapped to each target; only those targets identified as having at least one PT item mapped to it was included. The table also presents the minimum and maximum number of items mapped to one target, by grade. On average, 1.5 – 2 ELA/literacy PT items were mapped to a single target.

Table 5.G.23. Average, Minimum, and Maximum Number of ELA/literacy PT Items Mapped to Each Target, Averaged across Reviewers, by Grade and Claim[39]

Grade	# PTs	Avg # Items per PT	# of Reviewers	Avg # PT Items Per Target	Min Avg PT items Per PT Per Target	Max Avg PT items Per Target
3	3	4	10	1.8	1	4
4	3	4	10	1.5	1	3
5	3	4	10	2.0	1	3
6	3	4	8	1.6	1	4
7	3	4	10	1.6	1	3
8	3	4	10	1.6	1	3
11	6	4	17	1.5	1	3

Table Read: At grade 3, there were 3 PTs, with an average of 4 items per PT. These items were rated by 10 reviewers across workshops. Reviewers identified an average of 1.8 items per target, with a minimum of 1 item and a maximum of 4 items mapped to an individual target.

G.CR-2 (PT): Do the reviewers agree with the intended mapping of items to targets and grade-level standards as identified by the item developers?

For all grades, reviewers typically rated most of the ELA/literacy items within a PT as being fully aligned to the intended target (see Table 5.G.24). This was especially true for grades 6 – 8 and 11, where reviewers typically rated more than 90% of the items within a performance task as being fully aligned to its intended target. When comparing the alignment of items to targets and the alignment of items to evidence statements (Connection D), it is apparent that target alignment is better. This could be due to various factors. First, reviewers provided independent ratings of evidence statements to each item while they verified the intended target ratings. Second, only one target was intended per item, whereas as many as five evidence statements were intended per item. Therefore, the task of conducting the alignment to evidence statements was likely a more difficult task for reviewers.

[39] Averages based on only targets with at least one item mapped, as indicated by reviewers.

Table 5.G.24. Average Percentage of ELA/Literacy PTs Rated as Fully Aligned, Partially Aligned, or Not Aligned to Intended Target, by Grade

					Target Verification Rating		
Grade	# PTs	# of Reviewers	# PT items	# Items per PT	Fully Aligned %(n)	Partially Aligned %(n)	Not Aligned %(n)
3	3	10	12	4	81.7% (3)	15.0% (1)	3.3% (0)
4	3	10	12	4	77.8% (3)	20.6% (1)	1.7% (0)
5	3	10	12	4	98.3% (4)	0.0% (0)	1.7% (0)
6	3	8	12	4	91.3% (4)	3.8% (0)	5.0% (0)
7	3	10	12	4	90.0% (4)	1.7% (0)	8.3% (0)
8	3	10	12	4	95.0% (4)	0.0% (0)	5.0% (0)
11	6	17	24	4	97.1% (4)	2.9% (0)	0.0% (0)

Table Read: For grade 3, there were 10 reviewers who rated a total of 12 ELA/literacy items associated with 3 PTs, with 4 items per PT. On average, 81.7% of the items (or approximately 3 items), within each PT were rated as being fully aligned to its intended target.

As presented in Table 5.G.25, across grades and claims, reviewers rated the majority of ELA/literacy items within a PT as being fully aligned to their intended grade-level standard(s). The percentages were lower at grade 8, where an average of 58.3% of the items within a PT were rated as being fully aligned and 33.9% of the items within a PT were rated as not being aligned to the intended grade-level standard(s).

Table 5.G.25. Average Percentage of ELA/Literacy PTs Rated as Fully Aligned, Partially Aligned, or Not Aligned to Intended Grade-level Standard(s), by Grade

						Grade-level Standard Verification Rating		
Grade	# PTs	# of Reviewers	# PT Items	Avg # Items per PT	# of items Excluded*[40]	Fully Aligned %(n)	Partially Aligned %(n)	Not Aligned %(n)
3	3	10	12	4	0	86.7% (3)	13.3% (1)	0.0% (0)
4	3	10	11	4	1	84.4% (3)	10.6% (0)	5.0% (0)
5	3	10	11	4	0	78.3% (3)	16.7% (1)	5.0% (0)
6	3	8	11	4	0	73.8% (3)	19.6% (1)	6.7% (0)
7	3	10	11	4	0	85.0% (3)	11.7% (0)	3.3% (0)
8	3	10	8	3	2	58.3% (2)	13.9% (0)	27.8% (1)
11	6	17	23	4	1	78.7% (3)	17.2% (1)	4.2% (0)

Table Read: For grade 3, there were 10 reviewers who rated a total of 12 ELA/literacy items across 3 PTs. Each PT included 4 items. No items were excluded due to errors in metadata. On average, reviewers rated 86.7% of the ELA/literacy items within a PT as fully aligned to their intended grade-level standard(s) (an average of 3 items per PT), 13.3% of the items as partially aligned to their intended grade-level standard(s) (an average of 1 item), and 0% of the items as not aligned to their intended grade-level standard(s) (an average of 4 items).

[40] Some items were removed due to errors in grade-level standard meta-data.

DOK Distribution

Analyses were conducted to examine the distribution of DOK levels between the ELA/literacy PT items and the Content Specifications, and to address the following question:

- G.DD-1: How does the distribution of DOK of the items identified by the reviewers compare with the distribution identified in the Content Specifications?

Pairwise Agreement

Based on Table 5.G.26, the average item pairwise agreement among reviewers' ratings of DOK levels for ELA/literacy PT items ranged from 70% (grade 5) to 93.3% (grade 3).

Table 5.G.26. Average Pairwise Comparison of Reviewer Ratings for DOK Levels of ELA/Literacy PT Items, by Grade

Grade	# PTs	# of Reviewers	# PT items	Avg # of Items per PT	Avg Pairwise Agreement
3	3	10	12	4	93.3%
4	3	10	12	4	88.3%
5	3	10	12	4	70.0%
6	3	8	12	4	78.1%
7	3	10	12	4	90.0%
8	3	10	12	4	83.3%
11	6	17	24	4	91.3%

Table Read: At grade 3, 10 reviewers across two workshops rated items from 3 PTs. Each PT included 4 items, for a total of 12 items. The average pairwise agreement among reviewers' ratings for the DOK level of the ELA/literacy PT items was 93.3%

Findings

D.DD-1 (CAT): How does the distribution of DOK of the items identified by the reviewers compare with the distribution identified in the Content Specifications?

Reviewers mostly agreed with the DOK level of the ELA/literacy PT items, as indicated in the Content Specifications; however, reviewers were slightly more likely to rate an item at DOK level 4 than what was indicated in the Content Specifications, this was particularly true for grades 6 and 7 (see Table 5.G.27). Additionally, reviewers rated a very small percentage of items per PT at DOK levels and 2, whereas the Content Specifications indicated these items were at DOK levels 3 and 4.

Table 5.G.27. Distribution of ELA/Literacy PTs across DOK Levels, Average Percentage of Items Rated per PT, and Average Percentage of Items per PT as Indicated in Content Specifications

Grade	# of PTs	# of Reviewers	# PT Items	# Items per PT	DOK 1		DOK 2		DOK 3		DOK 4	
					Avg Items per Rater %(n)	Avg Items Intended %	Avg Items per Rater %(n)	Avg Items Intended %	Avg Items per Rater %(n)	Avg Items Intended %	Avg Items per Rater %(n)	Avg Items Intended %
3	3	10	12	4	0.0% (0)	0.0%	1.7% (0)	0.0%	55.0% (2)	58.3%	43.3% (2)	41.7%
4	3	10	12	4	0.0% (0)	0.0%	1.7% (0)	0.0%	46.7% (2)	50.0%	51.7% (2)	50.0%
5	3	10	12	4	1.7% (0)	0.0%	1.7% (0)	0.0%	51.7% (2)	58.3%	45.0% (2)	41.7%
6	3	9	12	4	1.7% (0)	0.0%	2.1% (0)	0.0%	37.6% (2)	50.0%	58.6% (2)	50.0%
7	3	10	12	4	0.0% (0)	0.0%	0.0% (0)	0.0%	28.3% (1)	33.3%	71.7% (3)	66.7%
8	3	10	12	4	0.0% (0)	0.0%	0.0% (0)	0.0%	31.7% (1)	33.3%	68.3% (3)	66.7%
11	6	17	12	4	0.0% (0)	0.0%	1.7% (0)	0.0%	30.6% (1)	33.3%	67.7% (3)	66.7%

Table Read: For grade 3, there were 3 ELA/literacy PTs rated with 4 items each, for a total of 12 items. Ten reviewers provided ratings each for one or two PTs across two workshops. On average, reviewers believed 0% of the items within a PT were at DOK level 1 and the Content Specifications indicated the same percentage of items at this level. Reviewers believed an average of 1.7% of items within a PT were at DOK level 2 while the Content Specifications indicated 0% of items were at this level. Reviewers rated an average of 55% of items (average 2 items per PT, per rater) within a PT at DOK level 3, compared to the Content Specifications indicating 58.3% of the items at this level. Reviewers rated an average of 43.3% of the items at DOK level 4, compared to the Content Specifications indicating 41.7% of the items at this level.

DOK Consistency

Analyses were conducted to examine the consistency of DOK levels between ELA/literacy PT items and the Content Specifications, and to address the following question:

- G.DC-1: Does the DOK of the item identified by the reviewers fall within the DOK distribution of the aligned target identified in the Content Specifications?

The DOK ratings used for this analysis were the same as those used to analyze the DOK distribution of the ELA/literacy PT items and, therefore, are not duplicated here. The reader is referred to the earlier report section that describes the ELA/literacy PT item DOK distribution for information about reviewer pairwise agreement.

Findings

D.DC-1 (PT): Is the cognitive complexity required in the items consistent with the cognitive complexity required in each assessment target?

Across grades, reviewers believed the DOK level for the vast majority of the ELA/literacy PT items fell within the range of DOK levels for the intended target. Table 5.G.27 shows that the reviewers rated the highest percentage of grade 3 (96.3%) PT items and the lowest percentage of PT items at grade 5 (86.5%) as falling within the range of the target. Reviewers rated essentially no PT items as having a higher or lower DOK level than that indicated by the Content Specifications.

Table 5.G.28. Average Percentage of ELA/Literacy PT Items Rated as Having DOK Levels Consistent and Inconsistent with Intended Range of the Intended Target

Grade	# of PTs	# of Reviewers	# of PT Items	# Items per PT	Consistent	Inconsistent	
					Avg PT items falling within the range of the intended Target %(n)	Avg PT items with DOKs higher than the highest of the intended Target range %(n)	Avg PT items with DOKs lower than the lowest of the intended Target range %(n)
3	3	10	12	4	96.3% (4)	1.3% (0)	2.5% (0)
4	3	10	12	4	93.8% (4)	2.5% (0)	3.8% (0)
5	3	10	12	4	86.3% (3)	6.3% (0)	7.5% (0)
6	3	9	12	4	87.5% (4)	6.9% (0)	5.6% (0)
7	3	10	12	4	95.0% (4)	5.0% (0)	0.0% (0)
8	3	10	12	4	90.0% (4)	10.0% (0)	0.0% (0)
11	6	17	24	4	95.6% (4)	1.5% (0)	2.9% (0)

Mathematics PT Items

Table 5.G.29 presents the number of reviewers who rated mathematics PT items. These reviewers are the same as those who rated the mathematics CAT items. At Workshop 4, each group rated one PT and at Workshop 5 each group rated 2 PTs. Each mathematics PT included 6 individual items. The analyses for mathematics PTs were based on the ratings provided for the items associated with each PT. While we were able to include a large, representative sample of CAT items, this was not the case with PTs; therefore, caution should be used when interpreting the findings for mathematics PT items.

Table 5.G.29. Numbers of Mathematics Reviewers and Mathematics PT Items Rated, by Workshop and Overall

Grade/ Group	# of Workshop 4 Reviewers	# of Workshop 5 Reviewers	Total Reviewers	# Workshop 4 PTs	# Workshop 5 PTs	# Items Per PT	Total PT Items
3	5	5	10	1	2	6	18
4	5	5	10	1	2	6	18
5	5	5	10	1	2	6	18
6	5	5	10	1	2	6	18
7	5	4	9	1	2	6	18
8	4	4	8	1	2	6	18
11 - Group 1	4	5	9	1	2	6	18
11 - Group 2	3	4	7	1	2	6	36

Content Representation

Analyses were conducted to examine the representation of mathematics content between the PT items and the Content Specifications, and to address the following questions:

- G.CR-1: How are the summative assessment items distributed across targets, grade-level standards, and mathematical practices?
- G.CR-2: Do the reviewers agree with the intended mapping of items to targets and grade-level standards, as identified by the item developers?
- G.CR-3: Do the reviewers agree with the intended mapping of items to mathematical practices, as identified by the item developers?

Pairwise Agreement

Tables 5.G.30, 5.G.31, and 5.G.32 present the pairwise agreement for reviewers' ratings of mathematics PT items and targets, reviewers' ratings of mathematics PT items and grade-level standards, and reviewers' ratings of mathematics PT items and mathematical practices, respectively. For each item, the pairwise agreement was calculated by determining the percent of pairs of

reviewers who agreed on their ratings. These percentages were then averaged to obtain the values presented.

The average pairwise agreement among reviewer ratings of mathematics PT item mappings to targets is presented in Table 5.G.30. The average pairwise agreement among reviewers' ratings for PT items within a claim ranged from 67.6% (grade 7) to 95.6% (grade 11).

Table 5.G.30. Average Pairwise Comparison of Mathematics PT Item and Target Ratings among Reviewers, by Grade

Grade	# Reviewers	# PTs	# items per PT	# PT Items	Avg Pairwise Agreement
3	10	3	6	18	81.7%
4	10	3	6	18	92.2%
5	10	3	6	18	87.2%
6	10	3	6	18	94.4%
7	8	3	6	18	67.6%
8	8	3	6	18	79.6%
11	18	6	6	36	95.6%

Table Read: At grade 3, 10 reviewers across two workshops rated items from 3 PTs. Each PT included 6 items, for a total of 18 items. The average pairwise agreement among reviewers for rating the mathematics PT items and targets was 81.7%.

Table 5.G.31 presents the average pairwise agreement among reviewer ratings of mathematics PT item mappings to grade-level standards, by grade. The average pairwise agreement among reviewers' ratings for PT items within a claim ranged from 70% (grade 5) to 97.2% (grade 7).

Table 5.G.31. Average Pairwise Comparison of Mathematics PT Item and Grade-level Standard Ratings among Reviewers, by Grade

Grade	# Reviewers	# PTs	# items per PT	# PT Items	Avg Pairwise Agreement
3	10	3	6	18	78.3%
4	10	3	6	18	82.2%
5	10	3	6	18	70.0%
6	10	3	6	18	66.7%
7	8	3	6	18	97.2%
8	8	3	6	18	80.6%
11	18	6	6	36	77.0%

Table Read: At grade 3, 10 reviewers across two workshops rated items from 3 PTs. Each PT included 6 items, for a total of 18 items. The average pairwise agreement among reviewers for rating the mathematics PT items and grade-level standards was 78.3%.

Table 5.G.32 presents the average pairwise agreement among reviewer ratings of mathematics PT item mappings to mathematical practices. The first pairwise agreement among reviewers reflects whether the primary mathematical practice the reviewers identified matched an intended mathematical practice. The second pairwise agreement among reviewers reflects whether at least one of the mathematical practices that the reviewers identified (primary or additional) matched an intended mathematical practice. As shown, the pairwise agreement was fairly similar regardless of whether a primary mathematical practice matched or at least one of the mathematical practices matched. The percentages across grades ranged between 65% – 82%.

Table 5.G.32. Average Pairwise Comparison of Mathematics PT Item and Mathematical Practice Ratings among Reviewers, by Grade

Grade	# Reviewers	# PTs	Avg # items per PT	# PT Items	Avg Pairwise Agreement for Primary Match Ratings	Avg Pairwise Agreement for At Least One Match Ratings
3	10	3	6	17	65.3%	70.0%
4	10	3	6	18	80.0%	77.8%
5	10	3	6	18	67.8%	73.3%
6	10	3	4	13	70.8%	81.5%
7	8	3	6	18	72.6%	66.9%
8	8	3	4	12	69.4%	81.9%
11	18	6	6	36	77.1%	71.8%

Table Read: At grade 3, 10 reviewers across two workshops rated items from 3 PTs. Each PT included 6 items, for a total of 17 items. The average pairwise agreement among reviewers' ratings for matching the primary mathematical practice was 65.3%. The average pairwise agreement among reviewers' ratings for including at least one of the same mathematical practices was 70.0%.

Findings

G.CR-1 (PT): How are the summative assessment items distributed across assessment targets?

Table 5.G.33 presents the average number of mathematics PT items mapped to each target. The last two columns of this table present the average minimum number of items mapped to any one target within the claim and the average maximum number of items assigned to a target within the claim. Targets included in this analysis were those for which the reviewers identified as having at least one PT item mapped to it. Typically, reviewers rated an average of 1.4 – 1.8 PT items as mapping to a single target.

Table 5.G.33. Average, Minimum, and Maximum Number of Mathematics PT Items Mapped to Each Target, Averaged across Reviewers, by Grade and Claim[41]

Grade	# PTs	Avg # Items per PT	# of Reviewers	Avg # PT Items Per Target	Min Avg PT items Per PT Per Target	Max Avg PT items Per Target
3	3	6	10	1.6	1	4
4	3	6	10	1.8	1	5
5	3	6	10	1.7	1	3
6	3	6	10	1.7	1	4
7	3	6	8	1.4	1	4
8	3	6	8	1.7	1	4
11	6	6	18	1.6	1	6

Table Read: At grade 3 there were 3 PTs, with an average of 6 items per PT, and 10 reviewers rated these items. Reviewers identified an average of 1.6 mathematics PT items per target, with a minimum of 1 mathematics PT item at each target and maximum of 4 mathematics PT items mapped to an individual target.

G.CR-2 (PT): Do the reviewers agree with the intended mapping of items to targets and grade-level standards as identified by the item developers?

Reviewers on average rated most of the mathematics items within PTs as being fully aligned to the target that had been identified by the item writers (see Table 5.G.33). The lowest average percentage for items within a PT that were rated as being fully aligned was at grade (76.4%) (see Table 5.G.34).

Table 5.G.34. Average Percentage of Math PTs Rated as Fully Aligned, Partially Aligned, or Not Aligned to Intended Target, by Grade

					Target Verification Rating		
Grade	# PTs	# of Reviewers	# PT Items	# Items per PT	Fully Aligned %(n)	Partially Aligned %(n)	Not Aligned %(n)
3	3	10	18	6	84.4% (5)	4.4% (0)	11.1% (1)
4	3	10	18	6	95.6% (6)	1.1% (0)	3.3% (0)
5	3	10	18	6	93.3% (6)	5.6% (0)	1.1% (0)
6	3	10	18	6	94.4% (6)	3.3% (0)	2.2% (0)
7	3	8	18	6	76.4% (5)	12.5% (1)	11.1% (1)
8	3	8	18	6	88.9% (5)	8.3% (1)	2.8% (0)
11	6	18	36	6	93.2% (6)	6.8% (0)	0.0% (0)

Table Read: For grade 3, there were 10 reviewers who rated a total of 18 items associated with 3 mathematics PTs, with 6 items per PT. On average 84.4% of items (5 items) within each PT were rated as fully aligned; 4.4% items (0 items) within each PT were rated as partially aligned; and 11.1% of items (1 item) within each PT were rated as not aligned to the intended target.

[41] Averages based on only targets with at least one item mapped, as indicated by reviewers.

On average, reviewers rated the majority of mathematics items within a PT as being fully aligned to the mapped grade-level standard. As Table 5.G.34 shows, reviewers rated only a small percentage of mathematics items within PTs (0% – 18.9%) as not being aligned to the intended grade-level standard.

Table 5.G.35. Average Percentage of Mathematics PTs Rated as Fully Aligned, Partially Aligned, or Not Aligned with Intended Grade-level Standard, by Grade

					Grade-level Standard Verification Rating		
Grade	# PT	# of Reviewers	# PT Items	# Items per PT	Fully Aligned %(n)	Partially Aligned %(n)	Not Aligned %(n)
3	3	10	18	6	75.6% (5)	5.6% (0)	18.9% (1)
4	3	10	18	6	87.8% (5)	7.8% (0)	4.4% (0)
5	3	10	18	6	71.1% (4)	18.9% (1)	10.0% (1)
6	3	10	18	6	70.0% (4)	28.9% (2)	1.1% (0)
7	3	8	18	6	98.6% (6)	1.4% (0)	0.0% (0)
8	3	8	18	6	88.9% (5)	8.3% (1)	2.8% (0)
11	6	18	36	6	81.1% (5)	10.3% (1)	8.6% (0)

Table Read: For grade 3, there were 10 reviewers who rated a total of 18 mathematics items across 3 PTs. Each PT included 6 items. On average, reviewers rated 75.6% of the mathematics items within a PT as fully aligned to their intended grade-level standard (average of 5 items per PT), 5.6% of the items partially aligned to their intended grade-level standard (average of 0 items per PT), and 18.9% of the items within a PT rated as not aligned to their intended grade-level standard.

G.CR-3(PT): Do the reviewers agree with the intended mapping of items to mathematical practices as identified by the item developers?

Across grades, reviewers tended to agree with the item writers about the mapping of mathematical practices to the mathematics PT items (see Table 5.G.36). The lowest agreement occurred for grades 6 and 7, where reviewers rated less than half the PT items' primary mathematical practice as matching the mathematical practice identified by the item writers. There was an increase in agreement of reviewer ratings for identifying at least one mathematical practice that had also been identified by the item writers. The highest percentage for reviewers not agreeing with the mathematical practices identified by the item writers was at these same grades (grade 6, 42.2% and grade 7, 43.1%).

Table 5.G.36. Average Percentage of Mathematics Items within a PT with Reviewer Identified Mathematical Practices Mapped to Intended Mathematical Practices Identified by Item Writers, by Grade

Grade	# of PTs	# of Reviewers	# PT Items	# Items per PT	Mathematical Practice Mapping				
					Avg # MPs rated per PT item	Avg # MPs intended per PT item	Reviewer Primary Mathematical Practice matched an Intended Mathematical Practice	At Least One Rating Matched At Least One Intended Mathematical Practices	No Agreement with Intended Mathematical Practice(s)
3	3	10	17	6	1.5	2.9	67.8% (4)	75.6% (4)	24.4% (1)
4	3	10	18	6	1.5	2.7	75.6% (5)	86.7% (5)	13.3% (1)
5	3	10	18	6	2.2	3.2	78.9% (5)	84.4% (5)	15.6% (1)
6	3	10	13	4	1.7	2.9	43.3% (3)	57.8% (3)	42.2% (1)
7	3	8	18	6	1.7	2.4	44.4% (3)	56.9% (3)	43.1% (3)
8	2	8	12	6	1.8	2.9	70.8% (4)	89.6% (5)	10.4% (1)
11	6	18	36	6	2.2	2.0	51.5% (3)	64.8% (4)	35.2% (2)

Table Read: At grade 3, 10 reviewers across 2 workshops rated 17 items within 3 PTs, for 6 items per PT. Reviewers identified an average of 1.5 mathematical practices to each mathematics PT item, compared to an average of 2.9 mathematical practices identified by item writers. Reviewers identified the same mathematical practice as did the item writers for the item's primary mathematical practice for an average of 67.8% of the items (or an average of 4 items per reviewer, per PT). At least one of the mathematical practices the reviewers identified matched the mathematical practice identified by the item writer for an average of 75.6%% of the items per PT (or an average of 4 items per reviewer, per PT). Reviewers did not agree with the mathematical practice identified by the item writer for 24.4% of the mathematics PT items (or an average of 1 item per reviewer).

DOK Distribution

Analyses were conducted to examine the distribution of DOK levels between the mathematics PT items and the Content Specifications, and to address the following question:

- G.DD-1: How does the distribution of DOK of the items identified by the reviewers compare with the distribution identified in the Content Specifications?

To address this question, reviewers were asked to verify the DOK level of the mathematics PT items that was indicated in the Content Specifications. If the reviewers disagreed with the DOK level indicated, they provided a DOK level they believed to be more appropriate to that item. In those instances, the alternate DOK level(s) provided by the reviewers were used in this analysis.

Pairwise Agreement

Table 5.G.37 presents the average item pairwise agreement among reviewers' DOK ratings for mathematics PT items. Across grades and claims, reviewer agreement was generally high. The exception was at grade 6, where the average pairwise agreement was 68.3%.

Table 5.G.37. Average Pairwise Comparison of Mathematics PT Item DOK Ratings among Reviewers, by Grade

Grade	# Reviewers	# PTs	# Items per PT	# PT Items	Avg Pairwise Agreement
3	10	3	6	18	95.6%
4	10	3	6	18	90.0%
5	10	3	6	18	82.2%
6	10	3	6	18	68.3%
7	8	3	6	18	96.3%
8	8	3	6	18	94.4%
11	18	6	6	36	80.0%

Table Read: At grade 3, 10 reviewers rated items from 3 PTs and each PT included 6 items, for a total of 18 items. The average pairwise agreement among reviewers for mathematics PT item DOK level ratings was 95.6%.

Findings

D.DD-1 (PT): How does the distribution of DOK of the items identified by the reviewers compare with the distribution identified in the Content Specifications?

As can be seen in Table 5.G.38, reviewers tended to agree with how the mathematics PT items were distributed in terms of their DOK level, as indicated in the Content Specifications. Across grades, reviewers rated most of the mathematics PT items at DOK levels 2 and 3. Except at grade 3, reviewers identified a very small percentage of items within a PT at DOK level 1 and/or DOK level 4 than what was indicated in the Content Specifications.

Table 5.G.38. Distribution of Mathematics PT Items across DOK Levels, Average Percentage of Items per PT Rated, and Average Percentage of Items per PT as Indicated in Content Specifications

Grade	# of PTs	# of Reviewers	# PT Items	# PT Items per PT	DOK 1		DOK 2		DOK 3		DOK 4	
					Avg PT Items per Rater %(n)	Avg PT Items Intended %	Avg PT Items per Rater %(n)	Avg PT Items Intended %	Avg PT Items per Rater %(n)	Avg PT Items Intended %	Avg PT Items per Rater %(n)	Avg PT Items Intended %
3	3	10	18	6	0.0% (0)	0.0%	47.8% (3)	50.0%	52.2% (3)	50.0%	0.0% (0)	0.0%
4	3	10	18	6	0.0% (0)	0.0%	58.9% (4)	66.7%	35.6% (2)	33.3%	5.6% (0)	0.0%
5	3	10	18	6	4.4% (0)	0.0%	44.4% (3)	50.0%	48.9% (3)	50.0%	2.2% (0)	0.0%
6	3	10	18	6	10.0% (1)	5.6%	65.6% (4)	83.3%	20.0% (1)	5.6%	4.4% (0)	0.0%
7	3	8	18	6	8.3% (1)	5.6%	58.3% (4)	61.1%	33.3% (2)	33.3%	0.0% (0)	0.0%
8	3	8	18	6	1.4% (0)	0.0%	72.2% (4)	72.2%	26.4% (2)	27.8%	0.0% (0)	0.0%
11	6	18	18	6	1.1% (0)	0.0%	27.5% (2)	22.2%	70.3% (4)	77.8%	1.1% (0)	0.0%

Table Read: For grade 3, there were 3 mathematics PTs with 6 items each, for a total of 12 items. These items were rated by 10 reviewers, with each reviewer rating one or two PTs. Reviewers and the Content Specifications both believed 0.0% of the items within a PT were at DOK level 1. Reviewers rated an average of 47.8% of items within a PT at DOK level 2, compared to 50% of items indicated being at that level by the Content Specifications. Reviewers rated an average of 52.2% of items (average 3 items per PT, per reviewer) within a PT at DOK level 3, compared to 50% of the items indicated at that level in the Content Specifications. Reviewers and the Content Specifications both believed 0.0% of the items within a PT were at DOK level 4.

DOK Consistency

Analyses were conducted to examine the consistency of DOK levels between mathematics PT items and the Content Specifications, and to address the following question:

- G.DC-1: Does the DOK of the item identified by the reviewers fall within the DOK distribution of the aligned target identified in the Content Specifications?

Pairwise Agreement

Because DOK ratings used for DOK consistency are consistent with those used above for DOK distribution, mathematics PT item DOK pairwise comparisons are presented in the Connection G, DOK distribution section.

Findings

D.DC-1 (PT): Is the cognitive complexity required in the items consistent with the cognitive complexity required in each assessment target?

Reviewers' ratings for the DOK levels of the mathematics PT items were generally consistent with the DOK range of their mapped target. As shown in Table 5.G.39, average reviewers' ratings were lowest at grade 11 (84.7%) for the DOK level of the mathematics items falling within the range of the intended target and highest at grade 8 (94.8%).

Table 5.G.39. Math.DC-1. Average Percentage of Mathematics PT Items Rated as Having DOK Levels Consistent and Inconsistent with Intended Range of Mapped Target

Grade	# of PTs	# of Reviewers	# PT Items	# of Items per PT	Consistent	Inconsistent	
					Avg PT items falling within the range of the intended Target %(n)	Avg PT items with DOKs higher than the highest of the intended Target range %(n)	Avg PT items with DOKs lower than the lowest of the intended Target range %(n)
3	3	10	18	6	88.3% (5)	8.3% (0)	3.3% (0)
4	3	10	18	6	94.2% (6)	2.5% (0)	3.3% (0)
5	3	10	18	6	92.5% (6)	1.7% (0)	5.8% (0)
6	3	10	18	6	86.7% (5)	7.5% (0)	5.8% (0)
7	3	8	18	6	87.5% (5)	6.3% (0)	6.3% (0)
8	3	8	18	6	94.8% (6)	0.0% (0)	5.2% (0)
11	6	18	36	6	84.7% (5)	13.9% (1)	1.4% (0)

Table Read: At grade 3, there were 3 mathematics PTs with 6 items each, for a total of 18 individual items. There were 10 Reviewers who each rated one or two PTs. Reviewers rated an average of 88.3% of the mathematics PT items (or 5 items per reviewer, per target) within a PT as falling within the range of the intended target. Reviewers rated 8.3% of the items (0 items) as falling above the DOK range of the intended target and 3.3% of the items (0 items) as falling below the DOK range of the intended target.

CHAPTER 6 –SUMMARY OF FINDINGS

Summary of Findings for ELA/Literacy

This study examined the content representation, DOK distribution, DOK consistency, as well as the agreement between the reviewers to the Content Specifications or item developer's ratings, and the agreement among reviewers on their ratings across Connections A through E and G.

Summary of Findings for Connection A: Alignment between Content Specifications and CCSS

- For content representation of Connection A, reviewers aligned more of the grade-level standards to the targets than what was intended. On average, they identified, 11.3 unique grade-level standards per target (SD=7.5) compared to 4.7 unique grade-level standards (SD=3.4) identified in the Smarter Balanced Content Specifications. Reviewers holistic rating to indicate how well identified standards collectively represented the content and knowledge required in the target ranged from 3.5 to 4.0, mostly to fully aligned.

- Reviewers generally believed that the targets represented the content and knowledge required in the grade-level standards. The exceptions were the grade-level standards in the Language and Speaking and Listening strands for grades 3 through 5. For these strands, reviewers rated less than half of the grade-level standards as being fully represented by the targets. The majority of the comments provided by the reviewers regarding this alignment were related to the lack of focus of Speaking in the targets, for which Smarter Balanced does not assess on the summative assessment.

- Overall, there were only a few ELA/literacy targets across claims and grades that did not have grade-level standards with at least 50% reviewer agreement and thus, were not included in the analysis. A fairly large average percentage of the grade-level standards per target rated as matching the intended mapping. Where there wasn't 100%, most of those grade-level standards per target were believed by the reviewers to fall within the intended strand, as specified in the Content Specifications.

- Targets were well represented by the grade-level standards when the reviewers' task was to identify grade-level standards aligned to each target. Reversing the task (i.e., aligning targets to standards) resulted in weaker content representation.

- Targets were well represented by the grade-level standards when the reviewers' task was to identify grade-level standards aligned to each target (Workshop 1). Reversing the task resulted in weaker content representation (Workshop 2). These results likely suggest that the broad nature of the targets made it more difficult to align the targets to specific standards. Workshop 2 activities permitted reviewers to rate a grade-level standard as not represented by any targets; however, most reviewers found targets that represented at least a small amount of the content and knowledge required in the grade-level standards. This resulted in a higher number of grade-level standards being aligned to each target, thus the average percentage of grade-level standards that matched the intended mapping inherently decreased for Workshop 2.

- The overall pairwise agreement in identifying grade-level standards aligned to each target between reviewers and the intended mapping as identified in the Content Specifications

(across all grades, claims, targets, and reviewers) was 36.4%. The rather low agreement was likely a result of the high number of grade-level standards that reviewers identified for each target compared to the number of grade-level standards identified in the Content Specifications. As shown in Table 5.A.10, however, reviewer agreement with the intended mapping increased when computing the average percent of reviewers per target that agreed with at least 50% of the intended standards. This suggests that while there was low overall agreement in identifying exactly what was intended, reviewers generally agreed with at least 50% of the intended standards.

- The overall pairwise agreement among reviewers in identifying DOK levels for each ELA/literacy target (across all grades, claims, targets, and reviewers) was 64.5%

- The overall pairwise agreement in identifying grade-level standards aligned to each target between reviewers and the intended mapping as identified in the Content Specifications (across all grades, claims, targets, and reviewers) was 47.3%. Reviewers usually agreed less often with the intended DOK level of targets in Claim 2 (Writing). This was likely because reviewers believed these targets required higher cognitive demand than what was intended. Except for grades 8 and 11, reviewers agreed most often with the intended DOK level of targets in Claim 1 (Comprehend Literary and Informational Texts). For grade 11, reviewers agreed most often with the intended DOK level of targets in Claim 3 (Speaking and Listening) and for grade 8, reviewers agreed most often with the intended DOK level of targets in Claim 4 (Research and Inquiry).

- When the DOK consistency definition was relaxed to requiring only one DOK level for each grade-level standard mapped to a target to fall within the range of the intended target DOK, the percentage of targets with DOK consistency increased from the more restrictive definition requiring an exact match of the DOK range. This suggests that of the grade-level standards with multiple DOK levels, the reviewers believed that part of the cognitive demand required in the grade-level standard matched that of the intended target, yet they believed there were some portions of the grade-level standards that fell outside that DOK range. For those targets that had inconsistent DOK levels, the general pattern remained that the grade-level standards required higher cognitive demand than what was intended by the target for Claims 1, 3, and 4, and lower levels of cognitive demand for Claim 2.

Summary of Findings for Connection B: Alignment between Evidence Statements and Content Specifications

- For content representation for Connection B, reviewers provided a holistic rating to indicate how well evidence statements collectively represented the content and knowledge required in the target, Reviewers generally rated the targets as being well-represented by the grade-level standards they identified. The mean alignment rating across grades and claims ranged from 3.5 to 4.0. Moreover, they rated the majority of targets as being fully aligned to their collective set of evidence statements (mean percentage of 75.6 – 100%). No clear patterns of alignment emerged across grades and claims. The reviewers generally identified more Claim 2 (Writing) and Claim 3 (Speaking and Listening) targets as being fully represented by the evidence statements; however, across grades, all of the targets were at least partially represented by their collective set of evidence statements.

- We had an expectation that the majority of evidence statements would be rated as 'partially-aligned' to the targets to which they were mapped. This outcome was supported by the data. Reviewers rated the majority of evidence statements as partially aligned to their targets (60.0% – 100%), indicating that an individual evidence statement most often reflected only some of the content and knowledge required in the target to which it was aligned. Reviewers' ratings for evidence statements being fully aligned to their target were typically much less, ranging from 0.0% – 40.0%. Some 'fully aligned' ratings were expected as some targets have only one or two evidence statements aligned to it.

- For the performance task evidence statements, reviewers generally rated targets as being well-represented by the grade-level standards they identified. The mean alignment rating across grades and claims ranged from 3.5 to 4.0.

- Reviewers across grades and claims rated the majority of ELA/literacy PT targets as being fully aligned to their collective set of evidence statements (mean percentage of evidence statements ranged from 85.4% – 100%).

Summary of Findings for Connection C: Alignment between Test Blueprint and Content Specifications

- Reviewers felt that ELA/Literacy blueprints were mostly to fully representative of the content and knowledge that Smarter Balanced outlines to be assessed in the Content Specifications.

Summary of Findings for Connection D: Alignment between Evidence Statements and Item/Task Pools

- When identifying evidence statements for each of the items, reviewers identified most of the possible evidence statements. We examined the extent to which reviewers' ratings of the evidence statements that they identified for each CAT item agreed with the evidence statements that were indicated by the item writers. This agreement was examined in two ways. One way examined the extent to which the evidence statements identified by the reviewers' matched exactly the evidence statements indicated by the item writer. The second way examined the extent to which at least one of the evidence statements that the reviewers' identified matched at least one evidence statement that was indicated by the item writers. For both types of agreement, reviewers' average item-level pairwise agreement was between approximately 57 – 85%; on average, more than half the reviewers agreed with one another that the intended evidence statement mapped to its respective item.

- Across grades and claims, reviewers and item writers generally agreed on the average number of evidence statements that were mapped to items. Typically, more than half the items were found to have an exact match between the evidence statements identified by the reviewers and those identified by the item writers; the exception was at Claim 2 (Writing) for all grades. In some cases, a slightly higher percentage of items were mapped to all of the intended evidence statements even though reviewers could have included additional evidence statements. For about one-third of the grade and claims, the percentage of items where at least one item matched at least one of the intended evidence statements resulted in a slightly higher match percentage compared to when reviewers selected all the intended evidence statements.

Summary of Findings for Connection G: Alignment between Item/Task Pools and Content Specifications

- The majority of reviewers agreed that the ELA/literacy CAT items fully aligned to their intended targets. Grades 3 and 4 had the lowest percentage of items rated as being fully aligned to the targets; however, across grades and claims, at least 87% of the items were rated as being fully aligned to their intended targets. Across grades and claims, reviewers rated the majority of ELA/literacy CAT items as fully aligned to their intended grade-level standards. The percentages were generally higher for the full alignment of items in Claims 2 (Writing) and 3 (Speaking and Listening). The exception was grade 11, which had the lowest average percentage of items (76.6%) fully aligned to their intended grade-level standard.

- Reviewers typically rated most of the ELA/literacy items within a PT as being fully aligned to the intended target. This was especially true for grades 6 – 8 and 11, where reviewers typically rated more than 90% of the items within a performance task as being fully aligned to its intended target.

- Across grades and claims, reviewers rated the majority of ELA/literacy items within a PT as being fully aligned to their intended grade-level standard(s). The percentages were lower at grade 8, where an average of 58.3% of the items within a PT were rated as being fully aligned and 33.9% of the items within a PT were rated as not being aligned to the intended grade-level standard(s).

- The average pairwise agreement among reviewers for the CAT items ranged from 72.8% and 97.9%, and the average item pairwise agreement among reviewers' ratings of DOK levels for ELA/literacy PT items ranged from 70% (grade 5) to 93.3% (grade 3).

- Agreement among reviewers' ratings and the intended was typically high across grades and claims. The lowest pairwise agreement among reviewers was for grade 11, Claim 3 (Speaking and Listening) items, with an average agreement of 63.8%. There was an average of 70% (grade 4) to 82.5% (grade 8) pairwise agreement among reviewer ratings for the ELA/literacy PT item and grade-level standards mapping, as verified by the reviewers.

- Reviewers agreed with the Content Specifications about the distribution of items across DOK levels. However, for DOK level 4, reviewers believed a smaller mean percentage of items were at a DOK level 4 than was intended by the Content Specifications. This was particularly true for grades 6, 7, 8, and 11.

- Reviewers mostly agreed with the DOK level of the ELA/literacy PT items, as indicated in the Content Specifications; however, reviewers were slightly more likely to rate an item at DOK level 4 than what was indicated in the Content Specifications, this was particularly true for grades 6 and 7. Additionally, reviewers rated a very small percentage of items per PT at DOK levels and 2, whereas the Content Specifications indicated these items were at DOK levels 3 and 4.

Summary of Findings for Mathematics

Summary of Findings for Connection A: Alignment between Content Specifications and CCSS

- The holistic rating of the targets as being well-represented by the identified grade-level standards had a mean alignment rating across grades and claims ranged from 3.3 to 3.8, on a scale of 0 to 4.

- Using an independent identification method (where reviewers had access to the full eligible pool of grade-level standards), resulted in reviewers aligning more of the grade-level standards to the targets than what included in the Content Specifications. Across all grades, targets, and reviewers, reviewers identified on average, 10.1 unique grade-level standards per Claim 1 (Concepts and Procedures) target (SD=6.0) and 48.0 unique grade level standards per claim for Claims 2 – 4 (Claim 2, Problem Solving; Claim 3, Communicating Reasoning; Claim 4, Modeling and Data Analysis) (SD=15.8) compared to 3.5 unique grade-level standards (SD=1.3) identified in the Smarter Balanced Content Specifications for Claim 1 targets and 49 unique standards (SD=29.37) per claim for Claims 2 – 4.

- Across grades and claims, the mathematics targets were generally represented well by their intended grade-level standards, especially for Claim 1. The lower percentages of 'fully aligned' targets in Claims 2 – 4 were not necessarily unexpected. Further, across grades and mathematics domains, reviewers strongly believed that most of the grade-level standards were fully aligned and were well represented by the content and knowledge required in the targets. The exception was the High School Trigonometric Functions (F-TF) domain; reviewers rated an average of approximately 10% of the grade-level standards for this domain as being fully aligned; however, the reviewers rated the remaining approximately 90% of the grade-level standards as being mostly or somewhat aligned to that domain.

- Across Claim 1 (Concepts and Procedures) targets, all targets had at least one grade-level standard with at least 50% reviewer agreement and thus, all Claim 1 targets were retained in the analysis. A fairly large average percentage of the grade-level standards per Claim 1 target were rated as matching the intended mapping. Where there wasn't 100%, most of those grade-level standards per target were believed by the reviewers to fall within the intended domains and to a lesser extent, fall within the intended clusters, as specified in the Content Specifications. This pattern remains for Claims 2 – 4 as well. Across grades, Claim 3 (Communicating Reasoning) targets generally had the weakest representation by the grade-level standards at the cluster level.

- The task of identifying targets aligned to each standard was more difficult than identifying grade-level standards that represented the content and knowledge required in the target. The average percentage of grade-level standards per target that matched the intended mapping was approximately 50% across grades for Claim 1 (Concepts and Procedures), with grade 3 having the lowest percentage of target grade-level standards per target that matched the intended mapping. As expected, Claims 2 – 4, across all grades, had low percentages of grade-level standards that matched the intended mapping. However, when the match of the grade-level standards to the intended clusters and domains was examined, the percentages substantially increased.

- Across all grades and claims, the average percentages of targets aligned to each mathematical practice were generally high. The notable exception was for grade 5, Claim 1 (Concepts and Procedures) where six of the eight mathematical practices had average percentages of aligned targets that were less than 60%. In addition, grade 11 Mathematical Practice 8 (Look for and express regularity in repeated reasoning) also had lower mean percentages of aligned targets across all claims. The mean percentages ranged from 22.1% for Claim 1 to 50.0% for Claim 3 (Communicating Reasoning). Lower percentages of alignment in Claim 1 targets were not unexpected as Claim 1 targets were designed to be more aligned with the grade-level standards than with the mathematical practices.

- The overall pairwise agreement in identifying grade-level standards aligned to each target between reviewers and the intended mapping as identified in the Content Specifications (across all grades, claims, targets, and reviewers) was 36.6%. The pairwise agreement at the cluster-level, however, is substantially higher (overall agreement of 64.9%). In addition, the overall pairwise agreement among reviewers in identifying mathematics grade-level standards aligned to each target (across all grades, claims, targets, and reviewers) was 51.3%. The moderate agreement rate was likely due to a combination of the relatively high number of grade-level standards that reviewers identified for each target and the fact that reviewers conducted a blind rating where they were permitted to choose from a lengthy list of eligible grade-level standards.

- Generally, the reviewers from grades 3 and 4 indicated targets required multiple levels of cognitive demand compared to the Content Specifications, which specified fewer levels. The reverse was true for grades 7 and 8, and grades 5, 6, and high school reviewers indicated similar numbers of DOK levels compared to the specifications.

- Using each independent DOK level, across grades for Claims 2, 3, and 4, the cognitive demand indicated by the reviewers and specifications was fairly similar. For Claim 1 (Concepts and Procedures), however, reviewers across grades indicated higher mean percentages of targets as requiring a higher cognitive demand that what was intended.

- The overall pairwise agreement between reviewers in identifying DOK ratings for each target (across all grades, claims, targets, and reviewers) was 63.7%, and the overall pairwise agreement in identifying grade-level standards aligned to each target between reviewers and the intended mapping as identified in the Content Specifications (across all grades, claims, targets, and reviewers) was 57.1%.

- There was no real pattern in the percentage of targets that had DOK consistency with all of the mapped grade-level standards. The percentage of targets that had DOK consistency with their grade-level standards ranged quite widely from 11.1% to 90%. Upon further investigation, the reason why so many targets had DOK inconsistency was because the reviewers rated the grade-level standards as requiring higher levels of cognitive demand than what was intended by the targets.

- When the DOK consistency definition was relaxed to requiring only one DOK level for each grade-level standard mapped to a target to fall within the range of the intended target DOK, the percentage of targets with DOK consistency substantially increased.

Summary of Findings for Connection B: Alignment between Evidence Statements and Content Specifications

- Reviewers generally rated the targets as being well-represented by the grade-level standards they identified. The mean alignment rating across grades and claims ranged from 3.5 to 4.0, on a scale of 0 to 4.

- We expected the majority of evidence statements would be rated as 'partially aligned' to the targets to which they were mapped. Reviewers rated the majority of evidence statements as partially aligned to their targets (71.4% – 98.0%), indicating that an individual evidence statement most often reflected only some of the content and knowledge required in the target to which it was aligned (which was intended by Smarter Balanced). Reviewers' ratings for evidence statements being fully aligned to their target were typically much less, ranging from 1.1% – 27.8%. Some 'fully aligned' ratings were expected as some targets only have one or two evidence statements aligned to it.

- For all grades except grade 3, the majority of mathematics CAT evidence statements (71.0% – 79.9%) were rated as having DOK levels within the range of the intended targets. The average percentage of grade 3 Claim 1 (Concepts & Problems) evidence statements with DOK levels within the range of the intended target was 49.0%.

- Especially for grade 3, reviewers believed that the DOK for the evidence statement was higher than that for its intended target. When the DOK consistency criterion was relaxed to only require at least one evidence statement's DOK level (since evidence statements could have multiple DOK levels) to match that of the intended target, the DOK consistency of the evidence statements with their intended targets increased across grades.

Summary of Findings for Connection C: Alignment between Test Blueprint and Content Specifications

- Reviewers felt that mathematics blueprints were mostly to fully representative of the content and knowledge that Smarter Balanced outlines to be assessed in the Content Specifications.

Summary of Findings for Connection D: Alignment between Evidence Statements and Item/Task Pools and Connection G: Alignment between Item/Task Pools and Content Specifications

- The majority of reviewers rated the mathematics CAT items as being fully aligned to their intended target. This was especially true for items in Claim 1 (Concepts and Procedures). The percentages were generally lower for the alignment of items in Claims 2 (Problem Solving), 3 Communicating Reasoning), and 4 (Modeling and Data Analysis); however, at least 81.3% of reviewers rated the items as fully aligned to their intended target.

- Across grades and claims, reviewers rated the majority of mathematics CAT items as being fully aligned to their intended grade-level standards. The lowest percentage of mathematics CAT items that were rated as being fully aligned to their intended grade-level standard was for grade 7, Claim 2 (Problem Solving, 79%).

- Reviewers typically identified one-third or fewer of the mathematics CAT items as matching the primary mathematical practice as that intended by the item writers. The mapping agreement increased, however, when examining the average percentage of items where at

least one of the mathematical practices identified by the reviewer matched at least one mathematical practice that was identified by the item writer.

- Reviewers tended to agree with the item writers about the mapping of mathematical practices to the mathematics PT items. The lowest agreement occurred for grades 6 and 7, where reviewers rated less than half the PT items' primary mathematical practice as matching the mathematical practice identified by the item writers. There was an increase in agreement of reviewer ratings for identifying at least one mathematical practice that had also been identified by the item writers. The highest percentage for reviewers not agreeing with the mathematical practices identified by the item writers was at these same grades (grade 6, 42.2% and grade 7, 43.1%).

- The average agreement was generally high. The average percentages ranged from a low of 71.5% (grade 7, Claim 2) to 98.7% (grade 3, Claim 1, Concepts and Procedures).

- The average pairwise agreement among reviewers for items within a claim was generally high, with the only average pairwise agreement below 80% occurring for Claim 4 (Modeling and Data Analysis) at grades 3 and 4.

- The pairwise agreement was fairly similar regardless of whether a primary mathematical practice matched or at least one of the mathematical practices matched. The percentages across grades and claims ranged between 50% – 80%.

- The average pairwise agreement among reviewers' ratings for PT items within a claim ranged from 67.6% (grade 7) to 95.6% (grade 11).

- The average pairwise agreement among reviewers' ratings for PT items within a claim ranged from 70% (grade 5) to 97.2% (grade 7).

- The pairwise agreement was fairly similar regardless of whether a primary mathematical practice matched or at least one of the mathematical practices matched. The percentages across grades ranged between 65% – 82%.

- Across grades and claims, reviewers generally agreed with the DOK levels that were indicated in the Content Specifications. Reviewers consistently believed there was a slightly higher percentage of Claim 1 (Concepts and Procedures) items at DOK level 1 than was indicated in the Content Specifications.

- Reviewers tended to agree with how the mathematics PT items were distributed in terms of their DOK level, as indicated in the Content Specifications. Across grades, reviewers rated most of the mathematics PT items at DOK levels 2 and 3. Except at grade 3, reviewers identified a very small percentage of items within a PT at DOK level 1 and/or DOK level 4 than what was indicated in the Content Specifications.

- Across grades and claims, reviewer agreement was generally high. The average agreement ranged from 71.9% (grade 5, Claim 3, Communicating Reasoning) to 97.4% (grade 7, Claim 2, Problem Solving).

- The average item pairwise agreement among reviewers' DOK ratings for mathematics PT items ranged from 69.3% to 96.3%. All of the grades had 80% or higher agreement except for grade 6, where the average pairwise agreement was 68.3%.

- Across grades and claims, reviewers rated at least half of the mathematics CAT items (53.3% – 100%) as falling within the range of DOK level for the intended target. For grades 3, 4, and 5, reviewers consistently believed fewer mathematics Claim 4 items fell within the range of DOK level of the intended target and consistently believed these items' DOK level was lower than the range specified for the intended target. For grades 6, 7, 8, and 11, reviewers consistently believed fewer mathematics Claim 3 items fell within the range of the DOK level of the intended target and consistently believed these items' DOK level was higher than the range specified for the intended target.

- The average reviewers' DOK ratings were lowest at grade 11 (84.7%) for the DOK level of the mathematics items falling within the range of the intended target and highest at grade 8 (94.8%).

Overall Alignment Summary

This alignment investigation was complex and included several elements and connections not usually included in typical alignment studies. Additionally, this study employed a two-way methodology to examine the alignment of the CCSS and Content Specifications. When all of the connections are considered, we find the overall alignment results acceptable for ELA/Literacy and mathematics. Reviewer agreement increased as they worked from standards through Content Specifications to evidence statements and then to items. Additionally, reviewer alignment agreement increased when the level of analysis was broadened, for example analyzing results at the cluster level instead of the target level.

Reviewers provided a holistic rating to indicate how well evidence statements collectively represented the content and knowledge required in the target. Reviewers generally rated the targets as being well represented by the grade-level standards they identified. Moreover, reviewers rated the majority of targets as being fully aligned to their collective set of evidence statements. Across grades, all of the targets were at least partially represented by their collective set of evidence statements.

In both ELA/Literacy and mathematics, reviewers felt that the blueprints were mostly to fully representative of the content and knowledge that Smarter Balanced outlines to be assessed in the Content Specifications.

When identifying evidence statements for each of the items, reviewers identified most of the possible evidence statements. Reviewers' average item-level pairwise agreement was, on average, moderate to high, more than half the reviewers agreed with one another that the intended evidence statement mapped to its respective item. Across grades and claims, reviewers and item writers generally agreed on the average number of evidence statements that were mapped to items. Most importantly, the majority of reviewers agreed that the CAT and PT items fully aligned to their intended targets and the data reflect that the reviewers had high agreement with each other as well as the item writers.

APPENDIX A: GLOSSARY

CAT: computer adaptive testing that adjusts to a student's ability basing the difficulty of future questions on previous answers. CAT aims to provide more accurate measurement of student achievement, particularly for high and low-performing standards.

CAT algorithm: programmatic logic that selects the items to be administered to students based on the designated specifications.

Claims: broad statements of the assessment system's learning outcomes, each of which requires evidence that articulates the types of data/observations that will support interpretations of competence towards achievement of the claims; claims serve as the fundamental drivers for the design of Smarter Balanced's English language arts (ELA)/literacy and mathematics summative assessments.

Cluster: a grouping of related mathematics standards within a given domain (e.g., grade 4, "Generalize understanding of place value for multi-digit numbers"); clusters are an effective means of communicating the focus and coherence of the mathematics standards because they provide an appropriate gain size for following the contours of important progressions in the standards across grades.

Common Core State Standards (CCSS): a set of high-quality academic standards in mathematics and English language arts (ELA)/literacy). These learning goals outline what a student should know and be able to do at the end of each grade.

Content Specifications: "bridge documents" document developed by Smarter Balanced that outlines the Consortium's interpretation and priority of the knowledge and skills identified in the Common Core State Standards that are intended to be measured by the ELA/literacy and mathematics assessments.

Depth of knowledge (DOK): the level of cognitive demand or cognitive complexity required by a standard, target, or item.

Domain: a larger grouping of related mathematics standards (e.g., Operations and Algebraic Thinking, Number and Operations—Base Ten).

Item specifications: documents that provide guidance specific to writing Smarter Balanced items; separate documents exist for English language arts (ELA)/literacy and mathematics. Each document includes a table for each claim and target combination expected to be addressed by the summative assessment.

Evidence statements: description of the specific knowledge and skills that an item or task elicits from students.

Mathematical practice: a balanced combination of procedure and understand that describes the ways in which developing student practitioners of mathematics increasingly ought to engage with the subject matter as they grow in mathematical maturity and expertise throughout the elementary,

middle, and high school years (e.g., “Make sense of problems and persevere in solving them,” “Attend to precision”).

Performance task: involves significant interaction of students with stimulus materials and/or engagement in a problem solution, ultimately leading to an exhibition of the students’ application of knowledge and skills, often in writing or spoken language; stimuli include a variety of information forms (e.g., readings, video clips, data) as well as assignment or problem situation.

Strand: a larger grouping of related English language arts (ELA)/literacy standards. (e.g., Reading, Writing, Speaking and Listening, Language).

Target: detailed information about the knowledge, skills, and abilities to be assessed by the items and tasks within each claim as well as the depth of knowledge (cognitive demand) required; targets represent the prioritized content for summative assessment.

Test blueprint: series of documents that together describe the content and structure of an assessment; these documents define the total number of items/tasks for any given assessment component, the standards measured, the item types, and the point values for each.

APPENDIX B: SUMMARY OF REVIEWER COMMENTS

Table B-1. Summary of Reviewer Comments Regarding Alignment Workshops

Comment Category	Count	Category Description
Workshop Experience-Positive	59	Comments in this category either thanked the staff for the opportunity to be part of the work or said that the experience was positive in some way (e.g., great, fun, amazing). Specifically, the staff, logistics, materials, and training were mentioned.
Materials-Improved materials	55	Comments in this category specified what and how the materials could be improved (e.g., binders, tabs on the binders, wording of instructions). Many comments related to the formatting of specific Excel worksheets and the size of the worksheets. A relatively large number of workshop 1 ELA participants were confused by the 3-2-1 rating and the Excel forms. Several comments in this category also related to the laptops. These commenters would have preferred two screens to do the work and more room at the table.
Training-Improved training	50	Comments in this category related to the large group presentation and training. Workshop one commenters voiced a preference for a shorter initial presentation. They said that there was too much information being provided at one time. Commenters from later workshops commented that the presentation was too vague. Other training related comments were diverse. One commenter would have preferred to have had the manual before the workshop. Other commenters said they felt confused or overwhelmed by their tasks. They would have preferred training in smaller sections.
Staff-Positive	29	Comments in this category complimented the staff. Many commenters said that the facilitators were knowledgeable and helpful.
PD/Learning Experience-Positive	11	Comments in this category noted that the experience was beneficial for professional development. The participants who wrote this type of comment (of whom all were teachers or coaches) were pleased that the workshop deepened their understanding of the CCSS and the Smarter Balanced claims and targets.
Task load-Heavy	11	Comments in this category noted that the task-load was heavy. These comments often noted how long the participant thought the task would take. Five of these comments were from workshop 2 participants.
Other	10	Comments in this category did not fit into one of the categories above. For example, one comment raised an issue with the logistics and another related to the accommodations at the hotel.

APPENDIX C: ITEM, EVIDENCE STATEMENT, AND REVIEWER SAMPLING

Table C-1. Number of Mathematics Items and Performance Tasks Sampled for Workshops 3, 4, and 5 for Connections B, D, and G

Grade	Items					Performance Tasks
	50% of Claim 1	50% of Claim 2	50% of Claim 3	50% of Claim 4	100% of All Evidence Statements	Number of Performance Tasks/Items
3	425	27	80	40	30	4/24
4	425	27	80	40	44	4/24
5	425	27	80	40	32	4/24
6	425	27	80	40	50	4/24
7	425	27	80	40	35	4/24
8	425	27	80	40	42	4/24
11	1,600	101	303	152	54	8/48
Total	4,150	263	783	392	286	32/192

Table C-2. Individual Workshop Design for Workshops 3, 4, and 5 for Connections B, D, and G – Mathematics[1]

Groups	Reviewers	Items		Performance Tasks	
Workshop Groups (Grade-level)	# of Reviewers per Workshop	# of Items per Reviewer	# of Evidence Statements per Reviewer	# of Performance Tasks/Items per Reviewer – Wkshp 3-4	# of Performance Tasks/Items per Reviewer – Wkshp 5
3 – 4	5	200	73	2/12	4/24
5 – 6	5	200	82	2/12	4/24
7 – 8	5	200	77	2/12	4/24
HS-1	5	190	54	2/12	2/12
HS-2	5	190	54	2/12	2/12
Total per workshop	25	982	286[2]	10/60	16/96
Total for 3 workshops	75	2947	286	—	—

[1]Totals may reflect rounding error.
[2]Both high school groups (HS-1 and HS-2) reviewed the same 54 evidence statements.

Table C-3. Number of ELA/Literacy Items and Performance Tasks to be Sampled for Workshops 3, 4, and 5 for Connections B, D, and G

Grade	Items					Performance Tasks
	50% of Claim 1	50% of Claim 2	50% of Claim 3	50% of Claim 4	100% of All Evidence Statements	Number of Performance Tasks/Items
3	220	110	99	66	72	4/16
4	220	110	99	66	73	4/16
5	220	110	99	66	72	4/16
6	220	110	99	66	77	4/16
7	220	110	99	66	75	4/16
8	220	110	99	66	89	4/16
11	772	385	347	231	75	8/32
Total	2092	1045	941	627	533	32/128

Table C-4. Individual Workshop Design for Workshops 3, 4, and 5 for Connections B, D, and G – ELA/Literacy[1]

		Items		Performance Tasks	
Workshop Groups (Grade-bands)	# of Reviewers per Workshop	Total # of Items per Reviewer	# of Evidence Statements per Reviewer	# of Performance Tasks/Items per Reviewer – Wkshp 3-4	# of Performance Tasks/Items per Reviewer – Wkshp 5
3 – 4	5	330	145	4/12	4/16
5 – 6	5	330	149	4/12	4/16
7 – 8	5	330	164	4/12	4/16
HS-1	5	289	75	2/8	2/8
HS-2	5	289	75	2/8	2/8
Total per workshop	25	1568	533[2]	16/52	12/64
Total for 3 workshops	75	4705	533	—	—

[1]Totals may reflect rounding error.

[2]Both high school groups (HS-1 and HS-2) reviewed the same 75 evidence statements.

APPENDIX D: ALIGNMENT ANALYSIS DETAILS

Connection A: Alignment of Content Specifications to CCSS

Criterion: Content Representation

The content representation (CR) criteria examine how well the content in the CCSS are represented by the assessment targets. The CR investigations are focused on the following six questions:

Question A.CR-1. Do the grade-level standards collectively reflect the content and skills required by the target?

Question A.CR-2. Do the targets collectively reflect the content and skills required by the grade-level standard?

Question A.CR-3. Do the individual grade-level standards reflect the content and skills required by the intended targets?

Question A.CR-4: Do the individual targets reflect the content and skills required by the intended grade-level standard?

Question A.CR-5. Does each mathematical practice reflect skills required by the intended target?

Question A.CR-6. Do the reviewers agree with the intended mapping of targets and grade-level standards as identified in the content specifications?

Details of the analyses and data available for each of these questions are described below. When results from each analysis are complete, they will be assembled together and used collectively as the basis for an overall judgment about the degree to which the CCSS are represented by the targets.

Content Representation Questions

Question A.CR-1. Do the grade-level standards collectively reflect the content and skills required by the target?

Analysis: Compute the mean percentage of targets that were rated holistically as (a) fully-aligned (target was adequately measured across all aligned grade-level standards), (b) mostly-aligned, (c) somewhat-aligned, and (d) small portion aligned

Step 1. For each reviewer, compute the percentage of targets that were rated holistically as (a) fully-aligned, (b) mostly-aligned, (c) somewhat-aligned, and (d) small portion aligned to the full set of grade-level standards

Step 2. For each claim, compute the average percentage for each alignment rating (e.g., fully-aligned, mostly-aligned) across reviewers

Available Data: Reviewers' holistic target coverage ratings (how well the target was represented by all of the grade-level standards identified by reviewers as being aligned to that target (4-point scale)

Question A.CR-2. Do the targets collectively reflect the content and skills required by the grade-level standard?

Analysis: Compute the mean percentage of grade-level standards that were rated holistically as (a) fully-aligned (grade-level standard was adequately measured across all aligned targets), (b) mostly-aligned, (c) somewhat-aligned, and (d) small portion aligned

Step 1. For each reviewer, compute the percentage of grade-level standards that were rated holistically

Step 2. For each domain (strand for ELA), compute the average percentage for each alignment rating (e.g., fully-aligned, mostly-aligned) across reviewers

Available Data: Reviewers' holistic grade-level standard coverage rating (how well the standard was represented by all of the targets identified by reviewers as being aligned (4-point scale)

Question A.CR-3. Do the individual grade-level standards reflect the content and skills required by the intended targets?

Analysis: Compute the mean percentage of grade-level standards aligned to a target that (a) match the intended mappings as specified by the content specifications, (b) fall within the intended domain /cluster (math) or strand (ELA), and (c) fall outside the intended domain/cluster (math) or strand (ELA)

Step 1. For a single target identify grade-level standards that at least 50% of the reviewers agreed the grade-level standards align to that target (fully or partially)

Step 2. For each target, compute the percentage of grade-level standards that were rated as aligned to a target (fully or partially) that (a) match the intended mappings as specified by the content specifications, (b) fall within the intended domain /cluster (math) or strand (ELA), and (c) fall outside the intended domain/cluster (math) or strand (ELA)

Step 3. For each claim, compute the average percentage of grade-level standards across targets that match the intended mapping and fall within and outside the intended domain(s)

Available Data: Reviewers' independent identification of aligned grade-level standards and mathematical practices to each target and degree of alignment (2-point scale); Intended mapping between targets and grade level standards as identified in the content specifications

Question A.CR-4: Do the individual targets reflect the content and skills required by the intended grade-level standard?

Analysis: Compute the mean percentage of grade-level standards aligned to a target that (a) match the intended mappings as specified by the content specifications, (b) fall within the intended domain /cluster (math) or strand (ELA), and (c) fall outside the intended domain/cluster (math) or strand (ELA), using data from workshop 2 where reviewers were given a list of grade-level standards and asked to identify targets to why they align.

Step 1. For each reviewer/grade-level standard combination, reshape the rating form to reflect each row as a reviewer/target/grade-level standard (similar to that of CR-3).

Step 2. For a single grade-level standard identify targets that at least 50% of the reviewers agreed the grade-level standards align to the targets

Step 3. For each grade-level standard, compute the percentage of targets that were rated as aligned to a grade-level standard that (a) match the intended mappings as specified by the content specifications, (b) fall within the intended domain /cluster (math) or strand (ELA), and (c) fall outside the intended domain/cluster (math) or strand (ELA)

Step 4. For each claim, compute the average percentage of targets across grade-level standards that match the intended mapping and fall within and outside the intended domain(s)

Available Data: Reviewers' independent identification of aligned targets to each grade-level standard; Intended mapping between targets and grade level standards as identified in the content specifications

Question A.CR-5. Does each mathematical practice reflect skills required by the intended target?

Analysis: Compute the mean percentage of targets that align to each mathematical practice

Step 1. For each reviewer, compute the mean percentage of targets that align to each mathematical practice

Step 2. Compute the average percentage of targets across reviewers

Available Data: Reviewers' independent identification of aligned grade-level standards and mathematical practices to each target and degree of alignment (2-point scale)

Question A.CR-6. Do the reviewers agree with the intended mapping of targets and grade-level standards as identified in the content specifications?

Analysis: Compute the pairwise percent agreement between reviewers' mappings of targets and grade-level standards/mathematical practices and the intended mapping as identified in the content specifications.

Step 1. Compute the pairwise agreement by counting the grade-level standards that the reviewer agreed on with the intended specs and divide by the greater number of grade-level standards identified between the two ratings.

Step 2. For each target, compute agreement by averaging the pairwise agreement between all comparisons.

Step 3. For each claim, average the pairwise agreement across targets

Available Data: Reviewers' independent identification of aligned grade-level standards and mathematical practices to each target and degree of alignment (2-point scale); Reviewers' independent identification of aligned targets to each grade-level standard; Intended mapping between targets and grade level standards as identified in the content specifications

Criterion: DOK Distribution

The DOK distribution (DD) criteria examines the reviewers' DOK distribution of the targets compared to the DOK distribution identified in the Smarter Balanced content specifications. The DD investigations are focused on the following three questions:

Question A.DD-1. Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the content specifications (using the max DOK level)?

Question A.DD-2. Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the content specifications (using the each independent DOK level)?

Question A.DD-3. Do the reviewers agree with the intended target DOK levels as identified in the content specifications?

Details of the analyses and data available for each of these questions are described below. When results from each analysis are complete, they will be assembled together and used collectively as the basis for an overall judgment about the degree to which the reviewers DOK distribution of the targets reflects the DOK distribution identified in the Smarter Balanced content specifications.

DOK Distribution Questions

Question A.DD-1. Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the content specifications (using the max DOK level)?

Analysis: Compute and compare the mean percentage of targets at each DOK level using the highest DOK level for a single target, based on the reviewers' ratings and the DOK levels identified in the content specifications

Step 1. For each reviewer, compute the max DOK level for each target

Step 2. For each reviewer, compute percentage of targets at each of the DOK levels by claim

Step 3. For each claim, average the percentages of targets at each of the DOK levels for each DOK level across reviewers

Available Data: Reviewers' independent DOK ratings for targets (yes/no ratings for each DOK level); Intended mapping between targets and grade level standards as identified in the content specifications

Question A.DD-2. Does the DOK distribution of the targets identified by the reviewers match that of the distribution identified in the content specifications (using the each independent DOK level)?

Analysis: Compute and compare the mean percentage of targets at each DOK level using each level independently (e.g., mean percentage of targets at DOK level 1) based on the reviewers' ratings and the DOK levels identified in the content specifications

Step 1. For each reviewer, compute percentage of targets at each DOK level by claim

Step 2. For each claim, average the percentage of targets at each DOK level across reviewers (Note, because reviewers can rate more than one DOK level for each target, the total percentage across DOK levels within a claim will equal more than 100%)

Available Data: Reviewers' independent DOK ratings for targets (yes/no ratings for each DOK level); Intended mapping between targets and grade level standards as identified in the content specifications

Question A.DD-3. Do the reviewers agree with the intended target DOK levels as identified in the content specifications?

Analysis: Compute the pairwise percent agreement between reviewers' target DOK ratings and intended target ratings as identified in the content specifications

Step 1. Compute the pairwise agreement by counting the DOK levels that the reviewer agreed on with the intended specs and divide by the greater number of DOK levels rated between the two ratings.

Step 2. For each target, compute agreement by averaging the pairwise agreement between all comparisons.

Step 3. For each claim, average the pairwise agreement across targets

Available Data: Reviewers' independent identification of aligned grade-level standards and mathematical practices to each target and degree of alignment (2-point scale); Intended mapping between targets and grade level standards as identified in the content specifications.

Criterion: DOK Consistency

The DOK consistency (DC) criterion examines the degree to which the cognitive complexity required by the targets is consistent with that required in the grade-level standards. The DC investigation focuses on a single question (see below). Details of this analysis and data available for this question are described below. Results from this analysis will be used as the basis for an overall judgment about the degree to which the cognitive complexity required by the targets is consistent with the CCSS grade-level standards.

Question A.DC-1. Is the cognitive complexity required in the targets consistent with the cognitive complexity required in each targets' mapped grade-level standards/mathematical practices?

Analysis: Compute the mean percentage of targets that have grade level standards/mathematical practices with DOK ratings that (a) fall within the range of the intended target DOK(s), (b) have the highest DOK as greater than the highest DOK of the intended target, and (c) have the lowest DOK as lower than the lowest DOK of the intended target

Step 1. For each target and for each reviewer, identify the reviewers' grade-level standards that match the intended mapping identified in the content specs

Step 2. For the standards that match the intended mapping, determine if at least 50% of the reviewers agreed

Step 2a. If at least 50% of the reviewers agreed with the intended mapping, compute the percentage of targets that (a) have DOK consistency with the standards' consensus DOK and (b) do not have DOK consistency with the standards. Additionally, compute the percentage of targets that do not have at least 50% reviewer agreement with the intended mapping

Step 2b.Average the percentage of targets by claim

Available Data: Reviewers' consensus DOK ratings for each grade-level standard and mathematical practice (yes/no ratings for each DOK level); Reviewers' independent identification of aligned grade-level standards and mathematical practices to each target and degree of alignment (2-point scale); Intended DOK ratings of targets as identified in the content specifications; Intended mapping between targets and grade level standards as identified in the content specifications

Connection B: Alignment of Evidence Statements to Content Specifications

Criterion: Content Representation

The content representation (CR) criteria examine how well the targets are represented by the evidence statements. The CR investigations are focused on the following two questions:

Question B.CR-1. Do the evidence statements collectively reflect the content and skills required by the target?

Question B.CR-2. Do the individual evidence statements reflect the content and skills required by the intended targets?

Details of the analyses and data available for each of these questions are described below. When results from each analysis are complete, they will be assembled together and used collectively as the basis for an overall judgment about the degree to which content in the evidence statements is consistent with the content in the targets.

Content Representation Questions

Question B.CR-1. Do the evidence statements collectively reflect the content and skills required by the target?

Analysis: Compute the mean percentage of targets that are fully reflected by all of the aligned evidence statements (holistic target coverage rating)

Step 1. For each reviewer, compute the percentage of targets that are fully, partially, and not reflected by the combination of all aligned evidence statements (Note: This is accomplished using reviewers' holistic ratings)

Step 2. For each claim, average the target percentages across reviewers

Available Data: Reviewers' holistic target coverage ratings (how well the target was collectively represented by the set of evidence statements for each target) (3-point scale)

Question B.CR-2. Do the individual evidence statements reflect the content and skills required by the intended targets?

Analysis: Compute the mean percentage of evidence statements within each target that fully, partially, and do not reflect the content and knowledge required in the target

Step 1. For each reviewer, compute the percentage of ESs that fully, partially, and do not reflect the content and knowledge required by each target

Step 2. For each claim, average the evidence statement percentages across reviewers

Available Data: Reviewers' verification of individual evidence statements (how well the target was represented by each individual evidence statement) (3-point

Criterion: DOK Consistency

The DOK consistency (DC) criterion examines the degree to which the cognitive complexity required by the assessment targets is consistent with that required in the evidence statements. The DC investigation focuses on the following single question:

Question B.DC-1. Do reviewers' evidence statement DOK ratings align with the DOK levels specified for the targets to which they are mapped (as indicated in the content specifications)?

Details of this analysis and data available for this question are described below. Results from this analysis will be used as the basis for an overall judgment about the degree to which the cognitive complexity required by the assessment targets is consistent with that required in the evidence statements.

DOK Consistency Question

Question B.DC-1. Do reviewers' evidence statement DOK ratings align with the DOK levels specified for the targets to which they are mapped (as indicated in the content specifications)?

Analysis: Compute the mean percentage of evidence statements that (a) fall within the range of the intended target DOK(s), (b) have the highest DOK as greater than the highest DOK of the intended target, and (c) have the lowest DOK as lower than the lowest DOK of the intended target

Step 1. For each reviewer, compute the percentage of evidence statements with DOK ratings (a) that fall within the range of the DOK of the intended target, (b) that have the highest DOK as greater than the highest DOK of the intended target, and (c) that have the lowest DOK as lower than the lowest DOK of the intended target

Step 2. For each claim, average the percentages of evidence statements across reviewers for each of the three percentages in Step 1

Available Data: Reviewers' independent DOK ratings for targets (yes/no ratings for each DOK level); Intended DOK ratings of targets as identified in the content specifications

Connection C: Alignment of Test Blueprint to Content Specifications

Criterion: Content Representation

The content representation (CR) criterion examines how well the content specifications are represented by the draft blueprints.[1] The CR investigation focuses on the following single question:

> Question C.CR-1. To what degree are the content specifications represented in the daft blueprints?

Details of the analyses and data available for each of these questions are described below. When results from each analysis are complete, they will be assembled together and used collectively as the basis for an overall judgment about the degree to which draft test blueprints represent the Smarter Balanced content specifications.

Content Representation Question

Question C.CR-1. To what degree are the content specifications represented in the daft blueprints?

Analysis: Compute means, medians and standard deviations of representativeness ratings and summarize panelist comments by grade and content area

Step 1. Aggregate reviewers' responses and calculate means, medians, and standard deviations for each grade and content area[2]

Step 2. Summarize reviewers' comments.

Available Data: Reviewers' independent ratings of representativeness on a 4-point Likert scale and their comments.

[1] The draft blueprints included in this alignment study were dated November 2013.

[2] Reviewers provided a holistic rating, using a 4-point Likert scale where 1 was anchored at *Not at all* and 4 at *Fully representative*, to answer "What is the overall extent to which the blueprint represents the summative content specifications?"

Connection D: Alignment of Item/Task Pools to Evidence Statements[3]

Criterion: Content Representation

The content representation (CR) criteria examine how well the content in the evidence statements is represented by the items. The CR investigations are focused on the following two questions:

Question D.CR-1. How are the summative assessment items distributed across evidence statements?

Question D.CR-2. Do the reviewers agree with the intended mapping of items to evidence statements as identified by the item developers?

Details of the analyses and data available for each of these questions are described below. When results from each analysis are complete, they will be assembled together and used collectively as the basis for an overall judgment about the degree to which content in the evidence statements is consistent with the content expressed in the items.

Content Representation Questions

Question D.CR-1. How are the summative assessment items distributed across evidence statements?

Analysis: Compute the mean percentage of items aligned to each evidence statement

Step 1. For each reviewer, compute the percentage of items aligned to each evidence statement for each reviewer and claim

Step 2. Average the percentages of items aligned to each evidence statement across reviewers

Available Data: Reviewers' independent identification of aligned evidence statement for each item

Question D.CR-2. Do the reviewers agree with the intended mapping of items to evidence statements as identified by the item developers?

[3] For Connection D, all analyses will be conducted and reported out separately for CAT items and performance tasks (PTs). The analysis steps presented are for CAT items. The PT item analyses will involve averaging first by grade, performance task, and rater, then within performance task, then averaged across the grade.

Analysis: Compute the percent agreement between reviewers' mappings of evidence statements and items and the intended mapping as identified by the item developers[4]

Step 1. For each item, compute the pairwise percentage agreement by counting the number of reviewers that agreed with the intended evidence statement to item mapping divided by the total number of comparisons
Step 2. For each claim, average the pairwise percentage agreement across items

Available Data: Reviewer independent item evidence statement ratings; Item meta-data intended evidence statement ratings

Criterion: DOK Consistency

The DOK consistency (DC) criterion examines the degree to which the cognitive complexity required by the items is consistent with that required in the evidence statements. The DC investigation focuses on the following single question:

Question D.DC-1. Is the cognitive complexity required in the items consistent with the cognitive complexity required in each evidence statement?

Details of this analysis and data available for this question are described below. Results from this analysis will be used as the basis for an overall judgment about the degree to which the cognitive complexity required by the items is consistent with that required in the evidence statements.

DOK Consistency Question

Question D.DC-1. Is the cognitive complexity required in the items consistent with the cognitive complexity required in each evidence statement?

Analysis: Because an ES does not have an intended DOK, these analyses use the DOK range of each ES as identified by the reviewers. Compute the mean percentage of items that have DOK ratings that (a) fall within the range of the identified ES DOK(s), (b) have the highest DOK greater than the highest DOK of the identified ES, and (c) have the lowest DOK lower than the lowest DOK of the identified ES.

Step 1. For each reviewer, compute the percentage of items that have DOK ratings that (a) fall within the range of the identified ES DOK(s), (b) have the highest DOK as greater than the highest DOK of the identified ES, and (c) have the lowest DOK as lower than the lowest DOK of the identified ES

Step 2. For each claim, average the percentage of items across reviewers for each of the three percentages in Step 1

Available Data: Reviewer item DOK verification ratings; Reviewer independent evidence statement DOK Ratings

[4] This analysis will be completed only for ELA/literacy since no evidence statements are available in the math item metadata.

Connection E: Alignment of CAT Algorithm to Test Blueprint

Criterion: DOK Consistency

The DOK consistency (DC) criterion examines how well does the DOK requirements outlined in the test blueprint are reflected in the CAT algorithm specifications. The DC investigation is focused on the following question:

> Question E.DC-1. How well are the DOK requirements outlined in the test blueprint reflected in the CAT algorithm specifications?

Details of this document review are described below. When results from this review are complete, they will be assembled together and used collectively as the basis for an overall judgment about the degree to which the DOK requirements in the test blueprint are included in the CAT algorithm.

DOK Consistency Question

Question E.DC-1. How well are the DOK requirements outlined in the test blueprint reflected in the CAT algorithm specifications?

> Analysis: Review algorithm specifications against the DOK requirements in the test blueprint
>
> Available Documents: Draft test blueprints; CAT algorithm specifications

Criterion: Content Representation

The content representation (CR) criterion examines how well does the content requirements outlined in the test blueprint are reflected in the CAT algorithm specifications. The CR investigation is focused on the following question:

> Question E.CR-1. How well are the content requirements outlined in the test blueprint reflected in the CAT algorithm specifications?

Details of this document review are described below. When results from this review are complete, they will be assembled together and used collectively as the basis for an overall judgment about the degree to which the content requirements in the test blueprint are included in the CAT algorithm.

Content Representation Question

Question E.CR-1. How well are the content requirements outlined in the test blueprint reflected in the CAT algorithm specifications?

> Analysis: Review algorithm specifications against the content requirements in the test blueprint
>
> Available Documents: Draft test blueprints; CAT algorithm specifications

Connection G: Alignment of Items/Performance Tasks to Content Specifications[5]

Criterion: Content Representation

The content representation (CR) criteria examine how well the content in the targets, grade-level standards, and mathematical practices are represented by the items. The CR investigations focus on the following three questions:

Question G.CR-1. How are the summative assessment items distributed across targets, grade-level standards, and mathematical practices?

Question G.CR-2. Do the reviewers agree with the intended mapping of items to targets, grade-level standards, and mathematical practices as identified by the item developers? [6]

Question G.CR-3. Do the reviewers agree with the intended mapping of items to mathematical practices as identified by the item developers?

Details of the analyses and data available for each of these questions are described below. When results from each analysis are complete, they will be assembled together and used collectively as the basis for an overall judgment about the degree to which content in the targets, grade-level standards, and mathematical practices is consistent with the content expressed in the items.

Content Representation Questions

Question G.CR-1. How are the summative assessment items distributed across targets, grade-level standards, and mathematical practices?

Analysis: Compute the mean percentage of items aligned to each target, grade-level standard, and mathematical practice

Step 1. For each reviewer, compute the percentage of items aligned to each target, grade-level standard, and mathematical practice (when applicable) for each claim
Step 2. Average the percentages of items aligned to each target, grade-level standard, and mathematical practice statement across reviewers

Available Data: Reviewer target verification ratings; Reviewer independent mathematical practice ratings

Question G.CR-2. Do the reviewers agree with the intended mapping of items to targets and grade-level standards as identified by the item developers?

[5] For Connection G, all analyses will be conducted and reported out separately for CAT items and performance tasks (PTs). The analysis steps presented are for CAT items. The PT item analyses will involve averaging first by grade, performance task, and rater, then within performance task, then averaged across the grade.
[6] In traditional alignment, one would evaluate whether there is a sufficient number of items associated with each claim/CCSS. However, because there are no test forms, we instead evaluate whether the items are being written to the intended target/CCSS. The CAT algorithm will use the intended target information to create forms, so this will provide evidence that the information being pulled to create the test forms is valid.

Analysis: Compute percentage of items fully aligned, partially aligned, or not aligned to the intended target and grade-level standard.

Step 1. For each reviewer, compute by claim the percentage of items with targets (and in separate analysis, GS) fully, partially, or not-aligned
Step 2. For each claim, average the percentage of items with targets (and in separate analysis, GS) fully, partially, or not-aligned across reviewers

Available Data: Reviewers' alignment ratings (fully, partial, or not-aligned) of the items to the intended targets and grade-level standards.

Question G.CR-3. Do the reviewers agree with the intended mapping of items to mathematical practices as identified by the item developers?

Analysis: Compute the pairwise agreement among reviewers' ratings of mathematical practices to items

Step 1. For each target, compute the pairwise percentage agreement by counting the number of pairs of reviewers who agree in their mathematical practices to item mappings, divided by the total number of pairs of reviewers
Step 2. For each claim, average the pairwise percentage agreement across reviewers

Available Data: Reviewers' independent identification of aligned mathematical practice for each math item; Intended mathematical practices from item meta-data

Criterion: DOK Distribution

The DOK distribution (DD) criterion examines the reviewers' DOK distribution of the items compared to the DOK distribution identified in the item metadata. The DD investigation focuses on the following single question:

Question G.DD-1. How does the distribution of DOK of the items identified by the reviewers compare with the distribution identified in the content specifications?

Details of the analysis and data available for this question are described below. The results will be used as the basis for an overall judgment about the degree to which the reviewer's item DOK levels reflect the DOK distributions identified in the Smarter Balanced content specifications.

DOK Distribution Question

Question G.DD-1. How does the distribution of DOK of the items identified by the reviewers compare with the distribution identified in the content specifications?

Analysis: Compute and compare the mean percentage of items at each DOK level based on the reviewers' ratings and the DOK levels identified in the content specifications

Step 1. For each reviewer, compute by claim the percentage of items that are at each DOK level

Step 2. For each claim, average the percentage of items that are at each DOK level across reviewers

Available Data: Reviewers' verification or independent identification of item DOK levels; DOK levels from item meta-data

Criterion: DOK Consistency

The DOK consistency (DC) criterion examines the degree to which the cognitive complexity required by the items is consistent with that required in the targets. The DC investigation focuses on the following single question:

Question G.DC-1. Does the DOK of the item identified by the reviewers fall within the DOK distribution of the aligned target identified in the content specifications?

Details of this analysis and data available for this question are described below. Results from this analysis will be used as the basis for an overall judgment about the degree to which the cognitive complexity required by the items is consistent with that required in the targets.

DOK Consistency Question

Question G.DC-1. Does the DOK of the item identified by the reviewers fall within the DOK distribution of the aligned target identified in the content specifications?

Analysis: Compute the mean percentage of item ratings that are below, at, and above the intended DOK of the targets (as indicated in the content specifications)

Step 1. For each reviewer, compute the percentage of item ratings at, below, and above the DOK range of the intended target

Step 2. For each claim, average the percentage of items that are at, below, and above the DOK range across reviewers

Available Data: Reviewer item DOK verification ratings; Content specifications intended DOK ratings

APPENDIX E:
LIST OF EXCLUDED MATHEMATICS AND ELA/LITERACY COMMON CORE STATE STANDARDS

Table E-1 presents the Mathematics CCSSs that were excluded from this alignment study. All exclusions for mathematics were at grade 11.

Table E-1. List of Excluded Mathematics Common Core State Standards.

Grade	Standards Level	Excluded CCSS
11	1	HSF-BF.A.1C
11	1	HSF-BF.B.4B
11	1	HSF-BF.B.4C
11	1	HSF-BF.B.4D
11	1	HSF-IF.C.7D
11	1	HSN-VM.B.4A
11	1	HSN-VM.B.4B
11	1	HSN-VM.B.4C
11	1	HSN-VM.B.5A
11	1	HSN-VM.B.5B
11	1	HSS-MD.B.5A
11	1	HSS-MD.B.5B
11	2	HSA-APR.C.5
11	2	HSA-APR.D.7
11	2	HSA-REI.C.8
11	2	HSA-REI.C.9
11	2	HSF-BF.B.5
11	2	HSF-TF.A.3
11	2	HSF-TF.A.4
11	2	HSF-TF.B.6
11	2	HSF-TF.B.7
11	2	HSF-TF.C.9
11	2	HSG-C.A.1
11	2	HSG-C.A.2
11	2	HSG-C.A.3
11	2	HSG-C.A.4
11	2	HSG-C.B.5
11	2	HSG-CO.D.12
11	2	HSG-CO.D.13

Table E-1. (Continued)

Grade	Standards Level	Excluded CCSS
11	2	HSG-GMD.A.2
11	2	HSG-GMD.B.4
11	2	HSG-GPE.A.1
11	2	HSG-GPE.A.2
11	2	HSG-GPE.A.3
11	2	HSG-GPE.A.7
11	2	HSG-GPE.B.4
11	2	HSG-GPE.B.5
11	2	HSG-GPE.B.6
11	2	HSG-SRT.D.10
11	2	HSG-SRT.D.11
11	2	HSG-SRT.D.9
11	2	HSN-CN.A.1
11	2	HSN-CN.A.2
11	2	HSN-CN.A.3
11	2	HSN-CN.B.4
11	2	HSN-CN.B.5
11	2	HSN-CN.B.6
11	2	HSN-CN.C.7
11	2	HSN-CN.C.8
11	2	HSN-CN.C.9
11	2	HSN-VM.A.1
11	2	HSN-VM.A.2
11	2	HSN-VM.A.3
11	2	HSN-VM.B.4
11	2	HSN-VM.B.5
11	2	HSN-VM.C.10
11	2	HSN-VM.C.11
11	2	HSN-VM.C.12
11	2	HSN-VM.C.6
11	2	HSN-VM.C.7
11	2	HSN-VM.C.8
11	2	HSN-VM.C.9
11	2	HSS-CP.B.6
11	2	HSS-CP.B.7
11	2	HSS-CP.B.8
11	2	HSS-CP.B.9
11	2	HSS-MD.A.1

Table E-1. (Continued)

Grade	Standards Level	Excluded CCSS
11	2	HSS-MD.A.2
11	2	HSS-MD.A.3
11	2	HSS-MD.A.4
11	2	HSS-MD.B.5
11	2	HSS-MD.B.6
11	2	HSS-MD.B.7
11	3	HSG-C.A
11	3	HSG-C.B
11	3	HSG-CO.D
11	3	HSG-GMD.B
11	3	HSG-GPE.A
11	3	HSG-GPE.B
11	3	HSG-SRT.D
11	3	HSN-CN.A
11	3	HSN-CN.B
11	3	HSN-CN.C
11	3	HSN-VM.A
11	3	HSN-VM.B
11	3	HSN-VM.C
11	3	HSS-CP.B
11	3	HSS-MD.A
11	3	HSS-MD.B
11	5	HSG-GPE
11	5	HSN-CN
11	5	HSN-VM
11	5	HSS-MD

Table E-2 presents the ELA/Literacy CCSSs that were excluded from this alignment study, by grade and strand.

Table E-2. List of Excluded ELA/Literacy Common Core State Standards

Grade	Strand	Excluded CCSS
3	RF	3.RF.3
3	RF	3.RF.3.a
3	RF	3.RF.3.b
3	RF	3.RF.3.c
3	RF	3.RF.3.d
3	RF	3.RF.4
3	RF	3.RF.4.a
3	RF	3.RF.4.b
3	RF	3.RF.4.c
3	RI	3.RI.10
3	RL	3.RL.10
3	RL	3.RL.8
3	SL	3.SL.1
3	SL	3.SL.4
3	SL	3.SL.5
3	SL	3.SL.6
3	W	3.W.10
3	W	3.W.6
3	W	3.W.7
3	W	3.W.9
4	L	L.4.3c
4	RF	RF.4.3
4	RF	RF.4.3a
4	RF	RF.4.4
4	RF	RF.4.4a
4	RF	RF.4.4b
4	RF	RF.4.4c
4	RI	RI.4.10
4	RL	RL.4.10
4	RL	RL.4.8
4	SL	SL.4.1
4	SL	SL.4.4
4	SL	SL.4.5
4	SL	SL.4.6
4	W	W.4.10
4	W	W.4.6

Table E-2. (Continued)

Grade	Strand	Excluded CCSS
4	W	W.4.7
5	RF	RF.5.3
5	RF	RF.5.3a
5	RF	RF.5.4
5	RF	RF.5.4a
5	RF	RF.5.4b
5	RF	RF.5.4c
5	RI	RI.5.10
5	RL	RL.5.10
5	RL	RL.5.8
5	SL	SL.5.1
5	SL	SL.5.4
5	SL	SL.5.5
5	SL	SL.5.6
5	W	W.5.10
5	W	W.5.6
5	W	W.5.7
6	RH	RH.6-8.10
6	RI	RI.6.10
6	RL	RL.6.10
6	RL	RL.6.8
6	RST	RST.6-8.10
6	SL	SL.6.1
6	SL	SL.6.4
6	SL	SL.6.5
6	SL	SL.6.6
6	W	W.6.10
6	W	W.6.6
6	W	W.6.7
6	WHST	WHST.6-8.10
6	WHST	WHST.6-8.6
6	WHST	WHST.6-8.7
7	RH	RH.6-8.10
7	RI	RI.7.10
7	RL	RL.7.10
7	RL	RL.7.8
7	RST	RST.6-8.10
7	SL	SL.7.1
7	SL	SL.7.4

Table E-2. (Continued)

Grade	Strand	Excluded CCSS
7	SL	SL.7.5
7	SL	SL.7.6
7	W	W.7.10
7	W	W.7.6
7	W	W.7.7
7	WHST	WHST.6-8.10
7	WHST	WHST.6-8.6
7	WHST	WHST.6-8.7
8	RH	RH.6-8.10
8	RI	RI.8.10
8	RL	RL.8.10
8	RL	RL.8.8
8	RST	RST.6-8.10
8	SL	SL.8.1
8	SL	SL.8.4
8	SL	SL.8.5
8	SL	SL.8.6
8	W	W.8.10
8	W	W.8.6
8	W	W.8.7
8	WHST	WHST.6-8.10
8	WHST	WHST.6-8.6
8	WHST	WHST.6-8.7
11	RH	RH.11-12.10
11	RI	RI.11-12.10
11	RL	RL.11-12.10
11	RL	RL.11-12.8
11	RST	RST.11-12.10
11	SL	SL.11-12.1
11	SL	SL.11-12.4
11	SL	SL.11-12.5
11	SL	SL.11-12.6
11	W	W.11-12.10
11	W	W.11-12.6
11	W	W.11-12.7
11	WHST	WHST.11-12.10
11	WHST	WHST.11-12.6
11	WHST	WHST.11-12.7

APPENDIX F: CLAIM 1 EMPHASIS BREAKOUT TABLES FOR MATHEMATICS

Note: there were no additional and supporting targets identified in the specifications for Grade 11.

Table F -1. A.Math.CR.PWA-1. Pairwise Percent Agreement among Reviewers' Mapping of Targets and Grade-level Standards by Emphasis

		Descriptives			Agreement
Grade	Claim	Avg # of Reviewers n	# of Targets	Avg # of Pairs	Pairwise Agreement %
Major Targets					
3	1	5	10	10	55.5%
4	1	5	6	10	63%
5	1	5	5	10	64.5%
6	1	5	6	10	54.2%
7	1	5	4	10	66.1%
8	1	5	6	10	75.1%
11	1	4	16	6	62.2%
Additional & Supporting Targets					
3	1	5	1	10	44.8%
4	1	5	6	10	61.9%
5	1	5	6	10	78.9%
6	1	5	4	10	67.5%
7	1	5	5	10	59.5%
8	1	5	4	10	71.4%

Table F-2. A.CR-1: Mean Percentage of ELA/Literacy Targets at Each Holistic Rating (Collectively Reflected by the Grade-Level Standards) by Emphasis

Grade	Claim	# of Targets in Claim	Holistic Target Rating				
			Fully-aligned	Mostly-aligned	Somewhat-aligned	Small-portion aligned	Not-aligned at all
			% (n)	% (n)	% (n)	% (n)	% (n)
Major Targets							
3	1	10	98.0% (9.6)	2.0% (0.2)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
4	1	6	100.0% (6.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
5	1	5	88.0% (4.4)	12.0% (0.6)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
6	1	6	72.7% (4.2)	20.7% (1.2)	0.0% (0.0)	6.7% (0.4)	0.0% (0.0)
7	1	4	95.0% (3.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	5.0% (0.2)
8	1	6	76.7% (4.4)	20.0% (1.2)	0.0% (0.0)	0.0% (0.0)	3.3% (0.2)
11	1	16	70.3% (11.3)	28.1% (4.5)	1.6% (0.3)	0.0% (0.0)	0.0% (0.0)
Additional & Supporting Targets							
3	1	1	100.0% (1.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
4	1	6	100.0% (5.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
5	1	6	71.7% (4.0)	28.3% (1.6)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
6	1	4	75.0% (3.0)	20.0% (0.8)	5.0% (0.2)	0.0% (0.0)	0.0% (0.0)
7	1	5	96.0% (4.8)	4.0% (0.2)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
8	1	4	90.0% (3.6)	10.0% (0.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)

Table F-3. A.Math.CR-3.GD-1 Comparison of Reviewer and Specifications Target CCSS Ratings Descriptives by Emphasis

Grade	Claim		Reviewer Descriptives			Specifications Descriptives		
		Total number of targets in claim N	Avg # of grade-level standards per target n	Minimum # of grade-level standards per target n	Maximum # of grade-level standards per target n	Avg # of grade-level standards per target n	Minimum # of grade-level standards per target n	Maximum # of grade-level standards per target n
Major Targets								
3	1	10	9.9	1	23	3.5	2	8
4	1	6	7.2	3	16	3.7	3	4
5	1	5	6	3	19	4.2	2	6
6	1	6	5	2	13	3.8	2	5
7	1	4	5.9	1	12	3.5	3	4
8	1	6	5	1	11	4.2	3	6
11	1	16	6.1	2	18	3.3	2	8
Additonal & Supporting Targets								
3	1	1	6.2	2	16	3	3	3
4	1	6	5.8	2	19	3.3	2	6
5	1	6	2.9	2	6	2.5	2	3
6	1	4	3.9	2	7	4	3	5
7	1	5	4.4	2	10	3.8	3	5
8	1	4	2.9	1	5	3.3	2	5

*Note: Due to the structure of the specifications for Claims 2 - 4, analyses for these claims occurred at the claim level (rather than the target level)

Table F-4. A.CR-3: Mean Percentage of Mathematics Grade-level Standards Aligned to Intended Targets (Workshop 1) by Emphasis

Grade	Claim		>= 50% Reviewer Agreement Descriptives		Content Representation		
		Total number of targets in claim N	Number of targets included in analysis[1] n	Avg # of grade-level standards per target with 50% reviewer agreement n	Avg % of grade-level standards per target that matched the intended mapping % (n)	Avg % of grade-level standards per target that fell within the intended domain % (n)	Avg % of grade-level standards per target that fell within the intended cluster % (n)
Major Targets							
3	1	10	10	9.0	55.6% (3.1)	92.5% (7.7)	67.5% (4.5)
4	1	6	6	6.7	61.7% (3.7)	81.7% (5.5)	73.4% (4.8)
5	1	5	5	5.0	67.2% (3.2)	100.0% (5.0)	100% (5.0)
6	1	6	6	4.0	79.9% (2.8)	100.0% (4.0)	100% (4.0)
7	1	4	4	6.8	62.6% (3.5)	100.0% (6.8)	100% (6.8)
8	1	6	6	5.2	84.2% (4.0)	100.0% (5.2)	100% (5.2)
11	1	16	16	4.5	74.9% (3.0)	90.9% (3.8)	89.6% (3.7)
Additional & Supporting Targets							
3	1	1	1	4.0	75.0% (3.0)	75.0% (3.0)	75% (3.0)
4	1	6	6	4.5	78.9% (3.3)	78.9% (3.3)	78.8% (3.3)
5	1	6	6	2.5	100.0% (2.5)	100.0% (2.5)	100% (2.5)
6	1	4	4	4.0	91.7% (3.8)	100.0% (4.0)	100% (4.0)
7	1	5	5	4.2	92.5% (3.6)	100.0% (4.2)	100% (4.2)
8	1	4	4	2.8	100.0% (2.8)	100.0% (2.8)	100% (2.8)

*Note: Due to the structure of the specifications for Claims 2 - 4, analyses for these claims occurred at the claim level (rather than the target level)

Table F-5. A.CR-4: Mean Percentage of Mathematics Grade-level Standards Aligned to Intended Targets Based on Reviewers Identifying Targets Aligned to Each Grade-level Standard by Emphasis (Workshop 2)

Grade	Claim	>= 50% Reviewer Agreement Descriptives			Content Representation		
		Total number of targets in claim N	Number of targets included in analysis[1] n	Avg number of grade-level standards per target with 50% reviewer agreement n	Avg % of grade-level standards per target that matched the intended mapping % (n)	Avg % of grade-level standards per target that fell within the intended domain % (n)	Avg % of grade-level standards per target that fell within the intended cluster % (n)
Major Targets							
3	1	10	10	15.6	21.5% (3.0)	60.1% (8.5)	32.9% (4.5)
4	1	6	6	9.2	51.1% (3.7)	72.8% (6.0)	60.8% (4.8)
5	1	5	5	9.4	48.5% (4.2)	85.8% (7.8)	77.4% (5.0)
6	1	6	6	8.2	60.8% (4.0)	84.0% (6.8)	79.8% (4.0)
7	1	4	4	7.0	60.8% (3.5)	100.0% (7.0)	100% (6.8)
8	1	6	6	5.0	82.6% (3.8)	100.0% (5.0)	100% (5.2)
11	1	16	16	5.7	64.8% (3.3)	91.1% (4.9)	81.4% (3.7)
Additional & Supporting Targets							
3	1	1	1	11.0	27.3% (3.0)	27.3% (3.0)	27.2% (3.0)
4	1	6	6	7.8	50.2% (3.2)	63.3% (4.5)	43.7% (3.3)
5	1	6	6	6.3	58.8% (2.5)	62.9% (2.7)	58.7% (2.5)
6	1	4	4	4.0	91.7% (3.8)	100.0% (4.0)	100% (4.0)
7	1	5	5	4.6	90.0% (3.6)	100.0% (4.6)	100% (4.2)
8	1	4	4	3.3	100.0% (3.3)	100.0% (3.3)	100% (2.8)

*These data are from Workshop 2 (A.CR-3 was from Workshop 1). Reviewers identified targets aligned to each standard

*Note: Due to the structure of the specifications for Claims 2 - 4, analyses for these claims occurred at the claim level (rather than the target level)

Table F-6 A.CR-4.Supp-1 Comparison of Mean Percentage of Grade-level Standards Aligned to Intended Targets by Emphasis (Workshops 1. vs 2)

Grade	Claim	CR-3 (Workshop 1)		CR-4 (Workshop 2)		Difference (CR4-CR3)		
		Avg number of grade-level standards per target with 50% reviewer agreement	Avg % of grade-level standards per target that matched the intended mapping	Avg number of grade-level standards per target with 50% reviewer agreement	Avg % of grade-level standards per target that matched the intended mapping	Avg number of grade-level standards per target with 50% reviewer agreement	Avg % of grade-level standards per target that matched the intended mapping	
		n	% (n)	n	% (n)	n	%	n
Major Targets								
3	1	9.0	55.6% (3.1)	15.6	21.5% (3.0)	6.6	-34.1%	-0.1
4	1	6.7	61.7% (3.7)	9.2	51.1% (3.7)	2.5	-10.6%	0
5	1	5.0	67.2% (3.2)	9.4	48.5% (4.2)	4.4	-18.7%	1
6	1	4.0	79.9% (2.8)	8.2	60.8% (4.0)	4.2	-19.1%	1.2
7	1	6.8	62.6% (3.5)	7.0	60.8% (3.5)	0.2	-1.8%	0
8	1	0%	84.2% (4.0)	5.0	82.6% (3.8)	5.0	-1.6%	-0.2
11	1	4.5	74.9% (3.0)	5.7	64.8% (3.3)	1.2	-10.1%	0.3
Additional & Supporting Targets								
3	1	4.0	75.0% (3.0)	11.0	27.3% (3.0)	7.0	-47.7%	0
4	1	4.5	78.9% (3.3)	7.8	50.2% (3.2)	3.3	-28.7%	-0.1
5	1	2.5	100.0% (2.5)	6.3	58.8% (2.5)	3.8	-41.2%	0
6	1	4.0	91.7% (3.8)	4.0	91.7% (3.8)	0.0	0.0%	0
7	1	4.2	92.5% (3.6)	4.6	90.0% (3.6)	0.4	-2.5%	0
8	1	2.8	100.0% (2.8)	3.3	100.0% (3.3)	0.5	0.0%	0.5

Table F-7. A.CR-5: Mean Percentage of Mathematics Targets Aligned to Each Mathematical Practice by Emphasis

Grade	Mathematical Practice	Claim 1			
		Aligned		Not Aligned	
		Major Targets % (n)	Additional & Supporting Targets % (n)	Major Targets % (n)	Additional & Supporting Targets % (n)
3	1	100.0% (9.8)	100.0% (1.0)	0.0% (0.0)	0.0% (0.0)
	2	98.0% (9.6)	100.0% (1.0)	2.0% (0.2)	0.0% (0.0)
	3	87.6% (8.6)	60.0% (0.6)	12.4% (1.2)	40.0% (0.4)
	4	100.0% (9.8)	100.0% (1.0)	0.0% (0.0)	0.0% (0.0)
	5	91.3% (9.0)	100.0% (1.0)	8.7% (0.9)	0.0% (0.0)
	6	100.0% (9.8)	100.0% (1.0)	0.0% (0.0)	0.0% (0.0)
	7	100.0% (9.8)	100.0% (1.0)	0.0% (0.0)	0.0% (0.0)
	8	100.0% (9.8)	100.0% (1.0)	0.0% (0.0)	0.0% (0.0)
4	1	100.0% (6.0)	100.0% (5.6)	0.0% (0.0)	0.0% (0.0)
	2	90.0% (5.4)	92.0% (5.4)	10.0% (0.6)	8.0% (0.5)
	3	60.0% (3.6)	64.0% (3.8)	40.0% (2.4)	36.0% (2.1)
	4	80.0% (4.8)	82.0% (4.8)	20.0% (1.2)	18.0% (1.1)
	5	80.0% (4.8)	82.7% (4.8)	20.0% (1.2)	17.3% (1.0)
	6	80.0% (4.8)	93.3% (5.4)	20.0% (1.2)	6.7% (0.4)
	7	93.3% (5.6)	83.3% (4.8)	6.7% (0.4)	16.7% (1.0)
	8	83.3% (5.0)	83.3% (4.8)	16.7% (1.0)	16.7% (1.0)
5	1	92.0% (4.6)	70.0% (3.4)	8.0% (0.4)	30.0% (1.5)
	2	60.0% (3.0)	57.3% (2.8)	40.0% (2.0)	42.7% (2.1)
	3	28.0% (1.4)	38.3% (2.0)	72.0% (3.6)	61.7% (3.2)
	4	80.0% (4.0)	67.3% (3.6)	20.0% (1.0)	32.7% (1.7)
	5	56.0% (2.8)	57.3% (3.0)	44.0% (2.2)	42.7% (2.2)
	6	52.0% (2.6)	58.0% (3.0)	48.0% (2.4)	42.0% (2.2)
	7	60.0% (3.0)	56.7% (3.0)	40.0% (2.0)	43.3% (2.3)
	8	52.0% (2.6)	38.3% (2.0)	48.0% (2.4)	61.7% (3.2)
6	1	96.7% (5.8)	86.7% (3.4)	3.3% (0.2)	13.3% (0.5)
	2	76.7% (4.6)	90.0% (3.6)	23.3% (1.4)	10.0% (0.4)
	3	50.0% (3.0)	61.7% (2.4)	50.0% (3.0)	38.3% (1.5)
	4	73.3% (4.4)	65.0% (2.6)	26.7% (1.6)	35.0% (1.4)
	5	56.7% (3.4)	50.0% (2.0)	43.3% (2.6)	50.0% (2.0)
	6	73.3% (4.4)	65.0% (2.6)	26.7% (1.6)	35.0% (1.4)
	7	53.3% (3.2)	75.0% (3.0)	46.7% (2.8)	25.0% (1.0)
	8	43.3% (2.6)	50.0% (2.0)	56.7% (3.4)	50.0% (2.0)

Table F-7. (Continued)

Grade	Mathematical Practice	Claim 1			
		Aligned		Not Aligned	
		Major Targets % (n)	Additional & Supporting Targets % (n)	Major Targets % (n)	Additional & Supporting Targets % (n)
7	1	100.0% (4.0)	100.0% (5.0)	0.0% (0.0)	0.0% (0.0)
	2	90.0% (3.6)	96.0% (4.8)	10.0% (0.4)	4.0% (0.2)
	3	55.0% (2.2)	88.0% (4.4)	45.0% (1.8)	12.0% (0.6)
	4	90.0% (3.6)	100.0% (5.0)	10.0% (0.4)	0.0% (0.0)
	5	85.0% (3.4)	92.0% (4.6)	15.0% (0.6)	8.0% (0.4)
	6	95.0% (3.8)	84.0% (4.2)	5.0% (0.2)	16.0% (0.8)
	7	90.0% (3.6)	60.0% (3.0)	10.0% (0.4)	40.0% (2.0)
	8	85.0% (3.4)	56.0% (2.8)	15.0% (0.6)	44.0% (2.2)
8	1	100.0% (5.4)	100.0% (3.8)	0.0% (0.0)	0.0% (0.0)
	2	93.3% (5.6)	95.0% (3.8)	6.7% (0.4)	5.0% (0.2)
	3	80.0% (4.8)	75.0% (3.0)	20.0% (1.2)	25.0% (1.0)
	4	86.7% (5.2)	85.0% (3.4)	13.3% (0.8)	15.0% (0.6)
	5	83.3% (5.0)	90.0% (3.6)	16.7% (1.0)	10.0% (0.4)
	6	96.7% (5.8)	100.0% (4.0)	3.3% (0.2)	0.0% (0.0)
	7	80.0% (4.8)	85.0% (3.4)	20.0% (1.2)	15.0% (0.6)
	8	80.0% (4.6)	70.0% (2.8)	20.0% (1.1)	30.0% (1.2)
11	1	92.2% (14.8)	. % (.)	7.8% (1.3)	. % (.)
	2	82.8% (13.3)	. % (.)	17.2% (2.8)	. % (.)
	3	25.5% (4.0)	. % (.)	74.5% (11.7)	. % (.)
	4	56.3% (9.0)	. % (.)	43.8% (7.0)	. % (.)
	5	68.8% (11.0)	. % (.)	31.3% (5.0)	. % (.)
	6	65.6% (10.5)	. % (.)	34.4% (5.5)	. % (.)
	7	76.6% (12.3)	. % (.)	23.4% (3.8)	. % (.)
	8	22.1% (3.5)	. % (.)	77.9% (12.3)	. % (.)

*Note: Lower percentages of alignment for Claim 1 targets are not unexpected based on the design of the specifications.

Table F-8. A.CR-6: Pairwise Agreement between Reviewers' and Intended Mapping of Mathematics Targets and Grade-level Standards by Emphasis

Grade	Claim	Descriptives			Agreement			
		# of Reviewers % (n)	# of Targets	# of Ratings	Pairwise Agree-ment	Pairwise Agreement (Cluster-level)	Hit All Intended Standards (but noted others) Avg % (n Reviewers)	Hit At Least 50% of the Intended Standards Avg % (n Reviewers)
Major Targets								
3	1	4.9	10	49	43.2%	46.0%	71.0% (3.5)	100.0% (4.9)
4	1	5.0	6	30	59.6%	69.0%	56.7% (2.8)	100.0% (5.0)
5	1	5.0	5	25	56.4%	86.1%	28.0% (1.4)	92.0% (4.6)
6	1	5.0	6	30	57.5%	82.8%	30.0% (1.5)	80.0% (4.0)
7	1	5.0	4	20	55.0%	90.0%	55.0% (2.8)	95.0% (4.8)
8	1	5.0	6	30	79.4%	91.1%	20.0% (1.0)	93.3% (4.7)
11	1	4.0	16	64	59.2%	72.2%	64.1% (2.6)	100.0% (4.0)
Additional & Supporting Targets								
3	1	5.0	1	5	61.1%	66.7%	60.0% (3.0)	100.0% (5.0)
4	1	5.0	6	30	70.4%	70.6%	50.0% (2.5)	100.0% (5.0)
5	1	5.0	6	30	87.9%	90.0%	16.7% (0.8)	100.0% (5.0)
6	1	5.0	4	20	78.9%	97.5%	10.0% (0.5)	95.0% (4.8)
7	1	5.0	5	25	73.9%	84.3%	16.0% (0.8)	84.0% (4.2)
8	1	5.0	4	20	82.1%	95.0%	5.0% (0.3)	85.0% (4.3)

*Note: Due to the structure of the specifications for Claims 2 - 4, analyses for these claims occurred at the claim level (rather than the target level)
*Decimals in the # of Reviewers column indicate missing data

Table F-9. A.Math.DD.GD-1 Descriptive Comparison of Reviewer and Specifications Target DOK Ratings by Grade and Claim by Emphasis

Grade	Claim	Avg # of DOK Levels Indicated per Target	
		Reviewers	Specifications
Major Targets			
3	1	2.4	1.3
4	1	2.5	2.0
5	1	2.0	2.0
6	1	1.9	1.8
7	1	1.3	1.8
8	1	1.2	2.0
11	1	2.3	1.9
Additional & Supporting Targets			
3	1	2.6	2.0
4	1	2.3	2.0
5	1	1.6	1.3
6	1	1.8	1.8
7	1	1.2	1.8
8	1	1.2	2.0

Table F-10. A.Math.DD.PWA-1. Pairwise Percent Agreement Among Reviewers' Target DOK Ratings by Emphasis

		Descriptives			Agreement
Grade	Claim	Avg # of Reviewers n	# of Targets	Avg # of Reviewer Pairs n	Pairwise Agreement
Major Targets					
3	1	4.9	10	9.6	82.7%
4	1	5.0	6	10.0	82.2%
5	1	5.0	5	10.0	73.0%
6	1	5.0	6	10.0	61.1%
7	1	5.0	4	10.0	33.3%
8	1	5.0	6	10.0	42.5%
11	1	4.0	16	6.0	77.3%
Additional & Supporting Targets					
3	1	5.0	1	10.0	80.0%
4	1	5.0	6	10.0	82.8%
5	1	5.0	6	10.0	56.9%
6	1	5.0	4	10.0	62.1%
7	1	5.0	5	10.0	42.0%
8	1	5.0	4	10.0	35.0%

Table F-11. A.DD-1: Reviewers' Mean Percentage of Mathematics at Each DOK Level (Max) by Grade and Claim Compared to Content Specifications by Emphasis

Grade	Claim	DOK 1		DOK 2		DOK 3		DOK 4	
		Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)
Major Targets									
3	1	2.0% (0.2)	50.0% (5.0)	49.1% (4.8)	50.0% (5.0)	48.9% (4.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
4	1	0.0% (0.0)	0.0% (0.0)	53.3% (3.2)	100.0% (6.0)	46.7% (2.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
5	1	8.0% (0.4)	0.0% (0.0)	68.0% (3.4)	100.0% (5.0)	24.0% (1.2)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
6	1	10.0% (0.6)	0.0% (0.0)	50.0% (3.0)	100.0% (6.0)	40.0% (2.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
7	1	5.0% (0.2)	0.0% (0.0)	55.0% (2.2)	100.0% (4.0)	30.0% (1.2)	0.0% (0.0)	10.0% (0.4)	0.0% (0.0)
8	1	10.0% (0.6)	0.0% (0.0)	46.7% (2.8)	100.0% (6.0)	43.3% (2.6)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
11	1	1.6% (0.3)	0.0% (0.0)	34.4% (5.5)	93.8% (15.0)	62.5% (10.0)	6.3% (1.0)	1.6% (0.3)	0.0% (0.0)
Additional & Supporting Targets									
3	1	0.0% (0.0)	0.0% (0.0)	40.0% (0.4)	100.0% (1.0)	60.0% (0.6)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
4	1	0.0% (0.0)	0.0% (0.0)	60.0% (3.6)	83.3% (5.0)	40.0% (2.4)	16.7% (1.0)	0.0% (0.0)	0.0% (0.0)
5	1	20.0% (1.2)	16.7% (1.0)	56.7% (3.4)	83.3% (5.0)	23.3% (1.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
6	1	20.0% (0.8)	0.0% (0.0)	45.0% (1.8)	100.0% (4.0)	30.0% (1.2)	0.0% (0.0)	5.0% (0.2)	0.0% (0.0)
7	1	0.0% (0.0)	0.0% (0.0)	28.0% (1.4)	100.0% (5.0)	64.0% (3.2)	0.0% (0.0)	8.0% (0.4)	0.0% (0.0)
8	1	15.0% (0.6)	0.0% (0.0)	55.0% (2.2)	100.0% (4.0)	30.0% (1.2)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)

*Note: For each group (reviewers and specifications) the percentages across DOK levels are mutually exclusive.

Table F-12. A.DD-2: Reviewers' Mean Percentage of Mathematics Targets at Each DOK Level (Independent) by Grade and Claim Compared to Content Specifications by Emphasis

Grade	Claim	DOK 1		DOK 2		DOK 3		DOK 4	
		Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)	Reviewers % (n)	Specs % (n)
Major Targets									
3	1	96.0% (9.4)	80.0% (8.0)	98.0% (9.6)	50.0% (5.0)	48.9% (4.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
4	1	100.0% (6.0)	100.0% (6.0)	100.0% (6.0)	100.0% (6.0)	46.7% (2.8)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
5	1	84.0% (4.2)	100.0% (5.0)	92.0% (4.6)	100.0% (5.0)	24.0% (1.2)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
6	1	66.7% (4.0)	83.3% (5.0)	86.7% (5.2)	100.0% (6.0)	40.0% (2.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
7	1	15.0% (0.6)	75.0% (3.0)	60.0% (2.4)	100.0% (4.0)	40.0% (1.6)	0.0% (0.0)	10.0% (0.4)	0.0% (0.0)
8	1	26.7% (1.6)	100.0% (6.0)	53.3% (3.2)	100.0% (6.0)	43.3% (2.6)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
11	1	67.2% (10.8)	87.5% (14.0)	98.4% (15.8)	100.0% (16.0)	64.1% (10.3)	6.3% (1.0)	1.6% (0.3)	0.0% (0.0)
Additional & Supporting Targets									
3	1	100.0% (1.0)	100.0% (1.0)	100.0% (1.0)	100.0% (1.0)	60.0% (0.6)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
4	1	90.0% (5.4)	83.3% (5.0)	100.0% (6.0)	100.0% (6.0)	40.0% (2.4)	16.7% (1.0)	0.0% (0.0)	0.0% (0.0)
5	1	60.0% (3.6)	50.0% (3.0)	80.0% (4.8)	83.3% (5.0)	23.3% (1.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
6	1	65.0% (2.6)	75.0% (3.0)	75.0% (3.0)	100.0% (4.0)	35.0% (1.4)	0.0% (0.0)	5.0% (0.2)	0.0% (0.0)
7	1	4.0% (0.2)	80.0% (4.0)	40.0% (2.0)	100.0% (5.0)	72.0% (3.6)	0.0% (0.0)	8.0% (0.4)	0.0% (0.0)
8	1	25.0% (1.0)	100.0% (4.0)	60.0% (2.4)	100.0% (4.0)	30.0% (1.2)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)

*Note: For each group (reviewers and specifications) the percentages across DOK levels are not mutually exclusive since a target could have multiple DOK levels.

Table F-13. A.DD-3: Pairwise Percent Agreement between Reviewers' and Intended Mathematics Target DOK Ratings by Emphasis

Grade	Claim	Descriptives			Agreement
		Avg # of Reviewers n	# of Targets	# of Ratings	Pairwise Agreement %
Major Targets					
3	1	4.9	10	49	55.7%
4	1	5.0	6	30	84.4%
5	1	5.0	5	25	81.3%
6	1	5.0	6	30	72.2%
7	1	5.0	4	20	34.2%
8	1	5.0	6	30	40.0%
11	1	4.0	16	64	68.2%
Additional & Supporting Targets					
3	1	5.0	1	5	80.0%
4	1	5.0	6	30	86.7%
5	1	5.0	6	30	63.9%
6	1	5.0	4	20	65.8%
7	1	5.0	5	25	26.0%
8	1	5.0	4	20	42.5%

Table F-14. Math.DC-1a. Percentage of Mathematics Targets with DOKs Consistent with Intended Grade-Level Standards that Matched Intended Mapping for All Targets by Emphasis– All CCSS within Range

Grade	Claim	Descriptives			DOK Consistency				
					Consistent	Inconsistent			
		Total number of targets in claim	Number of targets included in analysis[1]	Avg % of grade-level standards per target with >= 50% reviewer agreement[2]	% of Targets that Have All Mapped CCSS Consistent[3]	% of Targets With All Mapped CCSS Inconsistent	Avg % of CCSS per Inconsistent Target who's Max dok Consensus > Specs	Avg % of CCSS per Target who's Min dok Consensus < Specs	Number of Targets With < 50% agreement
		N	n	% (n)	% (n)	% (n)	% (n)	% (n)	n
Major targets									
3	1	10	10	95.0% (3.10)	20.0% (2)	80.0% (8)	70.8% (2.12)	25.0% (0.75)	0
4	1	6	6	100.0% (3.67)	50.0% (3)	50.0% (3)	38.9% (1.33)	0.0% (0.00)	0
5	1	5	5	76.0% (3.00)	60.0% (3)	40.0% (2)	50.0% (1.50)	0.0% (0.00)	0
6	1	6	6	76.7% (2.67)	66.7% (4)	33.3% (2)	66.7% (1.50)	0.0% (0.00)	0
7	1	4	4	100.0% (3.50)	25.0% (1)	75.0% (3)	47.2% (1.67)	8.3% (0.33)	0
8	1	6	6	97.2% (4.00)	83.3% (5)	16.7% (1)	66.7% (2.00)	0.0% (0.00)	0
11	1	16	16	98.4% (3.25)	31.2% (5)	68.8% (11)	54.5% (1.64)	12.1% (0.27)	0
Additional & Supporting Targets									
3	1	1	1	100.0% (3.00)	100.0% (1)	0.0% (0)			0
4	1	6	6	100.0% (3.33)	83.3% (5)	16.7% (1)	0.0% (0.00)	100.0% (2.00)	0
5	1	6	6	100.0% (2.50)	33.3% (2)	66.7% (4)	75.0% (1.75)	8.3% (0.25)	0
6	1	4	4	91.7% (3.75)	50.0% (2)	50.0% (2)	45.0% (2.00)	37.5% (1.50)	0
7	1	5	5	95.0% (3.60)	0.0% (0)	100.0% (5)	57.0% (2.00)	0.0% (0.00)	0
8	1	4	4	90.0% (2.75)	100.0% (4)	0.0% (0)			0
11	1								

*Note: Due to the structure of the specifications for Claims 2 - 4, analyses for these claims were excluded

[1]Number of targets with at least one standard with 50% reviewer agreement

[2]Standards that matched the intended mapping with greater than or equal to 50% reviewer agreement

[3]Consistent was defined as the grade-level standard DOK levels falling entirely within the range of the intended target DOK levels

Table F-15. Math.DC-1b. Percentage of Mathematics Targets with DOKs Consistent with Intended Grade-Level Standards that Matched Intended Mapping for All Targets by Emphasis– All CCSS At Least One

Grade	Claim	Descriptives			DOK Consistency				
					Consistent	Inconsistent			
		Total number of targets in claim	Number of targets included in analysis[1]	Avg % of grade-level standards per target with >= 50% reviewer agreement[2]	% of Targets that Have All Mapped CCSS Consistent[3]	% of Targets With All Mapped CCSS Inconsistent	Avg % of CCSS per Inconsistent Target who's Max dok Consensus > Specs	Avg % of CCSS per Target who's Min dok Consensus < Specs	Number of Targets With < 50% agreement
		N	n	% (n)	% (n)	% (n)	% (n)	% (n)	n
Major targets									
3	1	10	10	95.0% (3.10)	100.0% (10)	0.0% (0)			0
4	1	6	6	100.0% (3.67)	100.0% (6)	0.0% (0)			0
5	1	5	5	76.0% (3.00)	100.0% (5)	0.0% (0)			0
6	1	6	6	76.7% (2.67)	83.3% (5)	16.7% (1)	100.0% (2.00)	0.0% (0.00)	0
7	1	4	4	100.0% (3.50)	25.0% (1)	75.0% (3)	47.2% (1.67)	8.3% (0.33)	0
8	1	6	6	97.2% (4.00)	83.3% (5)	16.7% (1)	66.7% (2.00)	0.0% (0.00)	0
11	1	16	16	98.4% (3.25)	100.0% (16)	0.0% (0)			0
Additional & Supporting Targets									
3	1	1	1	100.0% (3.00)	100.0% (1)	0.0% (0)			0
4	1	6	6	100.0% (3.33)	100.0% (6)	0.0% (0)			0
5	1	6	6	100.0% (2.50)	83.3% (5)	16.7% (1)	100.0% (2.00)	0.0% (0.00)	0
6	1	4	4	91.7% (3.75)	75.0% (3)	25.0% (1)	50.0% (2.00)	75.0% (3.00)	0
7	1	5	5	95.0% (3.60)	20.0% (1)	80.0% (4)	62.9% (2.25)	0.0% (0.00)	0
8	1	4	4	90.0% (2.75)	100.0% (4)	0.0% (0)			0
11	1								

*Note: Due to the structure of the specifications for Claims 2 - 4, analyses for these claims were excluded

[1]Number of targets with at least one standard with 50% reviewer agreement

[2]Standards that matched the intended mapping with greater than or equal to 50% reviewer agreement

[3]Consistent was defined as at least one of the grade-level standard DOK levels matched at least one DOK level of the intended target

APPENDIX G: CONNECTION B CLAIM 1 EMPHASIS BREAKOUT TABLES FOR MATHEMATICS

Note: there were no additional and supporting targets identified in the specifications for Grade 11.

Table G-1. B.CR-1: Mean Percentage of Mathematics Targets at Each Holistic Rating (Collectively Reflected by the Evidence Statements) by Emphasis

		Holistic Target Rating			
Grade	Claim	Total Number of Targets	Fully-aligned % (n)	Partially-aligned % (n)	Not-aligned % (n)
Major Targets					
3	1	10	81.0% (8.0)	19.0% (1.9)	0.0% (0.0)
4	1	6	87.9% (5.2)	12.1% (0.7)	0.0% (0.0)
5	1	5	68.2% (3.5)	31.8% (1.6)	0.0% (0.0)
6	1	6	86.7% (5.1)	13.3% (0.8)	0.0% (0.0)
7	1	4	80.8% (3.2)	19.2% (0.8)	0.0% (0.0)
8	1	6	93.6% (5.6)	6.4% (0.4)	0.0% (0.0)
11	1	16	79.5% (12.6)	20.5% (2.7)	0.0% (0.0)
Additional & Supporting Targets					
3	1	1	100.0% (1.0)	0.0% (0.0)	0.0% (0.0)
4	1	6	95.2% (5.7)	4.8% (0.3)	0.0% (0.0)
5	1	6	91.1% (5.4)	8.9% (0.5)	0.0% (0.0)
6	1	4	90.0% (3.3)	10.0% (0.4)	0.0% (0.0)
7	1	5	81.5% (4.1)	18.5% (0.9)	0.0% (0.0)
8	1	4	98.1% (3.9)	1.9% (0.1)	0.0% (0.0)

Table G-2. B.CR-2: Mean Percentage of Mathematics CAT Evidence Statements Aligned to Targets, by Grade and Claim by Emphasis

		Individual Evidence Statement Ratings			
Grade	Claim	Total Number of ES	Fully-aligned % (n)	Partially-aligned % (n)	Not-aligned % (n)
Major Targets					
3	1	28	26.3% (7.4)	73.0% (20.1)	0.8% (0.2)
4	1	24	27.7% (6.6)	72.3% (17.3)	0.0% (0.0)
5	1	18	2.6% (0.5)	95.9% (17.3)	1.5% (0.3)
6	1	32	1.0% (0.3)	97.7% (31.3)	1.3% (0.4)
7	1	19	4.9% (0.9)	92.7% (17.6)	2.4% (0.5)
8	1	27	3.1% (0.8)	95.7% (25.8)	1.1% (0.3)
11	1	54	10.1% (5.2)	89.6% (47.6)	0.4% (0.2)
Additional & Supporting Targets					
3	1	2	30.8% (0.6)	69.2% (1.4)	0.0% (0.0)
4	1	20	27.9% (5.6)	70.3% (14.0)	1.8% (0.4)
5	1	14	6.7% (0.9)	93.3% (13.1)	0.0% (0.0)
6	1	18	1.1% (0.2)	98.5% (17.7)	0.4% (0.1)
7	1	16	7.2% (1.2)	91.3% (14.5)	1.4% (0.2)
8	1	16	8.2% (1.2)	90.3% (13.5)	1.5% (0.2)

Table G-3. B.DC-1: Mean Percentage of Mathematics AT Evidence Statements with DOK Levels Consistent with the Intended Targets

		Consistent		Inconsistent	
Grade	Claim	ES Within Range of Intended Target % (n)	ES DOK Match at Least One Intended Target DOK % (n)	ES max DOK > max DOK of Intended Target % (n)	ES min DOK < min DOK of Intended Target % (n)
Major Targets					
3	1	48.7% (13.6)	96.2% (26.9)	44% (12.4)	12% (3.3)
4	1	83.3% (20.0)	100.0% (24.0)	17% (4.0)	0% (0.0)
5	1	87.8% (15.7)	98.1% (17.6)	12% (2.2)	0% (0.0)
6	1	77.9% (24.9)	97.9% (31.3)	20% (6.4)	3% (1.0)
7	1	76.1% (14.5)	87.0% (16.5)	11% (2.2)	13% (2.4)
8	1	78.0% (21.0)	84.3% (22.7)	22% (5.9)	0% (0.0)
11	1	74.0% (39.9)	90.5% (48.8)	21% (11.1)	6% (3.0)
Additional & Supporting Targets					
3	1	53.6% (1.1)	100.0% (2.0)	46% (0.9)	0% (0.0)
4	1	75.7% (15.1)	98.9% (19.8)	15% (2.9)	10% (1.9)
5	1	58.1% (8.1)	91.0% (12.7)	23% (3.2)	25% (3.5)
6	1	66.3% (11.9)	92.2% (16.6)	25% (4.5)	10% (1.7)
7	1	64.9% (10.4)	74.5% (11.9)	35% (5.5)	1% (0.2)
8	1	74.9% (11.2)	81.5% (12.2)	25% (3.8)	0% (0.0)

APPENDIX H: EVIDENCE STATEMENTS NOT MAPPED TO ANY SAMPLED ELA/LITERACY ITEMS

In this appendix we present the evidence statements (for ELA/literacy) that were not mapped to any of the CAT items sampled for this alignment study.

The vast majority of evidence statements had at least one, typically more, evidence statements mapped to them. Given the large number of evidence we see it as positive that the list provided below is so short. This indicates that, in general, the item writers are including items to represent a wide range of student knowledge.

Table H-1 provides the list of evidence statements not mapped to any sampled CAT items. There were three evidence statements at grade 3, three at grade 4, one at grade 7, 2 at grade 8, and two at grade 11. For each of these evidence statements, there were multiple other evidence statements within the target that were represented by sampled CAT items.

It should be noted that while we stratified our sample to ensure a representative coverage of items across assessment targets, we did not do so for evidence statements. Therefore, it is entirely possible that the evidence statements listed below were covered by items in the item pool that were not sampled.

Table H-1. List of ELA/Literacy Evidence Statements Not Mapped to Any Sampled CAT Items.

Grade	Claim	Target	Evidence Statement
3	2	9	12
3	2	9	13
3	2	9	19
4	2	9	4
4	2	9	5
4	2	9	6
7	1	10	4
8	1	3	4
8	1	10	4
11	1	10	4
11	2	2	2

APPENDIX I: ASSESSMENT TARGETS NOT MAPPED TO ANY SAMPLED MATHEMATICS ITEMS

In this appendix we present the assessment targets for mathematics that were not mapped to any of the CAT items sampled for this alignment study. All relevant ELA/literacy assessment targets were mapped to at least one item.

The vast majority of assessment targets had at least one, typically more, items mapped to them. Given the large number of assessment targets, we see it as positive that the list provided below is short. This indicates that, in general, the item writers are including items to represent a wide range of student knowledge.

Table I-1 presents the mathematics assessment targets that were not mapped to any of the sampled CAT items. At Grade 11, all assessment targets were represented by at least one item; at grade 4 and 7, two targets were not represented. Other grades had one target not represented by items. The number of targets presented below represents only a small percent of the total number of assessment targets.

One likely reason our study did not include items across all targets is that we only sought to include a representative sample of CAT items; some targets are best suited for performance task items. We did not include a representative sample of performance tasks in our study, only three per grade, so if particular content was to be covered by performance task items, we may have missed them in our study. We did determine that the targets listed in Table H-1 below were not represented by the few performance task items reviewers rated.

Table I-1. List of Mathematics Assessment Targets Not Mapped to Any Sampled CAT Items.

Grade	Claim	Target
3	4	G
4	4	B
4	4	G
5	4	G
6	4	G
7	2	B
7	4	G
8	4	G

APPENDIX J: LIST OF MATHEMATICAL PRACTICES

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.
8. Look for and express regularity in repeated reasoning.