



Stackable Instructionally- embedded Portable Science (SIPS) Assessments

Standard Setting Technical Report

By Daniel Lewis, Creative Measurement Solutions LLC

September 29, 2023



SIPS Standard Setting Technical Report was developed with funding from the U.S. Department of Education under the Competitive Grants for State Assessments Program CFDA 84.368A. The contents of this paper do not represent the policy of the U.S. Department of Education, and no assumption of endorsement by the Federal government should be made.

All rights reserved. Any or all portions of this document may be reproduced and distributed without prior permission, provided the source is cited as: Stackable, Instructionally-embedded, Portable Science (SIPS) Assessments Project. (2023). *SIPS Standard Setting Technical Report*. Lincoln, NE: Nebraska Department of Education.

Table of Contents

Introduction	1
Embedded Standard Setting and Assessment System Coherence	2
Coordination of Embedded Standard Setting Iterative Processes	3
PLD Development	3
Prompt Development & Prompt-PLD Alignment.....	3
ESS analyses	3
Vertical Articulation	4
Technical Report & Peer Review Evidence	4
PLD Development	5
SME Qualifications	5
PLD Development Process	5
Prompt-PLD Alignment	8
SME Qualifications	8
Training for Aligning Prompts to PLDs	8
SME Prompt-PLD Alignment Process	8
ESS analyses	10
Data	10
Initial ESS Cut Score Estimation	11
The Efficacy of SMEs’ Prompt-PLD Alignments.....	11
Correlations.....	12
Classification Agreement and Weighted Kappa.....	12
Summary of the Efficacy of the SME Prompt-PLD Alignments	19
Actionable Information: ESS-Inconsistent Prompt Review.....	19
Vertical Articulation	25
Policy Consideration: Response Probability	25
Initial and Vertically Articulated (Smoothed) Cut Scores and Associated Impact Data.....	25
Initial RP67 and RP50 cut scores and associated impact data	25
Vertically Articulated (Smooth) RP67 and RP50 cut scores and associated impact data	29
Technical Reporting: Validity & Peer Review Evidence	32
Standard Setting Validity Criteria from the Measurement Literature.....	32
Procedural Validity.....	32

Internal Validity.....	34
External Validity.....	34
Peer Review Standard Setting Critical Elements.....	34
Peer Review Standard Setting Validity Evidence	35
Summary	36
Integrating the EoUs: Methods for Reporting an End-Of-Year Summative Score or Performance Level ..	38
Performance Level Profiles	38
Empirical Data Analyses of Performance Level Profiles.....	38
Developing a SIPS PLD-Based Scale.....	40
Summary	42
References	43
Appendix A: SIPS Policy Level Descriptors	44
Appendix B: ESS Powerpoint Presentation	46
Appendix C: Detailed ESS Prompt Maps	53
Appendix D: Rosters of Inconsistent and Essentially Consistent Prompts	75

List of Exhibits

Exhibit 1. SIPS Embedded Standard Setting Iterative Process.....	2
Exhibit 2. Impact Data N-Counts.....	10
Exhibit 3. Initial SIPS Cut Scores for Grades 5 and 8.....	11
Exhibit 4. Correlation of SMEs’ Prompt-PLD Aligned Performance Level Ordinality and IRT RP Location .	12
Exhibit 5. Classification Agreement Crosstab	13
Exhibit 6. Kappa Interpretations	14
Exhibit 7. Agreement Rate and Weighted Kappa	15
Exhibit 8. Grade 5 EOU1 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels	15
Exhibit 9. Grade 5 EOU2 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels	16
Exhibit 10. Grade 5 EOU3 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels	16
Exhibit 11. Grade 5 EOU4 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels	17
Exhibit 12. Grade 8 EOU1 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels	17
Exhibit 13. Grade 8 EOU2 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels	18
Exhibit 14. Grade 8 EOU3 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels	18
Exhibit 15. Grade 8 EOU4 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels	19
Exhibit 16. Summary of Consistent, Essentially Consistent, and Inconsistent Prompts by Claim, Grade 520	
Exhibit 17. Summary of Consistent, Essentially Consistent, and Inconsistent Prompts by Claim, Grade 821	
Exhibit 18. Number and Percent of Prompts by Performance Level, Grade 5	22
Exhibit 19. Number and Percent of Prompts by Performance Level, Grade 8	23
Exhibit 20. Initial SIPS RP67 and RP50 Cut Scores across EoUs for Grades 5 and 8.....	26
Exhibit 21. SIPS Standard Errors.....	26
Exhibit 22. SIPS Grade 5 RP67 Initial Cut Score Impact Data	27
Exhibit 23. SIPS Grade 5 RP50 Initial Cut Score Impact Data	27
Exhibit 24. SIPS Grade 8 RP67 Initial Cut Score Impact Data	28
Exhibit 25. SIPS Grade 8 RP50 Initial Cut Score Impact Data	28
Exhibit 26. Vertically Articulated SIPS RP67 and RP50 Cut Scores.....	29
Exhibit 27. Vertical Articulation Adjustments to Cut Scores in Standard Error Units.....	29
Exhibit 28. SIPS Grade 5 RP67 Smoothed Cut Score Impact Data	30
Exhibit 29. SIPS Grade 5 RP50 Smoothed Cut Score Impact Data	30
Exhibit 30. SIPS Grade 8 RP67 Smoothed Cut Score Impact Data	31
Exhibit 31. SIPS Grade 8 RP50 Smoothed Cut Score Impact Data	31

Exhibit 32. Forms of Standard Setting Validity Evidence from the Literature 32

Exhibit 33. Examples of Evidence Supporting Peer Review Critical Element 6.2..... 35

Exhibit 34. ELPA21 Profiles of Proficiency..... 38

Exhibit 35. Grade 5 Cross-EoU Performance Level Profiles 39

Exhibit 36. Grade 8 Cross-EoU Performance Level Profiles 40

Introduction

This report describes the methods, analyses, and results supporting the Stackable, Instructionally-embedded, Portable Science Assessments (SIPS) project standard setting activities. Student performance on each SIPS End-of-Unit (EoU) assessment is reported in terms of four performance levels (Level 1, Level 2, Level 3, and Level 4).

Embedded Standard Setting (ESS) was employed to establish the SIPS cut scores. ESS (Lewis & Cook, 2020) is the logical extension of Principled Assessment Design (PAD) to standard setting. ESS transforms standard setting from a standalone workshop that typically occurs after test administration and just prior to score reporting to a set of processes that are an active part of the assessment development lifecycle. ESS processes directly contribute to the valid interpretation and use of test scores and improve test quality and the strength of validity arguments by maintaining a consistent focus on optimizing the evidentiary relationship between test prompts and the academic content standards reflected by the associated performance level descriptors (PLDs).

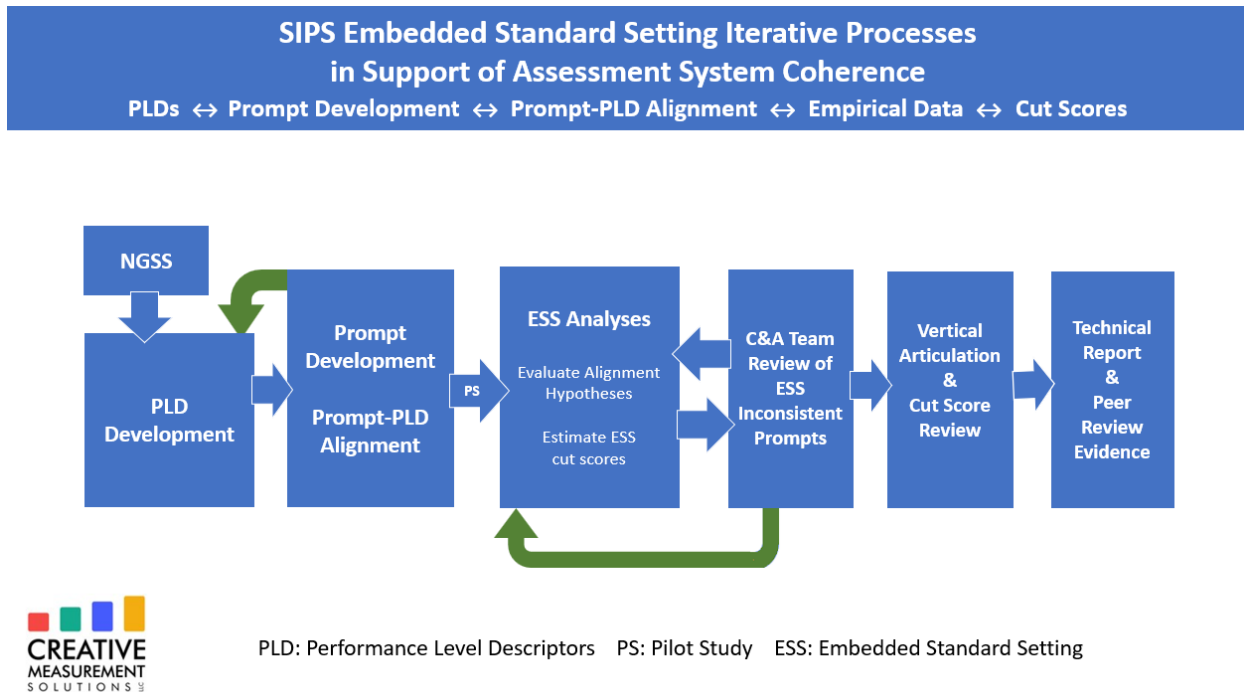
ESS is based on three big ideas:

1. PLDs are the fundamental component of standard setting. That is, the PLDs operationalize the policy goals of the sponsoring agency (as specified in the Policy PLDs) by articulating the knowledge, skills, and attributes (KSAs) of students at each performance level. The process of developing PLDs from the Next Generation Science Standards (NGSS) is represented by the first two boxes on the left in Exhibit 1.
2. Subject-Matter Expert (SME) alignment of test prompts to performance levels (Prompt-PLD alignment) are effectively the same judgments made during traditional prompt-based standard setting workshops (e.g., Bookmark, ID Matching, Modified Angoff Yes/No, etc.). Thus, the Prompt-PLD alignments resulting from SIPS SMEs' judgments during SIPS prompt development obviates the need for the judgments traditionally made by participants in a standard setting workshop.
3. When empirical data on test prompts are available from a pilot study, field test, or operational test administration, ESS cut scores emerge organically and analytically by optimizing the coherence of the SME Prompt-PLD alignments and empirical data. That is, ESS cut scores are estimated by optimizing the evidentiary relationship between test prompts and the NGSS articulated in the PLDs. In this case, data from the spring 2023 SIPS pilot study is used to support the estimation of ESS cut scores.

ESS is not a single activity—it is a set of iterative processes and analyses, as illustrated in Exhibit 1, that occur throughout the assessment development lifecycle. ESS advances the principled notion of assessment design based on evidentiary reasoning by requiring the alignment of each assessment prompt— more precisely, each within-prompt score point—to a performance level by the explicit linkage of the prompt to a specific PLD measurement target. Thus, the evidentiary chain runs not just from the NGSS to the test prompts, but first from NGSS to the PLDs, then from the PLDs to the test prompts, providing more precise interpretability of the measurement target evidenced by the prompts.

While ESS was developed to provide a practical approach to standard setting for assessments adhering to a PAD framework, its methods add value that extends well beyond the estimation of cut scores.

Exhibit 1. SIPS Embedded Standard Setting Iterative Process



Embedded Standard Setting encompasses the integrated and iterative set of processes and procedures that span the assessment lifecycle, supporting the coherence of the various assessment system elements described next and illustrated in Exhibit 1.

Embedded Standard Setting and Assessment System Coherence

Assessment system coherence refers to the interrelationship between the steps and processes engaged during assessment design and development working to preserve the chain of interpretability from the NGSS to PLD development to the realization of their interpretable operationalization through empirically identified cut scores and meaningful classifications. Assessment system coherence is manifested when the various assessment components form an internally consistent system. For example:

1. PLDs should clearly and comprehensively articulate the NGSS and reflect the content and rigor to fulfill the intent of the SIPS Theory of Action,
2. Prompts should provide evidence for the NGSS-based attributes of students as specified by the measurement targets in the various performance levels,
3. Prompts should be explicitly aligned to specific performance levels because they provide evidence for the NGSS claims and measurement targets of the associated level descriptors,
4. Empirical data should support SMEs' Prompt-PLD alignments, and
5. Cut scores should have empirical data supporting the evidentiary relationship between assessment prompts and the NGSS; that is, examinees in each performance level should have an appropriate likelihood of success on the prompts aligned to the claims and measurement targets in the associated level.

Assessment system coherence is supported by the application of PAD when the application appropriately employs the ESS iterative processes illustrated in Exhibit 1. A comprehensive application of PAD should, in fact, work to guarantee such coherence, and the ESS iterative processes ensure that the PAD process continues to do its work until said coherence is achieved.

Assessment system coherence results from the understanding that initial drafts of the various assessment elements—PLDs, the assessment prompts and tasks, SMEs’ Prompt-PLD alignments, and cut scores—often require iterative improvement and are only considered “final” once coherence is sufficiently supported by evidence. Cut scores are then imbued with the interpretations the assessment was developed to provide and ready for adoption by the sponsoring agency. By explicitly incorporating iterative processes in the assessment development lifecycle, we acknowledge that we not only are comfortable revisiting the various assessment elements when and if anomalies manifest, but explicitly plan for, manage, and document the iterative activities that provide evidence for assessment system coherence.

Next, we provide an overview of each element of the Embedded Standard Setting methodology and the SIPS standard setting design.

Coordination of Embedded Standard Setting Iterative Processes

Embedded Standard Setting iterative processes require coordination of activities that typically occur throughout the assessment development lifecycle, as well as ESS-specific processes. The coordinated ESS processes were conducted between September 2021 and July 2023 and include PLD development, Task and Prompt development, Prompt-PLD Alignment, ESS analyses, vertical articulation, and technical reporting. Each of these processes is described briefly below and in detail later in this section.

PLD Development

Performance Level Descriptors (PLDs) operationalize and articulate the NGSS by specifying the science knowledge, skills, and attributes (KSAs) expected of students in each performance level necessary to support the SIPS Theory of Action. The SIPS Curriculum and Assessment team developed unique PLDs for each of the four EoUs per grade in grades 5 and 8. The PLDs are available at <https://sipsassessments.org/resources/>.

Prompt Development & Prompt-PLD Alignment

The SIPS SMEs conducted Prompt-PLD alignments for each prompt and score point on each EoU. That is, for each of the three tasks in each EoU, each obtainable score point for each prompt was associated with a performance level based on alignment of (a) the measurement attributes and content characteristics of the score point (as reflected by the prompt and scoring rubric) and (b) the claims and measurement targets reflected by the associated PLDs.

ESS analyses

ESS analyses were conducted using Pilot Study data for each EoU resulting in (a) three unique cut scores defining the four levels of performance per EoU per grade, (b) evidence supporting the efficacy of the SMEs’ Prompt-PLD alignments, (c) impact data used to evaluate the reasonableness of the cut scores and to support vertical articulation, and (d) lists of ESS-Inconsistent prompts.

Vertical Articulation

Under ideal circumstances the estimation of initial ESS cut scores for each EoU results in a system of within-grade, across-EoU cut scores with impact data that is reasonable and supports the SIPS policy goals. That is, the proportion of students in each performance level should be appropriate when viewed across levels within an EoU and within each level across the EoUs. If they do not, then some statistical smoothing, referred to as vertical articulation, may be necessary to achieve this result. It is common to refine cut scores to support vertical articulation of cut scores either during a standard setting workshop or by policymakers and their technical advisors following a standard setting.

Data from the Pilot Study were not sufficient to recommend vertically articulated cut scores for adoption by states intent on using the SIPS assessments for their summative federal accountability science assessments. However, the adoption of SIPS cut scores may be considered following vertical articulation based on a more substantial field test conducted by the states and the smoothing of the cut scores based on Pilot Study data that may later be refined and validated. A detailed description of vertical articulation for SIPS is provided in the section under the heading, "Vertical Articulation."

Technical Report & Peer Review Evidence

Validity evidence is documented supporting the efficacy of the resulting system of cut scores. Methods of aggregating the profile of students' four EoU performance levels to a summative performance level to support federal accountability requirements are considered, investigated, and discussed.

PLD Development

Performance Level Descriptors (PLDs) operationalize and articulate the NGSS by specifying the science knowledge, skills, and attributes (KSAs) expected of students in each performance level necessary to support the SIPS Theory of Action. This section describes the process used by the SIPS SMEs to articulate and explicate the NGSS across four levels of science performance for each of the four EoUs per grade. The resulting PLDs are available at <https://sipsassessments.org/resources/>.

SME Qualifications

A combination of science subject matter experts and educational measurement experts collaborated in the development of the SIPS PLDs. Members of the SIPS team have extensive science expertise and experience in multiple areas of education, including as K-12 teachers, adjunct instructors at the university level, professional learning providers in both K-12 and higher education settings, and through positions in state-level science education leadership (i.e., senior content specialists, state assessment directors, and assistant state assessment directors). Members also have experience acting as both panelists and facilitators for science standard setting meetings as well as developers of state-level science assessment programs through the application of evidence-centered design (ECD) to design, develop, and implement NGSS-aligned assessments and to create performance level descriptors and ultimately cut scores for federal accountability and reporting purposes. Finally, the SIPS SMEs have extensive experience in the exploratory design of innovative assessments to produce both design approaches and early-stage tasks critical for establishing frameworks for researching and developing more extensive suites of innovative assessment tasks. As a result, the PLDs may be considered the product of collaboration among science experts, curriculum specialists, teachers, and policy makers.

PLD Development Process

The SIPS SMEs created a set of policy PLDs and range PLDs for state and organizational partner review. The policy level descriptors were created by modifying and adapting state partners' existing policy level science PLDs to act as high-level descriptions of expected performance in each performance level, 1 - 4. The policy PLDs are intended to provide information to educators and stakeholders about the overall meaning behind each performance level by describing the knowledge and skills expected of students in each performance level. They are not written to be specific to any given grade level. SIPS organizational and state partners reviewed the policy PLDs to make sure they reflect multi-dimensional science expectations for students at each level. The SIPS policy level descriptors are provided in Appendix A.

Once the policy PLDs were established, a small group of SIPS SMEs—the same SMEs tasked with designing the assessment framework and developing the EOU assessments—created the grade- and unit-specific range PLDs at grade 5 and grade 8 to support 1) alignment of the EoU assessments to NGSS expectations, 2) an explication of deep conceptual understanding and complex reasoning required of three-dimensional science, 3) a foundation for comparable score interpretation, and 4) a structure for the valid interpretations of scores. The SIPS range PLDs were created early in the ECD process to support the development of prompts and tasks along with ECD-based design tools (i.e., unpacking tools, design patterns, and task specifications and verification of alignment documents). The key to the development of the SIPS range PLDs was ensuring their alignment to the NGSS performance expectations (PEs) addressed by each EOU assessment and to the policy PLDs. Thus, for the task developers, the range PLDs define the construct that is being measured and describe what students should know and be able to do in relation to the construct.

The SIPS SMEs were tasked with considering how the domain for assessment would be defined, and consequently, how that domain would affect the types of claims that could be made about students. The SMEs explored multiple methods for incorporating the three dimensions of the NGSS PEs across four (4) performance levels. Each PE of the NGSS is a combination of three dimensions: a disciplinary core idea (DCI), a science and engineering practice (SEP), and a cross-cutting concept (CCC). The three-dimensional nature of the NGSS PEs requires different considerations about defining student performance than those typically used in defining performance on traditional standards. The SIPS SMEs consulted with state partners and a selection of technical advisory panel (TAP) experts and determined early in the design process that the dimensions of the NGSS should not be separated for PLD development purposes. The PEs as written are a small subset of all the possible ways that SEPs, DCIs, and CCCs can be combined. Combining the assessed DCIs with all the related SEPs and CCCs results in many, many possible dimension combinations. As defined by SIPS SMEs in collaboration with SIPS state partners, the EoU assessments are designed to measure two levels of transfer, close and proximal, in terms of time, place, and context relative to when instruction takes place. In terms of ‘close transfer,’ the SIPS EoU assessments elicit evidence of students’ ability to integrate the same dimension combinations as those represented by the PEs and in similar contexts or situations to those explored through instruction (e.g., terrestrial ecosystems). Regarding ‘proximal transfer,’ the SIPS EoU assessments also elicit evidence of students’ ability to flexibly combine the dimensions within the PEs in related but different contexts or situations to those explored through instruction (e.g., terrestrial vs. aquatic ecosystems).

Another key consideration to be made in PLD development, and subsequent test construction, was what aspect or aspects of quality would be used to order student performances. To this end and prior to drafting the first set of range PLDs, the SIPS SMEs consulted with state partners and TAP experts to develop a range PLD framework with differentiated indicators of performance including, but not limited to, breadth of content, cognitive complexity, degree of correctness, degree of challenge, sophistication of solution, and degree of independence (i.e., extent to which directions, background information, or other scaffolds are provided). This PLD framework is a general framework designed for universal application across grade levels and EOU assessments. To create a model framework based on these indicators, the SIPS SMEs used the PE topic bundles for Unit 1 at grade 5 and grade 8 to define PE-specific statements of how students might be expected to perform related to each indicator for each performance level. Organized as a series of tables by indicator, each row represents an aspect of the indicator, each column represents a degree of performance from less to more complex, and each cell represents a brief PE-specific statement of how students at the given performance level might be expected to perform on the assessment related to that aspect of the indicator. This PLD framework for Unit 1 at grades 5 and 8 was presented to state partners, organizational partners, and TAP experts as a resource to support their initial reviews of the draft policy and range PLDs. The SIPS SMEs applied revisions to the range PLD framework, policy PLDs, and Unit 1 range PLDs at grades 5 and 8 based on partner feedback and prior to full development of the range PLDs across units. The PLD framework was revised from six indicators to four indicators based on partner feedback.

The SIPS SMEs then applied the four indicators (i.e., breadth of content, cognitive complexity, sophistication of solution, and degree of independence) from the framework to draft the remaining unit-specific range-PLDs for units 2, 3, and 4 at grades 5 and 8. The SIPS SMEs provided the resulting PLDs to state partners, organizational partners, and TAP experts for review, prefacing the review with information describing the PLD development process and instructions for reviewing and suggesting revisions to the descriptors. Following this cyclical review process, the SIPS SMEs applied revisions to the range PLDs based on partner feedback and provided final drafts to the Nebraska Department of

Education for approval. It is this set of PLDs that was used to support the ESS analysis using Pilot Study data.

Prompt-PLD Alignment

The SIPS SMEs conducted Prompt-PLD alignments for each prompt and score point on each EoU assessment. That is, each obtainable prompt score point was associated with a performance level based on congruence of (a) the measurement attributes and content characteristics of the score point (as reflected by the prompt and scoring rubric) and (b) the claims and measurement targets reflected by the associated EoU assessment's PLDs.

SME Qualifications

See SME Qualifications in the PLD Development section of the report.

Training for Aligning Prompts to PLDs

The SIPS SMEs received training and guidelines to support their Prompt-PLD alignments as follows.

Training Materials

ESS Training slides shared with developers are provided in Appendix B. Training included a summary of the ESS methodology and procedures.

Polytomous Prompt-PLD Alignment Guidelines

Each polytomous prompt must be aligned to a performance level, though not necessarily a unique level. That is, multiple score points could be aligned to the same level if the KSAs associated with performance at each score point (as defined by the prompt and scoring rubric) is best associated with the same level.

The highest score point should clearly align to an evidence statement in one level of the PLDs, though not necessarily the highest level. The next lower point must also be aligned to a PLD level—the same level as the higher point or a lower level as best conforms with the KSAs reflected by the prompt and scoring rubric. While the highest score point should clearly align to an evidence statement in one level, that may not be true for lower score points. While it is preferable to have a clear performance level evidence statement associated with every score point, the Prompt-PLD alignment may be inferred for lower score points if necessary. When there is no clear evidence statement already in the PLDs, it provides an opportunity to improve them by adding an evidence statement at the appropriate performance level that corresponds to the scoring rubric.

Note that initial Prompt-PLD alignments are considered hypotheses. When data are available from the pilot study, there is an opportunity to evaluate the hypothesized alignments. At that time, a list of prompts and score points with PLD alignments that are inconsistent with empirical data are available and Prompt-PLD alignments and/or the PLDs may be revised based on a review and resolution process.

SME Prompt-PLD Alignment Process

The SIPS SMEs evaluated and mapped the alignment of each EOU assessment prompt and its associated score points with the range performance level descriptors in grades 5 and 8 using the following evaluation questions:

1. How well do the range PLDs represent the array of score points expressed by the rubric?
2. How well do the range PLDs capture the knowledge and skills measured by the EOU prompts?
3. How well do the sets of prompts that contribute to students' scores on each task reflect the knowledge and skills represented by the range PLDs?

Evaluation Question 1: How well do the range PLDs represent the array of score points expressed by the rubric?

This evaluation question focuses on the rubric and addresses the relationship between the range PLD descriptors at each of the four levels with the EOU scoring rubrics. In making ratings for Evaluation Question 1, the SMEs reviewed the PLDs and prompts for each EOU assessment and using a customized spreadsheet for each rating, identified the following for **every** available score point for each prompt as administered during the Pilot:

- a. Performance Expectation the prompt best reflects;
- b. Disciplinary Core Idea the prompt best reflects;
- c. Science and Engineering practice(s) the prompt best reflects;
- d. Crosscutting concept(s) the prompt best reflects;
- e. Performance level, based on the PLDs, that best reflects the prompt performance demands reflected in the rubric for the score point under consideration. In some instances, the reviewers may have indicated no alignment, as appropriate, for a given score point.

Evaluation Question 2: How well do the range PLDs capture the knowledge and skills measured by the EOU prompts?

Evaluation Question 2 targets the relationship between the prompts and the PLDs. In making decisions for this evaluation question, the SIPS SMEs reviewed the range PLDs and prompt-level SIPS Task Specifications to determine the degree to which the PLDs adequately capture knowledge, skills, and abilities of each prompt as defined in the task specifications.

Evaluation Question 3: How well do the sets of prompts that contribute to students' scores on each task reflect the knowledge and skills represented by the range PLDs?

Evaluation Question 3 addresses the relationship between the range PLDs and the sets of prompts on each task on which students' EOU assessment scores are based. Key sources of evidence collected to answer Evaluation Question 3 included the draft PLDs and each of the EOU Assessment Scoring Guides. The SIPS SMEs examined the relationship between the EOU assessment content, the scoring rubrics, and PLDs regarding the following aspects: a) breadth of content, b) cognitive complexity, c) sophistication of solution, and d) degree of independence.

ESS analyses

ESS analyses use empirical data from the SIPS Pilot Study to provide four key outcomes. First, initial ESS cut scores emerge analytically and organically by optimizing the coherence of the Prompt-PLD alignments and empirical data.

Second, is the information necessary to evaluate the efficacy of the SME’s Prompt-PLD alignments. Evaluation criteria include:

- a. the correlation of empirical prompt difficulty (IRT response probability location) and the ordinality of the SME Prompt-PLD alignment,
- b. agreement rates between SME Prompt-PLD alignments and the Empirical ESS Prompt-PLD alignments derived from the ESS cut scores, and
- c. weighted Kappa values that quantify the degree to which the SME Prompt-PLD alignments are concordant with the Empirical ESS prompt-PLD alignments.

Third, impact data—the proportion of students in each performance level—is estimated.

Fourth, lists of ESS-Inconsistent prompts are produced. These are prompts with alignment hypotheses that are not supported by empirical data. Resolution of the inconsistency may result from a review of the prompt with the goal of understanding the source of the inconsistency. Such inconsistencies may occur due to imprecise language, developmental disarticulations in the PLDs, construct irrelevant variance, etc.

Each of these outcomes are described in this section.

Data

ESS analyses were conducted using data from the 2022-23 SIPS Pilot Study to estimate item response theory prompt parameters and student theta scores. N-counts of examinees used to support the analyses and the estimation of impact data are provided in Exhibit 2.

Exhibit 2. Impact Data N-Counts

	N-Count			
Grade	EoU1	EoU2	EoU3	EoU4
5	237	412	270	253
8	92	61	161	41

The data were used to estimate IRT response probability (RP) scale locations for each prompt score point. The RP67 location is typically used for standard setting purposes; however, RP50 will also be investigated (see Lewis, Mitzel, Mercado, & Schulz, 2012 for a detailed discussion of response probabilities). The RP67 (RP50) location is the scale value at which a student has a .67 (or .50 for RP50) likelihood of success on a dichotomous item/prompt or a .67 (.50) likelihood of obtaining a given score point or higher on a polytomous prompt score point.

Initial ESS Cut Score Estimation

Embedded Standard Setting (Lewis & Cook, 2020) cut scores are estimated with CMS’ proprietary software, EmStanS (Lewis & Lee, 2020) by optimizing the coherence between the SME Prompt-PLD alignments and empirical data. That is, cut scores emerge organically and analytically from the empirically tested SME Prompt-PLD alignments by optimizing the evidentiary relationship between prompts and the claims and measurement targets articulated in the PLDs.

ESS cut scores were estimated using the ESS-Count algorithm described by Lewis & Cook (2020) and Lewis, Lee, and Choi (2021). ESS-Count can be expressed mathematically as:

$$\text{ESS-Count} \equiv \arg \min_c \sum_{i=1}^n I(\text{ESS-Inconsistent}) \quad (1)$$

Simply put, ESS-Count is the cut score c that minimizes the total number of inconsistent prompts on an EoU assessment. A prompt score point is called ESS-Inconsistent if $L_i^{(SME)} \neq L_i^{(RP|c)}$, i.e., the SME Prompt-PLD aligned level for prompt score point i is not equal to the ESS Empirical Prompt-PLD level based on the prompt’s IRT RP location relative to cut score candidate c . The binary indicator function, $I(\text{d-inconsistent})$, for a prompt is set to 1 if the prompt is ESS-inconsistent and 0 otherwise.

ESS cut scores are estimated by identifying the minimum value of ESS-Count for all cut score candidates across the test scale. The cut scores produced from the RP67 ESS-Count algorithm are provided in Exhibit 3. These cut scores are referred to as “initial” cut scores because they may be adjusted during vertical articulation. The initial cut scores are used to evaluate the efficacy of the SME’s Prompt-PLD alignments, described next.

Exhibit 3. Initial SIPS Cut Scores for Grades 5 and 8

Grade	EoU	Level 2	Level 3	Level 4
Grade 5	EoU1	0.0115	0.7564	2.6621
	EoU2	-0.5588	0.4675	1.4645
	EoU3	-1.5002	0.2763	1.6766
	EoU4	-0.2804	0.4862	2.6185
Grade 8	EoU1	-0.2749	1.8084	4.0000
	EoU2	-0.4366	0.7397	2.5317
	EoU3	-1.3670	-0.1578	1.1459
	EoU4	-1.0624	0.3550	3.0557

The Efficacy of SMEs’ Prompt-PLD Alignments

In this section, we examine criteria used to analyze the efficacy of SMEs’ Prompt-PLD alignments. First, we estimate the correlation of prompts’ performance level ordinality (Level 1 = 1, Level 2 = 2, Level 3 = 3, Level 4 = 4) and IRT RP location for each prompt score point. Then, we provide crosstabs and classification agreement rates between SMEs’ Prompt-PLD alignments and the Empirical ESS Prompt-

PLD alignments established by the initial ESS cut scores. Finally, we review the weighted Kappa values reflecting the concordance between the SMEs’ Prompt-PLD alignments and the Empirical ESS Prompt-PLD alignments.

Correlations

Exhibit 4 lists the correlations of prompts’ SME-aligned performance level ordinality and RP67 location by grade and EoU assessment. The column labeled “Correlations” is the standard Pearson correlation coefficient. However, because the IRT location is a continuous variable and performance level ordinality is an ordinal variable, the maximum correlation under perfect alignment is constrained to less than 1. We adjust for this to better interpret the magnitude of the correlation by estimating the “Maximum Correlation,” between the perfectly ordered Empirical ESS Prompt-PLD alignment and the RP67 locations. The ratio of the Correlation to Optimal Correlation is reported as the Adjusted Correlation.

The Adjusted Correlations for grade 5 are 0.81, 0.87, 0.77, and 0.64 for EoUs 1, 2, 3, and 4, respectively. The Adjusted Correlations for grade 8 are 0.71, 0.44, 0.77, and 0.72 for EoUs 1, 2, 3, and 4, respectively. These are moderate to good correlations supporting the efficacy of the SME Prompt-PLD correlations.

We note that the measurement literature widely reports the challenge of predicting item/prompt difficulty. A recent study by Schneider, Chen, & Nichols (2021) indicated that alignment to PLDs accounts for the greatest variance in the prediction of item/prompt difficulty.

Exhibit 4. Correlation of SMEs’ Prompt-PLD Aligned Performance Level Ordinality and IRT RP Location

GCA	EoU	Correlation	Maximum Correlation	Adjusted Correlation
Grade 5	EoU1	0.75	0.93	0.81
	EoU2	0.81	0.93	0.87
	EoU3	0.72	0.93	0.77
	EoU4	0.58	0.90	0.64
Grade 8	EoU1	0.66	0.94	0.71
	EoU2	0.40	0.91	0.44
	EoU3	0.72	0.93	0.77
	EoU4	0.65	0.90	0.72

Classification Agreement and Weighted Kappa

Crosstabs reflect the *classification agreement* between the SMEs’ Prompt-PLD alignments and the Empirical ESS Prompt-PLD alignments.

Establishing ESS Empirical Prompt-PLD Alignments

After each ESS cut score is estimated, prompts are classified into the following empirical performance levels if the prompt’s IRT RP location is:

- a. *Level 1*: below the ESS Level 2 cut score
- b. *Level 2*: at or above the ESS Level 2 cut score but below the Level 3 cut score

- c. *Level 3*: at or above the ESS Level 3 cut score but below the Level 4 cut score
- d. *Level 4*: at or above the ESS Level 4 cut score

Classification Agreement

Classification agreement is described in the following terms:

- a. *Agree*: The empirical performance level agrees with the SME-Aligned Level.
- b. *Disagree Adjacent*: The empirical performance level disagrees with the SME-Aligned Level, but they are adjacent levels.
- c. *Disagree Discrepant*: The empirical performance level disagrees with the SME-Aligned Level, and they are not adjacent levels.

Classification agreement is graphically represented as a crosstab in Exhibit 5.

Exhibit 5. Classification Agreement Crosstab

		SME Prompt-PLD Alignment			
		Level 1	Level 2	Level 3	Level 4
Empirical ESS Prompt-PLD Alignment	Level 1	Agree	Disagree: Adjacent	Disagree: Discrepant	Disagree: Discrepant
	Level 2	Disagree: Adjacent	Agree	Disagree: Adjacent	Disagree: Discrepant
	Level 3	Disagree: Discrepant	Disagree: Adjacent	Agree	Disagree: Adjacent
	Level 4	Disagree: Discrepant	Disagree: Discrepant	Disagree: Adjacent	Agree

ESS-Inconsistent Prompts

Prompt score points classified by the crosstabs as Disagree-Adjacent and Disagree-Discrepant are referred to as ESS-Inconsistent prompts. That is, an ESS-Inconsistent prompt’s alignment hypothesis is not supported by empirical data. An ESS-Inconsistent prompt does not preserve the interpretability of the test scores as required under the application of a PAD approach. If prompts are developed to provide evidence of an attribute associated with a specific performance level, and data indicates that the prompt is associated with a different level, then the PAD evidentiary validity argument is not supported by that prompt.

Under an ESS framework, we treat ESS-Inconsistency as a flag like other flags identified by traditional prompt analyses. For example, prompts are traditionally flagged for low or high p-values, anomalous point-biserial correlations, prompt bias, etc. ESS-Inconsistency is another prompt flag that can be considered when selecting prompts for test forms following field testing or identifying prompts in need of review and potential remediation. More specifically, ESS-Inconsistency identifies prompts that may be

remediated by SME review as described later in this section. Next, we describe concepts associated with prompt inconsistency.

ESS-Distance

Given a cut score, we define the Distance of an ESS-Inconsistent prompt as the minimum number of scale score points that the prompt’s IRT location must shift to place the prompt at a border of the SMEs’ Prompt-PLD Aligned level. The greater the ESS-Distance of an inconsistent prompt, the greater the magnitude of inconsistency.

Essentially Consistent Prompts

We say that a prompt is Essentially Consistent if the absolute value of its Distance is less than or equal to 1 Standard Error of Measurement (SEM) of the test. This arbitrary, but not capricious, metric is useful when subject matter experts are engaged in the exercise of resolving inconsistent prompts. That is, Essentially Consistent prompts have Distances that are so inconsequential that a SME would unlikely be able to identify a content-based rationale for the inconsistency; thus, attempts to resolve the inconsistency are not likely to be successful.

Weighted Kappa

In addition to classification agreement, we also provide the weighted Kappa statistic for each crosstab using quadratic weighting. The Kappa statistic is a value from 0 to 1 that indicates how two types of independent classifications of the same phenomenon (i.e., SME Prompt-PLD alignments and Empirical ESS Prompt-PLD alignments) compare to random classifications. Higher values indicate stronger agreement between the two independent classifications. The quadratic weighting penalizes disagreements that are discrepant more than disagreements that are adjacent. To aid in the interpretation of the Kappa values, Exhibit 6 Exhibit 6. Kappa Interpretations

displays the recommended ranges suggested by Landis and Koch (1977) and Exhibit 7 provides the Weighted Kappa values for the EoUs.

Exhibit 6. Kappa Interpretations

Kappa Value	Strength of Agreement
0	None
<0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

Crosstabs

Exhibit 6.

summarizes the agreement rates and weighted Kappas associated with the detailed crosstabs provided in **Error! Reference source not found.** through **Error! Reference source not found.**. There are two crosstabs per exhibit. The first crosstab provides the SMEs’ Prompt-PLD aligned levels for all prompts in

the given EoU crossed with the prompts' ESS Empirical Prompt-PLD alignment. The second crosstab results from treating Essentially Consistent prompts as consistent. That is, prompts with locations within one standard error of the ESS cut score associated with their SME Prompt-PLD alignments are classified as consistent. Exhibit 6.

reflects the agreement rates and weighted Kappas from the second table.

Grade 5

The grade 5 EoUs (see upper half of Exhibit 7) have agreement rates ranging from 52% for EoU4 to 76% for EoU2 and they have Weighted Kappas ranging from 0.67 for EoU4 to 0.88 for EoU2. The Kappa values are considered substantial to almost perfect according to the guidelines provided in Exhibit 6. Kappa Interpretations

Grade 8

The grade 8 EoUs (see lower half of Exhibit 7) have agreement rates ranging from 58% for EoU2 to 78% for EoU4 and they have Weighted Kappas ranging from 0.53 for EoU2 to 0.78 for EoU3. The grade 8 kappa values are considered moderate to substantial according to the guidelines provided in Exhibit 6.

Exhibit 7. Agreement Rate and Weighted Kappa

Grade 5	Agreement Rate	Weighted Kappa
EoU1	59%	0.75
EoU2	76%	0.88
EoU3	70%	0.75
EoU4	52%	0.67

Grade 8	Agreement Rate	Weighted Kappa
EoU1	63%	0.71
EoU2	58%	0.53
EoU3	63%	0.78
EoU4	78%	0.71

Exhibit 8. Grade 5 EOU1 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	6 16.2%	4 10.8%	0 0%	0 0%	10 27%
	Level 2	2 5.4%	5 13.5%	2 5.4%	0 0%	9 24.3%
	Level 3	1 2.7%	3 8.1%	8 21.6%	3 8.1%	15 40.5%
	Level 4	0 0%	0 0%	0 0%	3 8.1%	3 8.1%
	Total	9 24.3%	12 32.4%	10 27%	6 16.2%	37 100%

Agree	22	59%
Disagree	15	41%
Adjacent Disagreement	14	38%
Discrepant Disagreement	1	3%
Weighted Kappa	0.75	

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	6 16.2%	4 10.8%	0 0%	0 0%	10 27%
	Level 2	2 5.4%	5 13.5%	2 5.4%	0 0%	9 24.3%
	Level 3	1 2.7%	3 8.1%	8 21.6%	3 8.1%	15 40.5%
	Level 4	0 0%	0 0%	0 0%	3 8.1%	3 8.1%
	Total	9 24.3%	12 32.4%	10 27%	6 16.2%	37 100%

Agree	22	59%
Disagree	15	41%
Adjacent Disagreement	14	38%
Discrepant Disagreement	1	3%
Weighted Kappa	0.75	

Exhibit 9. Grade 5 EOU2 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	13 35.1%	3 8.1%	0 0%	0 0%	16 43.2%
	Level 2	0 0%	6 16.2%	2 5.4%	1 2.7%	9 24.3%
	Level 3	0 0%	1 2.7%	3 8.1%	1 2.7%	5 13.5%
	Level 4	0 0%	0 0%	2 5.4%	5 13.5%	7 18.9%
	Total	13 35.1%	10 27%	7 18.9%	7 18.9%	37 100%

Agree	27	73%
Disagree	10	27%
Adjacent Disagreement	9	24%
Discrepant Disagreement	1	3%
Weighted Kappa	0.86	

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	14 37.8%	2 5.4%	0 0%	0 0%	16 43.2%
	Level 2	0 0%	6 16.2%	2 5.4%	1 2.7%	9 24.3%
	Level 3	0 0%	1 2.7%	3 8.1%	1 2.7%	5 13.5%
	Level 4	0 0%	0 0%	2 5.4%	5 13.5%	7 18.9%
	Total	14 37.8%	9 24.3%	7 18.9%	7 18.9%	37 100%

Agree	28	76%
Disagree	9	24%
Adjacent Disagreement	8	22%
Discrepant Disagreement	1	3%
Weighted Kappa	0.88	

Exhibit 10. Grade 5 EOU3 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	4 7.4%	0 0%	0 0%	0 0%	4 7.4%
	Level 2	4 7.4%	17 31.5%	5 9.3%	0 0%	26 48.1%
	Level 3	1 1.9%	4 7.4%	12 22.2%	1 1.9%	18 33.3%
	Level 4	0 0%	0 0%	2 3.7%	4 7.4%	6 11.1%
	Total	9 16.7%	21 38.9%	19 35.2%	5 9.3%	54 100%

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	4 7.4%	0 0%	0 0%	0 0%	4 7.4%
	Level 2	4 7.4%	18 33.3%	4 7.4%	0 0%	26 48.1%
	Level 3	1 1.9%	4 7.4%	12 22.2%	1 1.9%	18 33.3%
	Level 4	0 0%	0 0%	2 3.7%	4 7.4%	6 11.1%
	Total	9 16.7%	22 40.7%	18 33.3%	5 9.3%	54 100%

Agree 37 69%
 Disagree 17 31%
 Adjacent Disagreement 16 30%
 Discrepant Disagreement 1 2%
 Weighted Kappa 0.73

Agree 38 70%
 Disagree 16 30%
 Adjacent Disagreement 15 28%
 Discrepant Disagreement 1 2%
 Weighted Kappa 0.75

Exhibit 11. Grade 5 EOU4 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	4 10%	3 7.5%	1 2.5%	0 0%	8 20%
	Level 2	5 12.5%	4 10%	2 5%	1 2.5%	12 30%
	Level 3	0 0%	3 7.5%	10 25%	5 12.5%	18 45%
	Level 4	0 0%	0 0%	0 0%	2 5%	2 5%
	Total	9 24.3%	12 32.4%	10 27%	6 16.2%	37 100%

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	5 12.5%	2 5%	1 2.5%	0 0%	8 20%
	Level 2	5 12.5%	4 10%	2 5%	1 2.5%	12 30%
	Level 3	0 0%	3 7.5%	10 25%	5 12.5%	18 45%
	Level 4	0 0%	0 0%	0 0%	2 5%	2 5%
	Total	10 25%	9 22.5%	13 32.5%	8 20%	40 100%

Agree 20 50%
 Disagree 20 50%
 Adjacent Disagreement 18 45%
 Discrepant Disagreement 2 5%
 Weighted Kappa 0.65

Agree 21 52%
 Disagree 19 48%
 Adjacent Disagreement 17 42%
 Discrepant Disagreement 2 5%
 Weighted Kappa 0.67

Exhibit 12. Grade 8 EOU1 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	11 23.9%	0 0%	0 0%	0 0%	11 23.9%
	Level 2	6 13%	13 28.3%	4 8.7%	1 2.2%	24 52.2%
	Level 3	0 0%	4 8.7%	4 8.7%	1 2.2%	9 19.6%
	Level 4	0 0%	0 0%	1 2.2%	1 2.2%	2 4.3%
	Total	17 37%	17 37%	9 19.6%	3 6.5%	46 100%

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	11 23.9%	0 0%	0 0%	0 0%	11 23.9%
	Level 2	6 13%	13 28.3%	4 8.7%	1 2.2%	24 52.2%
	Level 3	0 0%	4 8.7%	4 8.7%	1 2.2%	9 19.6%
	Level 4	0 0%	0 0%	1 2.2%	1 2.2%	2 4.3%
	Total	17 37%	17 37%	9 19.6%	3 6.5%	46 100%

Agree	29	63%
Disagree	17	37%
Adjacent Disagreement	16	35%
Discrepant Disagreement	1	2%
Weighted Kappa	0.7	

Agree	29	63%
Disagree	17	37%
Adjacent Disagreement	16	35%
Discrepant Disagreement	1	2%
Weighted Kappa	0.7	

Exhibit 13. Grade 8 EOU2 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	5 7.7%	4 6.2%	1 1.5%	0 0%	10 15.4%
	Level 2	5 7.7%	13 20%	10 15.4%	0 0%	28 43.1%
	Level 3	1 1.5%	4 6.2%	14 21.5%	0 0%	19 29.2%
	Level 4	0 0%	4 6.2%	1 1.5%	3 4.6%	8 12.3%
	Total	11 16.9%	25 38.5%	26 40%	3 4.6%	65 100%

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	6 9.2%	3 4.6%	1 1.5%	0 0%	10 15.4%
	Level 2	5 7.7%	15 23.1%	8 12.3%	0 0%	28 43.1%
	Level 3	1 1.5%	4 6.2%	14 21.5%	0 0%	19 29.2%
	Level 4	0 0%	4 6.2%	1 1.5%	3 4.6%	8 12.3%
	Total	12 18.5%	26 40%	24 36.9%	3 4.6%	65 100%

Agree	35	54%
Disagree	30	46%
Adjacent Disagreement	24	37%
Discrepant Disagreement	6	9%
Weighted Kappa	0.49	

Agree	38	58%
Disagree	27	42%
Adjacent Disagreement	21	32%
Discrepant Disagreement	6	9%
Weighted Kappa	0.53	

Exhibit 14. Grade 8 EOU3 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	1 3.3%	0 0%	0 0%	0 0%	1 3.3%
	Level 2	3 10%	8 26.7%	1 3.3%	0 0%	12 40%
	Level 3	0 0%	2 6.7%	6 20%	3 10%	11 36.7%
	Level 4	0 0%	0 0%	2 6.7%	4 13.3%	6 20%
	Total	4 13.3%	10 33.3%	9 30%	7 23.3%	30 100%

Agree	19	63%
Disagree	11	37%
Adjacent Disagreement	11	37%
Discrepant Disagreement	0	0%
Weighted Kappa	0.78	

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	1 3.3%	0 0%	0 0%	0 0%	1 3.3%
	Level 2	3 10%	8 26.7%	1 3.3%	0 0%	12 40%
	Level 3	0 0%	2 6.7%	6 20%	3 10%	11 36.7%
	Level 4	0 0%	0 0%	2 6.7%	4 13.3%	6 20%
	Total	4 13.3%	10 33.3%	9 30%	7 23.3%	30 100%

Agree	19	63%
Disagree	11	37%
Adjacent Disagreement	11	37%
Discrepant Disagreement	0	0%
Weighted Kappa	0.78	

Exhibit 15. Grade 8 EOU4 Crosstab of SME-Aligned Levels and ESS Empirical (Operational) Levels

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	1 2.4%	0 0%	0 0%	0 0%	1 2.4%
	Level 2	0 0%	14 34.1%	2 4.9%	0 0%	16 39%
	Level 3	1 2.4%	3 7.3%	14 34.1%	2 4.9%	20 48.8%
	Level 4	0 0%	0 0%	2 4.9%	2 4.9%	4 9.8%
	Total	2 4.9%	17 41.5%	18 43.9%	4 9.8%	41 100%

Agree	31	76%
Disagree	10	24%
Adjacent Disagreement	9	22%
Discrepant Disagreement	1	2%
Weighted Kappa	0.69	

		SME Aligned Level				Total
		Level 1	Level 2	Level 3	Level 4	
Operational Level	Level 1	1 2.4%	0 0%	0 0%	0 0%	1 2.4%
	Level 2	0 0%	15 36.6%	1 2.4%	0 0%	16 39%
	Level 3	1 2.4%	3 7.3%	14 34.1%	2 4.9%	20 48.8%
	Level 4	0 0%	0 0%	2 4.9%	2 4.9%	4 9.8%
	Total	2 4.9%	18 43.9%	17 41.5%	4 9.8%	41 100%

Agree	32	78%
Disagree	9	22%
Adjacent Disagreement	8	20%
Discrepant Disagreement	1	2%
Weighted Kappa	0.71	

Summary of the Efficacy of the SME Prompt-PLD Alignments

The results support the efficacy of the SMEs’ Prompt-PLD alignments. Correlations were moderate to good. Except for the grade 8 EoU2, Kappas are substantial to almost perfect. Given the challenge of predicting prompt difficulty, the agreement rates are sufficient to support the resulting cut scores.

Actionable Information: ESS-Inconsistent Prompt Review

The information in this section can be used to improve the coherence of the Prompt-PLD alignments.

and 17 summarize the number and percent of consistent, essentially consistent, and inconsistent prompts for each EoU by PE for grades 5 and 8, respectively. These tables can be reviewed by SMEs to determine the PEs that are more challenging to align to performance levels and to consider whether and how the alignments can be improved for the more challenging PEs. For example, 72.73% of the grade 5 PE labeled 5-PS1-3 (which is measured by 11 prompt score points) are inconsistent. SMEs may better understand the source of the inconsistency by comprehensively reviewing these prompt score points. PEs with high percentages of inconsistent prompts but which are measured by only a few score points should not be acted upon without more substantial evidence.

Exhibit 16. Summary of Consistent, Essentially Consistent, and Inconsistent Prompts by Claim, Grade 5

Grade/ Domain	Prompt Score Points	# Consistent	# Essentially Consistent	# Inconsistent	% Consistent	% Essentially Consistent	% Inconsistent
G5	206	133	4	69	64.56	1.94	33.50
EOU1	40	23		17	57.50	0.00	42.50
5-PS1-1	18	12		6	66.67	0.00	33.33
5-PS1-2	4	2		2	50.00	0.00	50.00
5-PS1-3	11	3		8	27.27	0.00	72.73
5-PS1-4	7	6		1	85.71	0.00	14.29
EOU2	47	36	1	10	76.60	2.13	21.28
5-LS2-1	4	3		1	75.00	0.00	25.00
5-PS3-1	6	6			100.00	0.00	0.00
5-LS1-1	11	9		2	81.82	0.00	18.18
5-LS2-1	22	16		6	72.73	0.00	27.27
5-PS3-1	4	2	1	1	50.00	25.00	25.00
EOU3	79	54	2	23	68.35	2.53	29.11
3-5-ETS1-2	4	3		1	75.00	0.00	25.00
3-5-ETS1-3	2	2			100.00	0.00	0.00
5-ESS2-1	2	1		1	50.00	0.00	50.00
5-ESS2-2	7	6		1	85.71	0.00	14.29
5-ESS3-1	10	5	1	4	50.00	10.00	40.00
3-5-ETS1-2	7	4	1	2	57.14	14.29	28.57
5-ESS2-1	19	16		3	84.21	0.00	15.79
5-ESS2-2	10	4		6	40.00	0.00	60.00
5-ESS3-1	18	13		5	72.22	0.00	27.78
EOU4	40	20	1	19	50.00	2.50	47.50
5-ESS1-1	6	1		5	16.67	0.00	83.33
5-ESS1-2	32	18	1	13	56.25	3.13	40.63
5-PS2-1	2	1		1	50.00	0.00	50.00

Exhibit 17. Summary of Consistent, Essentially Consistent, and Inconsistent Prompts by Claim, Grade 8

Grade/ Domain	Prompt Score Points	# Consistent	# Essentially Consistent	# Inconsistent	% Consistent	% Essentially Consistent	% Inconsistent
G8	219	138	6	75	63.01	2.74	34.25
EOU1	46	29		17	63.04	0.00	36.96
5-PS1-2	4	3		1	75.00	0.00	25.00
5-PS1-3	2	2			100.00	0.00	0.00
MS-PS2-1	11	7		4	63.64	0.00	36.36
MS-PS2-2	4	2		2	50.00	0.00	50.00
MS-PS2-4	5	4		1	80.00	0.00	20.00
MS-PS3-1	20	11		9	55.00	0.00	45.00
EOU2	88	50	5	33	56.82	5.68	37.50
5-PS3-1	3	2	1		66.67	33.33	0.00
MS-ESS1-2	6	4		2	66.67	0.00	33.33
MS-ESS1-3	14	9	1	4	64.29	7.14	28.57
5-LS2-1	7	2	1	4	28.57	14.29	57.14
5-PS3-1	2			2	0.00	0.00	100.00
MS-ESS1-1	21	10	1	10	47.62	4.76	47.62
MS-ESS1-2	11	6	1	4	54.55	9.09	36.36
MS-ESS1-3	11	9		2	81.82	0.00	18.18
MS-PS2-4	13	8		5	61.54	0.00	38.46
EOU3	38	23		15	60.53	0.00	39.47
3-5-ETS1-2	2	1		1	50.00	0.00	50.00
5-ESS2-1	2	1		1	50.00	0.00	50.00
5-ESS2-2	3	2		1	66.67	0.00	33.33
MS-LS4-4	1			1	0.00	0.00	100.00
5-ESS2-1	2	1		1	50.00	0.00	50.00
5-ESS3-1	7	4		3	57.14	0.00	42.86
MS-ESS1-4	5	4		1	80.00	0.00	20.00
MS-LS3-1	2			2	0.00	0.00	100.00
MS-LS4-1	3	2		1	66.67	0.00	33.33
MS-LS4-4	3	3			100.00	0.00	0.00
MS-LS4-6	8	5		3	62.50	0.00	37.50
EOU4	47	36	1	10	76.60	2.13	21.28
MS-ETS1-1	6	5		1	83.33	0.00	16.67
MS-PS4-2	3	3			100.00	0.00	0.00
5-ESS1-1	2	2			100.00	0.00	0.00
5-ESS1-2	4	2		2	50.00	0.00	50.00
MS-PS4-1	14	12		2	85.71	0.00	14.29
MS-PS4-2	21	15	1	5	71.43	4.76	23.81

Error! Reference source not found. and 19 provide the number of prompt score points per performance level for each EoU for grades 5 and 8, respectively. It is desirable to have enough score points at each level to support the reliability of the assessment across the range of performance and to adequately estimate the ESS cut scores. There are some EoUs with relatively modest numbers of prompts in a given level. For instance, Exhibit 19 indicates that there is only one prompt score point in Level 1 for EoU3 in grade 8 out of a total of 41 prompt score points. Subsequent prompt development efforts may be directed to provide additional prompt score points for levels with sparse coverage.

Exhibit 18. Number and Percent of Prompts by Performance Level, Grade 5

Grade/EoU/Level	# Prompt Score Points in Level	% of Total Points per EoU
G5		
EOU1	37	
Level1	10	27.0%
Level2	9	24.3%
Level3	15	40.5%
Level4	3	8.1%
EOU2	37	
Level1	16	43.2%
Level2	9	24.3%
Level3	5	13.5%
Level4	7	18.9%
EOU3	54	
Level1	4	7.4%
Level2	26	48.1%
Level3	18	33.3%
Level4	6	11.1%
EOU4	40	
Level1	8	20.0%
Level2	12	30.0%
Level3	18	45.0%
Level4	2	5.0%

Exhibit 19. Number and Percent of Prompts by Performance Level, Grade 8

Grade/EoU/Level	# Prompt Score Points in Level	% of Total Points per EoU
G8		
EoU1	46	
Level1	11	23.9%
Level2	24	52.2%
Level3	9	19.6%
Level4	2	4.3%
EoU2	65	
Level1	10	15.4%
Level2	28	43.1%
Level3	19	29.2%
Level4	8	12.3%
EoU3	30	
Level1	1	3.3%
Level2	12	40.0%
Level3	11	36.7%
Level4	6	20.0%
EoU4	41	
Level1	1	2.4%
Level2	16	39.0%
Level3	20	48.8%
Level4	4	9.8%

The tables provided in Appendix C: Detailed ESS Prompt Maps and Appendix D: Rosters of Inconsistent and Essentially Consistent Prompts can also support the review and resolution of ESS-Inconsistent prompts.

Appendix C: Detailed ESS Prompt Maps provides prompt-level information for all prompts for grades K through 6 including:

- prompt ID,
- order of difficulty (OOD),
- prompt IRT67 location (LOC),
- SME Prompt-PLD aligned level,
- ESS-Count and ESS-Weight associated with the prompt location, and
- empirical level.

Appendix D: Rosters of Inconsistent and Essentially Consistent Prompts provides rosters of inconsistent prompts (without the essentially consistent prompts) and essentially consistent prompts for each EoU assessment. These lists may be used to identify prompts for SME review to support improved alignment and prompt development. The information in the tables in Appendix D: Rosters of Inconsistent and Essentially Consistent Prompts includes:

- grade,
- prompt ID,
- order of difficulty (OOD),
- SME Prompt-PLD aligned level,
- ESS Empirical Prompt-PLD level,
- ordinal difference in SME-aligned and empirical levels,
- prompts' RP67 RP locations, and
- ESS-Distance associated with the prompt location.

Activities intended to review and resolve inconsistencies should consider prompts with the greatest magnitudes of ESS-Distance. Essentially consistent prompts need not be considered for resolution. Distance is in theta units.

Vertical Articulation

In this section we discuss considerations with respect to the vertical articulation (smoothing) of cut scores to support a coherent within-grade, cross-EoU assessment system. Under ideal circumstances the estimation of initial ESS cut scores for each EoU assessment results in a system of within-grade across-EoU cut scores with impact data that is reasonable and supports the SIPS policy goals. That is, the proportion of students in each performance level should be appropriate when viewed across levels within an EoU and within each level across the EoUs. If they do not, then some statistical smoothing, referred to as vertical articulation, may be necessary to achieve this result. It is common to refine cut scores to support vertical articulation of cut scores either during a standard setting workshop or by policymakers and their technical advisors following a standard setting.

Data from the Pilot Study were not sufficient to recommend cut scores for adoption by states intent on using the SIPS assessments for accountability purposes. However, the adoption of SIPS cut scores may be considered following vertical articulation based on a more substantial field test conducted by the states. We provide vertically articulated cut scores in this section based on Pilot Study data that may be refined following a more comprehensive field test.

Policy Consideration: Response Probability

The selection of an IRT response probability is a policy decision. RP67 is typically used for standard setting purposes because research suggests it reflects educators' notion of mastery of the content reflected by a prompt or prompt score point. It reflects a more rigorous expectation for student performance than other RP values that have been used for high stakes standard settings, such as RP50. RP67 results in higher, more rigorous cut scores than RP50.

Because the results of the SIPS assessments are currently subject to revision prior to operational use and the Pilot Study data was modest, we report results for both RP67 and RP50. See Lewis, Mitzel, Mercado, & Schulz (2012) for a detailed discussion of response probabilities.

Initial and Vertically Articulated (Smoothed) Cut Scores and Associated Impact Data

Next, we provide the initial RP67 and RP50 SIPS cut scores and associated impact data (percentage of students in each performance level) for each grade and EoU. We then provide vertically articulated RP67 and RP50 cut scores and associated impact data.

While there are no firm rules dictating acceptable levels of smoothing, it is common to report results in terms of the standard error of measurement. Three forms of the standard error are reported—the SEM of the SIPS assessments (SEM_{SIPS}), the SE of the ESS cut scores (SE_{ESS}) as described by Lewis, Lee, & Choi (2021), and the combined standard error ($SE_{SIPS+ESS}$) calculated as the square root of the sum of the squares of SEM_{SIPS} and SE_{ESS} . The standard error used to report the magnitude of adjustments made for vertical articulation in Exhibit 27 is $SE_{SIPS+ESS}$.

It is desirable to make as few adjustments as possible to achieve reasonable results, and to limit the magnitude of the adjustments to the degree possible. Only three adjustments of at least $1.5 SE_{SIPS+ESS}$ were required to achieve the smoothed results reported in **Error! Reference source not found.** and Exhibits 28 through 31.

Initial RP67 and RP50 cut scores and associated impact data

Exhibit 20 provides the initial ESS cut scores for RP67 and RP50 for each EoU assessment in grades 5 and 8.

Exhibit 20. Initial SIPS RP67 and RP50 Cut Scores across EoUs for Grades 5 and 8

Grade & RP	EoU	Level2	Level3	Level4
Grade 5 RP67	EoU1	0.0115	0.7564	2.6621
	EoU2	-0.5588	0.4675	1.4645
	EoU3	-1.5002	0.2763	1.6766
	EoU4	-0.2804	0.4862	2.6185
Grade 5 RP50	EoU1	-0.3655	0.4002	2.1684
	EoU2	-0.8226	0.0086	0.8856
	EoU3	-1.6696	-0.1320	1.1397
	EoU4	-0.7168	0.1369	2.0193
Grade 8 RP67	EoU1	-0.2749	1.8084	4.0000
	EoU2	-0.4366	0.7397	2.5317
	EoU3	-1.3670	-0.1578	1.1459
	EoU4	-1.0624	0.3550	3.0557
Grade 8 RP50	EoU1	-0.506	1.2081	4.0000
	EoU2	-0.9607	0.1782	1.9014
	EoU3	-1.8949	-0.5415	0.7875
	EoU4	-1.7309	-0.1434	2.4908

Error! Reference source not found. displays the standard errors of measurement of the assessments (SEM_{SIPS}), the standard error of the ESS cut scores (SE_{ESS}) estimated with the bootstrap methods described by Lewis, Lee, and Choi (2021), and the combined standard error ($SE_{SIP+ESS}$).

Exhibit 21. SIPS Standard Errors

Grade/EoU	SEM_{SIPS}			SE_{ESS}			$SE_{SIP+ESS}$		
	L2	L3	L4	L2	L3	L4	L2	L3	L4
G5 EoU1	0.27	0.29	0.31	0.46	0.51	0.6	0.53	0.59	0.68
G5 EoU2	0.30	0.35	0.42	0.17	0.45	0.42	0.35	0.57	0.60
G5 EoU3	0.28	0.26	0.38	0.31	0.2	0.53	0.42	0.33	0.65
G5 EoU4	0.28	0.28	0.35	0.57	0.36	0.58	0.63	0.46	0.68
G8 EoU1	0.27	0.38	0.38	0.4	1.01	0.38	0.48	1.08	0.53
G8 EoU2	0.24	0.26	0.31	0.39	0.59	0.87	0.46	0.64	0.92
G8 EoU3	0.37	0.34	0.42	0.34	0.28	0.6	0.50	0.44	0.73
G8 EoU4	0.35	0.31	0.35	0.4	0.36	0.48	0.53	0.47	0.59

Error! Reference source not found. and 23 provide impact data associated with the initial cut scores for the four grade 5 EoUs for RP67 and RP50, respectively. **Error! Reference source not found.** and 25 provide impact data associated with the initial cut scores for the four grade 8 EoUs for RP67 and RP50, respectively.

Exhibit 22. SIPS Grade 5 RP67 Initial Cut Score Impact Data

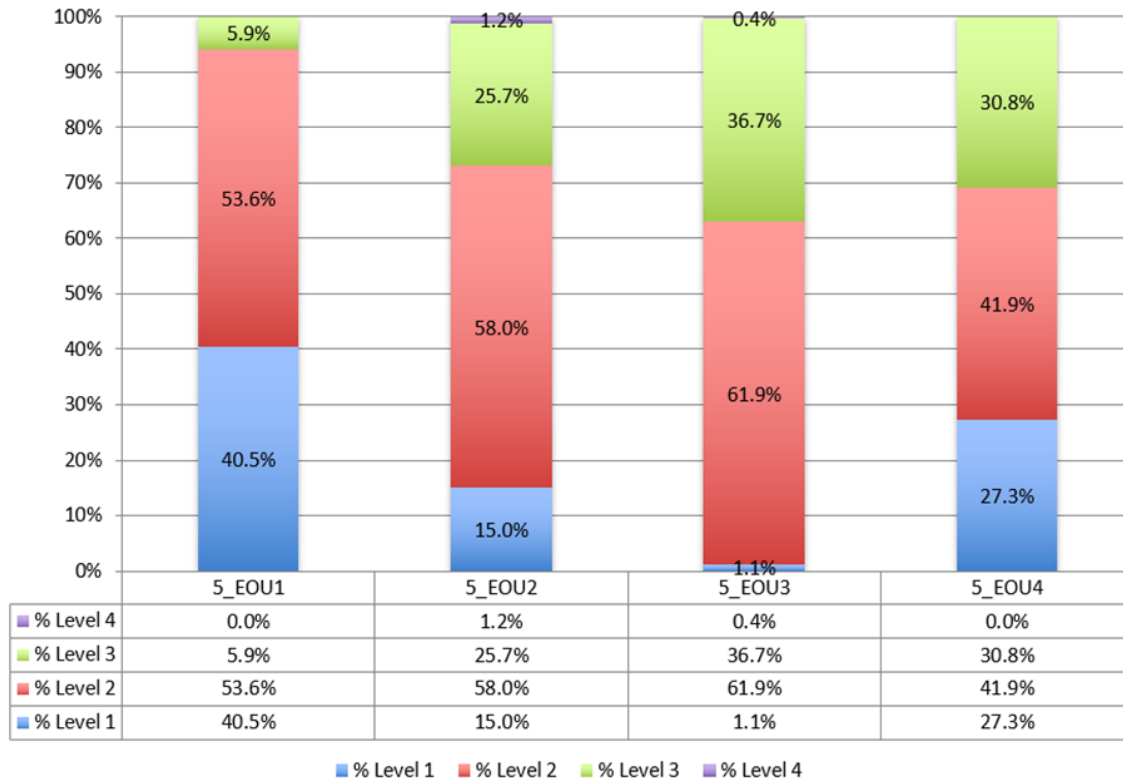


Exhibit 23. SIPS Grade 5 RP50 Initial Cut Score Impact Data

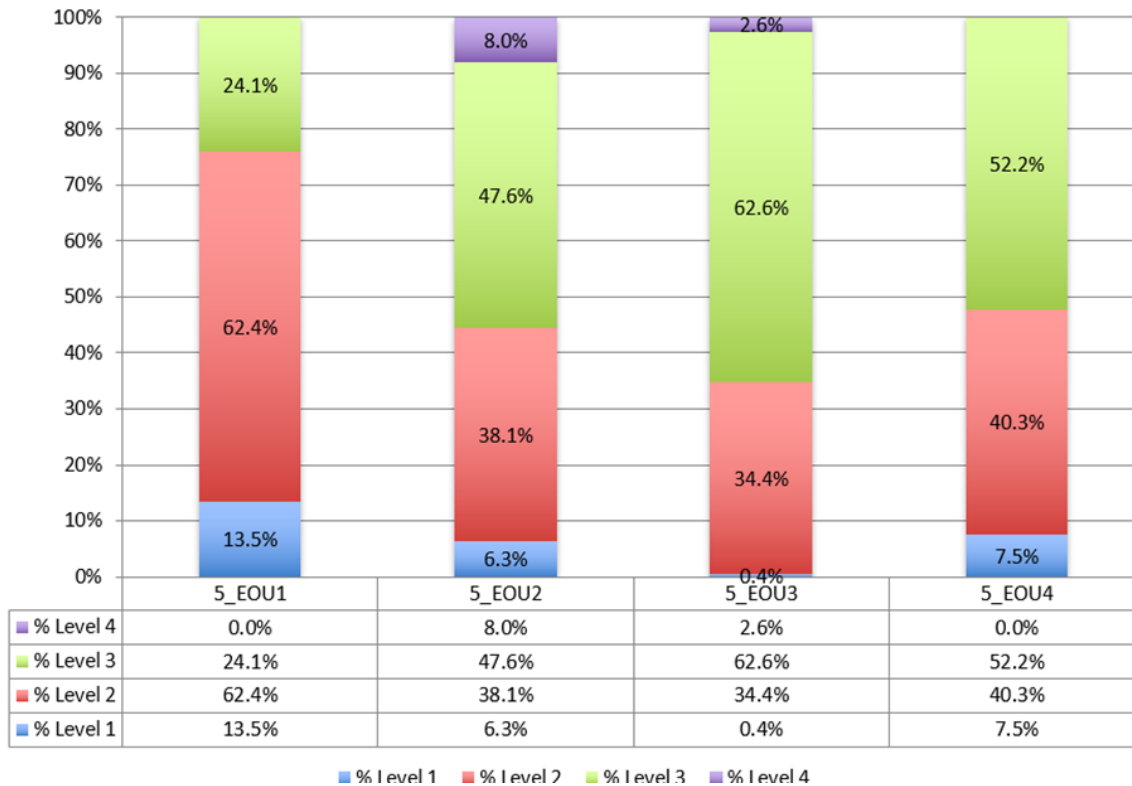


Exhibit 24. SIPS Grade 8 RP67 Initial Cut Score Impact Data

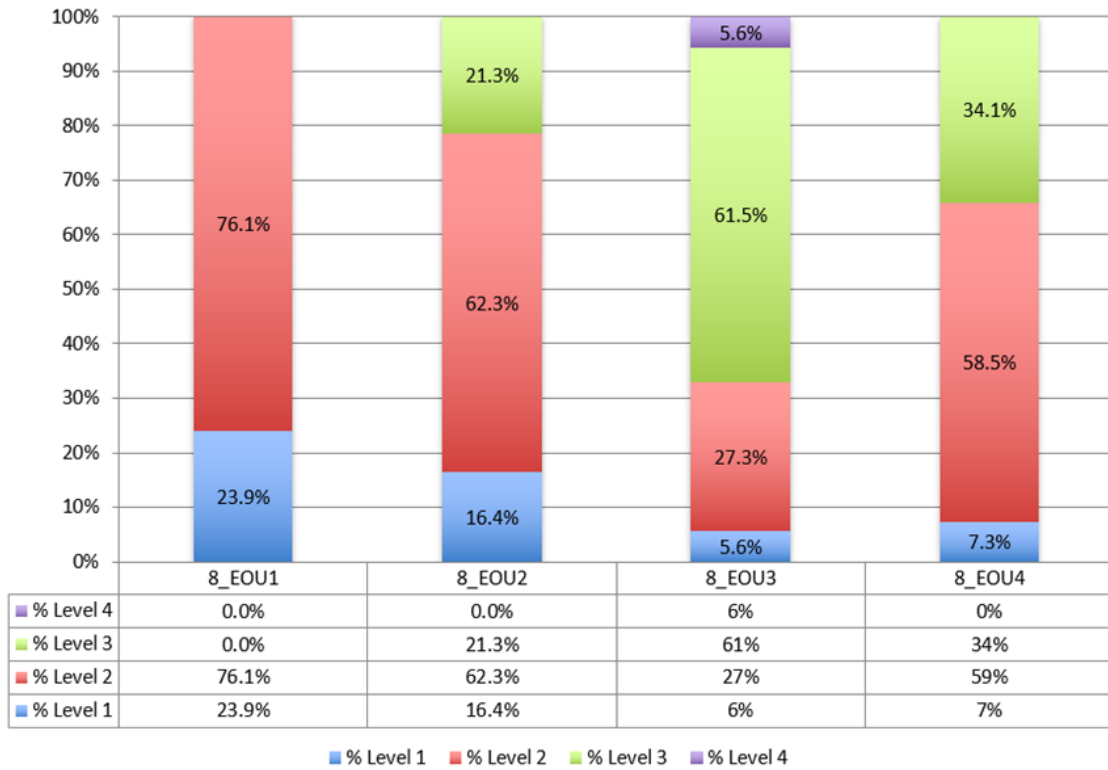
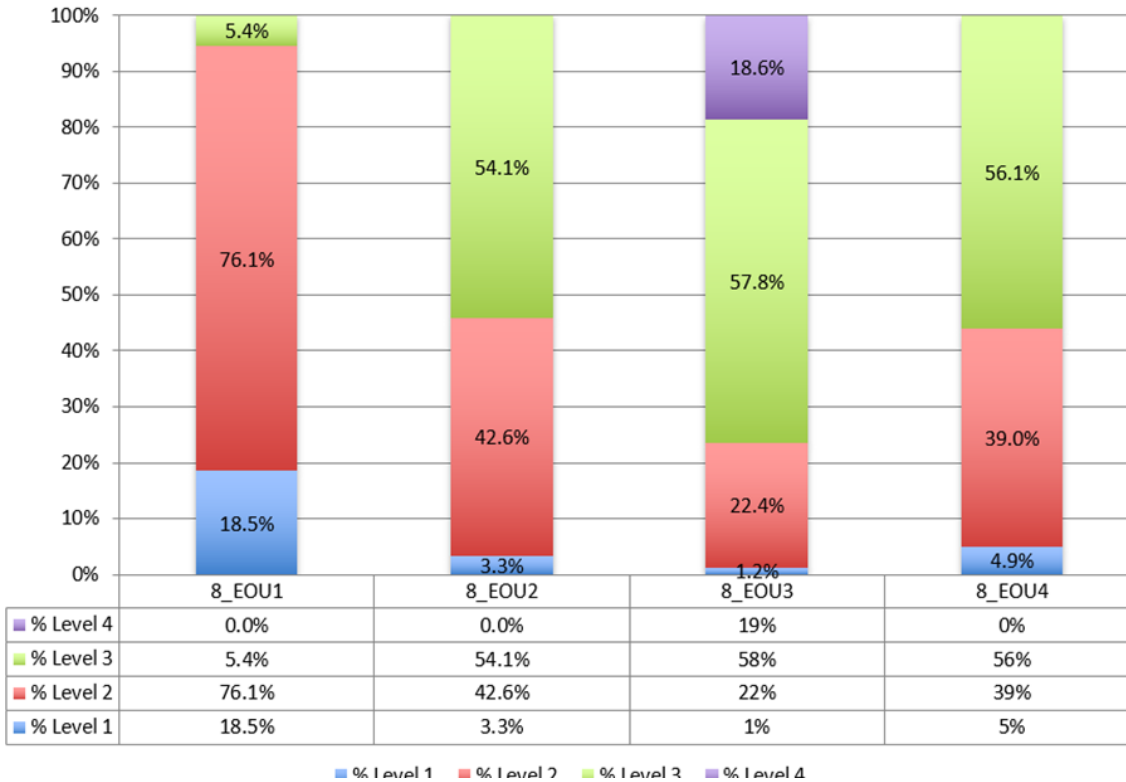


Exhibit 25. SIPS Grade 8 RP50 Initial Cut Score Impact Data



Vertically Articulated (Smooth) RP67 and RP50 cut scores and associated impact data

Error! Reference source not found. provides vertically articulated cut scores for RP67 and RP50 for each EoU and grade. **Error! Reference source not found.** provides adjustments to the initial cut scores to support the resulting vertical articulation.

Exhibit 26. Vertically Articulated SIPS RP67 and RP50 Cut Scores

Grade & RP	EoU	Level 2	Level 3	Level 4
Grade 5 RP67	EoU1	0.0115	0.4638	1.6483
	EoU2	-0.3863	0.4675	1.4645
	EoU3	-0.8720	0.2763	1.3516
	EoU4	-0.5978	0.2574	1.2587
Grade 5 RP50	EoU1	-0.3655	0.1076	1.1546
	EoU2	-0.6501	0.0086	0.8856
	EoU3	-1.0414	-0.1320	0.8147
	EoU4	-1.0342	-0.0919	0.6595
Grade 8 RP67	EoU1	-0.2749	0.7308	4.0000
	EoU2	-0.4366	0.4181	1.3791
	EoU3	-1.1163	0.1739	1.1459
	EoU4	-1.0624	0.0712	3.0557
Grade 8 RP50	EoU1	-0.5060	0.6693	4.0000
	EoU2	-0.6163	0.1782	1.2098
	EoU3	-1.6442	-0.0992	1.1526
	EoU4	-1.7309	-0.3799	2.4908

Exhibit 27. Vertical Articulation Adjustments to Cut Scores in Standard Error Units

Grade & RP	EoU	Level 2	Level 3	Level 4
Grade 5 RP67	EoU1	0	-0.5	-1.5
	EoU2	0.5	0	0
	EoU3	1.5	0	-0.5
	EoU4	-0.5	-0.5	-2.0
Grade 5 RP50	EoU1	0	-0.5	-1.5
	EoU2	0.5	0	0
	EoU3	1.5	0	-0.5
	EoU4	-0.5	-0.5	-2.0
Grade 8 RP67	EoU1	0	-1.0	0
	EoU2	0	-0.5	-1.25
	EoU3	0.5	0.75	0
	EoU4	0	-0.6	0
Grade 8 RP50	EoU1	0	-0.5	0
	EoU2	0.75	0	-0.75
	EoU3	0.5	1.0	0.5
	EoU4	0	-0.5	0

Error! Reference source not found. and 29 provide impact data associated with the vertically articulated cut scores for the four grade 5 EoUs for RP67 and RP50, respectively. **Error! Reference source not**

found. and 31 provide impact data associated with the vertically articulated cut scores for the four grade 8 EoUs for RP67 and RP50, respectively.

Exhibit 28. SIPS Grade 5 RP67 Smoothed Cut Score Impact Data

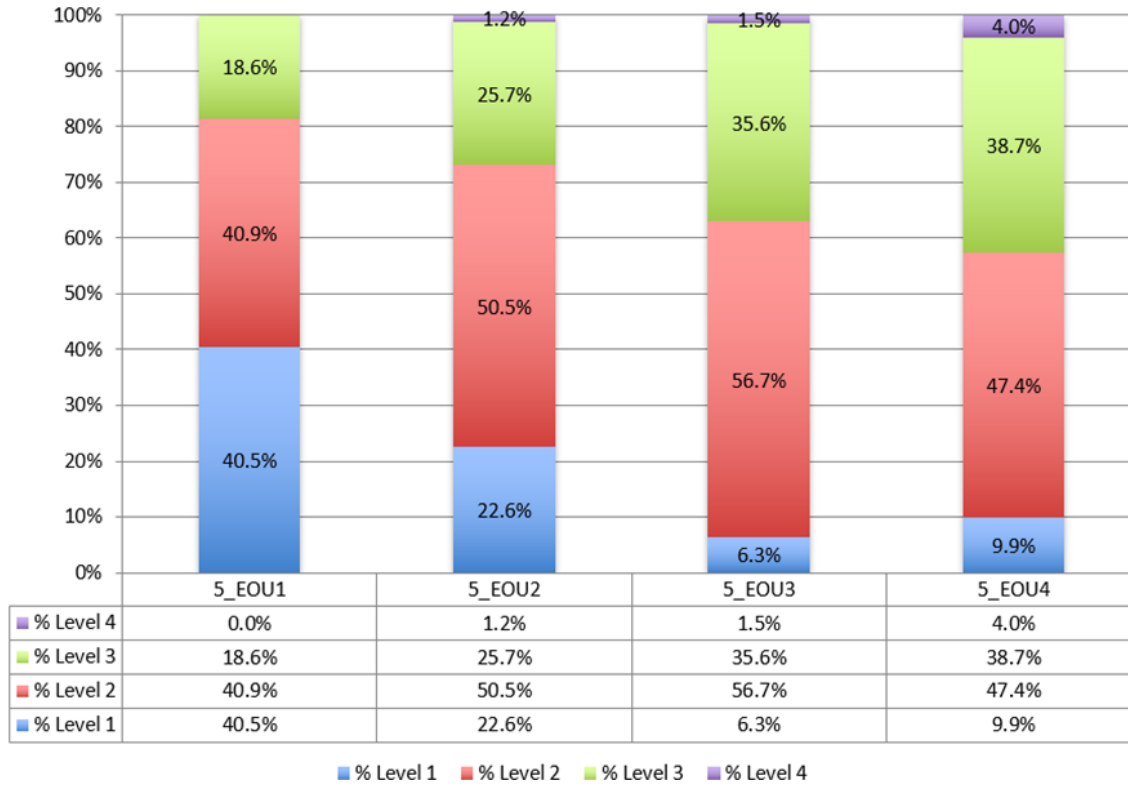


Exhibit 29. SIPS Grade 5 RP50 Smoothed Cut Score Impact Data

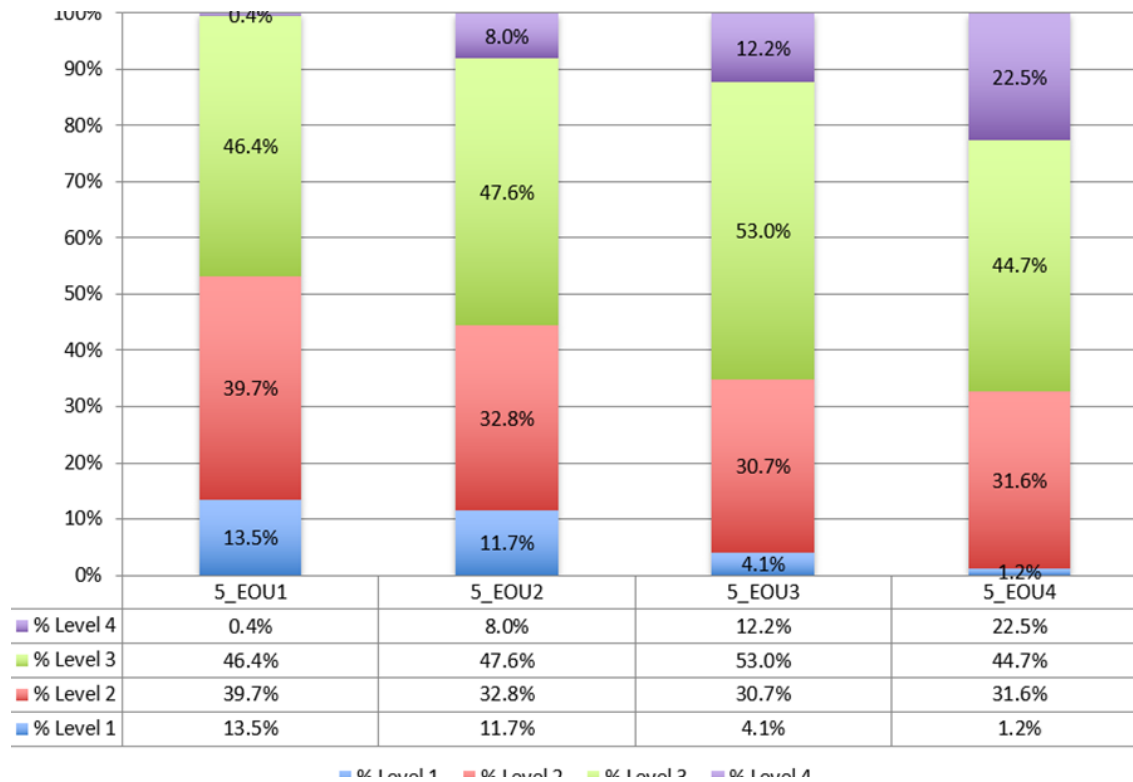


Exhibit 30. SIPS Grade 8 RP67 Smoothed Cut Score Impact Data

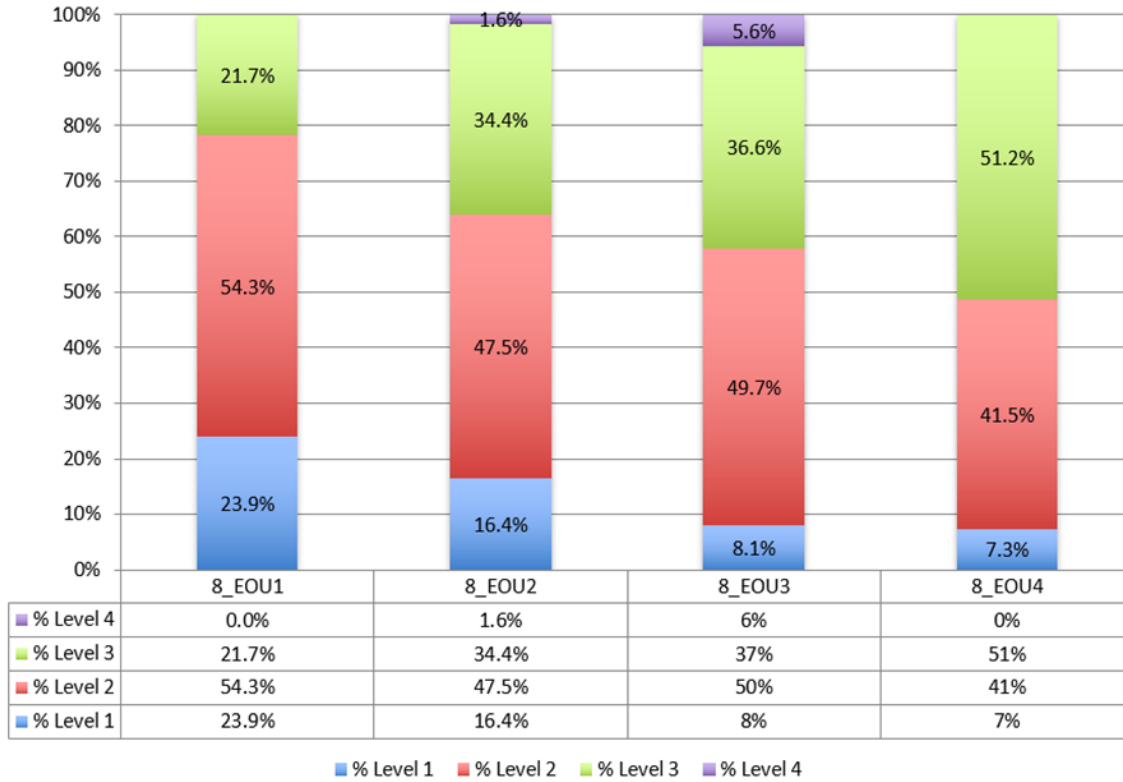
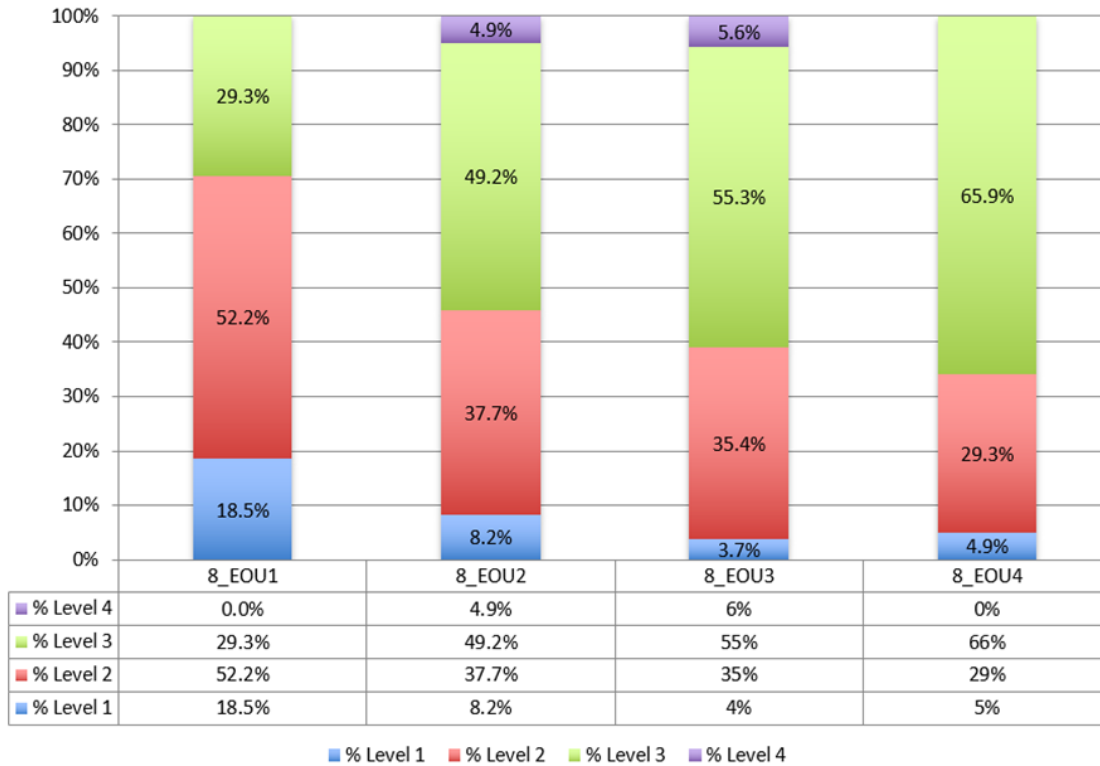


Exhibit 31. SIPS Grade 8 RP50 Smoothed Cut Score Impact Data



Technical Reporting: Validity & Peer Review Evidence

Two perspectives on validity evidence are provided here. First, the measurement literature provides validity criteria for the evaluation of standard setting processes (e.g., Cizek & Bunch, 2007; Kane, 2001; Hambleton, 2001). These criteria were reviewed and those relevant to Embedded Standard Setting are provided in **Error! Reference source not found.** Second, the USDOE (2018) provides peer review guidelines with respect to standard setting. The USDOE guidelines provide evaluation criteria they refer to as Critical Elements and examples of evidence for each Critical Element as described in Exhibit 33. Descriptions of the measurement literature and USDOE peer review evaluation criteria are provided next.

Standard Setting Validity Criteria from the Measurement Literature

There are several forms of validity evidence supporting standard setting including procedural, internal, and external validity. Exhibit 32 provides examples of specific evidence used to evaluate the relevant forms of validity evidence that have been suggested in the literature and which are appropriate for the evaluation of cut scores established under Embedded Standard Setting methodology.

Exhibit 32. Forms of Standard Setting Validity Evidence from the Literature

Forms of Validity	Validity Evidence
Procedural Validity	<ul style="list-style-type: none"> • Support for the SME-participants’ qualifications • Evidence that the participants understood the test and its intended use • Evidence that the participants understood the construct reflected by the PLDs and how items/prompts provide evidence for PLD evidence statements • Evidence that panelists were properly trained on the judgment task and were prepared to make the judgments • Evidence that the standard setting method was appropriately selected based on the test and the intended use of the cut scores • Evidence that the standard setting method was implemented as designed and if not, that the modifications were justified and appropriate • A design that incorporates iterative processes
Internal Validity	<ul style="list-style-type: none"> • The efficacy of Prompt-PLD alignment hypotheses is supported by data
External Validity	<ul style="list-style-type: none"> • Cut scores result in reasonable impact data • Placement level expectations are reasonable and consistent with expectations

Next, we summarize each form of validity and provide the associated evidence in support of the validity of the cut scores.

Procedural Validity

Support for the SMEs’ qualifications

Each SIPS SME has facilitated professional development for science teachers and they have participated in reviews of science assessment items in their content area. See SME Qualifications in the PLD Development section of the report.

Evidence that the participants understood the test and its intended use

As SIPS partners and developers of the SIPS EOU assessments, the SMEs understood the test and its intended use. Combined, the SMEs have decades of experience with state summative assessment programs, a deep understanding of *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Research Council, 2012), and the *Next Generation Science Standards* (NGSS Lead States, 2013). Throughout the duration of assessment development activities, the SIPS SMEs discussed their ratings, the knowledge and skills each identified as necessary to complete each of the prompts, and the challenges in matching a PLD descriptor to a specific rubric score. After these discussions, the SMEs reached a consensus on their ratings.

Evidence that the participants understood the construct reflected by the PLDs and how prompts provide evidence for PLD claims and targets

As SIPS partners and developers of the SIPS EoU assessments, SMEs understood the construct reflected by the PLDs. Each EoU assessment's PLDs and prompts were developed using a principled design approach (i.e., Evidence Centered Design). The SMEs developed Student Profiles and PLDs to reflect each other and the desired goals of each EoU assessment's associated curricular unit. Thus, the prompts were aligned by design to the Student Profiles and PLDs, and the data reported in the section under the heading, "The Efficacy of SMEs' Prompt-PLD Alignments," provides evidence that the resulting Prompt-PLD alignment hypotheses were supported by data.

Evidence that panelists were properly trained on the judgment task and prepared to make the judgments

The SMEs conducted the judgment task—the alignment of each EoU task and score point to a performance level. The SMEs were trained using the guidelines reported in the Prompt-PLD Alignment section of this report and were given the opportunity to ask questions. Evidence that they were able to follow the guidelines is provided by the data presented in this report in the section labeled, "The Efficacy of SMEs' Prompt-PLD Alignments." The reported correlations, weighted Kappas (which tended to be substantial to nearly perfect), and agreement rates all provide strong evidence that the SMEs were properly trained on the judgment task and prepared to make the judgments.

Evidence that the standard setting method was appropriately selected based on the test and the intended use of the cut scores

Embedded Standard Setting is an appropriate standard setting method for assessments (a) developed from inception to administration under a principled design framework, (b) with constructs that are well articulated and explicated by PLDs, and (c) with prompts that are aligned by qualified SMEs to the PLDs. The SIPS assessments meet all criteria and thus, ESS is an appropriate standard setting method for SIPS.

Evidence that the standard setting method was implemented as designed and if not, that the modifications were justified and appropriate

Embedded Standard Setting encompasses the integrated and iterative set of processes and procedures that span the assessment lifecycle, supporting the coherence of the various assessment system elements described next and illustrated in Exhibit 1.

) illustrates the ESS processes specified by Lewis (2021). The SIPS standard setting was conducted as intended for an assessment program utilizing PAD from inception.

A design that incorporates iterative processes

The green feedback loops illustrated in Embedded Standard Setting encompasses the integrated and iterative set of processes and procedures that span the assessment lifecycle, supporting the coherence of the various assessment system elements described next and illustrated in Exhibit 1.

) show the iterative nature of ESS processes. The modest nature of the Pilot Study described in this report was not sufficient to iterate to resolve ESS-Inconsistent prompts, which would provide the opportunity to refine the PLDs and/or the Prompt-PLD alignments. However, this is recommended when a more comprehensive field test is conducted by states considering the large-scale administration of the SIPS assessments. Thus, iterative processes are recommended before adopting SIPS cut scores.

Internal Validity

The efficacy of Prompt-PLD alignment hypotheses is supported by data

The reported correlations, weighted Kappas (which tended to be substantial to nearly perfect), and agreement rates reported in the section under header “The Efficacy of SMEs’ Prompt-PLD Alignments” all provide strong evidence for the efficacy of the Prompt-PLD alignments.

External Validity

Cut scores result in reasonable impact data and in accordance with expectations

Throughout the ESS process, SMEs were informed they would have an opportunity to evaluate the impact data associated with a set of ESS cut scores. The SMEs determined that performance standards were well-articulated when the impact data associated with a set of cut scores formed a reasonable, explainable pattern across grades. The SMEs inspected the impact data and were generally satisfied with the cut score recommendations. After discussions about the cut score recommendations for each EOU, the SMEs noted that various factors such as opportunity to learn or uneven implementation of curricula will likely cause shifts in prompt and task difficulty over the first few years of a testing program and thus, reliable data supporting cut score refinement will not emerge until after the first operational year or two of a testing program. Placement level expectations are reasonable and consistent with expectations.

Because the SIPS PLDs are hypothesized learning progressions articulated across the performance levels within each grade and SIPS instructional unit, it is understood that the placement level expectations too are hypothesized. The SIPS SMEs feel the placement level expectations are reasonable given the multi-dimensional nature of the assessed construct and the innovative ‘Chain-of-Sense-Making’ inherent throughout the prompts comprising any single task structure within a given EOU.

Peer Review Standard Setting Critical Elements

Federal peer review accountability guidelines associated with standard setting are provided in Critical Element Section 6.2—Performance Standards Setting (USDOE, 2018). The single Critical Element cited in this section follows:

“The State used a technically sound method and process that involved panelists with appropriate experience and expertise for setting Academic Performance standards..., such that cut scores are developed for every grade..., content domain...and/or composite for which Performance level scores are reported.”

Note that this Critical Element calls out the technical foundations of the method and the qualifications of the participants. The Embedded Standard Setting methodology has been reviewed and approved by state (Maine) and consortia (CAAELP) technical advisory committees for use for state summative federal accountability assessments. ESS methodology is technically sound as described in peer-reviewed journals (Lewis & Cook, 2020; Lewis, Graw, & Baker (in press)) and further detailed in numerous conference presentations.

The USDOE guidelines provide examples of evidence that may be included in the standard setting technical report to support this Critical Element for the assessments of interest. Exhibit 33 provides these examples.

Exhibit 33. Examples of Evidence Supporting Peer Review Critical Element 6.2

Peer review examples of evidence
<ul style="list-style-type: none">• A description of the standards-setting method and process used by the State;• The rationale for the method selected;• Documentation that the method used for setting cut scores allowed panelists to apply their knowledge and experience in a reasonable manner and supported the establishment of reasonable and defensible cut scores;• Documentation of the process used for setting cut scores and developing Performance level descriptors aligned to the State’s standards;• A description of the process for selecting panelists;• Documentation that the standards-setting panels consisted of panelists with appropriate experience and expertise;• If available, a summary of statistical descriptions and analyses that provides evidence of the reliability of the cut scores and the validity of recommended interpretations;• A technical report providing a description of the method used and results

Peer Review Standard Setting Validity Evidence

A description of the standards-setting method and process used by the State

ESS methodology is described in this section under the heading, “

ESS analyses.”

The rationale for the method selected

ESS was selected because it is the natural extension of PAD to standard setting and SIPS was developed using a PAD framework. Embedded Standard Setting is an appropriate standard setting method for assessments (a) developed from inception to administration under a principled design framework, (b) with constructs that are well articulated and explicated by PLDs, and (c) with items that are aligned by qualified SMEs to the PLDs. The SIPS assessments meet all criteria and thus, ESS is an appropriate standard setting method for SIPS.

Documentation that the method used for setting cut scores allowed panelists to apply their knowledge and experience in a reasonable manner and supported the establishment of reasonable and defensible cut scores

The information provided in the sections under the headings, “PLD Development” and “Prompt-PLD Alignment” demonstrate that the ESS processes required SME judgments that allow them to directly apply their knowledge and experience. The reasonableness of the manner and the validity of reasonable and defensible cut scores is evidenced by the results reported under the heading, “The Efficacy of SMEs’ Prompt-PLD Alignments,” in which the reported correlations, agreement rates, and weighted Kappas all support the efficacy of the SME judgments and defensibility of the cut scores. The SIPS SMEs felt that performance standards were reasonably well-articulated when the impact data associated with a set of cut scores formed a reasonable, explainable pattern across grades. While the pilot study data may suggest areas of instruction to be considered, without a sufficiently large nor representative sample of the target population, the results are considered ‘exploratory’ until sufficient data is available. The alignment across PLDs and EOUs should be viewed as an ongoing process in need of continual monitoring. As such, the SMEs will use the information provided by the ESS Study to improve the meaningful use and interpretation of students’ SIPS EOU assessment results.

Documentation of the process used for setting cut scores and developing performance-level descriptors aligned to the State’s standards

The process used to develop PLDs is described in detail under the heading, “PLD Development,” and the process used for setting cut scores is described under the headings, “ESS Analyses” and “Vertical Articulation.”

A description of the process for selecting panelists

The processes supporting ESS for the SIPS assessments—PLD development and Prompt-PLD alignments were conducted by the SIPS SMEs. It is recommended that states considering the adoption of SIPS cut scores conduct a review by panelists representative of the state’s educators and with appropriate teaching experience, knowledgeable about the NGSS and the state’s science curriculum, and student learning in science.

Documentation that the standards-setting panels consisted of panelists with appropriate experience and expertise

Documentation of the expertise and experience of the SIPS SMEs is provided in the section under the heading, “SME Qualifications.” Members of the SIPS team have extensive science expertise and experience in multiple areas of education, including as K-12 teachers, adjunct instructors at the university level, professional learning providers in both K-12 and higher education settings, and through positions in state-level science education leadership (i.e., senior content specialists, state assessment

directors, and assistant state assessment directors). Members also have experience acting as both panelists and facilitators for science standard setting meetings as well as developers of state-level science assessment programs through the application of evidence-centered design (ECD) to design, develop, and implement NGSS-aligned assessments and to create performance level descriptors and ultimately cut scores for federal accountability and reporting purposes. Finally, the SIPS SMEs have extensive experience in the exploratory design of innovative assessments to produce both design approaches and early-stage tasks critical for establishing frameworks for researching and developing more extensive suites of innovative assessment tasks. As a result, the PLDs may be considered the product of collaboration among science experts, curriculum specialists, teachers, and policy makers.

A summary of statistical descriptions and analyses that provides evidence of the reliability of the cut scores and the validity of recommended interpretations

A summary of the requested statistical descriptions is provided in the sections under the headings, “The Efficacy of SMEs’ Prompt-PLD Alignments” and “

ESS analyses.”

A technical report providing a description of the method used and results.

The current document summarizes the methodology and results of the ESS methodology and processes.

Summary

The data and evidence provide provisional support for the validity of the estimated SIPS cut scores. This technical report, while not formally structured in terms of a validity argument, presents one in terms of the singular focus on the following evidentiary chain of reasoning articulated throughout the report:

1. PLDs should explicate and articulate the content standards of interest—the NGSS—and map to intended interpretations as described in the PLD development section of the technical report.
2. Items should map to PLDs to operationalize and provide evidence for the claims and measurement targets articulated in the PLD evidence statements, as described in the Prompt-PLD alignment section of the technical report.
3. Cut scores should map to the appropriate items, which is supported by the ESS estimation of cut scores that optimize the coherence of the Prompt-PLD alignments and empirical data, as described in the ESS Analyses section of the technical report.

These evidentiary linkages are supported by design via the PAD and ESS processes. Inconsistent prompts, which degrade score interpretation, are identified in the technical report. Inconsistent prompts that degrade score interpretation are not specifically a reflection on the quality of the SIPS assessments. They exist under any item-based standard setting methodology (i.e., Bookmark, ID Matching, Yes-No Angoff, etc.) but go undetected under other approaches. ESS minimizes the degradation and offers opportunities to further mitigate degradation through iterative review and revision.

Given the relatively small case counts from the SIPS 2022-23 Pilot Study, the consistency status of items should be considered tentative until more substantial field- or operational-test administrations provide more reliable data to support subsequent analyses and item or PLD refinement. Continuing the application of PAD and ESS in this way will provide evidence supporting the mapping of SIPS assessment scores to the intended score interpretations. ESS is designed to optimize this evidentiary argument.

We close by noting that the data and analyses are based on the prototype EOUs that were administered in the pilot study. If prompts are revised based on the pilot study results, the standard setting analyses should be updated to reflect data from the updated prompts.

Integrating the EoUs: Methods for Reporting an End-Of-Year Summative Score or Performance Level

Performance levels have been estimated for each of the four EoUs in each grade. However, for SIPs to be considered as a science assessment that meets federal ESSA accountability requirements, a summative performance level is required. There are several methods that may be considered to aggregate students’ four EoU scores and/or EoU performance levels into an aggregate score and/or performance level and some require a common scale. The SIPS pilot study design was limited by participation and thus, a common scale was not developed. However, the SIPS pilot study was sufficient to estimate item response theory parameters for each EoU on a unique scale for each EoU in each grade. ESS analyses were conducted on each EoU to estimate EoU-specific cut scores, as described in this section. Next, we describe a few methods that may be used to support end-of-year summative reporting. First, we describe precedent and methodology for the use of performance level profiles and an exemplar rubric that may be considered to convert the four EoU scores in a grade into a summative performance level. We use the limited number of matched data cases from the Pilot Study to illustrate how individual profiles may be converted into a summative performance level. Second, we describe the development of a SIPS PLD-based Scale that is supported by the common qualitative interpretation of each EoU’s performance levels.

Performance Level Profiles

A rubric may be adopted that associates students’ four EoU performance level profiles with an overall performance level. For instance, ELPA21 reports five performance levels for each of four domains—Reading, Writing, Listening, and Speaking—and adopted the rubric presented in Exhibit 34 to aggregate the four performance levels into a summative Proficiency Determination.

Exhibit 34. ELPA21 Profiles of Proficiency

Rules	Profiles (examples)	Proficiency Determination
A profile of 4s and 5s meets assessment targets and indicates overall proficiency	4444 5555 4545 5454 4455 5544 4445 4454 4544 5444 5554 5545 5455 4555 4E44	Proficient
A profile with one or more domain scores above Level 2 that does not meet the requirements to be Proficient	3333 1333 3353 3233 2242 1234 1114 2232	Progressing
A profile of 1s and 2s indicates an “Emerging” level of proficiency	1122 1212 E222 2222	Emerging

Note. The order of the example profiles of the four domains is: 1) reading, 2) writing, 3) speaking, and 4) listening. “E” indicates an exempt test.

Empirical Data Analyses of Performance Level Profiles

The SIPS Pilot Study sample was modest—not all students in a grade took all four EoUs. However, 64 and 21 students took all four EoUs in grades 5 and 8, respectively. The cross-EoU performance level profiles for these matched cases are provided in Exhibits 35 and 36 for grades 5 and 8, respectively. The final columns of Exhibits 35 and 36 apply the following modification of the ELPA21 rubric shown in Exhibit 34.

We adapted the ELPA21 rubric to illustrate how the four individual EoU performance levels may be aggregated into a three-level summative performance level:

- Summative Level 3: Level 3 or 4 on all EoUs
- Summative Level 2: At least one EoU below Level 3 and above Level 1
- Summative Level 1: Level 1 on all EoUs

This rubric modification is provided for illustrative purposes only and is not intended to reflect SIPS or SIPS partner state policy; other rubrics are possible, including a four-level rubric.

We applied this adapted rubric to the matched data sets reported in Exhibits 35 and 36. Only one grade 5 student had a summative performance level other than Level 2. A single student achieved Level 3. All matched cases in grade 8 resulted in a summative performance level of Level 2.

Exhibit 35. Grade 5 Cross-EoU Performance Level Profiles

Grade 5				Summative Performance Level Based on Exemplar Rubric
Performance Level Profiles		Count	Percent	
EoU1	EoU2 EoU3 EoU4			
	1121	2	3.13%	Level 2
	1122	2	3.13%	Level 2
	1221	1	1.56%	Level 2
	1222	5	7.81%	Level 2
	1223	6	9.38%	Level 2
	2121	2	3.13%	Level 2
	2122	3	4.69%	Level 2
	2222	10	15.63%	Level 2
	2223	15	23.44%	Level 2
	2232	2	3.13%	Level 2
	2233	2	3.13%	Level 2
	2322	2	3.13%	Level 2
	2323	6	9.38%	Level 2
	2333	4	6.25%	Level 2
	3222	1	1.56%	Level 2
	3333	1	1.56%	Level 3
	Total	64	100.00%	

Exhibit 36. Grade 8 Cross-EoU Performance Level Profiles

Grade 8				Summative Performance Level Based on Exemplar Rubric
Performance Level Profiles		Count	Percent	
EoU1	EoU2			
	1121	1	4.76%	Level 2
	1122	3	14.29%	Level 2
	1222	1	4.76%	Level 2
	2222	1	4.76%	Level 2
	2232	6	28.57%	Level 2
	2233	8	38.10%	Level 2
	2333	1	4.76%	Level 2
	Total	21	100.00%	

Developing a SIPS PLD-Based Scale

Recall that we do not have a common scale across EoUs in a grade. However, a common scale can be developed in a few ways based on the following rationale: Each EoU has a unique set of PLDs that form the basis for the Task-PLD alignments and cut score estimation and each EoU’s PLD level reflects a common expectation for student performance relative to the EOU’s instructional unit. For example, Level 3 on each EOU reflects the target achievement for the associated curricular unit. Thus, each level is qualitatively comparable across EoUs. Averaging the level across EoUs provides an average student performance based on the four comparable EoU-specific targets.

Thus, each EoU performance level may be translated to a numerical value where Level 1 = 1, Level 2 = 2, Level 3 = 3, Level 4 = 4. An average of the four levels may be estimated to provide an overall score and students’ scores can be used to assign a summative performance level. For example, the following rubric may be used to assign the average level value to a summative level:

- Average from 1.0-1.5 = Summative Level 1
- Average from 1.51-2.5 = Summative Level 2
- Average from 2.51-3.5 = Summative Level 3
- Average from 3.51-4.00 = Summative Level 4

Next, we propose a possible refinement to the proposed SIPS PLD-Based Scale.

A PLD-Based Scale Refinement

A refinement to this performance-level-based scale may be useful and add precision by dividing the intervals between the SIPS EoU cut scores for a given EoU into say, three equal units (or some other logical division possibly suggested by the range of score points associated with each

EoU performance level). For example, a scale ranging from 1.1 to 4.3 might be developed as follows:

Range of Performance Level 1: 1.1 to 1.3, where

- 1.1 indicates the student is in the first third of the interval between the lowest obtainable score and the Level 2 cut score,
- 1.2 indicates the student is in the second third of the interval between the lowest obtainable score and the Level 2 cut score,
- 1.3 indicates the student is in the final third of the interval between the lowest obtainable score and just below the Level 2 cut score.

Range of Performance Level 2: 2.1 to 2.3, where

- 2.1 indicates the student is in the first third of the interval between the Level 2 and the Level 3 cut score,
- 2.2 indicates the student is in the second third of the interval between the Level 2 and the Level 3 cut score,
- 2.3 indicates the student is in the final third of the interval between the Level 2 cut score and just below the Level 3 cut score.

Range of Performance Level 3: 3.1 to 3.3, where

- 3.1 indicates the student is in the first third of the interval between the Level 3 and the Level 4 cut score,
- 3.2 indicates the student is in the second third of the interval between the Level 3 and the Level 4 cut score,
- 3.3 indicates the student is in the final third of the interval between the Level 3 cut score and just below the Level 4 cut score.

Range of Performance Level 4: 4.1 to 4.3, where

- 4.1 indicates the student is in the first third of the interval between the Level 4 cut score and the highest obtainable score,
- 4.2 indicates the student is in the second third of the interval between the Level 4 cut score and the highest obtainable score,
- 4.3 indicates the student is in the final third of the interval between the Level 4 cut score and the highest obtainable score.

By averaging students' PLD-based scale values across the four EoUs we can estimate a summative score which may be translated into a summative performance level using a policy-based rubric.

Summary

In this section, we proposed several methods that can be used to estimate summative performance levels and scores. Given the small sample sizes and limited number of matched cross-EoU cases, we recommend further analyses be conducted following a broader test administration that may be used to update the ESS cut scores and to evaluate the efficacy of the proposed summative scoring methods proposed here.

References


- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Brice, A. & Mix, D. (2021, May). Practical Application of Principled Assessment Design in Building Content Expertise While Developing and Refining PLDs Through Embedded Standard Setting. Presentation in Davis, L. (Organizer) *Creating Coherence: Integrating Principled Assessment Design, PLDs, and Standard Setting*. Organized session accepted for presentation at the 2021 annual meeting of the National Council on Measurement in Education.
- Cizek, G. J. (1996). Standard setting guidelines. *Educational Measurement: Issues and Practice*, 15(1), 12–21.
- Cizek, G., & Agger, C. (2012). Vertically Moderated Standard Setting. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). Routledge.
- Cizek, G., & Bunch, M. (2007). *A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.
- Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Jaeger, R. M. (1989). Selection of Judges for Standard Setting: What Kinds? How Many? Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- MEHRENS, W. A. (1986). Measurement Specialists: Motive to Achieve or Motive to Avoid Failure? *Educational Measurement: Issues and Practice*, 5(4), pp. 5–10.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–74.
- Lewis, D. Graw, M., & Baker, M. (in press). Embedded Standard Setting for Credentialing Exams. Manuscript accepted for publication in *Journal of Applied Testing Technology*, published by ATP.
- Lewis, D. & Cook, R. (2020). Embedded Standard Setting: Aligning standard setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practices*, 39(1), 8–21.
- Lewis, D. M., & Haug, C. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education*, 18(1), 11–34.
- Lewis, D. & Lee, S. (2020). EmStanS [Embedded Standard Setting computer software]. Creative Measurement Solutions LLC.
- Lewis, D. (2021, May). The Embedded Methods Validity Framework in Support of Assessment System Coherence. In D. Lewis (Organizer), *Developing an Alternate English Language Proficiency Assessment within a Principled Design Framework Symposium at the 2021 annual meeting of the National Council on Measurement in Education*, virtual.

- Lewis, D., Lee, S., & Choi, S. (2021, May). An Investigation of Two Embedded Standard Setting Cut Score Estimation Algorithms: ESS-Count & ESS-Weight. In D. Lewis (Organizer), *Embedded Standard Setting: Research & Advances*. Symposium at the 2021 annual meeting of the National Council on Measurement in Education, virtual.
- Schneider, M.C., Chen, J., Nichols, P.D. (2021). Using Principled Assessment Design and Item Difficulty Modeling to Connect Hybrid Adaptive Instructional and Assessment Systems: Proof of Concept. In: Sottolare, R.A., Schwarz, J. (eds) *Adaptive Instructional Systems. Adaptation Strategies and Methods*. HCII 2021. Lecture Notes in Computer Science(), vol 12793. Springer, Cham. https://doi.org/10.1007/978-3-030-77873-6_11
- USDOE (2018). *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process*. USDOE, author. Download available at [assessmentpeerreviewguidepubliccommentrevisionsjuly2018\(OGC\)\(8.28\)\(PDF \(ed.gov\)\)](https://www.ed.gov/sites/default/files/2018/07/assessmentpeerreviewguidepubliccommentrevisionsjuly2018(OGC)(8.28)(PDF%20ed.gov).pdf).

Appendix A: SIPS Policy Level Descriptors


SIPS Policy Level Descriptors			
Level 1	Level 2	Level 3	Level 4
When students are presented with units of instruction intended to promote sense-making of phenomena and/or the design of solutions to engineering problems by engaging in fair, equitable, authentic, and multiple opportunities to learn that integrate grade level-appropriate science and engineering practices (SEPs), disciplinary core ideas (DCIs), and crosscutting concepts (CCCs), student evidence of learning shows the absence of sense-making to describe phenomena and to identify solutions to engineering problems through integration of the NGSS dimensions.	When students are presented with units of instruction intended to promote sense-making of phenomena and/or the design of effective solutions to engineering problems by engaging in fair, equitable, authentic, and multiple opportunities to learn that integrate grade level -appropriate science and engineering practices (SEPs), disciplinary core ideas (DCIs), and crosscutting concepts (CCCs), student evidence of learning shows partial sense-making to describe phenomena and to design solutions to engineering problems through integration of the NGSS dimensions.	When students are presented with units of instruction intended to promote sense-making of phenomena and/or the design of effective solutions to problems by engaging in fair, equitable, authentic, and multiple opportunities to learn that integrate grade level-appropriate science and engineering practices (SEPs), disciplinary core ideas (DCIs), and crosscutting concepts (CCCs), student evidence of learning shows sense-making and thinking like a scientist and/or engineer to accurately explain phenomena and to design relevant solutions to engineering problems through integration of the NGSS dimensions.	When students are presented with units of instruction intended to promote sense-making of phenomena and/or the design of effective solutions to engineering problems by engaging in fair, equitable, authentic, and multiple opportunities to learn that integrate grade level -appropriate science and engineering practices (SEPs), disciplinary core ideas (DCIs), and crosscutting concepts (CCCs), student evidence of learning shows authentic sense-making and thinking like a scientist and/or engineer to fully and accurately explain phenomena and to design innovative solutions to engineering problems through the integration of the NGSS dimensions.
The student is making limited progress toward becoming an informed consumer of information to apply and transfer a limited understanding of three-dimensional science knowledge and skills in cross-disciplinary ways.	The student is working toward becoming an informed consumer of information and to apply and transfer an incomplete understanding of three-dimensional science knowledge and skills in cross-disciplinary ways.	The student is a critical consumer of information and accurately and meaningfully applies and transfers an adequate understanding of three-dimensional science knowledge and skills in cross-disciplinary ways.	The student is a critical consumer of information and accurately and meaningfully applies and transfers an exceptional understanding of three-dimensional science knowledge and skills in cross-disciplinary ways.
The student demonstrates limited progress toward preparedness for college, the workforce, and civic opportunities.	The student demonstrates some progress toward preparedness for college, the workforce, and civic opportunities.	The student demonstrates significant progress toward preparedness for college, the workforce, and civic opportunities.	The student demonstrates substantial progress toward preparedness for college, the workforce, and civic opportunities.
The student may need extensive instruction or reteaching of prior knowledge and/or key grade-level NGSS science concepts and skills.	The student may need additional instruction or reteaching of key grade-level NGSS science concepts and skills.		

Appendix B: ESS Powerpoint Presentation



SIPS Standard Setting Methodology

1



SIPS Standard Setting

- SIPS will utilize Embedded Standard Setting (ESS) Methodology
- Embedded Standard Setting Methodology was inspired by advances in principled assessment design and prompt alignment practice that results not only in the alignment of prompts to claims and measurement targets but also to specific performance levels. (Forte, 2019, 2018, 2017; Lewis & Forte, 2019)

Prompt Template			
Standard			
Claim			
...			
Performance Level Alignment			
Score Point	Performance Level	PLD Evidence Statement	Recommended PLD Edits
1	1		
2	2		
3	4		

2

Fundamental Standard Setting Judgments



Modified Angoff Yes/No

I am aligning this prompt (or prompt score point) to Level X because the Performance Level Descriptors indicate that a Level X examinee should have the KSAs to (more likely than not) respond successfully.

ID Matching Method

I am aligning this prompt (or score point) to Level X because the content knowledge, skills, and cognitive processes that this prompt (or prompt score point) requires most closely match the knowledge and skill expectations of the Level X PLD.

Embedded Standard Setting

I am developing this prompt to align to Level X by design. The prompt was developed to provide evidence that an examinee who responds successfully to this prompt (or prompt score point) exhibits the attributes of the specified evidence statement in the Level X PLD.



3

3

Prompt Alignment and ESS



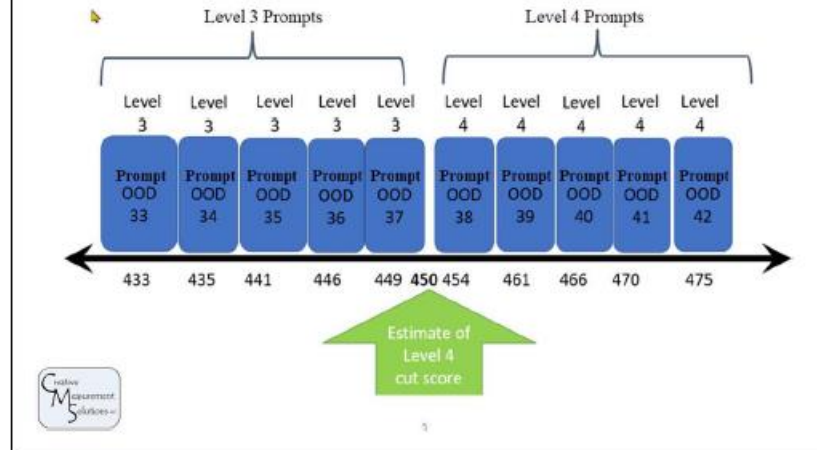
- Cut scores are organically and analytically estimated by identifying the scores that optimize the relationship between the Prompt-PLD alignments and empirical data.
- ESS cut scores optimize the evidentiary relationship between test prompts and the claims and measurement targets articulated in the ALDs.



4

4

Embedded Standard Setting

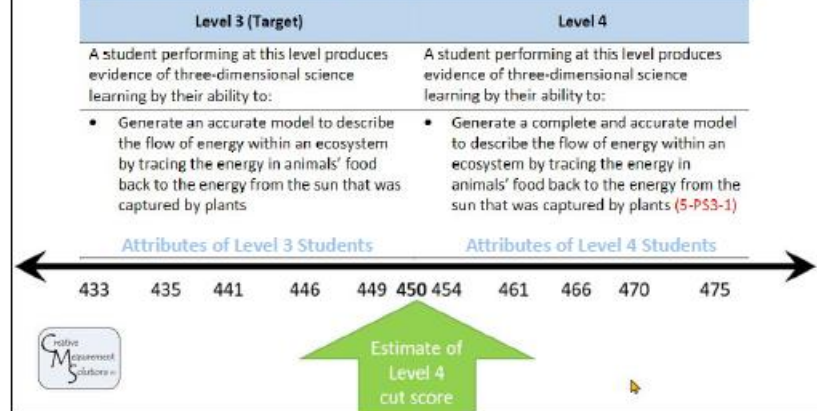


5

SIPS Grade 5 Unit 2 Range PLDs

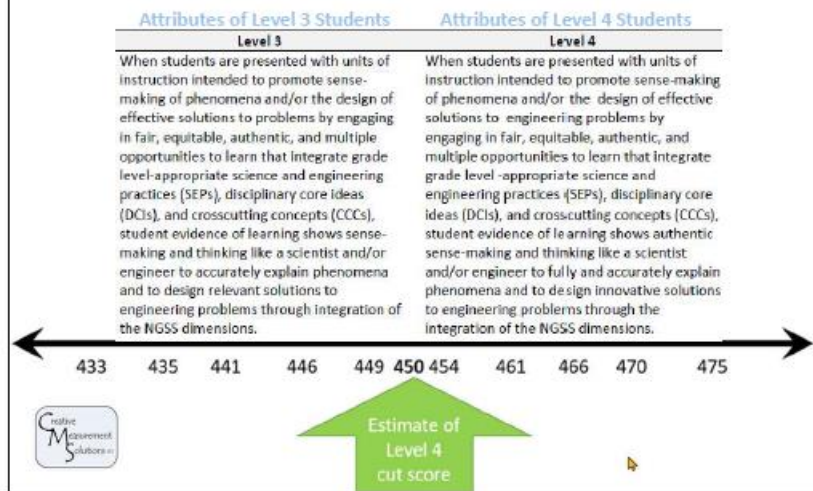


SIPS tasks require students to apply and transfer their science learning through engagement with science and engineering practices (SEPs) and application of the crosscutting concepts (CCCs) to demonstrate their understanding of disciplinary core ideas (DCIs) to make sense of and explain phenomena and/or to design solutions to phenomena-rooted engineering problems.



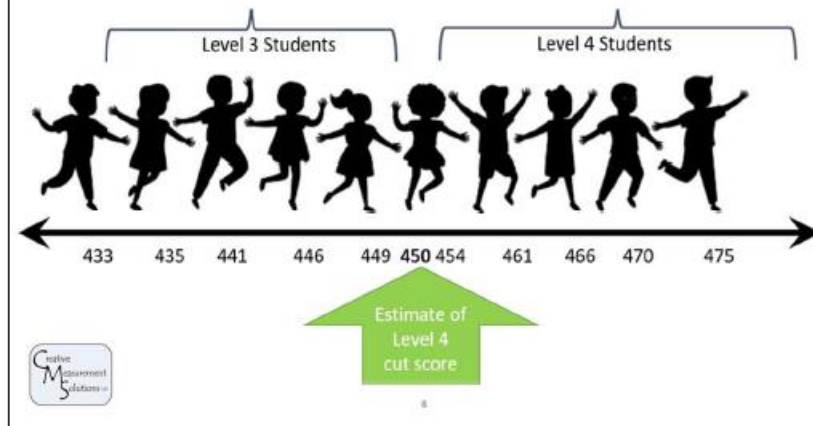
6

SIPS Policy Level Descriptors



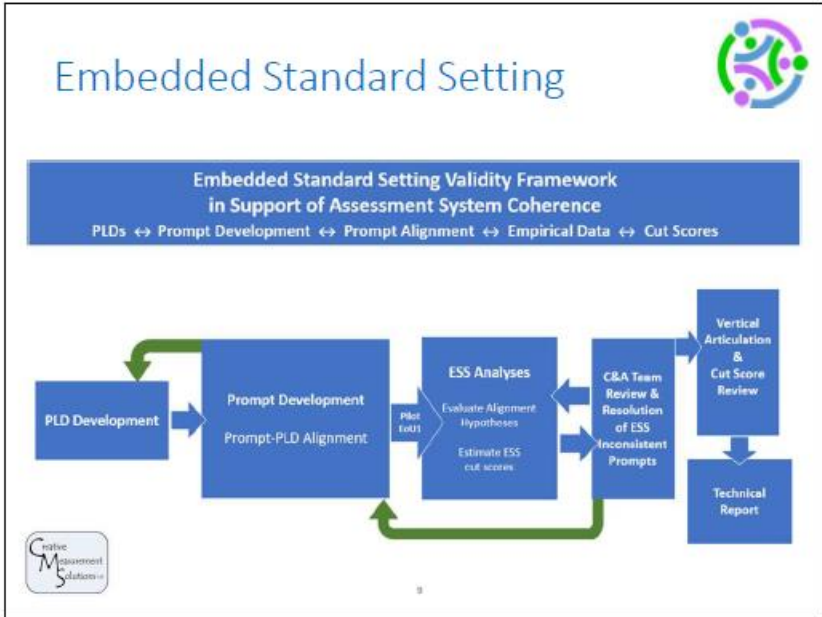
7

Embedded Standard Setting



8

Embedded Standard Setting

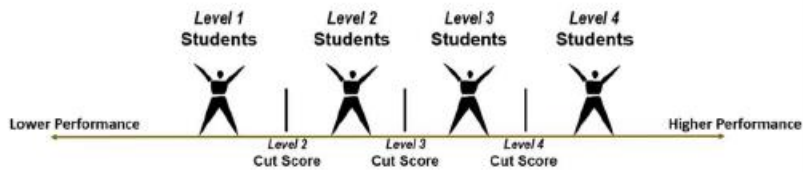


9

Timeline



- ESS cut scores will be estimated for each EOU assessment in each grade when sufficient data is collected.
- Following the 2022-23 pilot administration, we will have 3 cut scores establishing 4 performance levels for each EOU assessment for each grade.



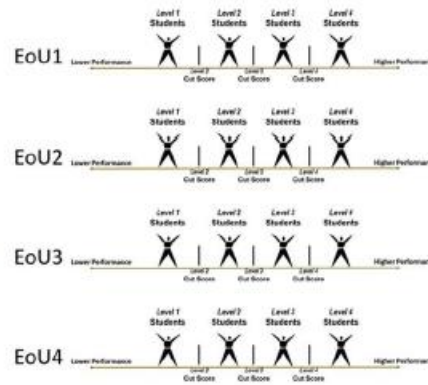
10

10

An Actionable Performance Scale



- Following the pilot study, we will examine the *system* of cut scores to support interpretation.
- Each EOU assessment is on its own scale—not a common scale.
- Performance levels have the same interpretation for each EOU assessment—they form an actionable SIPS performance scale.



11

11

Summative Proficiency



At the conclusion of the pilot study, we will discuss student profiles to establish a rubric associating EOU profiles with Overall Summary Proficiency Levels.

Overall	Eou1	EoU2	EoU3	EoU4
Level 3	Level 3	Level 3	Level 3	Level 3
Level 3?	Level 3	Level 3	Level 3	Level 2
?	Level 3	Level 3	Level 3	Level 1



12

12

Comments and Questions



13

Appendix C: Detailed ESS Prompt Maps

Table 1. Detailed ESS Prompt Maps: Grade 5 EOU1

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU1_T1_P1_1	1	-0.9651	Level1	9	7.91	21	23.27	31	45.62	Level1
EOU1_T2_P1_AB_1	2	-0.767	Level1	8	6.33	20	19.31	30	39.68	Level1
EOU1_T3_P2_AB_1	3	-0.7508	Level2	7	6.21	19	19	29	39.21	Level1
EOU1_T1_P3_1	4	-0.4911	Level2	8	4.66	18	14.32	28	31.94	Level1
EOU1_T2_P3_1	5	-0.4445	Level1	9	4.42	17	13.53	27	30.68	Level1
EOU1_T1_P3_2	6	-0.3295	Level2	8	3.96	16	11.69	26	27.69	Level1
EOU1_T1_P1_2	7	-0.2895	Level2	9	3.84	15	11.09	25	26.69	Level1
EOU1_T1_P4_1	8	-0.2285	Level1	10	3.72	14	10.24	24	25.22	Level1
EOU1_T3_P1_1	9	-0.1923	Level1	9	3.68	13	9.77	23	24.39	Level1
EOU1_T1_P2_1	10	-0.1704	Level1	8	3.68	12	9.5	22	23.91	Level1
EOU1_T2_P1_AB_2	11	0.0115	Level2	7	3.87	11	7.5	21	20.09	Level2
EOU1_T1_P3_3	12	0.0803	Level3	8	4	10	6.82	20	18.71	Level2
EOU1_T3_P2_AB_2	13	0.172	Level2	9	4.28	11	5.99	19	16.97	Level2
EOU1_T3_P1_2	14	0.1784	Level2	10	4.3	10	5.94	18	16.86	Level2
EOU1_T3_P3_1	15	0.1813	Level1	11	4.32	9	5.92	17	16.81	Level2
EOU1_T2_P3_2	16	0.3046	Level2	10	5.06	8	5.18	16	14.83	Level2
EOU1_T1_P4_2	17	0.4073	Level3	11	5.78	7	4.67	15	13.29	Level2

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU1_T3_P1_3	18	0.5786	Level2	12	7.15	8	3.98	14	10.9	Level2
EOU1_T2_P2_1	19	0.6764	Level1	13	8.03	7	3.69	13	9.62	Level2
EOU1_T1_P1_3	20	0.7564	Level3	12	8.83	6	3.53	12	8.66	Level3
EOU1_T2_P1_AB_3	21	0.8024	Level3	13	9.33	7	3.48	11	8.16	Level3
EOU1_T1_P1_4	22	1.0839	Level3	14	12.71	8	3.48	10	5.34	Level3
EOU1_T2_P1_C_1	23	1.1366	Level1	15	13.4	9	3.53	9	4.87	Level3
EOU1_T2_P2_2	24	1.1668	Level2	14	13.82	8	3.59	8	4.63	Level3
EOU1_T3_P1_4	25	1.3232	Level3	15	16.17	7	4.06	7	3.53	Level3
EOU1_T1_P2_2	26	1.366	Level2	16	16.85	8	4.23	6	3.28	Level3
EOU1_T1_P4_3	27	1.6712	Level4	17	22.04	7	5.76	5	1.75	Level3
EOU1_T3_P2_AB_3	28	1.695	Level3	18	22.47	8	5.9	6	1.65	Level3
EOU1_T2_P1_AB_4	29	1.7091	Level3	19	22.74	9	6	5	1.61	Level3
EOU1_T2_P3_3	30	1.8326	Level4	20	25.21	10	6.99	4	1.37	Level3
EOU1_T1_P2_3	31	1.836	Level4	21	25.28	11	7.02	5	1.36	Level3
EOU1_T3_P3_2	32	1.8575	Level2	22	25.75	12	7.24	6	1.36	Level3
EOU1_T3_P3_3	33	2.4022	Level3	23	38.28	11	13.23	5	1.91	Level3
EOU1_T2_P1_C_2	34	2.4423	Level3	24	39.24	12	13.71	4	1.99	Level3
EOU1_T1_P2_4	35	2.6621	Level4	25	44.74	13	16.57	3	2.65	Level4
EOU1_T2_P1_C_3	36	3.3515	Level4	26	62.66	14	26.22	4	5.4	Level4
EOU1_T3_P3_4	37	3.3928	Level4	27	63.78	15	26.84	5	5.61	Level4

Table 2. Detailed ESS Prompt Maps: Grade 5 EOU2

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU2_T3_P2_B_1	1	-3.5073	Level1	13	28.4	23	62.18	30	92.19	Level1
EOU2_T3_P2_A_1	2	-1.5532	Level1	12	4.95	22	19.19	29	35.53	Level1
EOU2_T2_P3_A_1	3	-1.4474	Level1	11	3.79	21	16.97	28	32.56	Level1
EOU2_T1_P1_1	4	-1.4441	Level1	10	3.76	20	16.9	27	32.47	Level1
EOU2_T1_P3_1	5	-1.3965	Level1	9	3.33	19	16	26	31.24	Level1
EOU2_T1_P2_1	6	-1.2073	Level1	8	1.81	18	12.59	25	26.51	Level1
EOU2_T2_P3_B_1	7	-1.1397	Level1	7	1.34	17	11.44	24	24.88	Level1
EOU2_T3_P1_1	8	-1.1312	Level1	6	1.29	16	11.31	23	24.69	Level1
EOU2_T2_P1_A_1	9	-1.0714	Level1	5	0.99	15	10.41	22	23.37	Level1
EOU2_T2_P3_A_2	10	-1.034	Level2	4	0.84	14	9.89	21	22.59	Level1
EOU2_T3_P3_1	11	-1.0147	Level2	5	0.78	13	9.64	20	22.2	Level1
EOU2_T2_P2_1	12	-0.9263	Level1	6	0.61	12	8.57	19	20.52	Level1
EOU2_T2_P1_B_1	13	-0.8562	Level1	5	0.54	11	7.8	18	19.26	Level1
EOU2_T1_P2_2	14	-0.8273	Level2	4	0.54	10	7.51	17	18.77	Level1
EOU2_T2_P1_A_2	15	-0.767	Level1	5	0.6	9	6.97	16	17.8	Level1
EOU2_T3_P2_A_2	16	-0.7444	Level1	4	0.64	8	6.79	15	17.47	Level1
EOU2_T2_P1_A_3	17	-0.5588	Level3	3	1.2	7	5.49	14	14.87	Level2
EOU2_T3_P2_A_3	18	-0.3178	Level3	4	2.16	8	4.05	13	11.73	Level2
EOU2_T1_P3_2	19	-0.1385	Level2	5	3.06	9	3.15	12	9.58	Level2

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU2_T3_P3_2	20	-0.0638	Level2	6	3.51	8	2.85	11	8.76	Level2
EOU2_T2_P1_A_4	21	-0.0617	Level4	7	3.52	7	2.84	10	8.74	Level2
EOU2_T3_P2_B_2	22	-0.0408	Level2	8	3.69	8	2.8	11	8.55	Level2
EOU2_T1_P1_2	23	0.0775	Level2	9	4.75	7	2.68	10	7.61	Level2
EOU2_T2_P3_B_2	24	0.1519	Level2	10	5.5	6	2.68	9	7.08	Level2
EOU2_T3_P1_2	25	0.417	Level2	11	8.41	5	2.95	8	5.49	Level2
EOU2_T2_P1_B_2	26	0.4675	Level3	12	9.02	4	3.05	7	5.24	Level3
EOU2_T1_P1_3	27	0.6895	Level3	13	11.91	5	3.72	6	4.35	Level3
EOU2_T3_P3_3	28	0.8154	Level3	14	13.67	6	4.22	5	3.98	Level3
EOU2_T3_P2_A_4	29	1.0401	Level4	15	17.04	7	5.34	4	3.53	Level3
EOU2_T2_P2_2	30	1.1772	Level2	16	19.23	8	6.17	5	3.39	Level3
EOU2_T2_P3_B_3	31	1.4645	Level4	17	24.12	7	8.18	4	3.39	Level4
EOU2_T2_P1_B_3	32	1.7544	Level4	18	29.34	8	10.5	5	3.68	Level4
EOU2_T1_P1_4	33	1.808	Level3	19	30.35	9	10.98	6	3.79	Level4
EOU2_T1_P3_3	34	1.8694	Level4	20	31.58	10	11.59	5	3.97	Level4
EOU2_T3_P1_3	35	2.1517	Level4	21	37.51	11	14.7	6	5.1	Level4
EOU2_T3_P3_4	36	2.4242	Level4	22	43.51	12	17.97	7	6.46	Level4
EOU2_T2_P2_3	37	2.5597	Level3	23	46.62	13	19.73	8	7.28	Level4

Table 3. Detailed ESS Prompt Maps: Grade 5 EOU3

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU3_T2_P1_A_1	1	-3.1219	Level1	9	16.37	30	75.07	49	144.32	Level1
EOU3_T1_P1_AD_1	2	-1.9538	Level1	8	7.03	29	41.19	48	88.25	Level1
EOU3_T2_P1_A_2	3	-1.783	Level1	7	5.83	28	36.41	47	80.22	Level1
EOU3_T1_P3_A_1	4	-1.5545	Level1	6	4.46	27	30.24	46	69.71	Level1
EOU3_T1_P2_B_1	5	-1.5002	Level2	5	4.19	26	28.83	45	67.27	Level2
EOU3_T1_P3_B_1	6	-1.4061	Level3	6	3.82	25	26.48	44	63.12	Level2
EOU3_T1_P1_AD_2	7	-1.2928	Level2	7	3.48	26	23.76	43	58.25	Level2
EOU3_T2_P4_A_1	8	-1.2673	Level1	8	3.42	25	23.17	42	57.18	Level2
EOU3_T3_P1_A_1	9	-1.2061	Level1	7	3.36	24	21.83	41	54.67	Level2
EOU3_T3_P1_B_1	10	-1.0733	Level2	6	3.36	23	19.04	40	49.36	Level2
EOU3_T1_P1_E_1	11	-1.0475	Level3	7	3.39	22	18.52	39	48.35	Level2
EOU3_T1_P2_A_1	12	-0.9377	Level1	8	3.61	23	16.43	38	44.18	Level2
EOU3_T3_P2_AB_1	13	-0.9131	Level3	7	3.68	22	15.99	37	43.27	Level2
EOU3_T2_P1_BC_1	14	-0.8098	Level2	8	4.1	23	14.24	36	39.55	Level2
EOU3_T1_P3_A_2	15	-0.7978	Level2	9	4.16	22	14.04	35	39.13	Level2
EOU3_T1_P2_A_2	16	-0.7926	Level2	10	4.19	21	13.97	34	38.96	Level2
EOU3_T3_P3_C_1	17	-0.6428	Level2	11	5.24	20	11.87	33	34.01	Level2
EOU3_T3_P3_AB_1	18	-0.6042	Level2	12	5.54	19	11.37	32	32.78	Level2
EOU3_T1_P2_A_3	19	-0.5198	Level2	13	6.3	18	10.35	31	30.16	Level2

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU3_T2_P1_A_3	20	-0.4815	Level2	14	6.69	17	9.93	30	29.01	Level2
EOU3_T1_P2_C_1	21	-0.4645	Level2	15	6.87	16	9.76	29	28.52	Level2
EOU3_T1_P1_AD_3	22	-0.4332	Level2	16	7.25	15	9.48	28	27.64	Level2
EOU3_T2_P2_1	23	-0.4296	Level2	17	7.3	14	9.45	27	27.55	Level2
EOU3_T2_P4_A_2	24	-0.4228	Level2	18	7.39	13	9.4	26	27.37	Level2
EOU3_T1_P3_C_1	25	-0.3073	Level3	19	9.12	12	8.71	25	24.48	Level2
EOU3_T2_P3_B_1	26	-0.2241	Level1	20	10.46	13	8.3	24	22.48	Level2
EOU3_T2_P4_BC_1	27	-0.1928	Level2	19	10.99	12	8.17	23	21.76	Level2
EOU3_T2_P2_2	28	-0.0616	Level2	20	13.35	11	7.78	22	18.88	Level2
EOU3_T1_P3_A_3	29	0.2287	Level3	21	18.86	10	7.2	21	12.78	Level2
EOU3_T2_P1_BC_2	30	0.2512	Level2	22	19.31	11	7.17	20	12.33	Level2
EOU3_T2_P2_3	31	0.2763	Level3	23	19.84	10	7.17	19	11.86	Level3
EOU3_T2_P3_A_1	32	0.3258	Level1	24	20.93	11	7.22	18	10.96	Level3
EOU3_T1_P1_E_2	33	0.3969	Level3	23	22.57	10	7.37	17	9.76	Level3
EOU3_T3_P2_AB_2	34	0.4502	Level3	24	23.85	11	7.53	16	8.9	Level3
EOU3_T3_P3_AB_2	35	0.4844	Level3	25	24.7	12	7.66	15	8.39	Level3
EOU3_T2_P3_A_2	36	0.5066	Level2	26	25.28	13	7.77	14	8.08	Level3
EOU3_T3_P3_C_2	37	0.5432	Level4	27	26.27	12	7.99	13	7.6	Level3
EOU3_T2_P4_BC_2	38	0.614	Level3	28	28.25	13	8.49	14	6.75	Level3
EOU3_T3_P4_AB_1	39	0.6614	Level3	29	29.62	14	8.87	13	6.23	Level3

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU3_T3_P1_B_2	40	0.7382	Level3	30	31.93	15	9.56	12	5.46	Level3
EOU3_T2_P3_B_2	41	0.7444	Level3	31	32.12	16	9.62	11	5.41	Level3
EOU3_T2_P4_A_3	42	0.89	Level2	32	36.78	17	11.22	10	4.24	Level3
EOU3_T1_P1_AD_4	43	0.9124	Level2	33	37.52	16	11.49	9	4.09	Level3
EOU3_T1_P3_B_2	44	0.9342	Level3	34	38.26	15	11.77	8	3.96	Level3
EOU3_T2_P2_4	45	1.0131	Level3	35	41.02	16	12.88	7	3.56	Level3
EOU3_T2_P3_A_3	46	1.0937	Level2	36	43.92	17	14.09	6	3.24	Level3
EOU3_T1_P2_C_2	47	1.414	Level3	37	55.77	16	19.21	5	2.28	Level3
EOU3_T1_P3_C_2	48	1.4831	Level3	38	58.4	17	20.39	4	2.14	Level3
EOU3_T3_P4_AB_2	49	1.6766	Level4	39	65.94	18	23.87	3	1.95	Level4
EOU3_T2_P4_BC_3	50	1.7224	Level4	40	67.78	19	24.74	4	1.95	Level4
EOU3_T1_P1_E_3	51	2.0778	Level3	41	82.35	20	31.85	5	2.3	Level4
EOU3_T2_P3_B_3	52	2.0884	Level3	42	82.79	21	32.07	4	2.32	Level4
EOU3_T3_P1_B_3	53	2.7857	Level4	43	112.78	22	47.41	3	4.41	Level4
EOU3_T1_P2_C_3	54	3.0784	Level4	44	125.66	23	54.14	4	5.59	Level4

Table 4. Detailed ESS Prompt Maps: Grade 5 EOU4

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU4_T3_P2_AB_1	1	-0.9976	Level2	9	8.16	19	21	32	46.85	Level1
EOU4_T3_P2_C_1	2	-0.8442	Level2	10	6.93	18	18.24	31	42.09	Level1
EOU4_T1_P1_A_1	3	-0.832	Level1	11	6.85	17	18.03	30	41.73	Level1
EOU4_T1_P1_B_1	4	-0.5065	Level2	10	4.9	16	12.83	29	32.29	Level1
EOU4_T2_P2_BC_1	5	-0.4733	Level3	11	4.73	15	12.33	28	31.36	Level1
EOU4_T1_P1_D_1	6	-0.4377	Level1	12	4.59	16	11.83	27	30.4	Level1
EOU4_T1_P1_A_2	7	-0.4272	Level1	11	4.56	15	11.69	26	30.12	Level1
EOU4_T2_P1_1	8	-0.3655	Level1	10	4.43	14	10.95	25	28.58	Level1
EOU4_T1_P1_B_2	9	-0.2804	Level2	9	4.35	13	10.02	24	26.54	Level2
EOU4_T1_P1_C_1	10	-0.1717	Level4	10	4.35	12	8.93	23	24.04	Level2
EOU4_T2_P3_1	11	-0.0813	Level1	11	4.44	13	8.12	24	22.05	Level2
EOU4_T3_P2_AB_2	12	-0.0272	Level3	10	4.55	12	7.68	23	20.91	Level2
EOU4_T3_P1_B_1	13	0.0949	Level1	11	4.91	13	6.83	22	18.47	Level2
EOU4_T1_P1_B_3	14	0.1125	Level3	10	4.98	12	6.72	21	18.14	Level2
EOU4_T1_P2_1	15	0.196	Level2	11	5.4	13	6.31	20	16.64	Level2
EOU4_T1_P1_A_3	16	0.2469	Level2	12	5.71	12	6.1	19	15.77	Level2
EOU4_T2_P2_A_1	17	0.3698	Level1	13	6.57	11	5.73	18	13.8	Level2
EOU4_T3_P1_A_1	18	0.3958	Level1	12	6.77	10	5.68	17	13.41	Level2
EOU4_T3_P3_1	19	0.4586	Level2	11	7.34	9	5.62	16	12.53	Level2

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU4_T2_P2_A_2	20	0.4669	Level1	12	7.42	8	5.62	15	12.43	Level2
EOU4_T3_P2_AB_3	21	0.4862	Level3	11	7.64	7	5.64	14	12.2	Level3
EOU4_T2_P2_BC_2	22	0.5302	Level3	12	8.16	8	5.73	13	11.71	Level3
EOU4_T3_P2_C_2	23	0.7603	Level3	13	11.15	9	6.42	12	9.41	Level3
EOU4_T1_P2_2	24	0.9148	Level3	14	13.32	10	7.03	11	8.02	Level3
EOU4_T1_P1_B_4	25	0.9843	Level4	15	14.36	11	7.38	10	7.46	Level3
EOU4_T1_P1_D_2	26	1.1278	Level3	16	16.66	12	8.24	11	6.46	Level3
EOU4_T3_P2_AB_4	27	1.2265	Level4	17	18.33	13	8.93	10	5.87	Level3
EOU4_T1_P1_C_2	28	1.2843	Level4	18	19.37	14	9.4	11	5.58	Level3
EOU4_T3_P1_B_2	29	1.2861	Level2	19	19.41	15	9.41	12	5.57	Level3
EOU4_T2_P3_2	30	1.3105	Level3	20	19.9	14	9.66	11	5.5	Level3
EOU4_T2_P2_BC_3	31	1.3216	Level4	21	20.13	15	9.78	10	5.48	Level3
EOU4_T3_P1_A_2	32	1.4296	Level2	22	22.51	16	11.07	11	5.37	Level3
EOU4_T2_P1_2	33	1.5949	Level3	23	26.31	15	13.22	10	5.37	Level3
EOU4_T3_P3_2	34	1.8766	Level2	24	33.07	16	17.17	9	5.65	Level3
EOU4_T2_P2_BC_4	35	1.8883	Level4	25	33.36	15	17.34	8	5.67	Level3
EOU4_T3_P1_B_3	36	2.0041	Level3	26	36.37	16	19.2	9	6.02	Level3
EOU4_T1_P2_3	37	2.2329	Level3	27	42.55	17	23.09	8	6.94	Level3
EOU4_T2_P3_3	38	2.3039	Level3	28	44.54	18	24.36	7	7.29	Level3
EOU4_T1_P1_D_3	39	2.6185	Level4	29	53.66	19	30.34	6	9.18	Level4

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU4_T1_P2_4	40	4	Level4	30	95.11	20	57.97	7	18.85	Level4

Table 5. Detailed ESS Prompt Maps: Grade 8 EOU1

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU1_T2_P4_B_1	1	-1.2642	Level1	17	16.26	34	53.69	43	80.96	Level1
EOU1_T3_P1_C_1	2	-1.0873	Level1	16	13.43	33	47.85	42	73.53	Level1
EOU1_T1_P1_B_1	3	-0.9865	Level1	15	11.91	32	44.63	41	69.4	Level1
EOU1_T3_P4_1	4	-0.9023	Level1	14	10.73	31	42.02	40	66.03	Level1
EOU1_T2_P2_1	5	-0.7827	Level1	13	9.18	30	38.43	39	61.37	Level1
EOU1_T3_P3_1	6	-0.7739	Level1	12	9.07	29	38.18	38	61.03	Level1
EOU1_T1_P2_A_1	7	-0.6293	Level1	11	7.48	28	34.13	37	55.68	Level1
EOU1_T2_P4_A_1	8	-0.4018	Level1	10	5.21	27	27.98	36	47.49	Level1
EOU1_T3_P1_AB_1	9	-0.3655	Level1	9	4.88	26	27.04	35	46.22	Level1
EOU1_T1_P1_A_1	10	-0.3407	Level1	8	4.68	25	26.42	34	45.38	Level1
EOU1_T1_P3_AB_1	11	-0.2996	Level1	7	4.4	24	25.43	33	44.02	Level1
EOU1_T2_P2_2	12	-0.2749	Level2	6	4.25	23	24.87	32	43.23	Level2
EOU1_T1_P2_A_2	13	-0.2715	Level2	7	4.23	22	24.79	31	43.13	Level2
EOU1_T3_P3_2	14	-0.0994	Level2	8	3.54	21	21.18	30	37.96	Level2
EOU1_T1_P1_A_2	15	-0.0682	Level2	9	3.45	20	20.55	29	37.06	Level2
EOU1_T1_P3_AB_2	16	-0.0533	Level2	10	3.42	19	20.27	28	36.64	Level2
EOU1_T1_P1_C_1	17	-0.0485	Level1	11	3.41	18	20.18	27	36.51	Level2
EOU1_T2_P3_A_1	18	0.2455	Level1	10	3.41	17	15.19	26	28.87	Level2
EOU1_T1_P3_AB_3	19	0.2597	Level3	9	3.43	16	14.96	25	28.51	Level2

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU1_T3_P1_AB_2	20	0.2941	Level2	10	3.5	17	14.44	24	27.69	Level2
EOU1_T3_P2_1	21	0.3012	Level1	11	3.52	16	14.34	23	27.53	Level2
EOU1_T3_P4_2	22	0.3214	Level2	10	3.6	15	14.08	22	27.08	Level2
EOU1_T1_P1_A_3	23	0.3255	Level3	11	3.62	14	14.03	21	26.99	Level2
EOU1_T2_P3_A_2	24	0.3861	Level2	12	3.98	15	13.36	20	25.78	Level2
EOU1_T1_P2_B_1	25	0.5073	Level1	13	4.83	14	12.15	19	23.48	Level2
EOU1_T1_P1_B_2	26	0.6101	Level2	12	5.65	13	11.23	18	21.63	Level2
EOU1_T2_P3_B_1	27	0.7206	Level1	13	6.65	12	10.34	17	19.75	Level2
EOU1_T2_P2_3	28	0.8533	Level3	12	7.98	11	9.41	16	17.63	Level2
EOU1_T3_P1_C_2	29	0.8547	Level2	13	7.99	12	9.41	15	17.61	Level2
EOU1_T2_P1_1	30	0.8719	Level1	14	8.2	11	9.32	14	17.37	Level2
EOU1_T1_P1_A_4	31	1.0203	Level4	13	10.13	10	8.73	13	15.44	Level2
EOU1_T3_P3_3	32	1.2658	Level3	14	13.56	11	7.99	14	12.49	Level2
EOU1_T3_P2_2	33	1.3046	Level2	15	14.15	12	7.91	13	12.06	Level2
EOU1_T2_P3_B_2	34	1.3236	Level2	16	14.45	11	7.89	12	11.87	Level2
EOU1_T1_P1_C_2	35	1.4135	Level2	17	15.98	10	7.89	11	11.07	Level2
EOU1_T3_P4_3	36	1.8084	Level3	18	23.09	9	8.29	10	7.91	Level3
EOU1_T1_P1_B_3	37	1.878	Level3	19	24.41	10	8.43	9	7.42	Level3
EOU1_T1_P3_AB_4	38	1.9309	Level4	20	25.47	11	8.59	8	7.1	Level3
EOU1_T2_P3_B_3	39	2.278	Level3	21	32.76	12	9.97	9	5.37	Level3

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU1_T2_P4_B_2	40	2.3593	Level2	22	34.54	13	10.38	8	5.04	Level3
EOU1_T1_P2_B_2	41	2.561	Level2	23	39.18	12	11.59	7	4.44	Level3
EOU1_T2_P1_2	42	2.5731	Level2	24	39.47	11	11.68	6	4.41	Level3
EOU1_T2_P4_A_2	43	2.7105	Level2	25	42.91	10	12.77	5	4.27	Level3
EOU1_T3_P2_3	44	2.725	Level3	26	43.29	9	12.91	4	4.27	Level3
EOU1_T3_P3_4	45	4	Level4	27	77.71	10	25.66	3	5.55	Level4
EOU1_T2_P1_3	46	4.5	Level3	28	91.71	11	31.16	4	6.55	Level4

Table 6. Detailed ESS Prompt Maps: Grade 8 EOU2

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU2_T2_P2_B_1	1	-1.043	Level2	11	9.5	36	55.93	62	108.42	Level1
EOU2_T2_P1_C_1	2	-0.946	Level2	12	8.53	35	52.54	61	102.5	Level1
EOU2_T3_P4_ABC_1	3	-0.9208	Level2	13	8.31	34	51.68	60	100.99	Level1
EOU2_T3_P2_AB_1	4	-0.9202	Level1	14	8.3	33	51.66	59	100.96	Level1
EOU2_T1_P1_A_1	5	-0.7779	Level3	13	7.31	32	47.11	58	92.7	Level1
EOU2_T3_P1_B_1	6	-0.7722	Level1	14	7.27	33	46.93	57	92.38	Level1
EOU2_T1_P3_AB_1	7	-0.6051	Level2	13	6.44	32	41.92	56	83.02	Level1
EOU2_T1_P4_1	8	-0.508	Level1	14	6.05	31	39.1	55	77.68	Level1
EOU2_T3_P1_C_1	9	-0.5062	Level1	13	6.04	30	39.05	54	77.58	Level1
EOU2_T3_P1_A_1	10	-0.4395	Level1	12	5.91	29	37.25	53	74.05	Level1
EOU2_T1_P2_AB_1	11	-0.4366	Level2	11	5.91	28	37.17	52	73.9	Level2
EOU2_T3_P2_AB_2	12	-0.3706	Level2	12	5.91	27	35.52	51	70.53	Level2
EOU2_T2_P3_A_1	13	-0.3696	Level3	13	5.91	26	35.5	50	70.48	Level2
EOU2_T1_P1_BC_1	14	-0.3183	Level3	14	6.01	27	34.32	49	67.97	Level2
EOU2_T3_P1_B_2	15	-0.3139	Level2	15	6.02	28	34.22	48	67.76	Level2
EOU2_T2_P1_D_1	16	-0.3031	Level1	16	6.07	27	34	47	67.25	Level2
EOU2_T2_P2_A_1	17	-0.3031	Level1	16	6.07	27	34	47	67.25	Level2
EOU2_T2_P3_B_1	18	-0.2289	Level2	14	6.51	25	32.59	45	63.91	Level2
EOU2_T1_P1_A_2	19	-0.028	Level3	15	7.92	24	28.97	44	55.07	Level2

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU2_T2_P2_B_2	20	-0.0167	Level3	16	8.01	25	28.78	43	54.58	Level2
EOU2_T2_P1_A_1	21	-0.0052	Level1	17	8.11	26	28.59	42	54.1	Level2
EOU2_T2_P2_C_1	22	0.1225	Level2	16	9.39	25	26.68	41	48.87	Level2
EOU2_T3_P3_A_1	23	0.1649	Level3	17	9.85	24	26.08	40	47.17	Level2
EOU2_T2_P3_A_2	24	0.2257	Level3	18	10.58	25	25.29	39	44.8	Level2
EOU2_T2_P2_A_2	25	0.2553	Level1	19	10.97	26	24.94	38	43.67	Level2
EOU2_T2_P1_D_2	26	0.2553	Level2	19	10.97	26	24.94	38	43.67	Level2
EOU2_T1_P3_AB_2	27	0.2983	Level3	19	11.61	24	24.51	36	42.13	Level2
EOU2_T3_P3_B_1	28	0.3225	Level2	20	12	25	24.29	35	41.28	Level2
EOU2_T3_P2_AB_3	29	0.3686	Level3	21	12.78	24	23.92	34	39.71	Level2
EOU2_T2_P3_C_1	30	0.3892	Level2	22	13.16	25	23.78	33	39.03	Level2
EOU2_T2_P1_B_1	31	0.3938	Level1	23	13.24	24	23.75	32	38.88	Level2
EOU2_T2_P3_B_2	32	0.4176	Level2	22	13.72	23	23.63	31	38.15	Level2
EOU2_T1_P3_C_1	33	0.5572	Level2	23	16.65	22	23.07	30	33.96	Level2
EOU2_T2_P2_B_3	34	0.6465	Level3	24	18.61	21	22.81	29	31.37	Level2
EOU2_T2_P3_B_3	35	0.7043	Level3	25	19.94	22	22.69	28	29.75	Level2
EOU2_T1_P4_2	36	0.7059	Level2	26	19.98	23	22.69	27	29.71	Level2
EOU2_T2_P1_C_2	37	0.7153	Level2	27	20.22	22	22.69	26	29.46	Level2
EOU2_T3_P1_C_2	38	0.7172	Level2	28	20.27	21	22.69	25	29.42	Level2
EOU2_T1_P2_AB_2	39	0.7397	Level3	29	20.87	20	22.73	24	28.88	Level3

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU2_T1_P1_BC_2	40	0.8075	Level3	30	22.77	21	22.94	23	27.32	Level3
EOU2_T3_P4_ABC_2	41	0.8204	Level3	31	23.15	22	22.99	22	27.03	Level3
EOU2_T2_P1_A_2	42	0.9614	Level2	32	27.38	23	23.69	21	24.07	Level3
EOU2_T3_P3_A_2	43	1.0048	Level3	33	28.72	22	23.96	20	23.2	Level3
EOU2_T2_P1_B_2	44	1.0906	Level2	34	31.47	23	24.56	19	21.57	Level3
EOU2_T3_P1_A_2	45	1.1283	Level2	35	32.71	22	24.86	18	20.89	Level3
EOU2_T2_P2_A_3	46	1.1381	Level1	36	33.05	21	24.95	17	20.73	Level3
EOU2_T2_P1_D_3	47	1.1381	Level3	36	33.05	21	24.95	17	20.73	Level3
EOU2_T3_P3_B_2	48	1.2708	Level3	36	37.82	21	26.41	15	18.74	Level3
EOU2_T3_P1_B_3	49	1.4573	Level2	37	44.72	22	28.64	14	16.13	Level3
EOU2_T1_P4_3	50	1.6444	Level3	38	51.83	21	31.08	13	13.69	Level3
EOU2_T1_P3_AB_3	51	1.6541	Level3	39	52.21	22	31.21	12	13.58	Level3
EOU2_T2_P3_B_4	52	1.6655	Level3	40	52.67	23	31.38	11	13.45	Level3
EOU2_T1_P4_4	53	1.7772	Level3	41	57.25	24	33.17	10	12.33	Level3
EOU2_T3_P3_A_3	54	1.8145	Level3	42	58.81	25	33.8	9	12	Level3
EOU2_T1_P1_BC_3	55	1.8479	Level3	43	60.25	26	34.41	8	11.73	Level3
EOU2_T2_P2_C_2	56	1.8941	Level3	44	62.28	27	35.28	7	11.41	Level3
EOU2_T2_P3_C_2	57	1.8956	Level3	45	62.35	28	35.31	6	11.4	Level3
EOU2_T1_P2_AB_3	58	2.5317	Level4	46	91.61	29	48.67	5	8.22	Level4
EOU2_T3_P4_ABC_3	59	2.7703	Level4	47	102.82	30	53.92	6	7.26	Level4

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU2_T1_P3_C_2	60	2.8776	Level2	48	107.98	31	56.39	7	6.94	Level4
EOU2_T2_P1_C_3	61	4.5	Level2	49	187.47	30	95.33	6	3.7	Level4
EOU2_T3_P1_B_4	62	4.5	Level2	49	187.47	30	95.33	6	3.7	Level4
EOU2_T3_P1_C_3	63	4.5	Level2	49	187.47	30	95.33	6	3.7	Level4
EOU2_T3_P3_A_4	64	4.5	Level3	49	187.47	30	95.33	6	3.7	Level4
EOU2_T3_P4_ABC_4	65	4.5	Level4	49	187.47	30	95.33	6	3.7	Level4

Table 7. Detailed ESS Prompt Maps: Grade 8 EOU3

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU3_T2_P1_AB_1	1	-1.6536	Level1	4	2.63	14	11.94	23	34.66	Level1
EOU3_T2_P2_1	2	-1.367	Level2	3	1.77	13	8.22	22	28.36	Level2
EOU3_T1_P1_A_1	3	-1.3546	Level2	4	1.75	12	8.07	21	28.09	Level2
EOU3_T3_P1_AB_1	4	-1.2122	Level2	5	1.61	11	6.5	20	25.25	Level2
EOU3_T1_P2_AB_1	5	-1.1651	Level1	6	1.61	10	6.03	19	24.35	Level2
EOU3_T2_P1_AB_2	6	-1.002	Level2	5	1.77	9	4.56	18	21.42	Level2
EOU3_T1_P2_C_1	7	-0.8168	Level3	6	2.14	8	3.08	17	18.27	Level2
EOU3_T2_P3_1	8	-0.7025	Level2	7	2.48	9	2.28	16	16.44	Level2
EOU3_T1_P1_B_1	9	-0.6977	Level1	8	2.5	8	2.25	15	16.37	Level2
EOU3_T1_P1_A_2	10	-0.6304	Level2	7	2.84	7	1.92	14	15.42	Level2
EOU3_T3_P1_AB_2	11	-0.5441	Level2	8	3.36	6	1.57	13	14.3	Level2
EOU3_T1_P2_AB_2	12	-0.4923	Level2	9	3.72	5	1.42	12	13.68	Level2
EOU3_T3_P2_AB_1	13	-0.4644	Level1	10	3.94	4	1.36	11	13.37	Level2
EOU3_T2_P1_AB_3	14	-0.1578	Level4	9	6.7	3	1.05	10	10.31	Level3
EOU3_T2_P2_2	15	-0.0609	Level2	10	7.67	4	1.05	11	9.44	Level3
EOU3_T3_P1_AB_3	16	0.0316	Level4	11	8.69	3	1.15	10	8.7	Level3
EOU3_T3_P3_AB_1	17	0.1405	Level2	12	9.99	4	1.36	11	7.93	Level3
EOU3_T2_P3_2	18	0.1958	Level3	13	10.71	3	1.53	10	7.6	Level3
EOU3_T1_P2_C_2	19	0.3044	Level3	14	12.23	4	1.96	9	7.06	Level3

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU3_T2_P1_AB_4	20	0.6738	Level4	15	17.78	5	3.81	8	5.58	Level3
EOU3_T1_P1_B_2	21	0.9494	Level3	16	22.18	6	5.47	9	4.75	Level3
EOU3_T1_P2_AB_3	22	1.0011	Level3	17	23.06	7	5.83	8	4.65	Level3
EOU3_T3_P3_AB_2	23	1.0806	Level3	18	24.49	8	6.46	7	4.57	Level3
EOU3_T3_P2_AB_2	24	1.1031	Level3	19	24.92	9	6.67	6	4.57	Level3
EOU3_T3_P1_AB_4	25	1.1459	Level4	20	25.78	10	7.09	5	4.61	Level4
EOU3_T2_P3_3	26	1.3441	Level4	21	29.94	11	9.27	6	5.01	Level4
EOU3_T1_P2_C_3	27	1.5952	Level4	22	35.46	12	12.29	7	5.76	Level4
EOU3_T1_P1_B_3	28	1.6837	Level4	23	37.5	13	13.44	8	6.12	Level4
EOU3_T3_P3_AB_3	29	1.8173	Level3	24	40.71	14	15.31	9	6.79	Level4
EOU3_T3_P2_AB_3	30	2.199	Level3	25	50.25	15	21.03	8	9.08	Level4

Table 8. Detailed ESS Prompt Maps: Grade 8 EOU4

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU4_T1_P2_A_1	1	-1.7611	Level1	2	2.37	19	33.06	37	92.13	Level1
EOU4_T2_P1_B_1	2	-1.0624	Level2	1	1.67	18	20.48	36	66.97	Level2
EOU4_T1_P3_A_1	3	-1.0076	Level2	2	1.67	17	19.55	35	65.06	Level2
EOU4_T1_P3_A_2	4	-0.6512	Level2	3	2.03	16	13.85	34	52.94	Level2
EOU4_T3_P2_AB_1	5	-0.5928	Level2	4	2.15	15	12.97	33	51.01	Level2
EOU4_T1_P2_A_2	6	-0.4051	Level2	5	2.71	14	10.34	32	45	Level2
EOU4_T3_P3_B_1	7	-0.2989	Level2	6	3.13	13	8.96	31	41.71	Level2
EOU4_T2_P2_A_1	8	-0.2634	Level2	7	3.31	12	8.54	30	40.65	Level2
EOU4_T1_P2_B_1	9	-0.1292	Level2	8	4.12	11	7.06	29	36.76	Level2
EOU4_T3_P2_C_1	10	-0.1265	Level3	9	4.14	10	7.03	28	36.68	Level2
EOU4_T3_P3_C_1	11	-0.0389	Level2	10	4.84	11	6.24	27	34.31	Level2
EOU4_T1_P1_1	12	0.0708	Level2	11	5.82	10	5.37	26	31.46	Level2
EOU4_T2_P2_B_1	13	0.1026	Level2	12	6.14	9	5.14	25	30.67	Level2
EOU4_T3_P3_A_1	14	0.1241	Level3	13	6.38	8	5.01	24	30.15	Level2
EOU4_T1_P3_B_1	15	0.1646	Level2	14	6.86	9	4.81	23	29.22	Level2
EOU4_T1_P2_A_3	16	0.1705	Level2	15	6.94	8	4.79	22	29.09	Level2
EOU4_T2_P1_C_1	17	0.1991	Level2	16	7.34	7	4.7	21	28.49	Level2
EOU4_T2_P1_A_1	18	0.355	Level3	17	9.68	6	4.39	20	25.37	Level3
EOU4_T2_P2_A_2	19	0.4298	Level3	18	10.88	7	4.32	19	23.95	Level3

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU4_T3_P3_B_2	20	0.453	Level3	19	11.27	8	4.32	18	23.53	Level3
EOU4_T2_P2_C_1	21	0.611	Level1	20	14.12	9	4.47	17	20.85	Level3
EOU4_T3_P3_B_3	22	0.7642	Level4	19	17.03	8	4.78	16	18.4	Level3
EOU4_T2_P2_B_2	23	0.876	Level2	20	19.26	9	5.12	17	16.72	Level3
EOU4_T3_P1_1	24	1.1016	Level2	21	24	8	6.02	16	13.56	Level3
EOU4_T2_P1_C_2	25	1.2465	Level3	22	27.19	7	6.74	15	11.68	Level3
EOU4_T3_P3_C_2	26	1.2627	Level3	23	27.56	8	6.84	14	11.48	Level3
EOU4_T2_P2_A_3	27	1.296	Level3	24	28.36	9	7.07	13	11.12	Level3
EOU4_T2_P2_C_2	28	1.3264	Level3	25	29.12	10	7.32	12	10.81	Level3
EOU4_T2_P1_A_2	29	1.4184	Level3	26	31.51	11	8.14	11	9.98	Level3
EOU4_T2_P1_B_2	30	1.4543	Level3	27	32.48	12	8.5	10	9.7	Level3
EOU4_T3_P2_C_2	31	1.5162	Level4	28	34.21	13	9.18	9	9.26	Level3
EOU4_T3_P2_AB_2	32	1.5945	Level3	29	36.48	14	10.12	10	8.79	Level3
EOU4_T1_P3_B_2	33	1.8431	Level3	30	43.94	15	13.36	9	7.55	Level3
EOU4_T3_P1_2	34	2.282	Level3	31	57.55	16	19.5	8	5.79	Level3
EOU4_T2_P1_B_3	35	2.2993	Level3	32	58.1	17	19.76	7	5.74	Level3
EOU4_T1_P1_2	36	2.5095	Level2	33	65.04	18	23.12	6	5.32	Level3
EOU4_T3_P3_A_2	37	2.8709	Level3	34	77.33	17	29.27	5	4.96	Level3
EOU4_T1_P3_B_3	38	3.0557	Level4	35	83.79	18	32.59	4	4.96	Level4
EOU4_T1_P2_B_2	39	3.2413	Level3	36	90.48	19	36.12	5	5.15	Level4

ID	OOD	LOC	Aligned Level	Level 2		Level 3		Level 4		Empirical Level
				Count	Weight	Count	Weight	Count	Weight	
EOU4_T2_P2_C_3	40	4	Level3	37	118.55	20	51.29	4	6.66	Level4
EOU4_T3_P3_C_3	41	4	Level4	37	118.55	20	51.29	4	6.66	Level4

Appendix D: Rosters of Inconsistent and Essentially Consistent Prompts

Table 9. Roster of Inconsistent Prompts: Grade 5 EOU1

GCA	Prompt ID	OOD	SME- Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G5	EOU1_T3_P2_AB_1	3	Level2	Level1	1	-0.7623	0.762
G5	EOU1_T1_P3_1	4	Level2	Level1	1	-0.5026	0.503
G5	EOU1_T1_P3_2	6	Level2	Level1	1	-0.341	0.341
G5	EOU1_T1_P1_2	7	Level2	Level1	1	-0.301	0.301
G5	EOU1_T1_P3_3	12	Level3	Level2	1	-0.6761	0.676
G5	EOU1_T3_P3_1	15	Level1	Level2	-1	1.1698	1.17
G5	EOU1_T1_P4_2	17	Level3	Level2	1	-0.3491	0.349
G5	EOU1_T2_P2_1	19	Level1	Level2	-1	1.6649	1.665
G5	EOU1_T2_P1_C_1	23	Level1	Level3	-2	2.1251	2.125
G5	EOU1_T2_P2_2	24	Level2	Level3	-1	1.4104	1.41
G5	EOU1_T1_P2_2	26	Level2	Level3	-1	1.6096	1.61
G5	EOU1_T1_P4_3	27	Level4	Level3	1	-0.9909	0.991
G5	EOU1_T2_P3_3	30	Level4	Level3	1	-0.8295	0.83
G5	EOU1_T1_P2_3	31	Level4	Level3	1	-0.8261	0.826
G5	EOU1_T3_P3_2	32	Level2	Level3	-1	2.1011	2.101

Table 10. Roster of Essentially Consistent Prompts: Grade 5 EOU1

GCA	Prompt ID	OOD	SME- Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
R0	ES_GK_SA_MC74_L4B	93	Level3	Level2	1	-0.3023	0.302

Table 11. Roster of Inconsistent Prompts: Grade G5 EOU2

GCA	Prompt ID	OOD	SME- Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G5	EOU2_T2_P3_A_2	10	Level2	Level1	1	-0.4752	0.475
G5	EOU2_T3_P3_1	11	Level2	Level1	1	-0.4559	0.456
G5	EOU2_T2_P1_A_3	17	Level3	Level2	1	-1.0263	1.026
G5	EOU2_T3_P2_A_3	18	Level3	Level2	1	-0.7853	0.785
G5	EOU2_T2_P1_A_4	21	Level4	Level2	2	-1.5262	1.526
G5	EOU2_T3_P2_A_4	29	Level4	Level3	1	-0.4244	0.424
G5	EOU2_T2_P2_2	30	Level2	Level3	-1	1.7097	1.71
G5	EOU2_T1_P1_4	33	Level3	Level4	-1	1.3435	1.344
G5	EOU2_T2_P2_3	37	Level3	Level4	-1	2.0952	2.095

Table 12. Roster of Essentially Consistent Prompts: Grade 5 EOU2

GCA	Prompt ID	OOD	SME- Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G5	EOU2_T1_P2_2	14	Level2	Level1	1	-0.2685	0.269

Table 13. Roster of Inconsistent Prompts: Grade 5 EOU3

GCA	Prompt ID	OOD	SME- Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G5	EOU3_T1_P3_B_1	6	Level3	Level2	1	-1.6824	1.682
G5	EOU3_T2_P4_A_1	8	Level1	Level2	-1	1.2329	1.233
G5	EOU3_T3_P1_A_1	9	Level1	Level2	-1	1.2941	1.294
G5	EOU3_T1_P1_E_1	11	Level3	Level2	1	-1.3238	1.324
G5	EOU3_T1_P2_A_1	12	Level1	Level2	-1	1.5625	1.562
G5	EOU3_T3_P2_AB_1	13	Level3	Level2	1	-1.1894	1.189
G5	EOU3_T1_P3_C_1	25	Level3	Level2	1	-0.5836	0.584
G5	EOU3_T2_P3_B_1	26	Level1	Level2	-1	2.2761	2.276
G5	EOU3_T2_P3_A_1	32	Level1	Level3	-2	2.826	2.826
G5	EOU3_T2_P3_A_2	36	Level2	Level3	-1	1.2303	1.23
G5	EOU3_T3_P3_C_2	37	Level4	Level3	1	-1.1334	1.133
G5	EOU3_T2_P4_A_3	42	Level2	Level3	-1	1.6137	1.614
G5	EOU3_T1_P1_AD_4	43	Level2	Level3	-1	1.6361	1.636
G5	EOU3_T2_P3_A_3	46	Level2	Level3	-1	1.8174	1.817
G5	EOU3_T1_P1_E_3	51	Level3	Level4	-1	1.4012	1.401
G5	EOU3_T2_P3_B_3	52	Level3	Level4	-1	1.4118	1.412

Table 14. Roster of Essentially Consistent Prompts: Grade 5 EOU3

Grade/ Domain	Prompt ID	OOD	SME- Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G5	EOU3_T1_P3_A_3	29	Level3	Level2	1	-0.0476	0.048

Table 15. Roster of Inconsistent Prompts: Grade 5 EOU4

GCA	Prompt ID	OOD	SME-Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G5	EOU4_T3_P2_AB_1	1	Level2	Level1	1	-0.7172	0.717
G5	EOU4_T3_P2_C_1	2	Level2	Level1	1	-0.5638	0.564
G5	EOU4_T2_P2_BC_1	5	Level3	Level1	2	-0.9595	0.96
G5	EOU4_T1_P1_C_1	10	Level4	Level2	2	-2.7902	2.79
G5	EOU4_T2_P3_1	11	Level1	Level2	-1	1.1991	1.199
G5	EOU4_T3_P2_AB_2	12	Level3	Level2	1	-0.5134	0.513
G5	EOU4_T3_P1_B_1	13	Level1	Level2	-1	1.3753	1.375
G5	EOU4_T1_P1_B_3	14	Level3	Level2	1	-0.3737	0.374
G5	EOU4_T2_P2_A_1	17	Level1	Level2	-1	1.6502	1.65
G5	EOU4_T3_P1_A_1	18	Level1	Level2	-1	1.6762	1.676
G5	EOU4_T2_P2_A_2	20	Level1	Level2	-1	1.7473	1.747
G5	EOU4_T1_P1_B_4	25	Level4	Level3	1	-1.6342	1.634
G5	EOU4_T3_P2_AB_4	27	Level4	Level3	1	-1.392	1.392
G5	EOU4_T1_P1_C_2	28	Level4	Level3	1	-1.3342	1.334
G5	EOU4_T3_P1_B_2	29	Level2	Level3	-1	1.7999	1.8
G5	EOU4_T2_P2_BC_3	31	Level4	Level3	1	-1.2969	1.297
G5	EOU4_T3_P1_A_2	32	Level2	Level3	-1	1.9434	1.943
G5	EOU4_T3_P3_2	34	Level2	Level3	-1	2.3904	2.39
G5	EOU4_T2_P2_BC_4	35	Level4	Level3	1	-0.7302	0.73

Table 16. Roster of Essentially Consistent Prompts: Grade 5 EOU4

GCA	Prompt ID	OOD	SME-Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G5	EOU4_T1_P1_B_1	4	Level2	Level1	1	-0.2261	0.226

Table 17. Roster of Inconsistent Prompts: Grade 8 EOU1

GCA	Prompt ID	OOD	SME- Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G8	EOU1_T1_P1_C_1	17	Level1	Level2	-1	1.2264	1.226
G8	EOU1_T2_P3_A_1	18	Level1	Level2	-1	1.5204	1.52
G8	EOU1_T1_P3_AB_3	19	Level3	Level2	1	-1.5487	1.549
G8	EOU1_T3_P2_1	21	Level1	Level2	-1	1.5761	1.576
G8	EOU1_T1_P1_A_3	23	Level3	Level2	1	-1.4829	1.483
G8	EOU1_T1_P2_B_1	25	Level1	Level2	-1	1.7822	1.782
G8	EOU1_T2_P3_B_1	27	Level1	Level2	-1	1.9955	1.995
G8	EOU1_T2_P2_3	28	Level3	Level2	1	-0.9551	0.955
G8	EOU1_T2_P1_1	30	Level1	Level2	-1	2.1468	2.147
G8	EOU1_T1_P1_A_4	31	Level4	Level2	2	-2.9797	2.98
G8	EOU1_T3_P3_3	32	Level3	Level2	1	-0.5426	0.543
G8	EOU1_T1_P3_AB_4	38	Level4	Level3	1	-2.0691	2.069
G8	EOU1_T2_P4_B_2	40	Level2	Level3	-1	1.5509	1.551
G8	EOU1_T1_P2_B_2	41	Level2	Level3	-1	1.7526	1.753
G8	EOU1_T2_P1_2	42	Level2	Level3	-1	1.7647	1.765
G8	EOU1_T2_P4_A_2	43	Level2	Level3	-1	1.9021	1.902
G8	EOU1_T2_P1_3	46	Level3	Level4	-1	1.5	1.5

Table 18. Roster of Essentially Consistent Prompts: Grade 8 EOU1

GCA	Prompt ID	OOD	SME- Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G8	ES_G4_SA_MC39_L4A	93	Level3	Level2	1	-0.1687	0.169

Table 19. Roster of Inconsistent Prompts: Grade 8 EOU2

GCA	Prompt ID	OOD	SME- Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G8	EOU2_T2_P2_B_1	1	Level2	Level1	1	-0.6064	0.606
G8	EOU2_T2_P1_C_1	2	Level2	Level1	1	-0.5094	0.509
G8	EOU2_T3_P4_ABC_1	3	Level2	Level1	1	-0.4842	0.484
G8	EOU2_T1_P1_A_1	5	Level3	Level1	2	-1.5176	1.518
G8	EOU2_T2_P3_A_1	13	Level3	Level2	1	-1.1093	1.109
G8	EOU2_T1_P1_BC_1	14	Level3	Level2	1	-1.058	1.058
G8	EOU2_T2_P1_D_1	16	Level1	Level2	-1	1.1335	1.134
G8	EOU2_T2_P2_A_1	17	Level1	Level2	-1	1.1335	1.134
G8	EOU2_T1_P1_A_2	19	Level3	Level2	1	-0.7677	0.768
G8	EOU2_T2_P2_B_2	20	Level3	Level2	1	-0.7564	0.756
G8	EOU2_T2_P1_A_1	21	Level1	Level2	-1	1.4314	1.431
G8	EOU2_T3_P3_A_1	23	Level3	Level2	1	-0.5748	0.575
G8	EOU2_T2_P3_A_2	24	Level3	Level2	1	-0.514	0.514
G8	EOU2_T2_P2_A_2	25	Level1	Level2	-1	1.6919	1.692
G8	EOU2_T1_P3_AB_2	27	Level3	Level2	1	-0.4414	0.441
G8	EOU2_T3_P2_AB_3	29	Level3	Level2	1	-0.3711	0.371
G8	EOU2_T2_P1_B_1	31	Level1	Level2	-1	1.8304	1.83
G8	EOU2_T2_P1_A_2	42	Level2	Level3	-1	1.2217	1.222
G8	EOU2_T2_P1_B_2	44	Level2	Level3	-1	1.3509	1.351
G8	EOU2_T3_P1_A_2	45	Level2	Level3	-1	1.3886	1.389
G8	EOU2_T2_P2_A_3	46	Level1	Level3	-2	2.5747	2.575
G8	EOU2_T3_P1_B_3	49	Level2	Level3	-1	1.7176	1.718
G8	EOU2_T1_P3_C_2	60	Level2	Level4	-2	3.1379	3.138
G8	EOU2_T2_P1_C_3	61	Level2	Level4	-2	4.7603	4.76
G8	EOU2_T3_P1_B_4	62	Level2	Level4	-2	4.7603	4.76
G8	EOU2_T3_P1_C_3	63	Level2	Level4	-2	4.7603	4.76
G8	EOU2_T3_P3_A_4	64	Level3	Level4	-1	2.9683	2.968

Table 20. Roster of Essentially Consistent Prompts: Grade 8 EOU2

GCA	Prompt ID	OOD	SME-Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G8	EOU2_T1_P3_AB_1	7	Level2	Level1	1	-0.1685	0.168
G8	EOU2_T2_P2_B_3	34	Level3	Level2	1	-0.0932	0.093
G8	EOU2_T2_P3_B_3	35	Level3	Level2	1	-0.0354	0.035

Table 21. Roster of Inconsistent Prompts: Grade 8 EOU3

GCA	Prompt ID	OOD	SME-Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G8	EOU3_T1_P2_AB_1	5	Level1	Level2	-1	1.2019	1.202
G8	EOU3_T1_P2_C_1	7	Level3	Level2	1	-0.659	0.659
G8	EOU3_T1_P1_B_1	9	Level1	Level2	-1	1.6693	1.669
G8	EOU3_T3_P2_AB_1	13	Level1	Level2	-1	1.9026	1.903
G8	EOU3_T2_P1_AB_3	14	Level4	Level3	1	-1.3037	1.304
G8	EOU3_T2_P2_2	15	Level2	Level3	-1	1.0969	1.097
G8	EOU3_T3_P1_AB_3	16	Level4	Level3	1	-1.1143	1.114
G8	EOU3_T3_P3_AB_1	17	Level2	Level3	-1	1.2983	1.298
G8	EOU3_T2_P1_AB_4	20	Level4	Level3	1	-0.4721	0.472
G8	EOU3_T3_P3_AB_3	29	Level3	Level4	-1	1.6714	1.671
G8	EOU3_T3_P2_AB_3	30	Level3	Level4	-1	2.0531	2.053

Table 22. Roster of Essentially Consistent Prompts: Grade 8 EOU3

GCA	Prompt ID	OOD	SME-Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G8	ES_G6_PSG08_MC1_RI2	27	Level2	Level1	1	-0.0845	0.085

Table 23. Roster of Inconsistent Prompts: Grade 8 EOU4

GCA	Prompt ID	OOD	SME- Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G8	EOU4_T3_P2_C_1	10	Level3	Level2	1	-0.4815	0.482
G8	EOU4_T2_P2_C_1	21	Level1	Level3	-2	2.6734	2.673
G8	EOU4_T3_P3_B_3	22	Level4	Level3	1	-2.2915	2.291
G8	EOU4_T2_P2_B_2	23	Level2	Level3	-1	1.521	1.521
G8	EOU4_T3_P1_1	24	Level2	Level3	-1	1.7466	1.747
G8	EOU4_T3_P2_C_2	31	Level4	Level3	1	-1.5395	1.539
G8	EOU4_T1_P1_2	36	Level2	Level3	-1	3.1545	3.155
G8	EOU4_T1_P2_B_2	39	Level3	Level4	-1	1.1856	1.186
G8	EOU4_T2_P2_C_3	40	Level3	Level4	-1	1.9443	1.944

Table 24. Roster of Essentially Consistent Prompts: Grade 8 EOU4

GCA	Prompt ID	OOD	SME- Aligned Level	Empirical Level	Level Difference	Distance	Absolute Distance
G8	EOU4_T3_P3_A_1	14	Level3	Level2	1	-0.2309	0.231