Stackable, Instructionally-embedded, Portable Science (SIPS) Assessments Project

Pilot Study Technical Report

September 2023

SIPS Pilot Study 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.

Table of Contents

Section 1. Introduction 1

Section 2. EOU Assessments 3

2.1 Summary of Intended Use 3

2.2 EOU Assessment Overview 3

2.3 EOU Task Design 5

2.4 EOU Task Rubric Development 6

2.5 EOU Development Summary 7

Section 3. Research Questions 8

Section 4. Sample Acquisition 10

4.1 Recruitment 10

4.2 Expectations for Participating Educators 10

4.3 Timeline 11

Section 5. Test Administration 13

5.1 Educator Survey 13

5.2 Student EOU Packets 13

5.3 Educator Workbook 13

5.4 Educator Scoring 14

Section 6. Data Analysis 16

6.1 Overview of the Data 16

6.2 RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning? 17

6.2.1: Can the assessments be administered within a single class period? 17

6.2.2: Are there patterns in the prompts that students skip? 19

6.2.3: Can educators score student responses on the EOU assessments reliably? 20

6.2.4: Do the EOU tasks allow students to demonstrate the full range of NGSS performance expectations? 21

6.2.5: Is performance on the EOU assessments associated statistically with other indicators of student learning (e.g., opportunity to learn (OTL), curriculum, or student performance on subsequent end-of-year (EOY) science assessments)? 22

6.3 RQ2: How well do latent variable measurement models fit the empirical EOU assessment data? . 24

6.4 RQ3: Overall, what do the EOU assessment results tell us about students' science learning? 25

6.4.1: What do the EOU assessment results tell us about student learning in terms of variation across student groups? 26

6.4.2: What do the EOU assessment results tell us about student learning in terms of variation in performance across instructional programs, instructional units, and instructional unit sequences? 27

6.4.3: What do the EOU assessment results tell us about student learning in terms of changes across administrations (i.e., growth)? 29

6.5 Using Data for Revisions 30

6.6 Conclusion 31

Section 7. SIPS Standard Setting 32

7.1 Introduction 32

7.2 Embedded Standard Setting and Assessment System Coherence 33

7.3 Coordination of Embedded Standard Setting Iterative Processes 34

PLD Development 34

Prompt Development & Prompt-PLD Alignment 34

ESS analyses 34

Vertical Articulation 34

Technical Report & Peer Review Evidence 35

7.4 Key Results of the SIPS Standard Setting Processes 35

The Efficacy of SIPS SMEs’ Hypothesized Item-Task Alignments 36

Establishing ESS Empirical Prompt-PLD Alignments 36

Classification Agreement and Weighted Kappa 36

Kappa Interpretations 37

Grade 5 37

Grade 8 37

Vertical Articulation of the SIPS Cut Scores and Investigation of Two IRT Response Probabilities 38

Standard Setting Validity Evidence 46

Section 8. Conclusions and Recommendations 48

8.1 Section Overview 48

8.2: Discussion related to RQ1: To what degree do the EOU assessments, generally, provide evidence of students’ three-dimensional science learning? 48

8.3: Discussion related to RQ2: How well did latent variable measurement models fit the empirical EOU assessment data? 49

8.4: Discussion related to RQ3: Overall, what do the EOU assessment results tell us about students’ science learning? 49

Cross-EOU Growth 50

8.5: Discussion related to use and reporting of the EOU results 52
Performance Level 52
A PLD-based SIPS Score 55
SIPS Summative Reporting 56
Using EOU Results to Inform Subsequent Units of Instruction 58
8.6: Discussion Related to Using the Data for Revisions to Tasks and PLDs 58
Revisions to Tasks 58
Revisions to PLDs 58
8.7 Conclusion 59
References 60
Appendix A. Post-Administration Survey 62
Appendix B. Data Tables for RQ1 (Section 6.2.2 Analyses) 66
RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning? 66
6.2.2: Are there patterns in the prompts that students skip? 66
Appendix C. Data Tables for RQ1 (Section 6.2.3 Analyses) 77
RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning? 77
6.2.3: Can educators score student responses on the EOU assessments reliably? 77
Appendix D. Data Tables for RQ1 (Section 6.2.4 Analyses) 82
RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning? 82
6.2.4: Do the EOU tasks allow students to demonstrate their full range of NGSS performance expectations? 82
Appendix E. Data Tables for RQ1 (Section 6.2.5 Analyses) 87
RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning? 87
6.2.5: Is performance on the EOU assessments associated statistically with other indicators of student learning (e.g., opportunity to learn (OTL), curriculum, or student performance on subsequent end-of-year (EOY) science assessments)? 87
Appendix F. Data Tables for RQ2 (Section 6.3 Analyses) 103
RQ2. How well do latent variable measurement models fit the empirical EOU assessment data? 103
Appendix G. Data Tables for RQ3 (Section 6.4.1 Analyses) 104
RQ3. Overall, what do the EOU assessment results tell us about students' science learning? 104
6.4.1 What do the EOU assessment results tell us about student learning in terms of variation across student groups? 104

Appendix H. Data Tables for RQ3 (Section 6.4.2 Analyses) 132

RQ3. Overall, what do the EOU assessment results tell us about students' science learning? 132

6.4.2 What do the EOU assessment results tell us about student learning in terms of variation in performance across instructional programs, instructional units, and instructional unit sequences? 132

Appendix I. Data Tables for Section 6.5 (Using Data for Revisions).................... 142

List of Exhibits

Exhibit 1.Components of a SIPS EOU Assessment Task 6

Exhibit 2. Overall Pilot Timeline 11

Exhibit 3. Example Sequence of Activities for Window A. 12

Exhibit 4. Scoring Workshop Attendance 14

Exhibit 5. Number of Educators and Students Included in the Sample, by EOU Assessment 16

Exhibit 6. Grade 5 EOU Timing Data: Means and Standard Errors 18

Exhibit 7. Grade 8 Timing Data: Means and Standard Errors 18

Exhibit 8. Percent Missing by Prompts for Grade 5 19

Exhibit 9. Percent Missing by Prompts for Grade 8 20

Exhibit 10. Overall Agreement Among Educators by Prompt 20

Exhibit 11. Agreement with Expert Rater by Prompt 21

Exhibit 12. Distributions of Scores for Each Unit 22

Exhibit 13. Number of Concepts with Significant Differences Related to Instruction on That Concept 23

Exhibit 14. Characteristics of Participating Students by Grade Level and EOU* 26

Exhibit 15. Gender Differences for Each EOU 27

Exhibit 16. Average Scores on EOUs by Achievement Level 27

Exhibit 17. EOU Performance Differences Related to Curriculum Materials by Grade 28

Exhibit 18. Correlations between EOU Scores, Grade 5 (N = 235) 29

Exhibit 19. Average EOU Percent Correct, Grade 5 29

Exhibit 20. Correlations between EOU Scores, Grade 8 (N = 21) 30

Exhibit 21. Average Percent Correct for Each EOU Assessment 30

Exhibit 22. Grade 5 Overall Flags 31

Exhibit 23. Grade 8 Overall Flags 31

Exhibit 24. SIPS Embedded Standard Setting Iterative Processes 33

Exhibit 25. Correlation of SMEs' Prompt-PLD Aligned Performance Level Ordinality and IRT RP Location36

Exhibit 26. Kappa Interpretations 37

Exhibit 27. Agreement Rate and Weighted Kappa 38

Exhibit 28. Initial SIPS RP67 and RP50 Cut Scores Across EOUs for Grades 5 and 8 39

Exhibit 29. SIPS Grade 5 RP67 Initial Cut Score Impact Data 40

Exhibit 30. SIPS Grade 5 RP50 Initial Cut Score Impact Data 40

Exhibit 31. SIPS Grade 8 RP67 Initial Cut Score Impact Data 41

Exhibit 32. SIPS Grade 8 RP50 Initial Cut Score Impact Data 41

Exhibit 33. Vertically Articulated SIPS RP67 and RP50 Cut Scores 42

Exhibit 34. Vertical Articulation Adjustments to Cut Scores in Standard Error Units 43

Exhibit 35. SIPS Grade 5 RP67 Smoothed Cut Score Impact Data 44

Exhibit 36. SIPS Grade 5 RP50 Smoothed Cut Score Impact Data 44

Exhibit 37. SIPS Grade 8 RP67 Smoothed Cut Score Impact Data 45

Exhibit 38. SIPS Grade 8 RP50 Smoothed Cut Score Impact Data 45

Exhibit 39. EOU Performance Level Profiles, Grade 5 51

Exhibit 40. EOU Performance Level Profiles, Grade 8 51

Exhibit 41. ELPA21 Profiles of Proficiency 53

Exhibit 42. EOU Performance Level Profiles and Possible Summative Performance Level Based on Rubric, Grade 5 54

Exhibit 43. EOU Performance Level Profiles and Possible Summative Performance Level Based on Rubric, Grade 8 54

Exhibit 44. EOU Performance Level Profiles and Possible Summative Performance Levels Based on Rubric and Averages, Grade 5 57

Exhibit 45. EOU Performance Level Profiles and Possible Summative Performance Levels Based on Rubric and Averages, Grade 8 57

Section 1. Introduction

The Stackable, Instructionally-embedded, Portable Science (SIPS) Assessments project, funded in 2020 by the US Department of Education's Competitive Grants for State Assessments (CGSA) program from the Office of Elementary and Secondary Education, brings together six states, five organizations, and a panel of experts to address states' needs for large-scale science assessments and the needs of educators, parents, and students for resources that support science learning throughout the school year. With coherence as the guiding principle, SIPS has developed curricular frameworks at grades 5 and 8 that include instructional resources, instructionally-embedded assessments, and end-of-unit (EOU) assessments for four units at Grade 5 and Grade 8 based on the three dimensions (i.e., Disciplinary Core Ideas (DCI), Science and Engineering Practices (SEP), and Crosscutting Concepts (CCC)) of the National Research Council's A Framework for K12 Science Education (2012), hereafter referred to as *The Framework*, and bundles of the Next Generation Science Standards (NGSS Lead States, 2013) performance expectations (PE). The focus of this report is on the end-of-unit (EOU) assessments.

The SIPS project set out to build a bank of stackable, instructionally-embedded, portable science assessment tasks to measure students' learning and support science instruction. We use the term stackable to indicate that the assessments can be used together in different orders. They are instructionally-embedded in the sense that they can be integrated into existing instruction. They are portable as they can be used with a variety of curriculum and in a variety of instructional settings. The project focused on creating these assessments for use in grades 5 and 8 as proof of concept in grades often tested statewide. The pilot study, which is described in detail in the following sections of this report, was designed and implemented to explore the quality of the SIPS EOU science assessments.

The SIPS EOU assessments, given at the end of each of four instructional units, are meant to provide a summative characterization of student learning in the prior unit, as well as to inform instruction within the upcoming unit. The SIPS curriculum and assessment team collaborated extensively with educators in six partner states (Alabama, Alaska, Montana, Nebraska, New York, and Wyoming) to develop the EOU assessments, as well as the curricular framework, instructional resources, and associated instructionally-embedded formative assessments. The SIPS EOU assessments were designed to be administered in class by educators and subsequently scored by educators at roughly ten-week intervals, at the end of pre-specified instructional units. Grade 5 and grade 8 educators from the partner states were invited to administer one or more EOU assessments based on the instructional scope and sequence of their science curricula. Each SIPS EOU assessment is composed of three tasks, each containing a number of prompts (i.e., test items), that participating educators administered to a classroom of approximately 20 students and scored using a prespecified scoring rubric.

The SIPS EOU assessments were developed using a principled assessment design framework (Mislevy & Haertel, 2006; Mislevy, Haertel, Risconscente, Rutstein & Ziker, 2017). Their purpose is to measure well-defined science constructs based on a clearly articulated theory of learning that builds toward the achievement of rigorous college and workforce readiness standards based on the *Framework* and the NGSS for curricula and instruction in a coherent and balanced system.

This report is organized into eight sections. Section 2 describes the design and development of the EOU assessments. Section 3 offers a summary of the research objectives of the pilot study, along with a list of the focused questions that influenced the design of the study. Section 4 describes the sample of grade 5 and grade 8 students who participated in this study, some of whom completed all four of the EOU assessments administered throughout the 2022-23 academic year. The subset of educators who participated in scoring these EOU assessments is also described in this section. Section 5 provides further detail on the administration and scoring of the EOU assessments. Section 6 provides detailed descriptions of the statistical and psychometric analyses the project team conducted to answer each of the research questions described in Section 3. With an understanding and accounting of the grade 5 and grade 8 students' responses and a view of the distribution of those scores, Section 7 introduces the approach to describing students' performance at various levels of proficiency by developing performance level descriptors associated with students' achievement on each of the EOUs using an Embedded Standard Setting method (Lewis & Cook, 2020). The report concludes in Section 8 by summarizing the results of the pilot study and discussing the potential for their use in science classrooms. In this final section, the report also speculates on the feasibility of conducting a larger-scale study (e.g., a field study) to further our understanding of the relationship among NGSS-aligned science curricula, a burgeoning collection of instructional approaches, and a promising array of novel and innovative classroom-based science assessments.

Section 2. EOU Assessments

2.1 Summary of Intended Use

The end-of-unit assessments were designed to work with instructionally-embedded formative assessments (also developed as part of the SIPS project) to support educators in determining where students are in their learning. Evidence from this collection of assessments helps support instruction by providing guidance to educators on next steps that would be appropriate for their students which in turn supports students with their learning. The application of the rubrics for the EOU assessments were designed to not only provide scores for students but also to provide educators with explanations for how students obtained those scores.

A given EOU assessment collects evidence on student proficiency toward the NGSS performance expectations (PEs) that were the focus of the unit. The evidence can be combined with additional evidence collected through the instructionally-embedded assessments to provide educators with snapshots of what students know and are able to do. This can support educators in determining next steps for students. For example, using the patterns found in student responses, the educator may decide to revisit topics before moving on to the next unit, or emphasizing certain topics in the next unit. The educator may also use the information to differentiate between students and determine if different students need different supports in subsequent instruction.

Looking across the four units at each grade, the EOU assessments were designed as a way to meet states' need for an end-of-year summative assessment. Instead of one larger test administered at the end of the year, or a point in time distal to when teaching and learning takes place, the scores from the individualized EOU assessments administered on a quarterly basis can be used to provide evidence of students' progress toward achieving proficiency of the targeted grade-level performance expectations. While the EOU assessments do not cover all possible topics of instruction, they are designed to cover critical aspects of the PEs that are the focus of instruction. If the state science standards are aligned to the NGSS standards, then the EOU assessments are appropriate for measuring students' ability related to the state standards.

2.2 EOU Assessment Overview

The EOU assessments were designed using an evidence-centered design (ECD) approach (Mislevy & Riconscente, 2006). ECD is a principled assessment design (PAD) approach that focuses on addressing these three questions: 1) What constructs do we want to measure, 2) What evidence is needed to make inferences about students' ability related to those constructs, and 3) How can tasks be designed to collect the desired evidence?

The goal of the SIPS project was to provide resources to support the claim that students are able to demonstrate proficiency in integrating Scientific and Engineering Practices with important Disciplinary Core Ideas and Crosscutting Concepts to scientifically investigate and understand natural phenomena and solve important science and engineering design problems. To begin development, this overall claim needed to be further defined. The first step to defining expectations for students was to determine bundles of PEs that could be taught and measured together and would meaningfully represent the scope of an instructional unit. For both the NGSS grade 5 and the middle school standards, the set of PEs was clustered into four unit bundles (refer to the "Claim, Measurement Target, and PE Bundle" for each unit on the SIPS Website Resources Page).

Each EOU assessment measures the key knowledge, skills, and abilities as represented by a thorough unpacking of the PEs (see “Unpacking Tools” for each unit on the SIPS Website Resources Page) within the associated unit bundle. Each PE is a combination of three dimensions: the disciplinary core ideas (DCI), science and engineering practices (SEPs), and cross-cutting concepts (CCC). Each of these dimensions is not unique to a given PE (e.g., the same scientific practice appears in multiple PEs), but the PE uniquely defines one combination of the three dimensions. When educators are teaching, they may decide to focus on the dimensions as combined in a given PE or they may decide to mix and match dimensions (or have students engage with only one or two dimensions at a given point in instruction). One of the first decisions that SIPS team needed to make was to determine how much variation in the combination of dimensions would be included in the EOU assessment tasks. The SIPS team, with input from the state partners, decided that students should be able to flexibly apply knowledge through the integration of the same combinations of dimensions within the PEs from the unit bundle, in the context of a phenomenon or phenomenon-rooted design problem based on the focal DCIs; and flexibly apply knowledge through the integration of new/different combinations of the dimensions represented by the PEs in the unit bundle, in the context of a phenomenon or phenomenon-rooted design problem based on the focal DCIs. Therefore, while a task on the EOU assessment may require students to apply a practice from one PE in the bundle with the core idea from another PE in the bundle, students would not be expected to engage with practices, disciplinary core ideas, or cross-cutting concepts that are not included in at least one of the PEs in the bundle.

As a key early step in the ECD process, the SIPS team collaborated with state partners to develop a set of performance level descriptors (PLDs). These descriptors organized multi-dimensional statements into levels representing different levels of student performance. The PLDs provide statements that are at a finer grain size than the overall claim and provide further insight into what is to be measured on the assessment.

Once the PLDs were developed, the SIPS team created design patterns. A design pattern was developed for each PE in the unit bundle. Each design pattern provides specification for the following:

- Knowledge, Skills, and Abilities (KSAs): Measurable statements that further specify how students engage with the PE;
- Student Demonstration of Learning: Expectations of students in relation to the KSAs;
- Work Product: A set of possible types of responses that students would produce when engaging with the KSAs;
- Task Features: Aspects that all tasks must have when measuring the PE;
- Variable Features: Aspects that can vary across tasks highlighting decisions that task developers must make when designing tasks;
- Assessment Boundaries: Clarifications on what is out of scope for the PE; and
- Technical Terms: Scientific terminology that is essential to the PE.

These design patterns provide a menu of options that task developers can use when designing tasks aligned to the PEs.

The design patterns and PLD documents provide guidance on what should be measured, as the PLD statements and the KSAs describe the concepts to measure that relate to the bundle of PEs. The design patterns also provided information on what evidence is needed to measure these concepts (through the demonstration of learning). Once the SIPS team established the design patterns, the next step was to determine how to measure these concepts. (Refer to the "Policy and Range Performance Level Descriptors" and "Design Patterns" for each unit on the SIPS Website Resources Page).

The EOU assessments had constraints on their design; specifically, they needed to be able to be completed in approximately one class period, and they needed to be administered as paper/pencil tasks. Keeping these constraints in mind, the SIPS team determined that each EOU assessment would consist of three tasks, each task using one scenario and/or phenomenon and a set of questions related to that phenomenon. Another critical design feature for measuring three-dimensional science standards is to engage students in a chain of sense-making. Therefore, the set of prompts within each task requires students to engage with different aspects of the scenario and increase in complexity with regard to the required response production. The SIPS team anticipated that each individual task would take students 10 to 15 minutes to complete, and consequently, determined that each EOU assessment would consist of three tasks. Further discussion of the task design is provided in the next section.

2.3 EOU Task Design

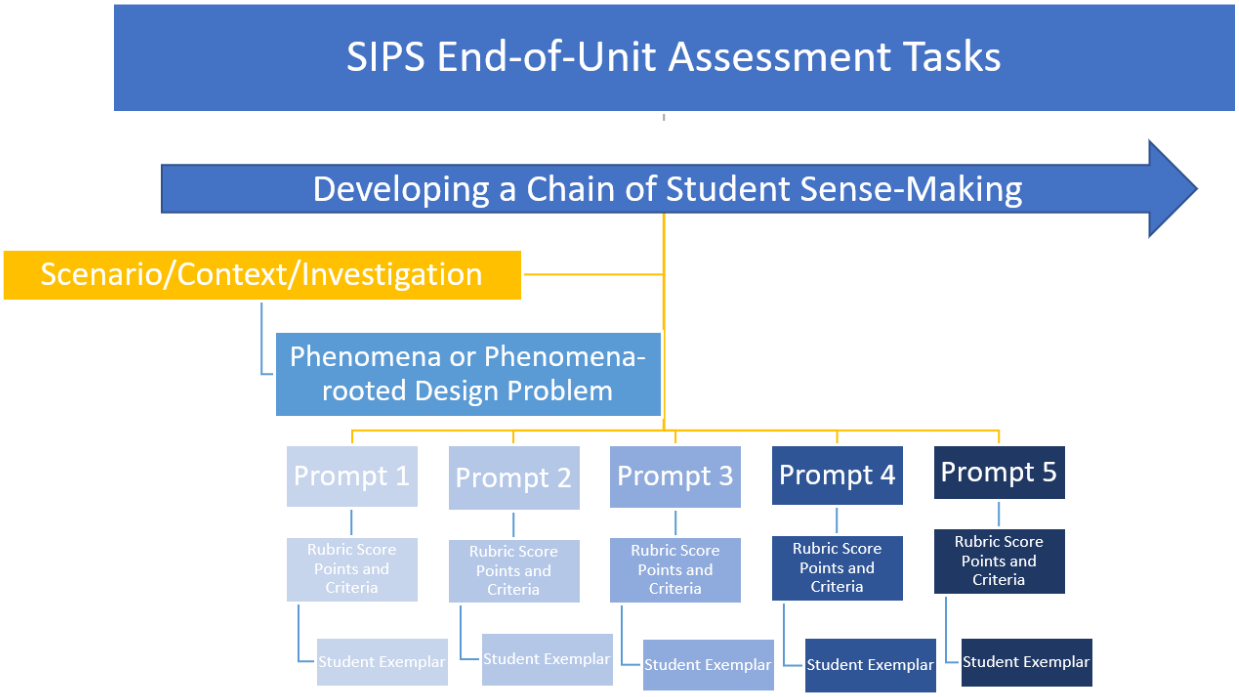
As noted previously, each EOU assessment consists of three tasks. To provide further specifications for each task as part of an ECD approach, the SIPS team created task specifications. Each task specification tool provides specification for the following:

- List of performance expectations covered in the task (each task covers one to two PEs);
- Information on the phenomenon or phenomenon-rooted design problem: Each task is rooted in a phenomenon or design problem related to the PEs;
- Scenario: Each task requires a scenario or situation which would make sense to students, be coherent and understandable to students, and provide enough context to allow students to engage meaningfully with the task;
- Variable Features: A list of features (or decision points) that could be modified to shift the complexity and/or focus of the task while still measuring the PEs;
- Chain of Sensemaking: An overview of the flow of the task, including the alignment of different sections to the KSAs;
- KSAs: A list of the KSAs that are targeted by the task, including any additional (not from the original set of design patterns) KSAs that are a cross between two PEs;
- Student Demonstration of Learning: A list of the expectations of students taken from the design patterns;
- Work Products: A list of the physical responses that students might produce;
- Application of Universal Design for Learning-based Guidelines: A set of guidelines to promote equity and inclusion in the task design; and
- SIPS Complexity Framework Components: A description of how the prompts for the task are designed to align with the degrees of sophistication represented by the complexity framework.

The task specification tool describes the design elements of the task and provides guidance to task developers. This information was used to further develop the tasks. Each task is aligned to one or two

PEs and is situated in a given phenomenon or design problem. The design problem or phenomenon is situated in an overall scenario and scaffolded such that students are provided a foundational context, the context is then problematized, and then students engage with the context through a series of prompts or questions. The scenario must make sense to students, be coherent and understandable, and provide enough context to allow students to engage meaningfully with the task. Each task includes rubrics that clearly define what is required of students and how evidence from students can be evaluated. Student exemplars are also included that provide a high-level response to each of the parts of the task. Exhibit 1 shows the components of a SIPS EOU assessment task.

Exhibit 1.Components of a SIPS EOU Assessment Task


SIPS End-of-Unit Assessment Tasks
Developing a Chain of Student Sense-Making
Scenario/Context/Investigation
Phenomena or Phenomena-rooted Design Problem
Prompt 1
Prompt 2
Prompt 3
Prompt 4
Prompt 5
Rubric Score Points and Criteria
Rubric Score Points and Criteria
Rubric Score Points and Criteria
Rubric Score Points and Criteria
Rubric Score Points and Criteria
Student Exemplar
Student Exemplar
Student Exemplar
Student Exemplar
Student Exemplar

While not every prompt has to cover every dimension in the PE cluster, every dimension within the unit's PE bundle must be aligned to at least one item on one task on the EOU assessment. The majority of items within the task must be either two or three dimensional.

Once tasks were developed, the SIPS team reviewed the tasks for alignment back to the task specification tool, ensuring coverage of the KSAs specified in the tool. Tasks were also reviewed for clarity, sense-making, accessibility and fairness, and the degree to which they require sense-making. Feedback was obtained from state partners as well as outside experts and included reviews of the tasks as well as the scoring rubrics (described below). The SIPS team applied revisions to the tasks based on this feedback.

2.4 EOU Task Rubric Development

The SIPS team developed a scoring rubric for each task to highlight aspects of the student response that demonstrate understanding of the concepts. The scoring rubrics include evaluative criteria to support the evaluation of evidence for each prompt (or a set of sub-prompts) within each task and were developed based on the student demonstration of learning from the task specification tool. The number of score points possible for each prompt or set of sub-prompts varied from one to four points depending on the expectations of students.

Rubrics were designed with the expectation that educators would be the main users of the rubrics. Each score point was defined to provide clear guidelines of the differences between student responses that fall in each score point. Rubrics also cover the range of possible student responses and are specific to the given prompts as this allows for more guidance for scorers.

Once the rubrics and tasks were developed, the SIPS team aligned them back to the PLD descriptors, ensuring that the tasks and rubrics are focused on aspects of the PLDs that are deemed important and that the set of tasks as a whole cover the critical aspects of the PLDs. The SIPS team applied revisions to either the tasks or the PLDs (as concepts of the PLDs changed throughout the development process).

2.5 EOU Development Summary

The EOU development process described above was used to produce four EOU assessments each at grade 5 and grade 8, each of which are intended to be administered after approximately 8 to 10 weeks of instruction (i.e., following each of the SIPS instructional units in each grade). Each assessment contains three multi-part tasks which are scenario/phenomena based and are designed in a way that students engage with sense-making as they move through the task.

To the extent possible, the task scenario is based on a phenomenon or design problem that occurs outside of the classroom and has local or global relevance. However, given variation in curricular and instructional resources used across states and districts, SIPS partners acknowledge that tasks address phenomena or phenomena-rooted design problems that may or may not have been addressed through instruction.

The tasks designed for each EOU are meant to be illustrative examples of (1) PE bundles and (2) task scenarios. Additional tasks can be designed using the SIPS design process to support use with other SIPS unit sequences or other curricula. While the EOUs were designed to be administered in the recommended order of the SIPS instructional units, if educators taught the instructional units in a different order, then the assessments may be administered in the sequence that best aligns with instruction. Scoring for these assessments would be the same regardless of the order in which they are administered.

Section 3. Research Questions

The release of the NGSS standards (NGSS Lead States, 2013) marked a shift in priorities for what students should know and be able to do related to science. The focus of the standards in how well students can apply knowledge affects how assessments should be developed to measure student ability (Pellegrino, 2013). While the use of an ECD approach can support claims about the validity of the assessment (Mislevy, 2007) it is still important to collect evidence that the assessment is valid for the intended purpose (AERA, et al., 2014).

The SIPS EOU assessments were designed to be usable in a classroom and to provide scores that could be used to make claims about students' proficiency related to three-dimensional science learning. To collect evidence about the validity of these assessments for this purpose, the pilot study was designed to focus on three overarching research questions. Each question was then further elaborated into a set of specific (testable) research questions. These questions are:

RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning?

- Can the assessments be administered within a single class period?
- Are there patterns in the prompts that students skip?
- Can educators reliably score student responses on the EOU assessments?
- Do the EOU tasks allow students to demonstrate the full range of NGSS performance expectations?
- Is performance on the EOU assessments associated statistically with other indicators of student learning (e.g., opportunity to learn (OTL), curriculum, or student performance on subsequent end-of-year (EOY) science assessments)?

RQ2: How well do latent variable measurement models fit the empirical EOU assessment data?

- Which measurement model(s) best fit the EOU data both within and across EOUs?
- Can a measurement model be used to create "empirical dimensions" that (i) are distinct, (ii) are interpretable, and (iii) correspond to aspects of the NGSS dimensions?
- Which models, and corresponding estimates, provide the most useful results in terms of:
    - Understanding three-dimensional science learning?
    - Informing the next unit of instruction?
    - Creating a single summative score to support federally required state systems of school identification and support?

RQ3: Overall, what do the EOU assessment results tell us about students' science learning?

- What do the EOU assessment results tell us about student learning in terms of variation across student groups (e.g., are students of multiple backgrounds being provided with equitable opportunities to learn)?
- What do the EOU assessment results tell us about student learning in terms of variation in performance across instructional programs, instructional units, and instructional unit sequences?

- What do the EOU assessment results tell us about student learning in terms of changes across administrations (i.e., growth)?

The pilot study was designed to gather sufficient data to analyze this extensive set of research questions. The next sections describe the sample of students and educators who participated in the study, what data were collected, and how these data were analyzed to address the research questions.

Section 4. Sample Acquisition

4.1 Recruitment

Recruitment for the SIPS pilot involved close collaboration between SIPS team members, state leads, district leads, and educators. The goal for recruitment was to have at least five classrooms of students per state take each SIPS EOU assessment in both grade 5 and grade 8. It was anticipated that not all classrooms recruited would take all four EOU assessments, and therefore the initial target was 20 classrooms per grade per state. The main requirement for educators to participate was teaching a curriculum aligned to three-dimensional science standards (e.g., NGSS standards or similar).

The overall recruitment approach was to have state leads connect the SIPS team to district leads, who would in turn connect the SIPS team to educators. The SIPS team followed state and district guidelines for how communication between SIPS, districts, and educators would be handled. The recruitment process was as follows:

1. State partners identified a set of district leads for their state.
2. SIPS invited the district leads to attend the District Leads Orientation Webinar, a one hour webinar that introduced the project.
3. District leads attended the webinar and afterwards signed an agreement giving educators in their district permission to participate.
4. SIPS worked with district leads to address any district Institutional Review Board (IRB) requirements.
5. Participating district leads identified and contacted educators and school leaders, using materials provided by SIPS.
6. Educators expressed interest in the assessment pilot by completing an electronic educator interest survey.
7. SIPS contacted interested educators to identify which assessments they intended to administer and to sign a Memorandum of Understanding (MoU) to commit to participating in the 2022-2023 school year.

At the end of the recruitment process, the SIPS team had a total of 121 educators from across four states that expressed initial interest in participating in the pilot. Of those 121 educators, 63 educators representing three of the six partner states participated in the study by administering one or more EOU assessments. See “6.1 Overview of the Data” for a summary of the number of educators and students that participated in each EOU assessment administration.

4.2 Expectations for Participating Educators

The main expectations for educators were to administer one or more of the EOU assessments and to provide scores for students on those assessments. To support educators in the administration and scoring of the assessments, educators were also asked to participate in several webinars. The list of activities educators were asked to complete for each EOU they administered is as follows:

- **Attend a Training and Orientation Webinar [approximately 1.5 hours]**. Prior to or early on in each window, educators were asked to participate in a short workshop that introduces them to the SIPS project and provides training on the use of the SIPS EOU assessment materials. Educators also received guidance on the use of the SIPS Understanding by Design (McTighe & Wiggins, 1998) Unit

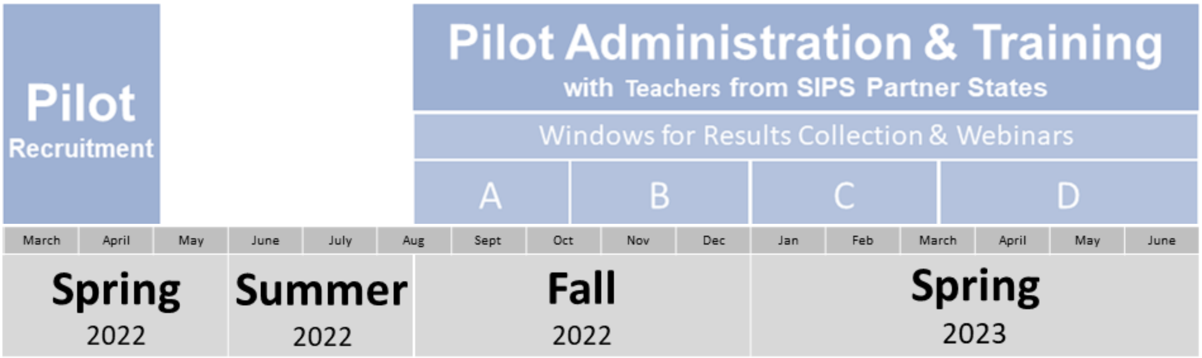
Map Stage 1 Learning Goals and their articulation across instructional segments to evaluate and potentially modify their instruction to ensure they are covering the PEs and providing opportunities for students to learn the knowledge, skills, and abilities measured by the EOU assessment. Note that once educators completed this initial training, they were provided a shortened training for subsequent EOU administrations that focused on the learning goals for the specified unit and the logistics and timeline for completing the piloting activities.

- **Administer the EOU Assessments.** Within the selected window, participating educators were asked to administer the EOU assessment at the end of their instruction on that given unit.
- **Score the EOU Assessments.** Within the selected window, participating educators were asked to attend a virtual scoring workshop and score and upload their anonymized student work.
  - **Upload Student Work.** Participating educators were asked to scan and **upload** all of their students' work on the EOU assessment prior to the scoring workshop to facilitate scoring consistency.
  - **Attend a Virtual Scoring Workshop.** Participating educators were asked to join a virtual **scoring workshop** aimed at scoring, with consistency, a single task from a given EOU assessment. Educators that administered more than one EOU assessment were required to attend a minimum of one scoring workshop.
  - **Score Remaining Student Work.** After attending the virtual scoring workshop, educators were asked to independently score the remaining tasks and upload the results.
- **Complete an Instructional Practices and Assessment Use Survey.** As part of the administration, educators were asked to complete an online survey that asked about the instruction leading up to the administration, including the number of days of instruction, whether the unit was completed, and potentially, the specific lessons completed. The survey also asked questions about the usefulness of the results and other materials in supporting various instructional next steps.

4.3 Timeline

The SIPS team recruited educators during early Spring 2022 and facilitated the pilot administration during the 2022-2023 school year, as shown in Exhibit 2.

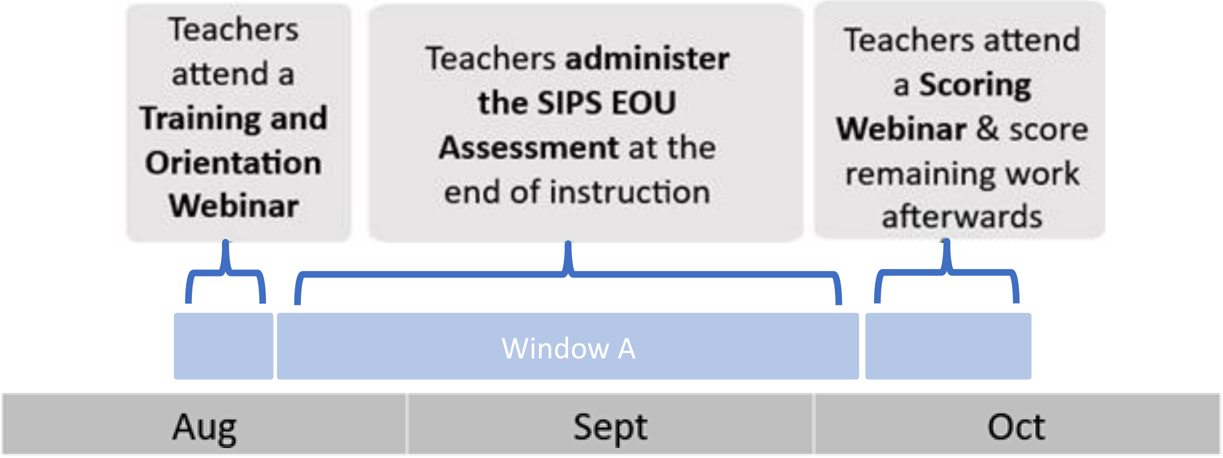
Exhibit 2. Overall Pilot Timeline



The pilot was organized into four administration windows with the expectation that educators would administer at most one EOU assessment in a given window. For educators who administered all four EOU assessments, the SIPS team anticipated there would be one assessment administered in each window. While the SIPS team provided an expected order for the administration of the assessments (EOU1, EOU2, EOU3, EOU4), educators were invited to administer one or more assessments in a different order if it better aligned to their instruction.

Within each window, educators were expected to attend two webinars, administer an EOU assessment, and score and upload students' work (see Exhibit 3). If educators planned to administer more than one EOU assessment, they were required to attend at least one scoring workshop.

Exhibit 3. Example Sequence of Activities for Window A.



Teachers attend a **Training and Orientation Webinar**
Teachers **administer the SIPS EOU Assessment** at the end of instruction
Teachers attend a **Scoring Webinar** & score remaining work afterwards
Window A
Aug
Sept
Oct

All data were collected by the end of June, 2023.

Section 5. Test Administration

As mentioned previously, educators participating in the pilot administration were asked to administer one or more EOU Assessments, score student work, scan and upload assessments, provide SIPS with scores on those assessments, and complete an educator survey. The educator activities produced three data sources: educator surveys, student EOU packets, and educator workbooks with student demographic information and scores for each prompt and task. Each of these data sources are discussed in more details below.

5.1 Educator Survey

Educators responded to a survey (see Appendix A) after each administration of an EOU assessment. The survey was designed to be completed after educators had scored student responses on the EOU assessment with the goal of obtaining feedback on the assessment and the assessment administration from the educators. The survey was organized into two parts. The first part focused on students' experiences with the EOU assessment and included questions related to how long it took students to finish the assessment, the degree of student engagement, the degree of challenge, and how well students performed on the assessment. In this section, educators were also able to provide specific feedback for improving tasks and prompts.

The second part focused on the context for the assessment administration. In this section, educators provided information about instruction prior to the administration, including the curriculum that was taught, the degree to which they covered the learning goals targeted in the assessment, and how similar this assessment was to other assessments or activities used in the classroom. This section also included an open-response prompt for educators to reflect on how meaningful student performance results were for informing teaching and learning.

5.2 Student EOU Packets

Educators were provided with packets that contained 30 copies of the EOU assessment with associated educator and student IDs. Educators were asked to distribute the assessment packets to students, making sure that if they administered more than one EOU assessment to their students in later administration windows, each student would receive the assessment with the same associated student ID.

Educators monitored the students taking the assessments and then collected the packets. They scanned the packets and uploaded them into educator-specific folders on Box, a secure file sharing solution. Educators were asked to upload the student packets within a week after their students had taken the assessment.

5.3 Educator Workbook

An individualized *Educator Workbook* was created for each educator and each EOU assessment. Each workbook included (a) one worksheet containing instructions and links for downloading the student packets along with an educator version, the *Assessment Scoring Guide*, which included scoring guides and student exemplars for each task, (b) a single worksheet containing a master roster of the educator's classroom for providing important demographic information about students (i.e., English learner status, IEP status, 504 plan status, proficiency level of previous state test performance in English language arts and mathematics), and (c) a worksheet for entering students' scores for each prompt on the EOU assessment. Students used this worksheet for indicating which students were present or not present when the assessment was administered, adding administration notes, and providing scores for each student and prompt based on the application of the scoring rubric to the student responses. To support the scoring of the student work, educators were asked to attend at least one scoring workshop for the first EOU assessment they administered to students.

In addition to this file, educators were given a student roster spreadsheet. This spreadsheet contained a list of all of the student IDs. Educators assigned each student in the class a student ID using the roster spreadsheet and kept this list in their personal files for reference for current and future EOU assessment administrations.

5.4 Educator Scoring

Educators were asked to score all of their student work for a classroom of approximately 20 students. After educators had administered the assessment and uploaded the scans of student work, they were invited to attend a scoring workshop.

Four scoring workshops were conducted during the 2022-2023 school year, during which educators engaged in collaborative scoring of student work through the application of the SIPS scoring rubrics and student exemplars. The SIPS team facilitated two scoring workshops for Administration Window A, which provided scoring guidance for the grade 5 Unit 1 and 2 EOU assessments and the grade 8 Unit 1 EOU assessment, and two scoring workshops for Administration Window B, which provided scoring guidance for two additional assessments: the grade 5 Unit 4 EOU assessment and the grade 8 Unit 2 EOU assessment. These workshops provided similar content but were held at different times to better accommodate educator schedules. Educators selected and attended the scoring workshop that best accommodated their schedule. Educators were asked to attend at least one scoring workshop to ensure familiarity and practice with the application of the SIPS scoring rubrics. Exhibit 4 provides the dates of the four scoring workshops, the EOU assessments they addressed, and the number of grade 5 and grade 8 educators in attendance.

Exhibit 4. Scoring Workshop Attendance

Window	Workshop Date	EOUs	# of Grade 5 Educators	# of Grade 8 Educators	Attendance by Workshop
A	October 19, 2022	G5 EOU1, EOU2 G8 EOU1	17	6	23
A	November 7, 2022	G5 EOU1, EOU2 G8 EOU1	17	3	20
B	December 3, 2022	G5 EOU1, EOU2, EOU4 G8 EOU1, EOU2	2	1	3
B	January 26, 2023	G5 EOU1, EOU2, EOU4 G8 EOU1, EOU2	24	4	28
		Totals	60	14	**74**

Overall, 40 of the 43 participating grade 5 educators attended at least one scoring workshop, and 16 of the 20 participating grade 8 educators attended at least one scoring workshop.

To facilitate the scoring workshops, the SIPS team organized participating educators into break-out rooms by grade level and EOU assessment. Within each break-out room, facilitators reviewed and discussed a selection of prompts and rubrics from the EOU assessment and presented a set of example responses for each score point (selected from the submitted student work) for discussion. After this training, educators independently scored a set of student work, with each educator in the session scoring the same set of student responses. These initial scores were recorded and saved. Educators then reconvened with the group and the facilitator led a discussion of the set of student work, the scores the educators' applied during their independent reviews, and any discrepancies in the scores. Educators were encouraged to share the reasoning for their scores with the goal of having the group come to agreement and gain consistency in their application of the scoring rubric. Following their discussion of student work for the first prompt, educators were then given a set of papers for a second prompt to score independently, which again were recorded and discussed as a group.

After the scoring workshop, educators were asked to complete scoring of their students' responses and enter the scores for each prompt in the *Educator Workbook*. To further support educator scoring, the SIPS team also prepared and disseminated scored and annotated student work for a selection of prompts from each EOU assessment (four each at grade 5 and 8). The SIPS team also facilitated Question and Answer, or Q&A, sessions with any educators seeking clarification about how to apply the scoring rubrics to evaluate their students' responses for the EOU assessments. These Q&A sessions were scheduled upon request by participating educators.

Section 6. Data Analysis

In this section, we provide an overview of the data and analysis used to address the research questions developed at the outset of the pilot study. The analyses were used to reflect on the pilot assessments and were often exploratory in nature. In addition, data from the pilot were used to identify revisions to the prototype assessment tasks themselves. With these revisions, the tasks and the EOU assessments were modified. A larger field study will need to be conducted to collect further information about the behavior of these updated assessment tasks. Further reflections on the revisions to the assessments and the research questions are presented in Section 8 of this report.

Specifically, this section will provide an overview of the data and the data cleaning process. We then address each of the three overarching research questions. As a reminder, these overall research questions are:

- RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning?
- RQ2: How well do latent variable measurement models fit the empirical EOU assessment data?
- RQ3: Overall, what do the EOU assessment results tell us about students' science learning?

The section will provide a discussion of data used for revisions to the tasks and general conclusions.

6.1 Overview of the Data

As mentioned above, the SIPS team collected student response data for all eight (four grade 5 and four grade 8) EOU assessments. We also gathered data via an online survey from nearly all the educators who participated in the pilot study which asked them to report on the science curricula they were using, reflect on the science concepts they taught, and record their impressions of the overall quality the assessments they administered. Each educator also provided student level demographic data (e.g., gender, prior achievement). Exhibit 5 shows the number of educators and students who took part in the study by assessment.

Exhibit 5. Number of Educators and Students Included in the Sample, by EOU Assessment

EOU Assessment	Number of Teachers	Number of Students
Grade 5 Unit 1	23	341
Grade 5 Unit 2	28	473
Grade 5 Unit 3	19	341
Grade 5 Unit 4	26	417
Grade 8 Unit 1	14	151
Grade 8 Unit 2	10	189
Grade 8 Unit 3	13	258
Grade 8 Unit 4	4	51

These data indicate that not all educators administered all four EOUs, and only a small subset of educators administered all four assessments for their grade level. In addition, not all students earned a score on all of the tasks within an EOU. As a step in the data cleaning process, the decision was made that if a student was missing a response to part of a task a total score for that task was not computed. In addition, if a student did not have a score for one of the tasks, a total score for that EOU was not computed. Hence, sample sizes vary in the analyses presented later in this section from the numbers of students in Exhibit 5.

If students did have scores for all prompts within a task, then a total score for the task was computed by summing up the scores on the prompts. The EOU score was generated by summing up the individual task scores, again provided students had a score on all three tasks. The scores from these assessments were merged with the educator survey data using the educator ID as the matching variable, meaning that educator responses on the survey were associated with each student in that educator's class. Note that no student would have been in multiple educators' classrooms. Data were converted to numeric value when appropriate, making sure that all missing data were coded appropriately.

In the next three subsections we discuss the analyses that were completed to address each of the three overarching research questions.

6.2 RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning?

Research question 1 was addressed using descriptive statistics and classical test theory analyses to examine the subset of questions that together offer evidence in support of RQ1. As specified earlier in Section 3, the RQ1 sub-questions are:

- Can the assessments be administered within a single class period?
- Are there patterns in the prompts that students skip?
- Can educators score student responses on the EOU assessments reliably?
- Do the EOU tasks allow students to demonstrate their full range of NGSS performance expectations?
- Is performance on the EOU assessments associated statistically with other indicators of student learning (e.g., opportunity to learn (OTL), curriculum, or student performance on subsequent end-of-year (EOY) science assessments)?

Data were drawn from the student assessments, the educator scoring workshops, and the educator surveys. More specific information on the analyses is described below, while reflections and discussion of these analyses are described in Section 8 of this report.

6.2.1: Can the assessments be administered within a single class period?

To address this question, timing data from the student assessment were collected and summarized. Teachers were asked to indicate the start time and end time for each of the assessment tasks they administered. These data were used to compute the total time taken for each task within each EOU. The average time was calculated across students by task. For Grade 5 the average time per task ranged from 20 minutes to 36 minutes (see Exhibit 6), with students taking the longest to complete tasks in Unit 3.

Exhibit 6. Grade 5 EOU Timing Data: Means and Standard Errors

Task Administration Time (min)		
EOU Assessment Task	Mean	SE
EOU1 Task 1	31.18	.72
EOU1 Task 2	31.11	.68
EOU1 Task 3	30.60	1.26
EOU2 Task 1	20.00	.76
EOU2 Task 2	25.85	1.05
EOU2 Task 3	23.09	.70
EOU3 Task 1	36.42	1.23
EOU3 Task 2	35.37	1.95
EOU3 Task 3	31.04	1.02
EOU4 Task 1	28.91	.90
EOU4 Task 2	25.64	.80
EOU4 Task 3	29.18	.79

For Grade 8, the average time per task ranged from 17 to 43 minutes, with students spending the most time on tasks for Unit 1 (see Exhibit 7).

Exhibit 7. Grade 8 Timing Data: Means and Standard Errors

Task Administration Time (min)		
EOU Assessment Task	Mean	SE
EOU1 Task 1	43.14	1.43
EOU1 Task 2	40.69	1.59
EOU1 Task 3	41.63	1.57
EOU2 Task 1	32.82	1.45
EOU2 Task 2	29.60	1.30
EOU2 Task 3	29.22	1.11
EOU3 Task 1	21.77	.82
EOU3 Task 2	26.48	1.10
EOU3 Task 3	25.10	.96
EOU4 Task 1	19.53	.94
EOU4 Task 2	17.45	1.12
EOU4 Task 3	27.39	1.01

During this prototyping study it was assumed that an EOU, whether designed for grade 5 or grade 8 students, could be administered within a 45-to-50-minute class period. The data presented in Exhibits 6 and 7 indicate that the tasks within an EOU would need to be shortened considerably to meet the goal of administering all three tasks for each EOU within a single class period.

6.2.2: Are there patterns in the prompts that students skip?

For this analysis, we calculated the percent of students who did not respond to each task (see Exhibit 8 and Exhibit 9). This analysis does not differentiate between students who did not attempt the prompt or skipped over it completely, or students who looked at the prompt but did not respond because they no longer had time to complete the prompt or the task during the class period. In addition to examining missing data by prompt, we also examined missing data by task and educator. This allowed us to see if there were classrooms for which the entire class skipped. See Appendix B for data tables that address analyses for subsection 6.2.2.

For Grade 5, we found that EOU2 had a relatively low proportion of missing responses, EOU 1 and EOU 3 had a range of missing responses depending on the prompt, and EOU 4 had a high proportion of missing responses across all prompts and all tasks. For EOU 1 we saw that one classroom skipped prompts from task 1, two classrooms skipped prompts from task 2, and two classrooms skipped prompts from task 3. For EOU 4 we found that there were three classrooms for which over half the students had missing scores for all the tasks. There were two classrooms that had missing scores for task 1, two classrooms with missing scores for task 2, and one classroom with missing scores for task 3. For EOU 2 and EOU 3 the missing was spread across classrooms (see Appendix B).

Exhibit 8. Percent Missing by Prompts for Grade 5

Grade 5	Min % Missing	Max % Missing	Average Percent Missing		
			Task 1	Task 2	Task 3
EOU 1	4.72	18.58	7.60	15.95	15.34
EOU 2	2.96	7.40	5.21	5.03	5.39
EOU 3	2.93	12.61	5.90	6.03	11.53
EOU 4	12.95	20.62	16.69	16.85	16.55

For Grade 8, we found low missing responses on task 1 and task 2 of EOU 4, but that EOU also had the lowest response rate. Otherwise, missing responses was fairly high across the EOUs and the tasks. For EOU 1, there were three classrooms where most students responded to all tasks and five classrooms for which only two or three students responded to all tasks and the pattern of missing responses varied across the other students. In EOU 2 there was one classroom where all students had missing data on task 1, one educator where all students had missing data on task 2, and two educators where all students had missing data on task 3. We saw similar behavior for EOU 3, where the same educators had missing data for all students on select tasks. In EOU 4 we did not see patterns of missing data by educators.

Exhibit 9. Percent Missing by Prompts for Grade 8

Grade 8	Min % Missing	Max % Missing	Average Percent Missing		
			Task 1	Task 2	Task 3
EOU 1	3.45	27.59	7.01	21.27	18.07
EOU 2	13.23	44.97	26.54	26.20	27.36
EOU 3	10.08	24.81	16.57	10.60	22.09
EOU 4	1.96	17.65	1.96	1.96	12.42

6.2.3: Can educators score student responses on the EOU assessments reliably?

Data for inter-rater reliability were collected during three of the four educator scoring workshops that took place during scoring windows 1 and 2. Educators scored a set of responses during the workshop and their scores were recorded before any adjudication took place. Cohen's kappa (Cohen, 1960) values were generated, both using all educators to determine the level of agreement among educators, and then also comparing the educators to the expert coders. In total, there were six prompts at Grade 5 and four prompts at Grade 8 used in the scoring sessions. The number of raters and the number of student responses varied by prompt (see Exhibit 10).

The degree of agreement varied depending on the prompt (see Exhibit 10). No prompt was deemed as having excellent inter-rater reliability, with Grade 5, EOU 1, prompt 2 having a very low rate of agreement. Grade 5, EOU 4 prompt 2 and Grade 8, EOU 2 prompt 2AB also had low inter-rater reliability. However, it should be noted that there were discussions about the differences after educators had scored which was designed to address issues in differences in scoring. The number of responses that educators scored after this discussion was limited (often just three or four responses). In addition, the number of student responses is small and contained several responses that were border-line responses.

Exhibit 10. Overall Agreement Among Educators by Prompt

Grade	End-Of-Unit	Prompt	# of Raters	# of Student Responses	Kappa	z	prob>z
5	1	1_AB	23	10	0.64	61.18	0.00
5	1	2	21	10	0.26	19.66	0.00
5	2	1_A	11	14	0.61	31.6	0.00
5	2	2_A	12	8	0.65	20.06	0.00
5	4	1_1B	2	17	0.73	4.73	0.00
5	4	2	2	17	0.35	2.44	0.01
8	1	1_A	10	10	0.76	30.67	0.00
8	1	1_B	10	10	0.63	22.12	0.00
8	2	2_AB	7	6	0.31	5.26	0.00
8	2	4	7	6	0.60	12.64	0.00

In addition to calculating the inter-rater reliability across raters, we also calculated the agreement of each rater with the expert scorer. The expert scorer was someone who was involved in the development of the tasks and the rubrics, and the score was considered the desired score. The agreement varied across prompts and raters (see Exhibit 11), with some prompts (e.g., grade 5, EOU2, prompt 2A) having fairly high agreement, and other prompts (e.g., grade 8, EOU2, prompt 2AB) having very low agreement across raters. Overall, while some tasks and rubrics were able to be scored reliably, others would need revisions to support consistency in scoring. Further analyses for Subsection 6.2.3 are provided in Appendix C.

Exhibit 11. Agreement with Expert Rater by Prompt

Grade	EOU	Prompt	# of Raters	# in Exact Agreement	# with Kappa Values Greater Than or Equal to .8	Average Percent Agreement	Average Kappa Value with Expert
5	1	1_AB	23	6	14	84.4%	0.79
5	1	2	21	1	1	54.3%	0.39
5	2	1_A	11	0	7	80.9%	0.75
5	2	2_A	12	7	8	89.8%	0.84
5	4	1_1B	2	0	0	73.5%	0.63
5	4	2	2	0	0	76.5%	0.68
8	1	1_A	10	2	6	87.0%	0.83
8	1	1_B	10	2	5	84.0%	0.77
8	2	2_AB	7	0	0	50.0%	0.30
8	2	4	7	1	2	75.0%	0.65

6.2.4: Do the EOU tasks allow students to demonstrate the full range of NGSS performance expectations?

To address this question, we examined the distribution of performance of students for each of the EOU assessments (see Exhibit 12). While this analysis only addresses part of this question, Section 7 provides additional information related to the demonstration of student learning across levels.

Overall, we found that scores were distributed across the range, although only two assessments (Grade 5 EOU 2 and Grade 8 EOU 3) had students who achieved the highest possible score points. Grade 5 EOU 3 and Grade 8 EOU 2 had the highest score at least 10 points below the maximum scores. The average for students on all assessments was close to 50% of the total possible points. This indicates that while students are able to show a range of performance on the assessments, the assessments may be asking students to demonstrate knowledge or ability beyond their current range. For visual representation of the distribution of scores, see Appendix D.

Exhibit 12. Distributions of Scores for Each Unit

Grade	EOU	N Students	Max Points	Mean	Std. Dev.	Min	Max	5%	25%	50%	75%	95%
5	1	228	37	16.4	5.4	3	32	7	13	17	20	25
5	2	389	37	24.1	5.6	5	37	15	21	25	28	33
5	3	256	54	28.9	6.9	9	44	16	25	29.5	34	38
5	4	235	40	20.7	7.0	4	36	10	15	21	26	32
8	1	80	45	25.5	8.5	6	40	8	21	27	31.5	37
8	2	63	58	27.2	8.5	4	43	14	23	28	33	39
8	3	149	30	18.6	5.8	1	30	8	16	19	23	27
8	4	41	41	17.8	8.0	0	30	5	11	19	24	30

6.2.5: Is performance on the EOU assessments associated statistically with other indicators of student learning (e.g., opportunity to learn (OTL), curriculum, or student performance on subsequent end-of-year (EOY) science assessments)?

For each unit, a set of concepts was developed that was associated with the unit. These concepts were covered as part of the EOU assessment. As part of the EOU-specific educator survey, educators indicated which of these set of concepts they included in their instruction before students engaged with the assessment. We hypothesized that students whose educators indicated they taught the material would do better on the assessment than those that did not. For each unit, the number of students that were in classrooms in which the educator indicated they included instruction on the concept was calculated. For those classrooms in which at least 30% of the students did not receive instruction on the concept, t-tests were performed to determine if there was a difference in how students performed based on if the educator indicated they had delivered instruction on the concepts. This testing does conflate classroom characteristics but is used as exploratory analyses, as data on the classroom are limited. Also note that for some of these concepts the same educator indicated the concepts were not taught, and so some of the analysis on differences based on classrooms uses the same students in each group. Further study would be needed to draw conclusions about the relationship between OTL and student performance.

Differences between groups were not found for all of the concepts where students differed in how much instruction they received. When differences were found they were also not always in favor of the students who educators indicated they received this instruction. For example, in Grade 4, Unit 1 there were four concepts that differed in whether or not educators indicated they had been taught. Of these concepts, statistically significant differences were found that indicated students who received instruction outperformed students who did not for two of these concepts while for the other two the students who did not receive instruction outperformed those that did (see Exhibit 13). For more details on the concepts and the differences see Appendix E.

Exhibit 13. Number of Concepts with Significant Differences Related to Instruction on That Concept

Grade	EOU	N Concepts	N Concepts where ≥ 30% of Students NOT Provided Direct Instruction	N Concepts with Significant Differences (p<.1) Favoring Students							
				Instructed in the Concept				Not Instructed in the Concept			
				Task 1	Task 2	Task 3	EOU	Task 1	Task 2	Task 3	EOU
5	1	11	3	1	3	1	3	0	0	0	0
5	2	10	1	1	0	0	0	0	0	1	0
5	3	11	5	1	1	0	0	1	3	5	1
5	4	11	6	2	2	2	2	2	2	4	2
8	1	11	4	2	2	2	2	2	2	2	2
8	2	11	2	1	1	1	1	0	0	0	0
8	3	23	8	5	8	6	6	1	0	0	0
8	4	11	2	0	2	0	0	0	0	0	0

6.3 RQ2: How well do latent variable measurement models fit the empirical EOU assessment data?

In this section, the latent variable measurement modeling analysis is discussed. In particular, we address the specific RQ2 sub questions:

- Which measurement model(s) best fit the EOU data both within and across EOUs?
- Can a measurement model be used to create "empirical dimensions" that are (i) distinct, (ii) interpretable and (iii) correspond to aspects of the NGSS dimensions?
- Which of the IRT models provides the most useful data for purposes of scaling the EOUs in both grades in terms of:
  - Understanding three-dimensional science learning?
  - Informing the next unit of instruction?
  - Creating a single summative score to support federally required state systems of school identification and support?

This section focuses on the psychometric analysis of the SIPS EOU assessment data, in terms of classical test theory (CTT) and item response theory (IRT) methods. In terms of CTT analysis, Cronbach's alpha (Cronbach, 1951) was calculated as were item (or prompt) level statistics, specifically item point-biserial correlations and p-values. For Grade 5, Cronbach's alpha ranged from 0.69 (for Unit 1) to 0.82 (for Unit 3), indicating fair reliability of the tasks. Point-biserial correlations and p-values were used to flag items for improvement and are discussed further in section 6.5.

In terms of IRT, each SIPS EOU assessment was scaled separately using the Rasch model (Bond & Fox, 2001; Rasch, 1960) for prompts that were scored dichotomously and the Partial Credit Model (Masters, 1982) for prompts that were scored polytomously. In addition to item parameter and student ability estimates (i.e., theta scores), item infit and outfit and test information were calculated. Both statistics are expressed as standardized values, typically with a mean of 0 and a standard deviation of 1, making them comparable across different prompts and different EOUs. In the Rasch model, infit and outfit statistics are useful tools for evaluating the quality of the EOU prompts because they assess how well individual prompts fit into the overall measurement model. They are essential for ensuring the reliability and validity of the EOUs.

The infit and outfit statistics are derived from the IRT model, and they offer insights into the relationships between students' abilities and their responses to the specific EOU prompts. More specifically, infit statistics measure the appropriateness of a prompt's difficulty relative to the students' abilities. A lower-than-expected infit value indicates that the prompt may be too easy for our sample of students, leading to a high probability of correct responses. Conversely, a higher-than-expected infit value suggests that the prompt may be too difficult. Ideally, infit values close to 1 indicate a good fit, implying that the prompt's difficulty matches the students' abilities. Values significantly greater or less than 1 may indicate misfit.

The Rasch model outfit statistics evaluate the fit of a prompt in a more general sense, reflecting how well a prompt performs across the entire student ability spectrum. An outfit value greater than 1 suggests that the prompt's performance is erratic and influenced by factors other than the students' abilities, such as guessing or misunderstanding the prompt. An outfit value less than 1 may indicate that the prompt is too predictable and does not sufficiently discriminate among students with different abilities. In sum, infit and outfit statistics in the Rasch model provide a valuable means of identifying problematic prompts within and across the prototype EOU assessments investigated in this pilot study. Further discussion of these fit statistics can be found in Section 6.5 of this report and in Appendix I.

Unlike typical statewide assessment programs used for accountability, scale scores do not play a major role in the SIPS project, and thus are not computed based on a theta to scale score conversion formula. Instead, theta estimates, along with an additional calculated statistic about item difficulty, are used within the Embedded Standards Setting (ESS) process. This additional calculated statistic is theta value associated with a 0.67 probably of scoring in a specific category, referred to as the RP67. The ESS process is explained in detail in Section 7.

For dichotomous items, the Rasch model characterizes the probability that a student with a latent trait or theta value (θ) will respond correctly to item *j as*

$$P(x_i = 1|\theta_p, b_i) = \frac{exp\ [\theta_p - b_i]}{1 + exp\ [\theta_p - b_i]} \quad \text{(Eq 6.1)}$$

Where:

- *P(xi=1|θp,bi)* is the conditional probability of a correct response for examinee *p* on item *I* given *θp*,
- *θp* is the students' level on the latent trait, and
- *bi* is the item difficulty.

For polytomous items, the Partial Credit Model characterizes the probability that a student with a latent trait will receive a rubric score of *h* as

$$P(x_i = 1|\theta_p, b_i, d_{iv}) = \frac{exp\ [\Sigma_{v=0}^{h}\ (\theta_p - b_i - d_{iv})]}{\Sigma_{c=0}^{m_i}\ exp\ [\Sigma_{v=0}^{c}\ (\theta_p - b_i - d_{iv})]} \quad \text{(Eq 6.2)}$$

where *P(xi=1|θpbi,div)* is the probability of examinee *p* obtaining a score of *h* on item *i*; m_i is the number of item score categories; b_i is the item location parameter; d_{iv} is the category parameter for item *i* and category *v*.

All item parameter estimations were conducted in the R package mirt (Chalmers, 2012) using the default expectation maximization algorithm with fixed quadrature points (see Bock & Aitkin, 1981). Theta estimates were created using both EAP sum score and EAP methods, with resulting estimates provided to support the ESS analysis described in Section 7 of this report.

6.4 RQ3: Overall, what do the EOU assessment results tell us about students' science learning?

To ensure that the results are reflecting on students' ability related to science as opposed to other characteristics, we want to examine the relationship between the assessment and other variables. In this next section we further explore these relationships through analysis geared at addressing these three questions:

- What do the EOU assessment results tell us about student learning in terms of variation across student groups?
- What do the EOU assessment results tell us about student learning in terms of variation in performance across instructional programs, instructional units, and instructional unit sequences?
- What do the EOU assessment results tell us about student learning in terms of changes across administrations (i.e., growth)?

6.4.1: What do the EOU assessment results tell us about student learning in terms of variation across student groups?

As part of the process of collecting data on students, educators provided characteristics of students related to their gender, English learner (EL) status, Individualized Education Program (IEP) status, 504 Plan status, prior English language arts (ELA) achievement and prior mathematics achievement (see Exhibit 14). While there were not enough students who were identified as EL, or having an IEP or 504 plan, analysis was conducted to examine differences related to gender, prior ELA achievement and prior mathematics achievement. For gender, educators identified students as either male or female or unknown. Teachers also reported on ELA and math achievement based on the state's achievement test, using a scale that ranged from 0 (lowest level) to 2 (highest level). For achievement, a level 3 was included if the educator was unclear about the appropriate level of students. Analysis was conducted that used a t-test to look for differences across gender, and ANOVA to examine differences across achievement level. This was used to examine how appropriate the assessment was for different groups of students.

Exhibit 14. Characteristics of Participating Students by Grade Level and EOU*

Grade	EOU	# EL	# 504	# IEP	# Males	# Females
5	1	9	7	60	162	177
5	2	18	15	67	236	235
5	3	9	32	32	119	135
5	4	10	34	46	206	211
8	1	1	21	17	85	61
8	2	0	11	22	96	92
8	3	0	16	21	129	125
8	4	0	0	2	12	11

*Some students have missing values for these indicators and are not included here.

Statistically significant differences were found between males and females ($p<0.05$) for 3 of the assessments. In all three of these assessments, females outperformed males (on average). These differences were not found on the other assessments, and therefore further exploration is needed to determine why differences based on gender may exist. Similar patterns held at the task level (see Appendix G).

Exhibit 15. Gender Differences for Each EOU

Grade	EOU	# of Males	# of Females	Mean (Males)	Mean (Females)	diff	t	df	pr(T>t)
5	1	109	119	15.63	17.18	-1.54	-2.18	226	0.99
5	2	192	195	24.27	23.98	0.29	0.50	385	0.31
5	3	90	103	28.58	29.11	-0.53	-0.53	191	0.70
5	4	109	126	20.28	21.04	-0.76	-0.83	233	0.80
8	1	48	32	25.35	25.78	0.43	0.22	78	0.59
8	2	33	30	25.18	29.50	-4.32	-2.08	61	0.98
8	3	72	73	17.22	19.66	-2.44	-2.55	143	0.99
8	4	12	10	16.00	16.50	-0.50	-0.15	20	0.56

The ANOVA analysis found statistically significant differences based on achievement levels for both mathematics and ELA on the EOU assessment except for Grade 8 EOU 4 (which had a sample size of only 40 students across the levels). Students at the higher level tended to have higher scores on average (see Exhibit 16). Assessments in Grade 8 were an exception, where students at Level 2 sometimes had lower scores. However, in these cases there were only two students who were identified as being at level 2. Similar patterns existed at the task level (see Appendix G for additional analyses).

Exhibit 16. Average Scores on EOUs by Achievement Level

Grade	EOU	ELA Achievement				Math Achievement			
		Level 0	Level 1	Level 2	Unknown	Level 0	Level 1	Level 2	Unknown
5	1	15.6	17.5	20.1	14.6	15.0	18.0	19.6	15.2
5	2	21.4	25.1	26.8	23.8	21.3	25.3	26.3	24.0
5	3	25.6	30.1	32.0	29.4	25.1	30.6	30.9	30.7
5	4	18.8	22.7	23.5	18.0	19.6	23.1	21.5	18.0
8	1	18.0	25.3	26.5	22.8	18.2	28.1	25.5	19.8
8	2	25.1	32.1	19.5	22.9	27.5	30.9	30.5	22.0
8	3	13.9	19.3	20.7	18.4	16.0	18.9	22.1	19.0
8	4	16.3	22.0	21.0	14.8	17.0	20.3	15.5	16.0

6.4.2: What do the EOU assessment results tell us about student learning in terms of variation in performance across instructional programs, instructional units, and instructional unit sequences?

Educators indicated on the unit survey what curriculum, or curricular materials, they used with their students. While we do not have information about what activities they specifically worked on with their students we can do some preliminary analysis to determine if there are differences across curriculum.

There was more variation in Grade 5 on what curriculum materials were used (see Exhibit 17). While for Unit 1 most educators tended to have students interact with curricular materials from the SAIL Garbage unit, for Unit 2 more educators used the BOCES: Deer, Deer Everywhere material. For both of these units a comparison was made between the most popular curricular materials and the other materials, as individual sample sizes for other curricular materials were low. Note that not all educators specified which curricular materials they used and so there was also a high number of students for which we did not know what materials they used, or they used something other than the most popular units. For Grade 5 units 3 and 4, there was more of an even spread in which curricular materials educators used. For these two units we used an ANOVA to determine if there were differences in student scores on the EOU assessments based on the curricular materials. For all four units, we found statistically significant differences (p<0.05) in scores on the EOU based on the curricular materials students used. See Appendix H for additional information on this analysis.

For Grade 8, the most popular curricular materials were the Open Sci-Ed materials, which were used across units. While for Unit 1 educators also used materials from Amplify and StemScopes, for the other units educators not using this often indicated that they just used other materials without specifying which ones (see Exhibit 17). While we did see a statistically significant difference in favor of classrooms using the OpenSciEd materials for Unit 2 we did not see this for the other units. However, sample sizes were small for some the other units and so caution must be taken when interpreting these results. Results were similar across tasks (see Appendix H for additional analyses).

Exhibit 17. EOU Performance Differences Related to Curriculum Materials by Grade

Grade	EOU	Most Popular Curricular Materials		Comparison Curricular Materials		Significant Difference
		Name	N Students	Description	N Students	
5	1	Sail: Garbage Unit	119	Combination of Amplify, Inspire, Mosa Mack and other	109	Yes, favoring SAIL
5	2	BOCES: Deer, Deer Everywhere	156	Combination of Inspire, Mosa Mack, NGSS and other	233	Yes, favoriting BOCES
5	3	Distributed across BOCES, Inspire, Mosa Mack, Mystery Science, NGSS and other				Yes, with Mosa Mack having the highest average score
5	4	Spread across Ambitious Science, Mosa Mack, Mystery Science, NGSS and other				Yes, with Mosa Mack having the highest average score
8	1	Open Sci-Ed	69	Combination of Amplify and STEMscopes	11	No
8	2	Open Sci-Ed	38	Other	25	Yes, favoring Open Sci-Ed
8	3	Open Sci-Ed	49	Other	100	No
8	4	Open Sci-Ed	32	Other	9	No

6.4.3: What do the EOU assessment results tell us about student learning in terms of changes across administrations (i.e., growth)?

We have limited data on students who took all four EOU assessments. It should also be noted that each EOU covered a different set of learning goals, and there were no overlapping items included between assessments. Thus, we are not able to directly measure growth across the assessments. In addition, total possible scores differed across each of the assessments, so direct comparison of scores is not appropriate. Section 7 does include analyses about how students performed in relation to the PLDs which provide more direct evidence to address this research question.

Grade 5

Examining the correlations between EOUs for Grade 5 we see that there are statistically significant (p<0.05; Spearman rho) correlations between all EOUs (see Exhibit 18). After converting scores on each EOU into percent correct scores (so the possible values are comparable across EOUs), we see that students did not necessarily do better on later assessments (see Exhibit 19). However, there is the possibility that students took the assessments out of order. In addition, concepts were different on each assessment and so doing better on later assessments might not directly measure growth of a students' science ability.

Exhibit 18. Correlations between EOU Scores, Grade 5 (N = 235)

	EOU1	EOU2	EOU3
EOU2	0.40 0.00	1	
EOU3	0.51 0.09	0.54 0.00	1
EOU4	0.80 0.00	0.77 0.00	0.85 0.00

Exhibit 19. Average EOU Percent Correct, Grade 5

EOU	N Students	Mean	Std. Dev.	Min	Max
1	228	0.44	0.15	0.08	0.86
2	389	0.65	0.15	0.14	1.00
3	256	0.55	0.13	0.17	0.83
4	235	0.52	0.17	0.10	0.90

Grade 8

Examining the correlations between EOUs for Grade 8, we see that there are statistically significant (p<0.1; Spearman rho) correlations between all EOUs (see Exhibit 20). Similar to Grade 5, we do not see that the percent correct increases for higher number assessments (see Exhibit 21). This again may be due to the fact that some students might have taken the assessments in a different order or could also be related to the fact that these assessments vary in the content they cover.

Exhibit 20. Correlations between EOU Scores, Grade 8 (N = 21)

	EOU1	EOU2	EOU3
EOU2	0.45 0.04	1	
EOU3	0.47 0.03	0.82 0.00	1
EOU4	0.41 0.07	0.59 0.01	0.76 0.00

Exhibit 21. Average Percent Correct for Each EOU Assessment

EOU	N Students	Mean	Std. Dev.	Min	Max
1	80	0.57	0.19	0.13	0.89
2	63	0.44	0.14	0.07	0.69
3	148	0.62	0.19	0.03	1.00
4	41	0.43	0.20	0.00	0.73

6.5 Using Data for Revisions

In addition to the research questions, data from the pilot study were used to inform revisions to the tasks. These data included qualitative data, in terms of feedback obtained from the educators with regard to the tasks, and quantitative data based on data analysis. The quantitative data review focused on examining the p-values (difficulty of the items), point-biserial correlations (item to total), and infit and outfit statistics of the Rasch parameters. Prompts were then flagged based on if the p-value was too high or too low, if the point-biserial correlations were too low, and if the infit or outfit statistics indicated poor fit.

Once prompts were flagged, the task was reviewed as a whole to determine what revisions were to be made. Additional considerations, particularly with the recognition that the tasks took students too long, included revisiting the measurement targets to ensure tasks were targeted to the intended skills and reviewing the wording and the complexity of tasks. This resulted in tasks that were not flagged also being included in revisions. Revisions included scenario language and/or graphic revisions, stem changes and/or graphic changes, and clarifying language within task sets to support students' understanding and production of complete and accurate evidence of their science learning. Revisions to the scoring rubrics included clarification of language, differentiating score point criteria, opportunities for holistic scoring (i.e., combining parts of a prompt), as well as updating to reflect changes to related prompts. All exemplar responses included in the SIPS EOU assessment scoring guides were updated as well to reflect any changes to the corresponding prompts.

Overall, in Grade 5, there were 57 prompts piloted. Of these, 16 were identified for review based on the parameters listed in the data review criteria above (see Exhibit 22). Further information on why prompts were flagged can be found in the Appendix I.

Exhibit 22. Grade 5 Overall Flags

Assessment Year Version	EOU			
	1	2	3	4
Scored Prompts	11	12	20	14
Flags	5	2	6	3
Revised Prompts	9	7	8	10

Overall, in Grade 8, there were 67 prompts field tested. Of these, 27 were identified for review based on the parameters listed in the data review criteria (see Exhibit 23). Note, a very small number of students completed the Unit 4 EOU in grade 8. Further information on why prompts were flagged can be found in the Appendix I.

Exhibit 23. Grade 8 Overall Flags

Assessment Year Version	EOU			
	1	2	3	4
Scored Prompts	17	23	10	17
Flags	6	7	3	11*
Revised Prompts	14	11	3	11

*n-count 50

6.6 Conclusion

The results presented in this section provide an overview of how the EOU assessments performed. Information about student performance on these tasks was used to revise tasks to better measure students' ability related to science learning. The next section further explores the question about how the assessments reflect on the performance level descriptors. Section 8 provides additional discussion of the research questions.

Section 7. SIPS Standard Setting

7.1 Introduction

This section describes the methods, analyses, and results supporting the SIPS Assessments Project standard setting activities. Student performance on each SIPS 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 PLDs.

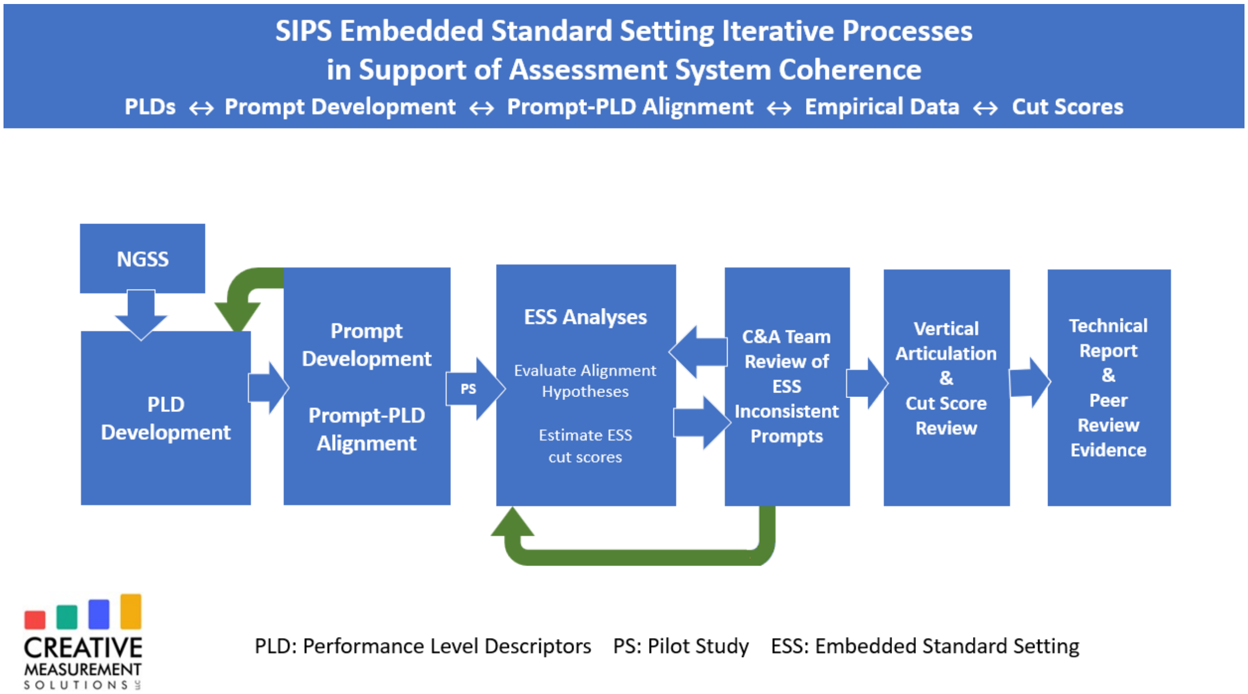
ESS is based on three big ideas:

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 KSAs of students in each performance level. The process of developing PLDs from the NGSS is represented by the first two boxes on the left in Exhibit 24.
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 24, 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 extend well beyond the estimation of cut scores.

Exhibit 24. SIPS Embedded Standard Setting Iterative Processes



SIPS Embedded Standard Setting Iterative Processes
in Support of Assessment System Coherence

PLDs ↔ Prompt Development ↔ Prompt-PLD Alignment ↔ Empirical Data ↔ Cut Scores

NGSS
PLD Development
Prompt Development
Prompt-PLD Alignment
PS
ESS Analyses
Evaluate Alignment Hypotheses
Estimate ESS cut scores
C&A Team Review of ESS Inconsistent Prompts
Vertical Articulation & Cut Score Review
Technical Report & Peer Review Evidence
CREATIVE MEASUREMENT SOLUTIONS

PLD: Performance Level Descriptors PS: Pilot Study ESS: Embedded Standard Setting

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 24.

7.2 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 24. 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.

7.3 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 KSAs expected of students in each performance level necessary to support the SIPS Theory of Action. The SIPS SMEs 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 of the SIPS Cut Scores and Investigation of Two IRT Response Probabilities."

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.

A detailed description of these five coordinated ESS iterative processes are described in detail in a stand-alone SIPS standard setting technical report (Stackable, Instructionally-embedded, Portable Science Assessments Standard Setting Technical Report, available at https://sipsassessments.org/resources/). Next, we provide a summary of the key findings in the full technical report.

7.4 Key Results of the SIPS Standard Setting Processes

Three sources of data were used to estimate and evaluate the SIPS cut scores including well-articulated PLDs, the alignment of each Task score point to a performance level (Item-Task alignments), and the empirical data provided by the SIPS Pilot Study. ESS analyses were conducted in four stages. First, we evaluated the efficacy of the SIPS SMEs' hypothesized Item-Task alignments for each EOU in each grade.

Second, we estimated ESS cut scores for each EOU and grade and examined them with respect to vertical articulation—the coherence and reasonableness of impact data, the percentage of students in each performance level within and across EOUs. When necessary, smoothing was applied to support an integrated system of coherent cut scores for the four EOUs in each grade.

Third, we examined smoothed, vertically articulated cut scores under two Item Response Theory (IRT) response probability values—RP67 and RP50. See Lewis, Mitzel, Mercado, & Schulz (2012) for a detailed discussion of IRT response probabilities.

Fourth, we considered validity evidence supporting the standard setting procedures using frequently cited sources of evidence from the measurement literature and peer review guidelines.

We summarize these four analyses here. A detailed description and discussion are provided in the stand-alone SIPS Standard Setting Technical Report.

The Efficacy of SIPS SMEs' Hypothesized Item-Task Alignments

Correlations

Exhibit 25(Exhibit 4 in the technical report) lists the correlations of prompts' SME-aligned performance level ordinality and RP67 location by grade and EOU. 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.

Exhibit 25. 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

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:

- *Level 1:* Below the ESS Level 2 cut score.
- *Level 2:* At or above the ESS Level 2 cut score but below the Level 3 cut score.
- *Level 3:* At or above the ESS Level 3 cut score but below the Level 4 cut score.
- *Level 4:* At or above the ESS Level 4 cut score.

Classification Agreement and Weighted Kappa

Classification agreement is described in the following terms:

- *Agree:* The empirical performance level agrees with the SME-Aligned Level.

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

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 26 (Exhibit 6 in the technical report) displays the recommended ranges suggested by Landis and Koch (1977).

Exhibit 26. 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

Kappa Interpretations

Grade 5

The grade 5 EOUs (see upper half of Exhibit 27 (Exhibit 7 in the technical report)) 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 26.

Grade 8

The grade 8 EOUs (see lower half of Exhibit 27(Exhibit 7 in the technical report)) 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 26.

Exhibit 27. 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

Vertical Articulation of the SIPS Cut Scores and Investigation of Two IRT Response Probabilities

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 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 an item 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 were 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. 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.

Initial RP67 and RP50 Cut Scores and Associated Impact Data

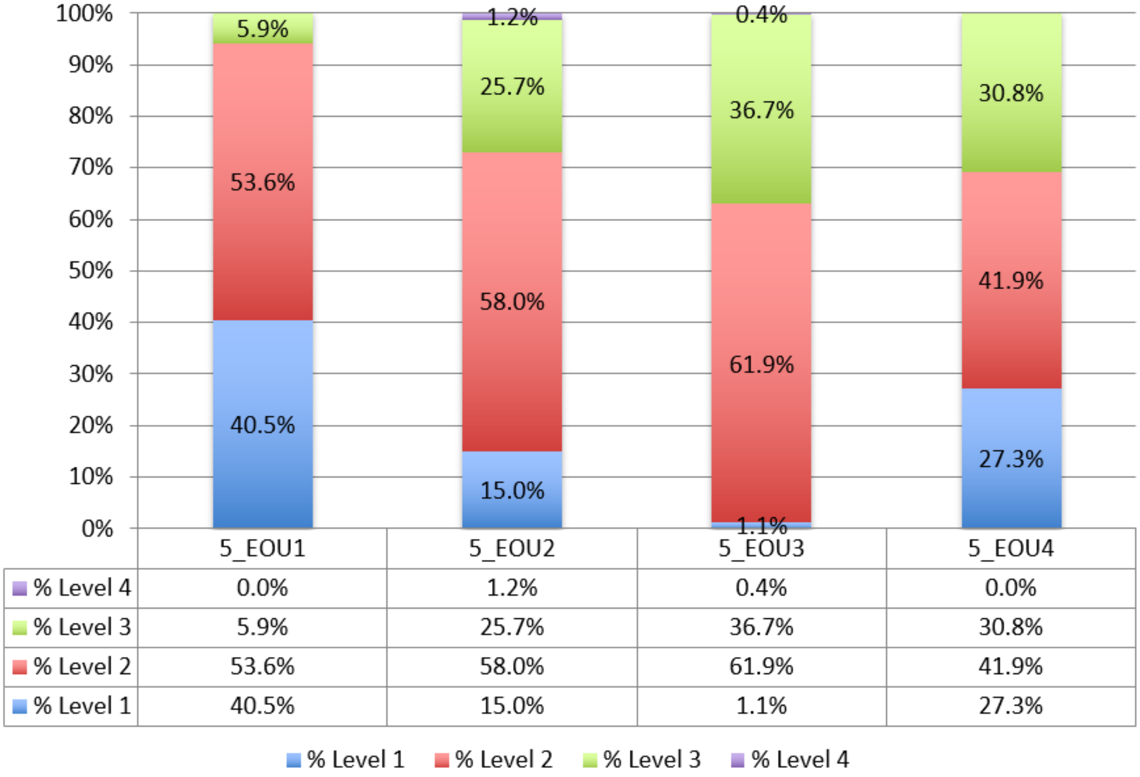
Exhibit 28(Exhibit 20 in the technical report) provides the initial ESS cut scores in the theta metric for RP67 and RP50 for each EOU in grades 5 and 8. Recall that the four EOUs in a grade are not on a common scale and thus, the cut scores cannot be directly compared.

Exhibit 28. Initial SIPS RP67 and RP50 Cut Scores Across EOUs for Grades 5 and 8

Grade & RP	EOU	Level2	Level3	Level4
Grade 5 RP67	EOU1	0.01	0.76	2.66
	EOU2	-0.56	0.47	1.46
	EOU3	-1.50	0.28	1.68
	EOU4	-0.28	0.49	2.62
Grade 5 RP50	EOU1	-0.37	0.40	2.17
	EOU2	-0.82	0.01	0.89
	EOU3	-1.67	-0.13	1.14
	EOU4	-0.72	0.14	2.02
Grade 8 RP67	EOU1	-0.27	1.81	4.00
	EOU2	-0.44	0.74	2.53
	EOU3	-1.37	-0.16	1.15
	EOU4	-1.06	0.36	3.06
Grade 8 RP50	EOU1	-0.51	1.21	4.00
	EOU2	-0.96	0.18	1.90
	EOU3	-1.89	-0.54	0.79
	EOU4	-1.73	-0.14	2.49

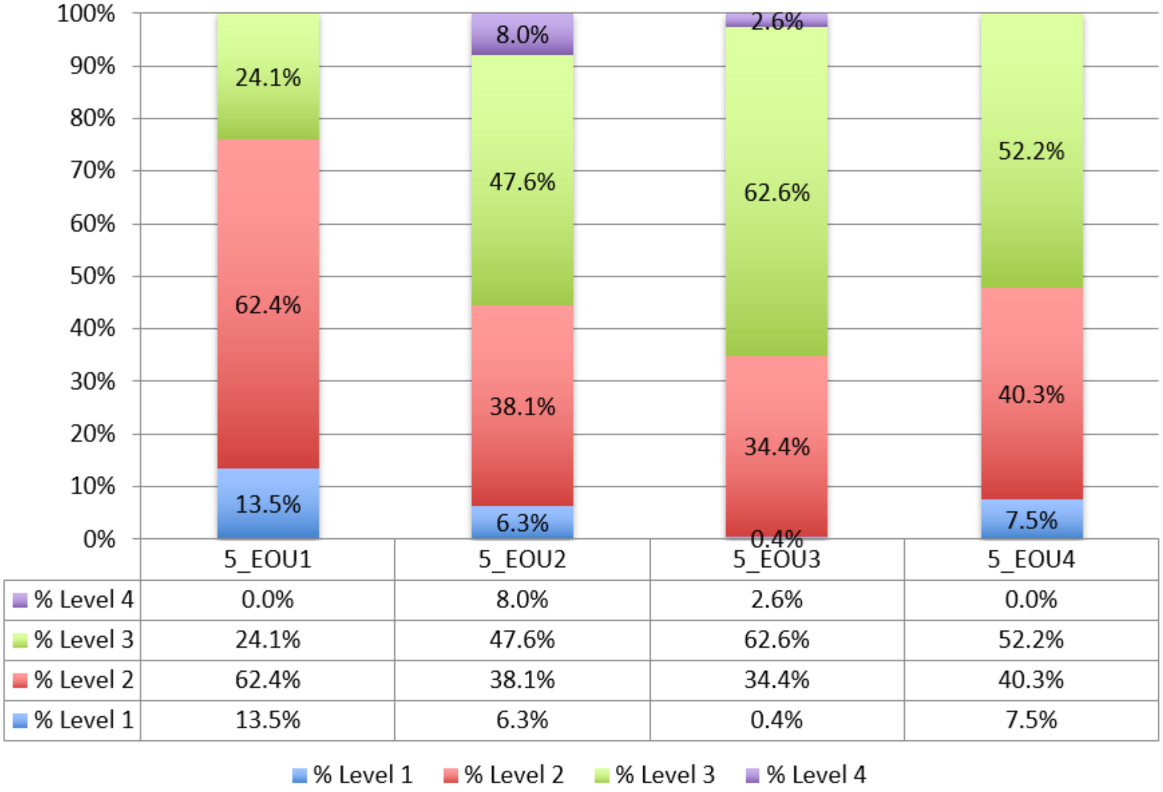
Exhibit 29and Exhibit 30 (Exhibits 22 and 23 in the technical report) provide impact data associated with the initial cut scores for the four grade 5 EOUs for RP67 and RP50, respectively. Exhibit 31and Exhibit 32 (Exhibits 24 and 25 in the technical report) provide impact data associated with the initial cut scores for the four grade 8 EOUs for RP67 and RP50, respectively.

Exhibit 29. SIPS Grade 5 RP67 Initial Cut Score Impact Data



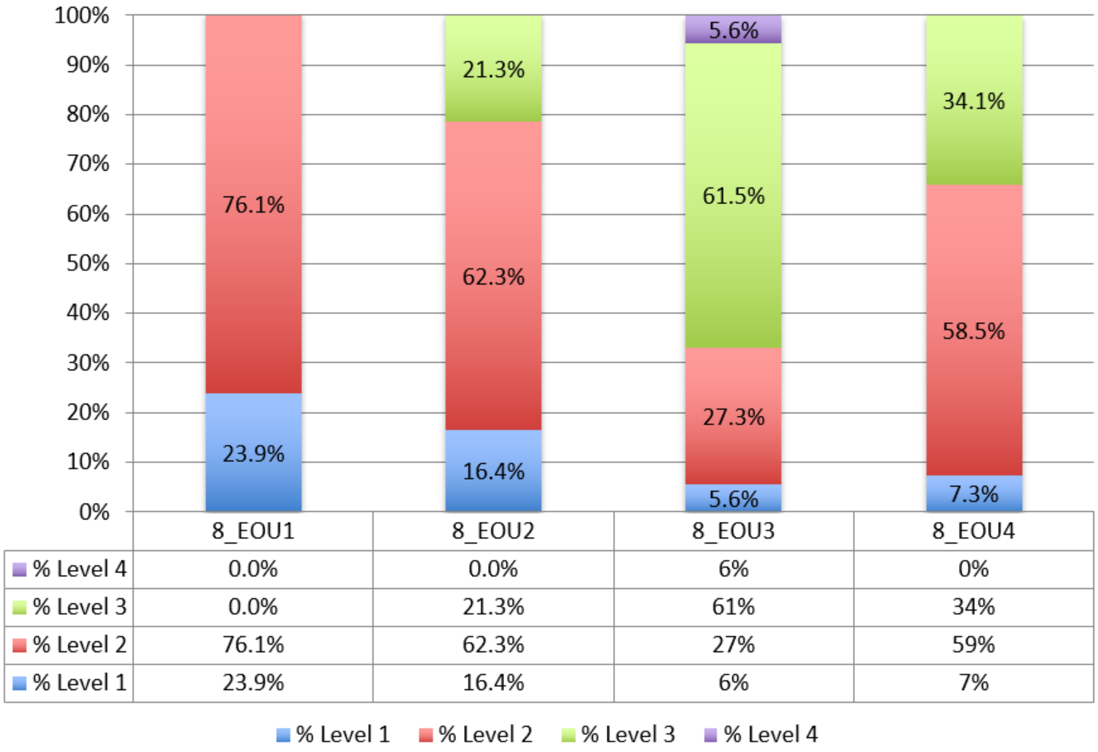
	5_EOU1	5_EOU2	5_EOU3	5_EOU4
🟪 % Level 4	0.0%	1.2%	0.4%	0.0%
🟩 % Level 3	5.9%	25.7%	36.7%	30.8%
🟥 % Level 2	53.6%	58.0%	61.9%	41.9%
🟦 % Level 1	40.5%	15.0%	1.1%	27.3%

Exhibit 30. SIPS Grade 5 RP50 Initial Cut Score Impact Data



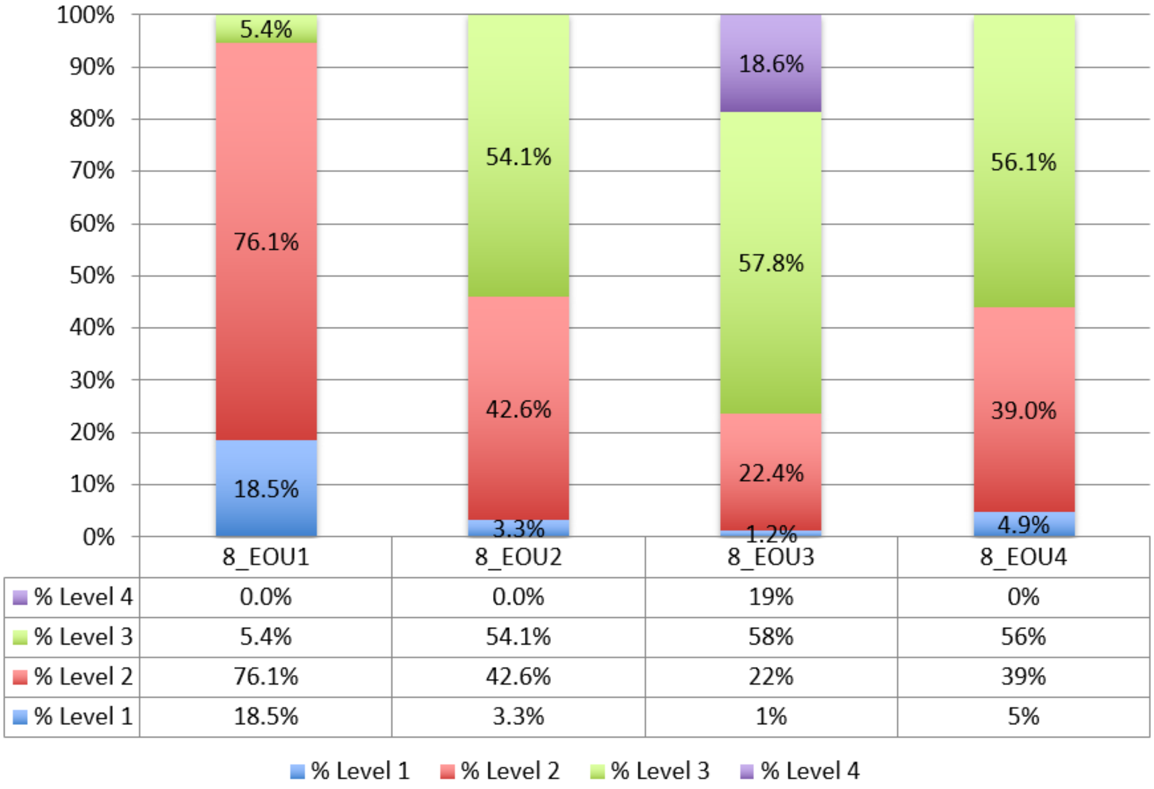
	5_EOU1	5_EOU2	5_EOU3	5_EOU4
🟪 % Level 4	0.0%	8.0%	2.6%	0.0%
🟩 % Level 3	24.1%	47.6%	62.6%	52.2%
🟥 % Level 2	62.4%	38.1%	34.4%	40.3%
🟦 % Level 1	13.5%	6.3%	0.4%	7.5%

Exhibit 31. SIPS Grade 8 RP67 Initial Cut Score Impact Data



	8_EOU1	8_EOU2	8_EOU3	8_EOU4
% Level 4	0.0%	0.0%	6%	0%
% Level 3	0.0%	21.3%	61%	34%
% Level 2	76.1%	62.3%	27%	59%
% Level 1	23.9%	16.4%	6%	7%

Exhibit 32. SIPS Grade 8 RP50 Initial Cut Score Impact Data



	8_EOU1	8_EOU2	8_EOU3	8_EOU4
% Level 4	0.0%	0.0%	19%	0%
% Level 3	5.4%	54.1%	58%	56%
% Level 2	76.1%	42.6%	22%	39%
% Level 1	18.5%	3.3%	1%	5%

Vertically Articulated (Smooth) RP67 and RP50 Cut Scores and Associated Impact Data

Exhibit 33(Exhibit 26 in the technical report) provides vertically articulated cut scores for RP67 and RP50 for each EOU and grade and Exhibit 34 (Exhibit 27 in the technical report) provides adjustments made to the initial cut scores to support the resulting vertical articulation.

Exhibit 33. Vertically Articulated SIPS RP67 and RP50 Cut Scores

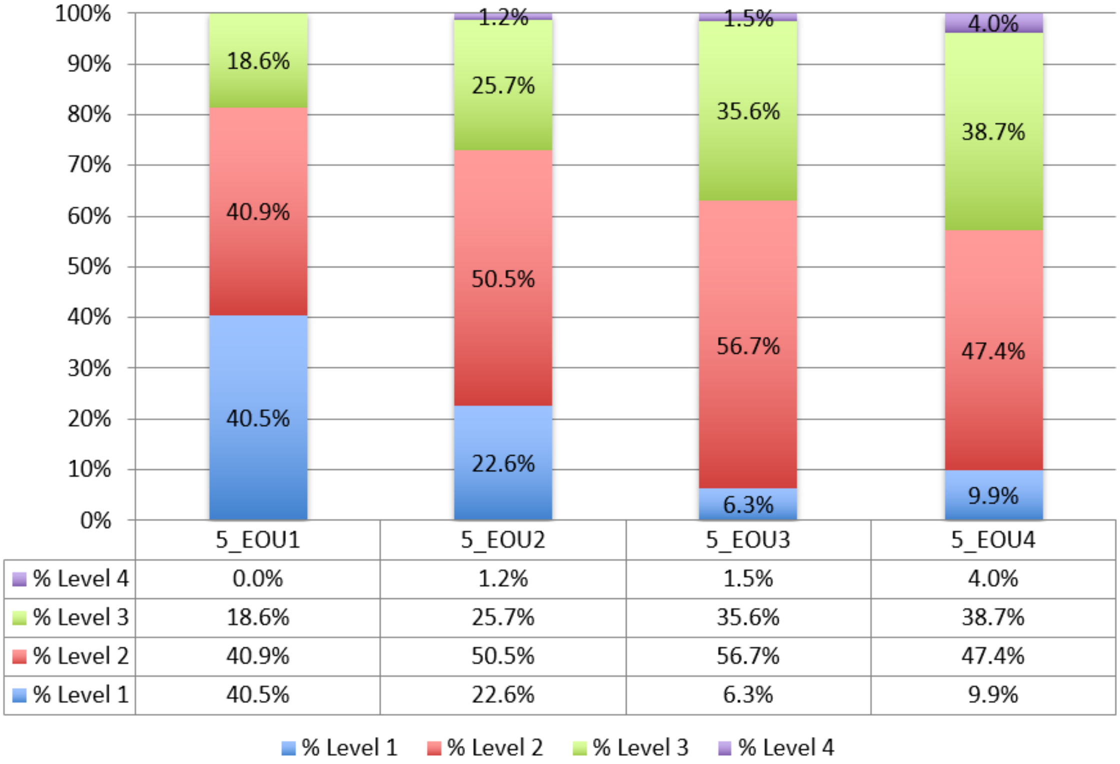
Grade & RP	EOU	Level2	Level3	Level4
Grade 5 RP67	EOU1	0.01	0.46	1.65
	EOU2	-0.39	0.47	1.46
	EOU3	-0.87	0.28	1.35
	EOU4	-0.60	0.26	1.26
Grade 5 RP50	EOU1	-0.37	0.11	1.15
	EOU2	-0.65	0.01	0.89
	EOU3	-1.04	-0.13	0.81
	EOU4	-1.03	-0.10	0.66
Grade 8 RP67	EOU1	-0.27	0.73	4.00
	EOU2	-0.44	0.42	1.38
	EOU3	-1.12	0.17	1.15
	EOU4	-1.06	0.07	3.06
Grade 8 RP50	EOU1	-0.51	0.67	4.00
	EOU2	-0.62	0.18	1.21
	EOU3	-1.64	-0.10	1.15
	EOU4	-1.73	-0.38	2.49

Exhibit 34. Vertical Articulation Adjustments to Cut Scores in Standard Error Units

Grade & RP	EOU	Level2	Level3	Level4
Grade 5 RP67	EOU1	0.0	-0.5	-1.5
	EOU2	0.5	0.0	0.0
	EOU3	1.5	0.0	-0.5
	EOU4	-0.5	-0.5	-2.0
Grade 5 RP50	EOU1	0.0	-0.5	-1.5
	EOU2	0.5	0.0	0.0
	EOU3	1.5	0.0	-0.5
	EOU4	-0.5	-0.5	-2.0
Grade 8 RP67	EOU1	0.0	-1.0	0.0
	EOU2	0.0	-0.5	-1.3
	EOU3	0.5	0.8	0.0
	EOU4	0.0	-0.6	0.0
Grade 8 RP50	EOU1	0.0	-0.5	0.0
	EOU2	0.8	0.0	-0.8
	EOU3	0.5	1.0	0.5
	EOU4	0.0	-0.5	0.0

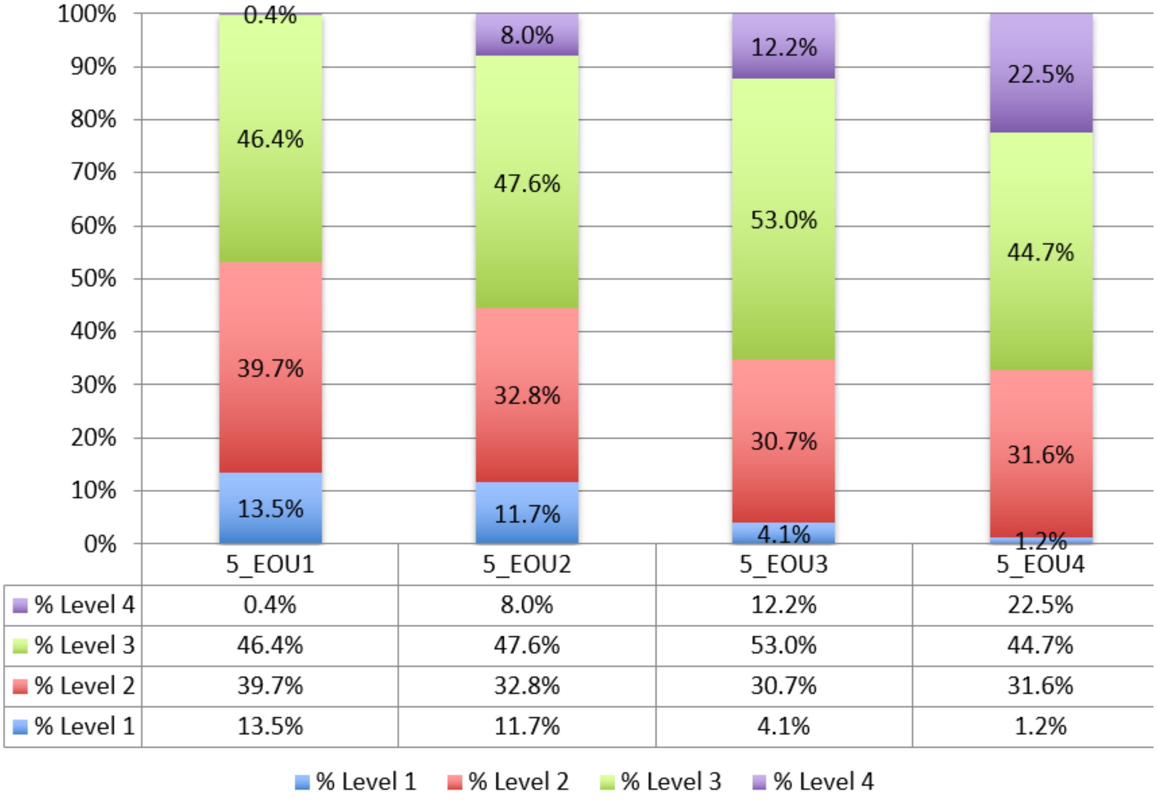
Exhibit 35and Exhibit 36 (Exhibits 28 and 29 in the technical report) provide impact data associated with the vertically articulated cut scores for the four grade 5 EOUs for RP67 and RP50, respectively. Exhibit 37 and Exhibit 38 (Exhibits 30 and 31 in the technical report) provide impact data associated with the vertically articulated cut scores for the four grade 8 EOUs for RP67 and RP50, respectively.

Exhibit 35. SIPS Grade 5 RP67 Smoothed Cut Score Impact Data



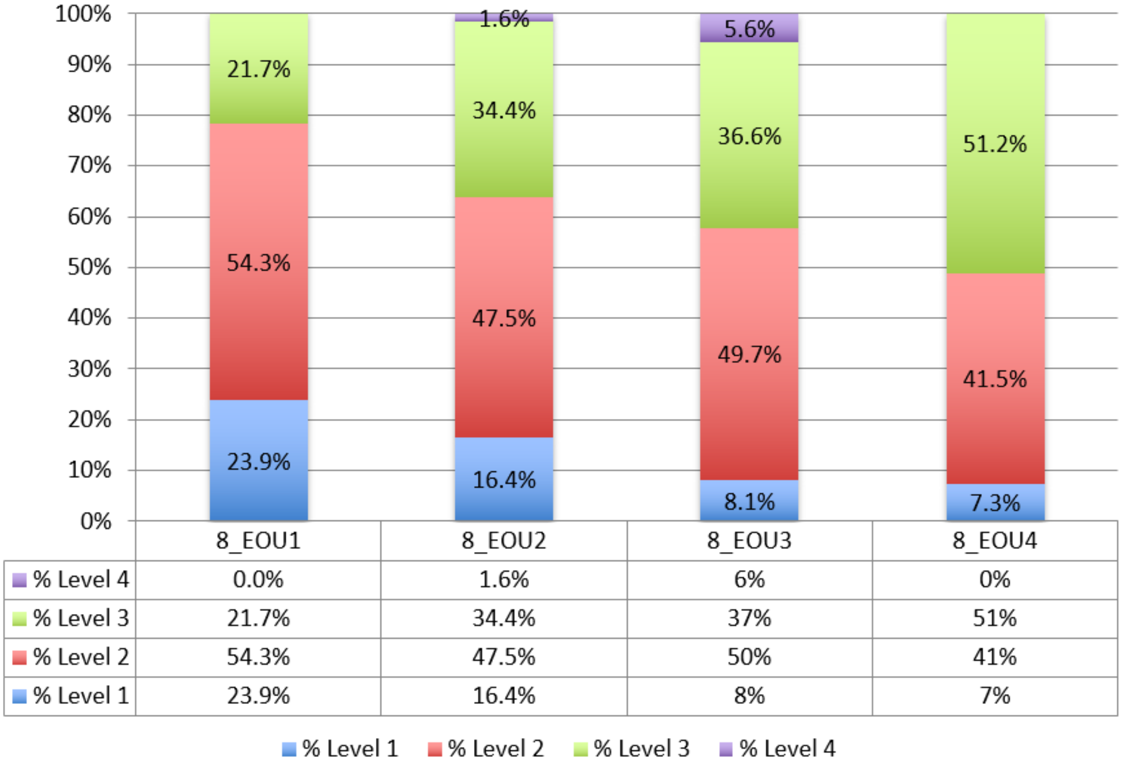
	5_EOU1	5_EOU2	5_EOU3	5_EOU4
🟪 % Level 4	0.0%	1.2%	1.5%	4.0%
🟩 % Level 3	18.6%	25.7%	35.6%	38.7%
🟥 % Level 2	40.9%	50.5%	56.7%	47.4%
🟦 % Level 1	40.5%	22.6%	6.3%	9.9%

Exhibit 36. SIPS Grade 5 RP50 Smoothed Cut Score Impact Data



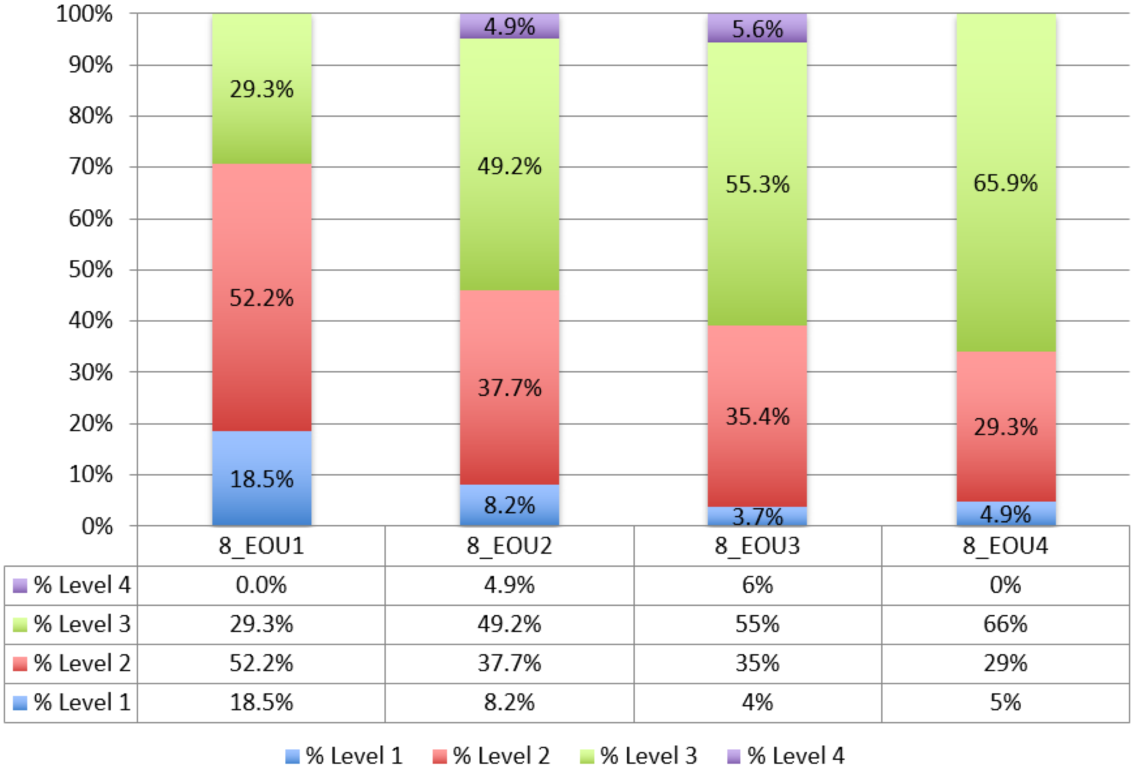
	5_EOU1	5_EOU2	5_EOU3	5_EOU4
🟪 % Level 4	0.4%	8.0%	12.2%	22.5%
🟩 % Level 3	46.4%	47.6%	53.0%	44.7%
🟥 % Level 2	39.7%	32.8%	30.7%	31.6%
🟦 % Level 1	13.5%	11.7%	4.1%	1.2%

Exhibit 37. SIPS Grade 8 RP67 Smoothed Cut Score Impact Data



	8_EOU1	8_EOU2	8_EOU3	8_EOU4
🟪 % Level 4	0.0%	1.6%	6%	0%
🟩 % Level 3	21.7%	34.4%	37%	51%
🟥 % Level 2	54.3%	47.5%	50%	41%
🟦 % Level 1	23.9%	16.4%	8%	7%

Exhibit 38. SIPS Grade 8 RP50 Smoothed Cut Score Impact Data



	8_EOU1	8_EOU2	8_EOU3	8_EOU4
🟪 % Level 4	0.0%	4.9%	6%	0%
🟩 % Level 3	29.3%	49.2%	55%	66%
🟥 % Level 2	52.2%	37.7%	35%	29%
🟦 % Level 1	18.5%	8.2%	4%	5%

Standard Setting Validity Evidence

A detailed summary of the validity evidence supporting the SIPS standard setting methodology is summarized in the full SIPS Standard Setting Technical Report using commonly cited forms of evidence from the measurement literature and peer review. We briefly summarize key elements of that evidence here.

The standard setting method is appropriate for the assessment of interest. 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.

The SMEs had backgrounds and training that qualified them and prepared them to make the required standard setting judgments. The SIPS 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 the technical report. 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 SME's 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. This provides evidence that the SMEs had backgrounds and training that qualified them and prepared them to make the required standard setting judgments.

The SMEs making the specified standard setting judgments understood the construct of interest and the assessments. As SIPS partners and developers of the SIPS EOUs, the SMEs understood the construct reflected by the PLDs. Each EOU's PLDs and prompts were developed using a backward design approach and PAD framework. The SMEs developed Student Profiles and PAD PLDs to reflect each other and the desired goals of each EOU'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 header "The Efficacy of SMEs' Prompt-PLD Alignments" provides evidence that the resulting Prompt-PLD alignment hypotheses were supported by data. This provides evidence that the SMEs making the ESS standard setting judgments understood the construct of interest and the assessments.

Summary

The pilot study had modest numbers and thus, the results should be considered preliminary and replicated when a more substantial field- or operational test administration is conducted. However, the data and evidence provided here, and the additional evidence detailed in the full SIPS Standard Setting Technical Report, provide provisional support for the validity of the estimated SIPS cut scores.

The full SIPS Standard Setting 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.

Section 8. Conclusions and Recommendations

8.1 Section Overview

In this section we offer reflections on each of the research questions we outlined earlier in Section 3. It is important to note at the outset that the current study was designed as a pilot study of a limited set of initial prototypes of each of the four end-of-unit (EOU) assessments administered to samples of 5th and 8th graders. As we did earlier in Section 6 of this report, we organized this section around the research questions that animated the general design of the pilot study. Our goal throughout was to collect information related to each of the guiding research questions to support, ultimately, the revisions to the prototypes and to learn more about how three-dimensional end-of-unit tasks could be used in practice by teachers.

8.2: Discussion related to RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning?

RQ1 focuses on collecting information related to whether it is appropriate to use the EOU assessments for measuring students' science learning. What we found is that while students were able to demonstrate science knowledge, there were some issues with the initial versions of the prototype assessments. In particular, if the plan was for each EOU to be administered in one class period, then substantial revisions will be needed to the EOUs when they are designed for a larger, more rigorous field study. Most tasks comprising each EOU took students more than 20 minutes to complete, which meant, for the most part, students could complete only two of the three EOU tasks in a class period.

While we expected to see some degree of missing responses from students, the number of missing responses by prompt (i.e., test item) was often much higher than we expected. Some of this may be because students simply ran out of time. We also found that a number of full classrooms skipped certain prompts or tasks within an EOU suggesting that there were certain science topics that students were not familiar with or were not able to engage with on the assessment as intended.

Overall, the prototype EOUs were challenging for students in our study. While there were two assessments for which students were able to achieve the highest possible points, for most assessments, students fell short. The prototype EOUs did provide information about where students stood with respect to the rubrics scoring scheme used, and they also allowed us to measure variation in students' achievement as we found prompts, tasks and EOU scores distributed across a range of performances. However, if these prototypes are presented as typical classroom assessments, where scores below 50% correct are often considered failing, then adjustments to the timing and difficulty levels of the prototypes will need to be made. These adjustments may be to the assessment tasks themselves and/or to how the scores are generated and reported.

Further study will be needed to determine how well the end-of-unit assessments were able to reflect students' opportunities to learn. Throughout the pilot study teachers reported on whether they taught a particular topic, but there was no information on how deeply they went into a topic or how the topic was taught. While we found some evidence of differences in scores based on if teachers indicated they taught a given concept or not, these differences did not always favor the students who received instruction related to this concept. However, this could be due to differences in the organization of classrooms, or to the degree or depth to which the concept was taught.

Finally, while teachers were able to provide scores on student work, further study is needed to determine the reliability of these scores, particularly if the goal is to compare students across classrooms. While data on scores from different teachers on the same set of students were collected, these data were limited, and we saw differences in the overall reliability of scoring depending on the prompt or task being scored.

Overall, we recommend collecting additional data using the revised EOUs to fully explore how well these assessments capture students' three-dimensional science learning. While the limited pilot study data indicate we were able to see differences between and among students, and that students were able to demonstrate their science knowledge, further information on how future iterations of the assessments will be used in the classroom should be gathered to guide additional explorations into the design and use of assessment tasks.

8.3: Discussion related to RQ2: How well did latent variable measurement models fit the empirical EOU assessment data?

As we noted earlier in Section 6, each of the prototype EOUs was scaled separately using the Rasch model, i.e., a one parameter IRT model. This modeling approach produced reasonable estimates of the items' difficulty parameters and student ability estimates. When using the Rasch model, item (or prompt) fit statistics are estimated which, in turn, proved useful for evaluating the measurement quality of the EOU prompts. Further, these fit statistics offered insights into the relationships among students' abilities and their responses to specific EOU prompts. More specifically, the fit statistics generated by the Rasch model measured the appropriateness of a prompt's difficulty relative to the students' abilities. Lower than expected values indicated that the prompt may have been too easy for our sample of students, leading to a high probability of correct responses. Conversely, a higher-than-expected value suggested that the prompt may be too difficult. This model fit information was shared with the designers of the prototypes as they worked to improve the measurement quality of the next iteration of the EOU assessments.

The Rasch model fit statistics allowed us to evaluate the fit of a prompt or task in a more general sense, i.e., reflecting how well a prompt performs across the entire student ability spectrum. The use of latent variable models, like the Rasch model, allowed us to identify prompts that performed erratically suggesting that students' performance on the prompt may have been influenced by factors other than the students' abilities, such as guessing or simply misunderstanding the prompt. With this approach we were also able to flag prompts that were too predictable and, therefore, did not discriminate sufficiently among students with different abilities. In sum, our approach to latent variable modeling provided rich information about the measurement characteristics of the prototype EOUs.

Unlike typical statewide assessment programs used for accountability purposes, IRT derived scale scores did not play a major role in this pilot, and thus were not computed based on a theta to scale score conversion formula. Instead, the theta estimates, along with the additional information about task difficulty were used to inform the Embedded Standards Setting (ESS) process. The ESS process is explained in detail in Section 7 of this report.

8.4: Discussion related to RQ3: Overall, what do the EOU assessment results tell us about students' science learning?

As part of the investigation into this research question we examined the relationship between student scores and additional variables, including gender, prior ELA and Math learning, and curricular materials. We found that three out of the eight EOU assessments had statistically significant differences based on gender (in favor of females), but the sample size for this was fairly low and so further study is needed to draw conclusions. We also found that scores on the assessment tended to increase as prior ELA and mathematics levels increased. While this could indicate a dependency between ELA and math ability and the science assessment, there is often overlap between the science practices and ELA skills (e.g., communicating information) as well as the science practices and mathematical practices (e.g., problem solving). Therefore, more exploration is warranted to determine if there is too much of a dependency between skills.

Our analysis found statistically significant differences between students who used different curricular materials at the 5th grade (and for the Grade 8 EOU 2 assessment). However, without further investigation between the differences among the different curricula materials it is not clear how to interpret these differences. Further investigation to determine if the differences are due to desirable characteristics (e.g., if different curricula cover different aspects on the assessment, we would expect different scores) or to characteristics we would want to address in the assessment (e.g., if different curricula use different representations and the assessment is too closely aligned to one specific representation).

We found that there were high correlations, ranging from Spearman's rho of 0.4 to 0.8, across the assessments. While each assessment covered a different set of NGSS PEs, there was overlap in the science practices and cross-cutting concepts across some of the units. To further explore these correlations, it would help to better understand the curricula and instructional contexts for each of the classes of students.

Cross-EOU Growth

The pilot study sample was modest—not all students in a grade took all four EOUs. Nonetheless, 64 5th graders and 21 8th students took all four EOUs. The cross-EOU performance level profiles for these matched cases are provided in Tables 8.1 and 8.2 for grades 5 and 8, respectively.

We assert that an increase in performance level from EOU to EOU reflects growth because (a) each EOU has a unique set of performance level descriptors (PLDs) that form the basis for the task-PLD alignments and cut score estimations and (b) each level of each EOU's PLDs reflects a common expectation for student performance relative to the EOU's instructional unit. For example, PLD level 3 reflects the minimal performance expected of all students following each instructional unit. Thus, each level is qualitatively comparable across the four EOUs.

The bold rows of Exhibit 39 and Exhibit 40 indicate profiles obtained by 5 or more matched cases. For grade 5, these profiles are 1222, 1223, 2222, 2223, and 2323. We observe that most of these matched profiles reflect some growth across EOUs. That is, the profile 1222 reflects growth from EOU1 to EOU2 that is maintained for the remaining EOUs. Profile 1223 reflects growth from EOU1 to EOU2, maintenance from EOU2 to EoOU3, and growth from EOU3 to EOU4. And profile 2223 reflects maintenance of performance from EOU1 to EOU3 and growth from EOU3 to EOU4. The remaining two grade 5 profiles reflect maintenance of performance (2222) and mixed performances (2323).

One of the two profiles with at least 5 cases for grade 8 reflects some growth and the other reflects mixed results. Profile 2233 reflects maintenance of performance from EOU1 to EOU2, growth from EOU2 to EOU3, and maintenance of that growth from EOU3 to EOU4. Profile 2232 reflects maintenance of performance from EOU1 to EOU2, growth from EOU2 to EOU3, and a decline in performance from EOU3 to EOU4.

Exhibit 39. EOU Performance Level Profiles, Grade 5

Profile	N	Percent
1121	2	3.13
1122	2	3.13
1221	1	1.56
1222	**5**	**7.81**
1223	**6**	**9.38**
2121	2	3.13
2122	3	4.69
2222	**10**	**15.63**
2223	**15**	**23.44**
2232	2	3.13
2233	2	3.13
2322	2	3.13
2323	**6**	**9.38**
2333	4	6.25
3222	1	1.56
3333	1	1.56
Total	64	100.00

Note: Rows in bold reflect profiles with 5 or more cases

Exhibit 40. EOU Performance Level Profiles, Grade 8

Profile	N	Percent
1121	1	4.76
1122	3	14.29
1222	1	4.76
2222	1	4.76
2232	**6**	**28.57**
2233	**8**	**38.10**
2333	1	4.76
Total	21	100.00

Note: Rows in bold reflect profiles with 5 or more cases

In summary, the calibration of each level of the PLDs to a common goal relative to the instructional unit supports the measurement of cross-EOU growth. The current study had a limited number of cases from which to evaluate the efficacy of the proposed growth metric—change in performance level from EOU to EOU. It is recommended that the efficacy of this approach be further evaluated when a more robust data set is available.

8.5: Discussion related to use and reporting of the EOU results

In the case of the pilot study, teachers scored their own students, and thus had access to student level data. However, no additional data were reported back to teachers about their students, and additional guidance on how this information could be used to inform subsequent units of instruction were not provided. This sub-section focuses on how data from the EOUs could be used to report back to teachers. First, we describe two different reporting metrics that might be used to summarize individual student performance for each EOU and aggregated across EOUs.

Performance Level

Students receive a reportable performance level based on each administered EOU and Exhibit 39 and Exhibit 40 illustrate individual student profiles that may be used for reporting individual student results from multiple EOUs. Profiles can be summarized at the individual student level by reporting performance level profiles in both tabular and graphical formats.

Performance level results can also be reported at the group level for each EOU. Group level performance level results are typically reported as the percentage of students in the group attaining each level. Multiple EOU administrations can be reported at the group level by reporting the percentage of students in the group achieving each level on each EOU in both tabular and graphical (e.g., stacked bar chart) formats.

Performance level reports for multiple EOU administrations over the course of the year can be supported via Performance Level Profiles. For example, 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 41 to aggregate the four performance levels into a summative Proficiency Determination.

Exhibit 41. 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.

We adapt 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 Exhibit 42 and Exhibit 43. 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 42. EOU Performance Level Profiles and Possible Summative Performance Level Based on Rubric, Grade 5

Profile	N	Percent	Possible Summative Performance Level
1121	2	3.13	2
1122	2	3.13	2
1221	1	1.56	2
1222	**5**	**7.81**	**2**
1223	**6**	**9.38**	**2**
2121	2	3.13	2
2122	3	4.69	2
2222	**10**	**15.63**	**2**
2223	**15**	**23.44**	**2**
2232	2	3.13	2
2233	2	3.13	2
2322	2	3.13	2
2323	**6**	**9.38**	**2**
2333	4	6.25	2
3222	1	1.56	2
3333	1	1.56	3
Total	64	100.00	

Note: Rows in bold reflect profiles with 5 or more cases

Exhibit 43. EOU Performance Level Profiles and Possible Summative Performance Level Based on Rubric, Grade 8

Profile	N	Percent	Possible Summative Performance Level
1121	1	4.76	2
1122	3	14.29	2
1222	1	4.76	2
2222	1	4.76	2
2232	**6**	**28.57**	**2**
2233	**8**	**38.10**	**2**
2333	1	4.76	2
Total	21	100.00	

Note: Rows in bold reflect profiles with 5 or more cases

A PLD-based SIPS Score

Recall that we do not have a common scale across EOUs in a grade. However, performance-level based scores can be reported for each EOU and aggregated across EOUS to support within-grade, cross-EOU score interpretation 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 and averaging the ordinality of the level across EOUs provides an average student performance based on the four comparable EOU-specific targets.

That is, each EOU performance level may be translated to a numerical value as follows:

- Level 1 = 1,
- Level 2 = 2,
- Level 3 = 3,
- Level 4 = 4

The average of these numerical values provides an aggregate score summarizing the profile of EOUs taken by a student.

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.

PLD-based scores can be averaged on individual student reports to summarize multiple EOU administrations. Group level scores can be reported as an average of the individual students' PLD-based scores.

SIPS Summative Reporting

End-of-Year summative performance can be summarized using either of the two reporting metrics described in this section—performance level profiles and PLD-based scores. First, the rubrics illustrated in Exhibit 42 associate the performance level profiles from the four annual EOU administrations with a single summative performance level that may be used for federal accountability purposes.

Second, the SIPS PLD-based scale can be used to associate the four annual EOU administrations with a single summative PLD-based score for federal accountability purposes. Specifically, after translating the performance levels from the four annual EOU administration into ordinal values, the average of the four values can be translated to a summative score and performance level using a rubric such as the following exemplar:

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

Exhibit 44 and Exhibit 45 demonstrate these two methods of summative reporting.

Exhibit 44. EOU Performance Level Profiles and Possible Summative Performance Levels Based on Rubric and Averages, Grade 5

Profile	N	Percent	Possible Summative Performance Level	
			Rubric	Averages
1121	2	3.13	2	1
1122	2	3.13	2	1
1221	1	1.56	2	1
1222	**5**	**7.81**	**2**	**2**
1223	**6**	**9.38**	**2**	**2**
2121	2	3.13	2	1
2122	3	4.69	2	2
2222	**10**	**15.63**	**2**	**2**
2223	**15**	**23.44**	**2**	**2**
2232	2	3.13	2	2
2233	2	3.13	2	2
2322	2	3.13	2	2
2323	**6**	**9.38**	**2**	**2**
2333	4	6.25	2	3
3222	1	1.56	2	3
3333	1	1.56	3	3
Total	64	100.00		

Note: Rows in bold reflect profiles with 5 or more cases

Exhibit 45. EOU Performance Level Profiles and Possible Summative Performance Levels Based on Rubric and Averages, Grade 8

Profile	N	Percent	Possible Summative Performance Level	
			Rubric	Averages
1121	1	4.76	2	1
1122	3	14.29	2	1
1222	1	4.76	2	2
2222	1	4.76	2	2
2232	**6**	**28.57**	**2**	**2**
2233	**8**	**38.10**	**2**	**2**
2333	1	4.76	2	3
Total	21	100.00		

Note: Rows in bold reflect profiles with 5 or more cases

Using EOU Results to Inform Subsequent Units of Instruction

Educators may use the PLDs to inform subsequent units of instruction. That is, educators should review the descriptor for a student's current level of performance on an EOU—this tends to describe the range of performance for students achieving that level. However, by examining the next higher level, the educator can observe the skills the student needs to acquire to advance to that higher level. While the subsequent unit of instruction may be quite different, the information obtained from such a review may provide insight into students' strengths and weaknesses to inform the next unit of instruction.

8.6: Discussion Related to Using the Data for Revisions to Tasks and PLDs

Revisions to Tasks

Across all four EOU assessments in both grades 5 and 8, prompts were flagged for revision based on their statistical data. The design team used that information, along with the qualitative information from teacher reviews to revise the tasks. All prompts were revisited to determine if revisions would be made, and the rationale behind each revision was documented. In general, revisions to the prompts included scenario language and/or graphic revisions, stem changes and/or graphic changes, and clarifying language within task sets to support students' understanding and production of complete and accurate evidence of their science learning. Revisions to the scoring rubrics included clarification of language, differentiating score point criteria, opportunities for holistic scoring (i.e., combining parts of a prompt), as well as updating to reflect changes to related prompts. All exemplar responses included in the SIPS EOU Scoring Guides were updated as well to reflect any changes to the corresponding prompts.

In response to the over-burdensome administration times reported by the participating educators, redundant prompts across tasks assessing similar concepts in both grades 5 and 8 were removed. Another revision, for example, involved the inclusion of partial graphs and models to scaffold the student response rather than the provision of a blank, undefined response space. Attention was given to the original number of total score points and how this total was impacted based on revisions to ensure sufficient representation of the KSAs on each task. An effort was made to maintain as much as possible the original number of prompt and task total points so as not to reduce the number of points possible too drastically.

Revisions to PLDs

The SIPS Performance Level Descriptors (PLDs) are hypothesized descriptions of the expected knowledge, skills, and other attributes (KSAs) associated with each performance level for each SIPS end-of-unit assessment. The KSAs associated with each level are EOU specific and are based on the desired outcomes of instruction aligned with each EOU's associated instructional curricula.

PLDs are hypothesized learning progressions articulated across the performance levels within each grade and SIPS instructional unit. Because they are hypothesized, we expect to refine them based on empirical data support, or lack thereof, for the various hypothesized PLD evidence statements. In this case, the empirical data resulted from the 2022-2023 SIPS pilot study in Grades 5 and 8.

PLD refinements should not be made without support from robust empirical data that is based on a sufficiently large and representative sample of the target population. When data are sufficient, then specific PLD evidence statements may be refined based on Embedded Standard Setting (ESS) results.

Specifically, the ESS list of inconsistent items provides information about items aligned to PLD evidence statements and levels such that SIPS pilot data does not support the hypothesized Item-PLD alignment.

Even when data are sufficient, changes to PLDs should be made judiciously—evidence based on a single item should not result in a change to the PLDs, especially evidence resulting from a pilot study in an inaugural year of a testing program. That is, we know that various factors such as opportunity to learn or uneven implementation of curricula will result in shifts in item difficulty over the first few years of a testing program and thus, reliable data supporting PLD refinement will not emerge until after the first operational year or two of a testing program. Even then, multiple sources of information should inform PLD refinement. Thus, several items reflecting a PLD evidence statement that consistently point to a common PLD refinement after data has stabilized may support a refinement to the PLDs but a single item in the inaugural year of a testing program does not.

In the case of the SIPS pilot study, neither of the desired empirical data characteristics are present—we have neither a sufficiently large nor representative sample of the target population. Therefore, the pilot study data may suggest refinements that can be considered when sufficient data are available, but we do not yet consider the data sufficient to refine PLDs at this time.

8.7 Conclusion

Overall, we found great potential in the prototype EOU assessment tasks for measuring students' science proficiency even though the set of prototypes investigated had issues with respect to their length and difficulty that would not make them appropriate for use as a classroom summative assessment in the current form.

Results from the pilot study suggested that the content development processes generally succeeded in presenting grade-and age-appropriate content to students in our sample. However, the lack of a sufficiently large sample and concerns about the representativeness of the sample warrant caution. Nevertheless, the empirical and qualitative review processes yielded rich information for improving the quality of the EOU tasks that was consistent with the larger SIPS project's theory of action which was to design assessment tasks that elicit evidence of students' science learning, and inform teaching and learning at specific points along an instructional pathway.

Given the deliberate iterative nature of this pilot study, further investigations will be needed to ensure that the next iteration(s) of the EOUs can be used by teachers to inform and improve the teaching and learning of science in their elementary and secondary school classrooms irrespective of variations in curricula and instructional design.

The method described for measuring growth and for reporting individual and summative EOU results are supported by the intentional development of PLDs that are comparable in terms of target performance across EOUs. The methods described for reporting individual EOU and summative cross-EOU end-of-year results should be investigated further when more robust data are available.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for educational and psychological testing*. Amer Educational Research Assn.

Bock, R.D., Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika 46, 443–459 (1981). https://doi.org/10.1007/BF02293801.

Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, 48(6), 1-29. doi:10.18637/jss.v048.i06.

Cohen, J. (1960). *A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 20 (1): 37–46.*

Cronbach L J. (1951). Coefficient Alpha and the internal structure of tests. Psychometrika 16, pp. 297-334.

Embretson, S.E., Reise, S.P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates Publishers.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159—74.

Lewis, D. & Cook, R. (2020). Embedded standard setting: Aligning standard-Setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice*, 39(1), pp. 8-21.

Lewis, D. M., Mitzel, H. C., Mercado, R. & Schulz, M. (2012). The Bookmark Standard Setting Procedure. Chapter in Setting Performance Standards: Concepts, Methods, and Perspectives, Second Edition. (ed: G. J. Cizek), Lawrence Erlbaum.

McTighe, J. & Wiggins, G. (1998). Understanding by Design. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Mislevy, R.J. (2007). Validity by design. *Educational Researcher, 36*(8), 463-469.

Mislevy, R. J., & Haertel, G. D. (2006). *Implications of evidence-centered design for educational testing. Educational Measurement: Issues and Practice,* 25(4), pp. 6–20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Lawrence Erlbaum.

Mislevy, R.J., Haertel, G., Riconscente, M., Rutstein, D.W. & Ziker, C. (2017). *Assessing model-based reasoning using evidence-centered design: A suite of research-based design patterns*. Cham, Switzerland, Springer.

National Research Council (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

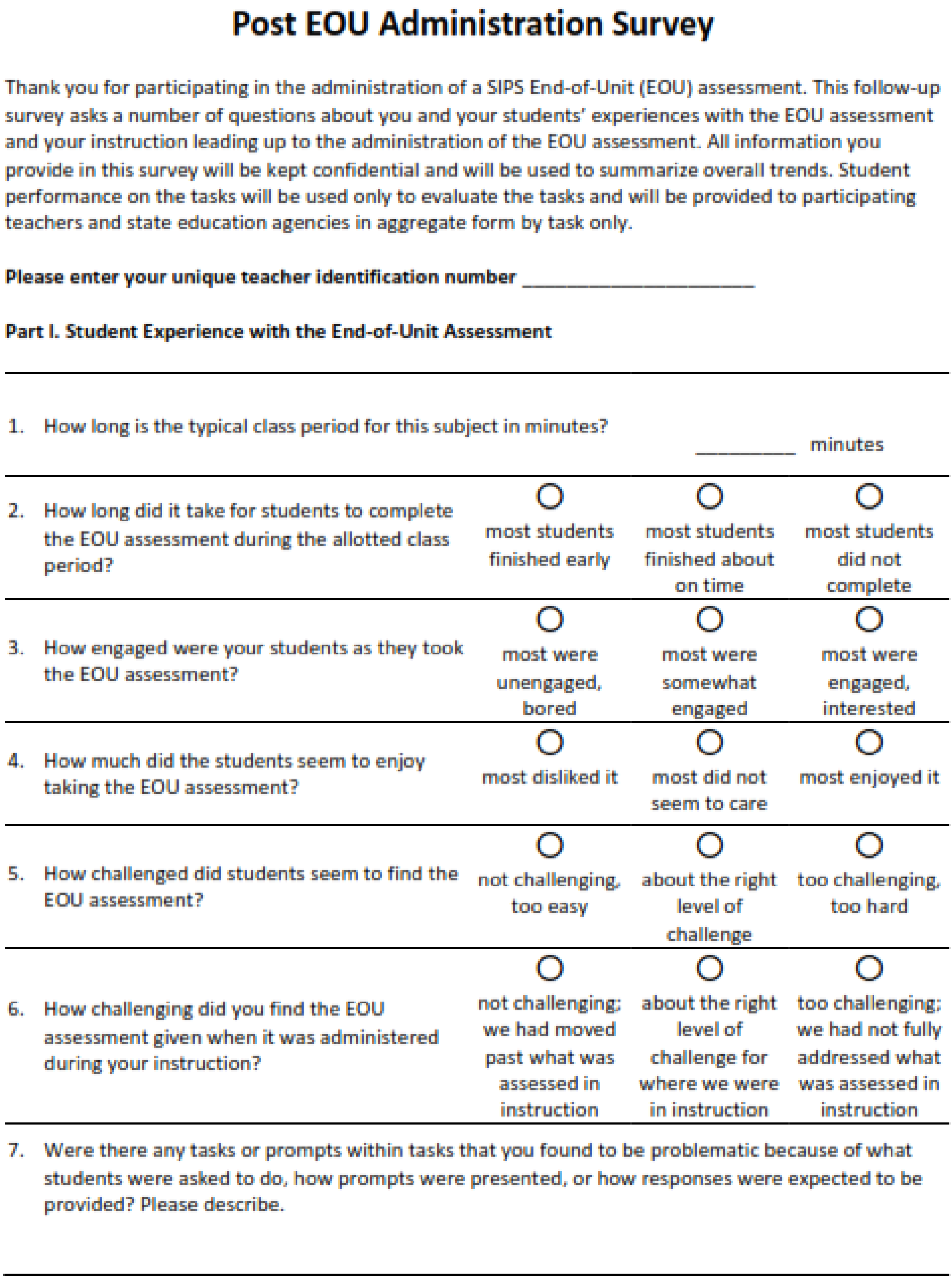
NGSS Lead States. (2013). Next Generation Science Standards: For states, by states. National Academies Press.

NGSS Lead States. 2013. *Next Generation Science Standards: For States, By State*s. Washington, DC: The National Academies Press.

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. Science, 340, 320-323.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometric Society, New York.

Appendix A. Post-Administration Survey



Post EOU Administration Survey

Thank you for participating in the administration of a SIPS End-of-Unit (EOU) assessment. This follow-up survey asks a number of questions about you and your students' experiences with the EOU assessment and your instruction leading up to the administration of the EOU assessment. All information you provide in this survey will be kept confidential and will be used to summarize overall trends. Student performance on the tasks will be used only to evaluate the tasks and will be provided to participating teachers and state education agencies in aggregate form by task only.

Please enter your unique teacher identification number _______________________

Part I. Student Experience with the End-of-Unit Assessment

1. How long is the typical class period for this subject in minutes? __________ minutes

2. How long did it take for students to complete the EOU assessment during the allotted class period?	○ most students finished early	○ most students finished about on time	○ most students did not complete
3. How engaged were your students as they took the EOU assessment?	○ most were unengaged, bored	○ most were somewhat engaged	○ most were engaged, interested
4. How much did the students seem to enjoy taking the EOU assessment?	○ most disliked it	○ most did not seem to care	○ most enjoyed it
5. How challenged did students seem to find the EOU assessment?	○ not challenging, too easy	○ about the right level of challenge	○ too challenging, too hard
6. How challenging did you find the EOU assessment given when it was administered during your instruction?	○ not challenging; we had moved past what was assessed in instruction	○ about the right level of challenge for where we were in instruction	○ too challenging; we had not fully addressed what was assessed in instruction

7. Were there any tasks or prompts within tasks that you found to be problematic because of what students were asked to do, how prompts were presented, or how responses were expected to be provided? Please describe.

8. Was there anything you particularly liked about the EOU assessment and hope to adopt in the future?

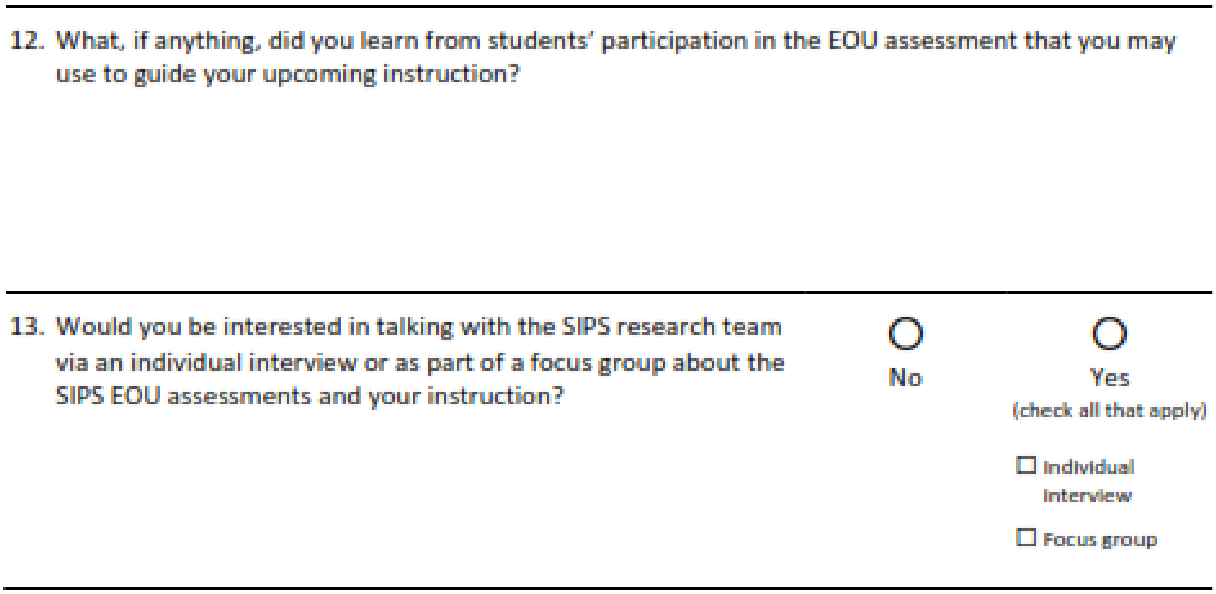
9. Was there anything you particularly disliked about the EOU assessment and would prefer not to use again?

Part 2. Instruction Immediately Prior the EOU Assessment Administration
Please help us understand where this EOU assessment fit in relation to what you had been teaching prior to its administration.

1. What science curriculum and unit did you teach prior to giving the EOU assessment (e.g., OpenSciEd Unit 6.1: Light & Matter)?			
2. When did you start teaching this unit (provide an approximate date, e.g., 8/29/2022)			
3. How many class periods were planned to complete this unit?			
4. Did you finish this unit before administering the EOU assessment?		○ Yes	○ No, we were about ________ % through the unit
5. Do you typically administer an assessment at the end of a science instructional unit? (For questions 6-10, please compare the EOU assessment to the type of assessment you typically administer at the end of a science unit)		○ Yes	○ No, (please skip to question 11)
6. Similarity in terms of length to administer	○ EOU takes less time	○ EOU takes about the same time	○ EOU takes more time
7. Similarity in terms of scenario, context, or situation	○ Typical assessment does not use a scenario	○ Both use a similar type of scenario	○ EOU uses a ☐ richer scenario than the typical ☐ weaker scenario than the typical

8. Similarity in terms of question type	○ Typical assessment uses only selected-response questions	○ Both use a combination of selected-response and open-ended questions	○ Typical assessment uses only open-ended questions
9. Similarity in terms of the complexity demands on students' reasoning	○ Typical assessment demands more complex reasoning	○ Both demand a similar degree of reasoning complexity	○ EOU demands more complex reasoning
10. Similarity in terms of time it would take to score student responses	○ EOU would take less time to score	○ EOU would take about the same time to score	○ EOU would take more time to score

11. For each of the following, please indicate:	This was a target of your current or most recent science instructional unit		This appeared to be a target of the EOU	
	Yes	No	Yes	No
- Develop a model to describe matter	○	○	○	○
- Plan aspects of an investigation to identify materials based on their properties	○	○	○	○
- Use observations and measurements as evidence to explain the identification of a material	○	○	○	○
- Use observations of the properties of matter to identify a substance	○	○	○	○
- Use observations and measurements as evidence to explain that matter is made of particles too small to be seen	○	○	○	○
- Use a model to represent amounts, relationships, or scales to classify materials based on their properties	○	○	○	○



12. What, if anything, did you learn from students' participation in the EOU assessment that you may use to guide your upcoming instruction?

13. Would you be interested in talking with the SIPS research team via an individual interview or as part of a focus group about the SIPS EOU assessments and your instruction?

○ No

○ Yes
(check all that apply)

□ Individual interview

□ Focus group

Appendix B. Data Tables for RQ1 (Section 6.2.2 Analyses)

RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning?

6.2.2: Are there patterns in the prompts that students skip?

Table 6.2.2.1: Percent of Students who did not Respond by Prompt for Grade 5, EOU 1

Grade	EOU	Task	Prompt	Missing
5	EOU1	1	1	4.72
5	EOU1	1	2	8.26
5	EOU1	1	3	9.73
5	EOU1	1	4	7.67
5	EOU1	2	1 - AB	13.27
5	EOU1	2	1 - C	18.37
5	EOU1	2	2	15.34
5	EOU1	2	3	16.81
5	EOU1	3	1	14.16
5	EOU1	3	2 - AB	13.27
5	EOU1	3	3	18.58

Table 6.2.2.2: Percent of Students who did not Respond by Prompt for Grade 5, EOU 2

Grade	EOU	Task	Prompt	Missing
5	EOU2	1	1	3.59
5	EOU2	1	2	7.40
5	EOU2	1	3	4.65
5	EOU2	2	1 - A	2.96
5	EOU2	2	1 - B	6.55
5	EOU2	2	2	4.65
5	EOU2	2	3 - A	4.65
5	EOU2	2	3 - B	6.34
5	EOU2	3	1	5.50
5	EOU2	3	2 - A	5.07
5	EOU2	3	2 - B	4.86
5	EOU2	3	3	6.13

Table 6.2.2.3: Percent of Students who did not Respond by Prompt for Grade 5, EOU 3

Grade	EOU	Task	Prompt	Missing
5	EOU3	1	1 - ABCD	2.93
5	EOU3	1	1 - E	4.99
5	EOU3	1	2 - A	3.52
5	EOU3	1	2 - B	6.16
5	EOU3	1	2 - C	5.87
5	EOU3	1	3 - A	6.74
5	EOU3	1	3 - B	7.62
5	EOU3	1	3 - C	9.38
5	EOU3	2	1 - A	2.93
5	EOU3	2	1 - BC	3.81
5	EOU3	2	2	3.81
5	EOU3	2	3 - A	5.28
5	EOU3	2	3 - B	7.04
5	EOU3	2	4 - A	9.68
5	EOU3	2	4 - BC	9.68
5	EOU3	3	1 - A	12.61
5	EOU3	3	1 - B	10.26
5	EOU3	3	2 - AB	10.85
5	EOU3	3	3 - AB	12.02
5	EOU3	3	3 - C	11.44
5	EOU3	3	4 - AB	12.02

Table 6.2.2.4: Percent of Students who did not Respond, by Prompt for Grade 5, EOU 4

Grade	EOU	Task	Prompt	Missing
5	EOU4	1	1 - A	13.43
5	EOU4	1	1 - B	16.79
5	EOU4	1	1 - C	18.47
5	EOU4	1	1 - D	17.75
5	EOU4	1	2	17.03
5	EOU4	2	1	13.67

5	EOU4	2	2 - A	15.59
5	EOU4	2	2 - BC	18.47
5	EOU4	2	3	19.66
5	EOU4	3	1 - A	12.95
5	EOU4	3	1 - B	15.35
5	EOU4	3	2 - AB	15.59
5	EOU4	3	2 - C	18.23
5	EOU4	3	3	20.62

Table 6.2.2.5: Number of Students Without Responses by Task and Educator for Grade 5, EOU 1

Teacher	Scores on all Tasks	Missing Score on Task 1 Only	Missing Score on Task 2 Only	Missing Score on Task 3 Only	Missing Score on Task 1 and Task 2	Missing Score on Task 1 and Task 3	Missing Score on Task 1 and Task 3	Missing all Scores	Total
C1	0	0	0	0	0	0	0	1	1
C10	3	1	0	0	1	4	1	1	11
C11	13	0	0	0	0	0	0	0	13
C12	12	2	1	0	0	0	0	0	15
C14	24	1	0	0	0	0	0	0	25
C15	28	1	0	0	1	0	0	0	30
C16	30	0	0	0	0	0	0	0	30
C17	14	1	0	0	0	0	0	0	15
C18	25	0	0	0	0	0	0	0	25
C19	12	0	0	2	1	0	1	0	16
C4	11	1	0	3	0	1	1	1	18
C5	2	0	3	4	2	0	0	2	13
C6	24	0	0	0	0	0	0	0	24
C66	10	1	0	0	1	0	1	1	14
C67	0	0	0	0	0	0	8	0	8
C68	0	0	0	11	0	1	0	0	12
C69	0	1	0	0	1	2	0	0	4
C7	10	0	0	3	0	0	2	0	15
C70	0	5	0	0	0	0	0	1	6

C71	0	0	20	0	0	0	1	0	21
C77	0	0	0	0	0	0	0	1	1
C8	5	2	0	3	0	1	2	4	17
C9	5	0	1	1	0	0	0	0	7
Total	228	16	25	27	7	9	17	12	341

Table 6.2.2.6: Number of Students without Responses by Task and Educator for Grade 5, EOU 2

Teacher	Scores on all Tasks	Missing Score on Task 1 Only	Missing Score on Task 2 Only	Missing Score on Task 3 Only	Missing Score on Task 1 and Task 2	Missing Score on Task 1 and Task 3	Missing Score on Task 1 and Task 3	Missing all Scores	Total
C10	5	2	0	0	0	0	3	1	11
C11	12	0	1	0	0	0	1	0	14
C12	14	0	0	0	1	0	0	0	15
C14	25	0	0	0	0	0	0	0	25
C15	29	0	0	1	0	0	0	0	30
C16	30	0	0	0	0	0	0	0	30
C17	16	0	0	1	2	0	1	1	21
C18	22	0	0	0	0	0	0	0	22
C20	11	1	2	1	0	0	1	0	16
C21	13	3	1	2	0	0	0	0	19
C22	13	0	2	1	1	1	0	1	19
C23	17	0	1	1	0	1	0	0	20
C24	15	0	1	1	0	0	0	0	17
C25	13	1	0	0	0	0	0	0	14
C26	5	0	3	2	0	0	0	0	10
C27	7	0	3	0	0	0	0	0	10
C28	11	1	1	2	0	1	0	0	16
C29	7	1	0	0	0	0	0	0	8
C31	6	0	0	1	1	0	0	0	8
C33	16	1	0	0	0	0	0	1	18
C36	11	1	2	0	1	0	0	0	15
C37	18	0	0	0	0	0	0	0	18

C4	12	1	2	0	1	0	2	1	19
C6	22	2	0	0	0	0	0	0	24
C66	12	2	1	0	0	0	0	0	15
C7	10	2	0	2	0	1	0	0	15
C8	17	0	0	0	0	0	0	0	17
C9	0	1	0	1	1	1	0	3	7
Total	389	19	20	16	8	5	8	8	473

Table 6.2.2.7: Number of Students without Responses by Task and Educator for Grade 5, EOU 3

Teacher	Scores on all Tasks	Missing Score on Task 1 Only	Missing Score on Task 2 Only	Missing Score on Task 3 Only	Missing Score on Task 1 and Task 2	Missing Score on Task 1 and Task 3	Missing Score on Task 1 and Task 3	Missing all Scores	Total
C10	1	1	0	3	1	1	0	4	11
C14	24	0	0	0	0	0	0	0	24
C15	13	4	1	11	0	0	1	0	30
C16	25	0	1	4	0	0	0	0	30
C17	19	2	0	0	0	0	0	0	21
C19	13	0	0	0	0	0	1	1	15
C22	14	3	0	0	0	1	0	2	20
C23	11	6	0	1	1	1	0	1	21
C24	16	0	1	0	0	0	0	0	17
C25	12	1	0	0	0	0	1	0	14
C26	8	0	0	1	1	0	0	0	10
C27	9	0	0	1	0	0	0	0	10
C29	8	0	0	0	0	0	0	0	8
C5	9	0	0	3	0	0	3	0	15
C6	24	0	0	0	0	0	0	0	24
C66	10	0	1	3	0	0	1	0	15
C7	10	2	0	1	0	1	1	0	15
C77	11	2	2	4	1	0	0	0	20
C81	19	0	0	1	1	0	0	0	21
Total	256	21	6	33	5	4	8	8	341

Table 6.2.2.8: Number of Students without Responses by Task and Educator for Grade 5, EOU 4

Teacher	Scores on all Tasks	Missing Score on Task 1 Only	Missing Score on Task 2 Only	Missing Score on Task 3 Only	Missing Score on Task 1 and Task 2	Missing Score on Task 1 and Task 3	Missing Score on Task 1 and Task 3	Missing all Scores	Total
C10	5	0	3	0	0	1	1	1	11
C11	9	3	1	0	0	0	0	0	13
C14	24	1	0	0	0	0	0	0	25
C15	28	0	1	1	0	0	0	0	30
C16	29	1	0	0	0	0	0	0	30
C17	17	2	2	0	0	0	0	0	21
C18	25	0	0	0	0	0	0	0	25
C19	14	0	0	0	0	0	2	0	16
C20	0	2	0	1	1	0	1	10	15
C21	15	1	1	1	0	1	0	0	19
C22	8	2	2	2	0	1	3	1	19
C23	5	6	2	3	0	2	1	1	20
C24	13	2	0	1	1	0	0	0	17
C25	6	2	1	2	0	1	1	1	14
C26	8	0	1	0	1	0	0	0	10
C27	6	0	1	2	0	0	1	0	10
C28	6	4	0	0	1	2	1	2	16
C29	6	0	2	0	0	0	0	0	8
C5	7	6	0	1	0	2	0	2	18
C67	0	0	0	9	0	1	0	0	10
C68	0	0	0	0	0	9	0	1	10
C69	0	2	0	0	1	0	0	1	4
C70	0	5	0	0	0	1	0	0	6
C71	0	0	20	0	4	0	2	0	26
C8	4	2	0	5	0	1	4	1	17
C9	0	0	0	0	0	2	1	4	7
Total	235	41	37	28	9	24	18	25	417

Table 6.2.2.9: Percent of Students who did not Respond by Prompt for Grade 8, EOU 1

Grade	EOU	Task	Prompt	Missing
8	EOU1	1	1 - A	15.17
8	EOU1	1	1 - B	3.45
8	EOU1	1	1 - C	4.14
8	EOU1	1	2 - A	4.14
8	EOU1	1	2 - B	7.59
8	EOU1	1	3 - AB	7.59
8	EOU1	2	1	9.66
8	EOU1	2	2	11.72
8	EOU1	2	3 - A	27.59
8	EOU1	2	3 - B	24.83
8	EOU1	2	4 - A	26.90
8	EOU1	2	4 – B	26.90
8	EOU1	3	1 – AB	12.41
8	EOU1	3	1 – C	13.79
8	EOU1	3	2	22.76
8	EOU1	3	3	20.00
8	EOU1	3	4	21.38

Table 6.2.2.10: Percent of Students who did not Respond by Prompt for Grade 8, EOU 2

Grade	EOU	Task	Prompt	Missing
8	EOU2	1	1A	20.11
8	EOU2	1	1BC	20.11
8	EOU2	1	2	21.69
8	EOU2	1	3AB	28.57
8	EOU2	1	3C	34.39
8	EOU2	1	4	34.39
8	EOU2	2	1 - A	19.05
8	EOU2	2	1 - B	24.34
8	EOU2	2	1 - C	23.38
8	EOU2	2	1 - D	24.34

8	EOU2	2	2-A	27.51
8	EOU2	2	2-B	25.93
8	EOU2	2	2-C	24.87
8	EOU2	2	3 - A	44.97
8	EOU2	2	3 - B	13.23
8	EOU2	2	3 - C	34.39
8	EOU2	3	1-A	26.46
8	EOU2	3	1-B	22.75
8	EOU2	3	1-C	25.93
8	EOU2	3	2	24.87
8	EOU2	3	3-A	29.63
8	EOU2	3	3-B	31.22
8	EOU2	3	4	30.69

Table 6.2.2.11: Percent of Students who did not Respond by Prompt for Grade 8, EOU 3

Grade	EOU	Task	Prompt	Missing
8	EOU3	1	1 - A	15.50
8	EOU3	1	1 - B	16.67
8	EOU3	1	2 - AB	16.28
8	EOU3	1	2 - C	17.83
8	EOU3	2	1 - AB	10.08
8	EOU3	2	2	10.47
8	EOU3	2	3	11.24
8	EOU3	3	1 - AB	18.60
8	EOU3	3	2 - AB	22.87
8	EOU3	3	3 - AB	24.81

Table 6.2.2.12: Percent of Students who did not Respond by Prompt for Grade 8, EOU 4

Grade	EOU	Task	Prompt	Missing
8	EOU4	1	1	1.96
8	EOU4	1	2 - A	1.96
8	EOU4	1	2 - B	1.96
8	EOU4	1	3 - A	1.96

8	EOU4	1	3 - B	1.96
8	EOU4	2	1 - A	1.96
8	EOU4	2	1 - B	1.96
8	EOU4	2	1 - C	1.96
8	EOU4	2	2 - A	1.96
8	EOU4	2	2 - B	1.96
8	EOU4	2	2 - C	1.96
8	EOU4	3	1	7.84
8	EOU4	3	2 - AB	5.88
8	EOU4	3	2 - C	9.80
8	EOU4	3	3 - A	15.69
8	EOU4	3	3 - B	17.65
8	EOU4	3	3 - C	17.65

Table 6.2.2.13: Number of Students without Responses by Task and Educator for Grade 8, EOU 1

Teacher	Scores on all Tasks	Missing Score on Task 1 Only	Missing Score on Task 2 Only	Missing Score on Task 3 Only	Missing Score on Task 1 and Task 2	Missing Score on Task 1 and Task 3	Missing Score on Task 1 and Task 3	Missing all Scores	Total
C42	37	0	0	0	0	0	0	0	37
C43	21	0	0	0	0	0	0	0	21
C45	11	0	1	0	0	0	0	0	12
C47	3	2	0	0	4	2	1	0	12
C49	2	0	7	0	2	0	1	4	16
C50	2	0	0	4	0	0	3	2	11
C51	2	0	10	0	0	0	4	0	16
C63	2	1	1	1	0	1	7	7	20
C72	0	0	0	0	0	0	0	1	1
C73	0	0	0	0	0	0	0	1	1
C74	0	0	0	0	0	0	0	1	1
C75	0	0	0	0	0	0	0	1	1
C76	0	0	0	0	0	0	0	1	1
C90	0	0	0	0	0	0	0	1	1

Total	80	3	19	5	6	3	16	19	151

Table 6.2.2.14: Number of Students without Responses by Task and Educator for Grade 8, EOU 2

Teacher	Scores on all Tasks	Missing Score on Task 1 Only	Missing Score on Task 2 Only	Missing Score on Task 3 Only	Missing Score on Task 1 and Task 2	Missing Score on Task 1 and Task 3	Missing Score on Task 1 and Task 3	Missing all Scores	Total
C43	26	0	0	0	1	0	0	0	27
C45	12	0	0	0	0	0	0	0	12
C47	10	1	0	1	1	0	0	1	14
C53	3	0	0	1	0	2	1	8	15
C54	3	4	1	1	3	1	1	7	21
C63	9	1	5	0	0	0	0	5	20
C72	0	0	17	0	4	0	1	2	24
C73	0	0	0	9	0	1	0	0	10
C74	0	0	0	15	0	0	2	1	18
C75	0	23	0	0	1	4	0	0	28
Total	63	29	23	27	10	8	5	24	189

Table 6.2.2.15: Number of Students without Responses by Task and Educator for Grade 8, EOU 3

Teacher	Scores on all Tasks	Missing Score on Task 1 Only	Missing Score on Task 2 Only	Missing Score on Task 3 Only	Missing Score on Task 1 and Task 2	Missing Score on Task 1 and Task 3	Missing Score on Task 1 and Task 3	Missing all Scores	Total
C101	35	0	0	2	0	0	1	0	38
C103	19	0	0	0	0	0	0	0	19
C43	26	1	0	1	0	0	0	0	28
C45	10	0	1	1	0	0	0	0	12
C49	8	1	0	2	0	3	1	1	16
C50	3	1	0	5	0	0	0	2	11
C51	13	0	0	3	0	0	0	0	16
C63	12	3	2	2	1	1	0	0	21
C72	0	0	7	0	0	0	9	1	17
C73	0	0	0	9	0	0	1	0	10

C74	0	0	0	15	0	1	1	0	17
C75	0	24	0	0	1	5	0	0	30
C99	23	0	0	0	0	0	0	0	23
Total	149	30	10	40	2	10	13	4	258

Table 6.2.2.16: Number of Students without Responses by Task and Educator for Grade 8, EOU 4

Teacher	Scores on all Tasks	Missing Score on Task 3 Only	Missing Score on Task 1 and Task 2	Total
C43	20	8	1	29
C45	12	0	0	12
C59	5	0	0	5
C60	4	1	0	5
Total	41	9	1	51

Appendix C. Data Tables for RQ1 (Section 6.2.3 Analyses)

RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning?

6.2.3: Can educators score student responses on the EOU assessments reliably?

Table 6.2.3.2: Agreement Between Rater and Expert Rater on Grade 5, EOU1 Prompt 1AB

	Agreement	Expected Agreement	Kappa	Std. err.	Z	Prob>Z
C7	100.00%	26.0%	1.00	0.17	5.79	0.00
C8	100.00%	26.0%	1.00	0.17	5.79	0.00
C9	100.00%	26.0%	1.00	0.17	5.79	0.00
C10	70.00%	24.0%	0.61	0.16	3.69	0.00
C12	100.00%	26.0%	1.00	0.17	5.79	0.00
C66	100.00%	26.0%	1.00	0.17	5.79	0.00
C13	70.00%	17.0%	0.64	0.13	4.74	0.00
C67	70.00%	21.0%	0.62	0.15	4.26	0.00
C68	90.00%	28.0%	0.86	0.17	5.02	0.00
C69	80.00%	25.0%	0.73	0.17	4.42	0.00
C70	80.00%	25.0%	0.73	0.17	4.42	0.00
C71	70.00%	19.0%	0.63	0.14	4.40	0.00
M1	50.00%	20.0%	0.38	0.14	2.61	0.01
M2	90.00%	28.0%	0.86	0.17	5.02	0.00
M3	90.00%	28.0%	0.86	0.17	5.02	0.00
M4	90.00%	28.0%	0.86	0.17	5.02	0.00
M5	50.00%	29.0%	0.30	0.18	1.66	0.05
M6	90.00%	28.0%	0.86	0.17	5.02	0.00
M7	90.00%	28.0%	0.86	0.17	5.02	0.00
M8	90.00%	28.0%	0.86	0.17	5.02	0.00
M9	80.00%	25.0%	0.73	0.17	4.42	0.00
M10	90.00%	28.0%	0.86	0.17	5.02	0.00
M11	100.00%	26.0%	1.00	0.17	5.79	0.00
Average	84.35%		0.79			

Table 6.2.3.3: Agreement Between Rater and Expert Rater on Grade 5, EOU1 Prompt 2

	Agreement	Expected Agreement	Kappa	Std. err.	Z	Prob>Z
C7	60.00%	24.00%	0.47	0.15	3.07	0.00
C8	70.00%	32.00%	0.56	0.20	2.73	0.00
C9	30.00%	26.00%	0.05	0.17	0.32	0.37
C10	30.00%	26.00%	0.05	0.17	0.32	0.37
C12	70.00%	32.00%	0.56	0.20	2.73	0.00
C66	50.00%	23.00%	0.35	0.15	2.29	0.01
C13	30.00%	20.00%	0.13	0.14	0.89	0.19
C67	30.00%	18.00%	0.15	0.13	1.12	0.13
C68	80.00%	31.00%	0.71	0.20	3.52	0.00
C70	60.00%	28.00%	0.44	0.18	2.42	0.01
C71	55.56%	32.10%	0.35	0.21	1.69	0.04
M1	20.00%	17.00%	0.04	0.12	0.30	0.38
M2	50.00%	22.00%	0.36	0.14	2.57	0.01
M3	50.00%	30.00%	0.29	0.19	1.47	0.07
M4	30.00%	20.00%	0.13	0.14	0.89	0.19
M5	50.00%	26.00%	0.32	0.16	1.99	0.02
M6	50.00%	27.00%	0.32	0.16	1.92	0.03
M7	50.00%	27.00%	0.32	0.16	1.92	0.03
M8	70.00%	29.00%	0.58	0.17	3.42	0.00
M9	80.00%	30.00%	0.71	0.20	3.63	0.00
M10	80.00%	31.00%	0.71	0.19	3.67	0.00
M11	100.00%	30.00%	1.00	0.20	5.08	0.00
Average	54.34%		0.39			

Table 6.2.3.4: Agreement Between Rater and Expert Rater on Grade 5, EOU2 Prompt 1_1

	Agreement	Expected Agreement	Kappa	Std. err.	Z	Prob>Z
E1	85.71%	26.53%	0.81	0.15	5.51	0.00
E2	78.57%	33.16%	0.68	0.17	4.07	0.00
E3	92.86%	29.08%	0.90	0.15	5.85	0.00
E4	92.86%	29.08%	0.90	0.15	5.85	0.00

E5	85.71%	26.53%	0.81	0.15	5.51	0.00
E6	90.00%	28.00%	0.86	0.18	4.78	0.00
E7	71.43%	19.90%	0.64	0.12	5.16	0.00
E8	50.00%	22.96%	0.35	0.13	2.62	0.00
E9	85.71%	24.49%	0.81	0.14	5.72	0.00
E10	64.29%	18.88%	0.56	0.12	4.63	0.00
E11	92.86%	27.04%	0.90	0.15	6.01	0.00
Average	80.91%		0.75			

Table 6.2.3.5: Agreement Between Rater and Expert Rater on Grade 5, EOU2 Prompt 2

	Agreement	Expected Agreement	Kappa	Std. err.	Z	Prob>Z
E1	100.00%	40.63%	1.00	0.27	3.65	0.00
E2	100.00%	40.63%	1.00	0.27	3.65	0.00
E3	100.00%	40.63%	1.00	0.27	3.65	0.00
E4	100.00%	40.63%	1.00	0.27	3.65	0.00
E5	100.00%	40.63%	1.00	0.27	3.65	0.00
E6	100.00%	40.63%	1.00	0.27	3.65	0.00
E7	100.00%	40.63%	1.00	0.27	3.65	0.00
E8	75.00%	40.63%	0.58	0.27	2.11	0.02
E9	50.00%	29.69%	0.29	0.20	1.44	0.07
E10	75.00%	40.63%	0.58	0.27	2.11	0.02
E11	87.50%	37.50%	0.80	0.25	3.20	0.00
Average	89.77%		0.84			

Table 6.2.3.6: Agreement Between Rater and Expert Rater on Grade 5, EOU4 Prompt 1_1B

	Agreement	Expected Agreement	Kappa	Std. err.	Z	Prob>Z
G1	70.59%	28.37%	0.59	0.15	4.04	0.00
G2	76.47%	30.10%	0.66	0.14	4.65	0.00
Average	73.53%		0.63			

Table 6.2.3.7: Agreement Between Rater and Expert Rater on Grade 5, EOU4 Prompt 1_2

	Agreement	Expected Agreement	Kappa	Std. err.	Z	Prob>Z
G1	82.35%	26.99%	0.76	0.14	5.33	0.00
G2	70.59%	25.95%	0.60	0.14	4.27	0.00
Average	76.47%		0.68			

Table 6.2.3.8: Agreement Between Rater and Expert Rater on Grade 8, EOU1 Prompt 1A

	Agreement	Expected Agreement	Kappa	Std. err.	Z	Prob>Z
C57	70.00%	24.00%	0.61	0.16	3.79	0.00
C51	100.00%	22.00%	1.00	0.16	6.08	0.00
C56	100.00%	22.00%	1.00	0.16	6.08	0.00
C61	80.00%	21.00%	0.75	0.16	4.78	0.00
C63	90.00%	22.00%	0.87	0.16	5.37	0.00
C76	80.00%	23.00%	0.74	0.16	4.52	0.00
G1	90.00%	22.00%	0.87	0.16	5.37	0.00
G2	80.00%	21.00%	0.75	0.16	4.72	0.00
G3	90.00%	23.00%	0.87	0.16	5.32	0.00
G4	90.00%	23.00%	0.87	0.16	5.32	0.00
Average	87.00%		0.83			

Table 6.2.3.9: Agreement Between Rater and Expert Rater on Grade 8, EOU1 Prompt 1B

	Agreement	Expected Agreement	Kappa	Std. err.	Z	Prob>Z
C57	90.00%	28.00%	0.86	0.19	4.64	0.00
C51	80.00%	32.00%	0.71	0.18	3.86	0.00
C56	90.00%	26.00%	0.86	0.18	4.75	0.00
C61	70.00%	30.00%	0.57	0.18	3.12	0.00
C63	70.00%	24.00%	0.61	0.17	3.57	0.00
C76	80.00%	30.00%	0.71	0.19	3.81	0.00
G1	70.00%	32.00%	0.56	0.18	3.10	0.00
G2	100.00%	28.00%	1.00	0.19	5.33	0.00
G3	100.00%	28.00%	1.00	0.19	5.33	0.00

G4	90.00%	28.00%	0.86	0.19	4.64	0.00
Average	84.00%		0.77			

Table 6.2.3.10: Agreement Between Rater and Expert Rater on Grade 8, EOU2 Prompt 2AB

	Agreement	Expected Agreement	Kappa	Std. err.	Z	Prob>Z
J1	66.67%	36.11%	0.48	0.29	1.66	0.05
J2	66.67%	27.78%	0.54	0.23	2.31	0.01
J3	50.00%	25.00%	0.33	0.20	1.66	0.05
J4	50.00%	30.56%	0.28	0.25	1.13	0.13
J5	33.33%	25.00%	0.11	0.21	0.52	0.30
J6	16.67%	27.78%	-0.15	0.20	-0.77	0.78
J7	66.67%	30.56%	0.52	0.25	2.10	0.02
Average	50.00%		0.30			

Table 6.2.3.11: Agreement Between Rater and Expert Rater on Grade 8, EOU2 Prompt 4

	Agreement	Expected Agreement	Kappa	Std. err.	Z	Prob>Z
J1	62.50%	26.56%	0.49	0.20	2.51	0.01
J2	100.00%	31.25%	1.00	0.23	4.44	0.00
J3	62.50%	28.13%	0.48	0.21	2.29	0.01
J4	75.00%	28.13%	0.65	0.21	3.12	0.00
J5	75.00%	31.25%	0.64	0.22	2.95	0.00
J6	62.50%	25.00%	0.50	0.19	2.68	0.00
J7	87.50%	31.25%	0.82	0.22	3.67	0.00
Average	75.00%		0.65			

Appendix D. Data Tables for RQ1 (Section 6.2.4 Analyses)

RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning?

6.2.4: Do the EOU tasks allow students to demonstrate their full range of NGSS performance expectations?

Figure 6.2.4.1: Score Distribution for Grade 5 EOU 1 (max possible score = 37)

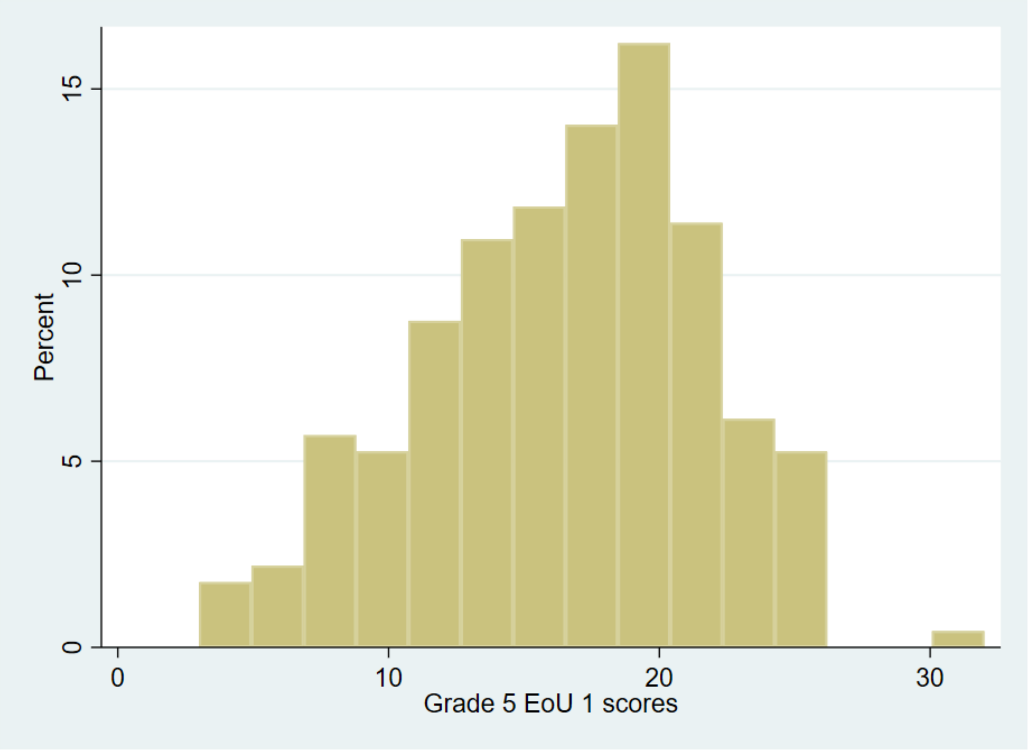


Figure 6.4.2.2: Score Distribution for Grade 5 EOU 2 (max possible score = 37)

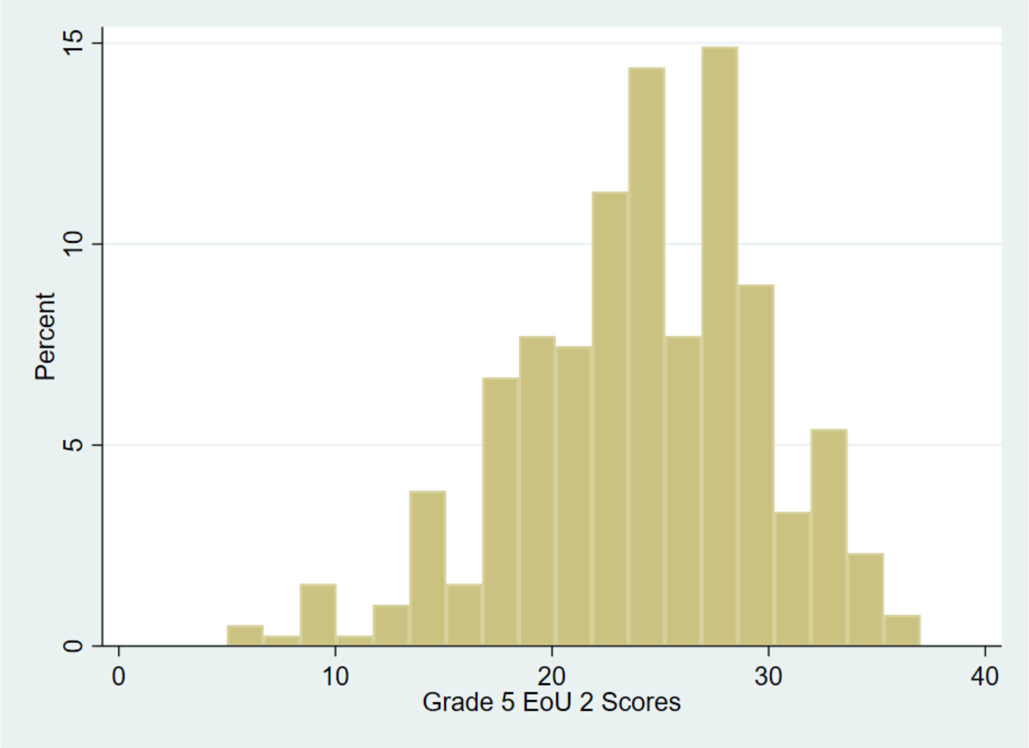


Figure 6.4.2.3: Score Distribution for Grade 5 EOU 3 (max possible score = 54)

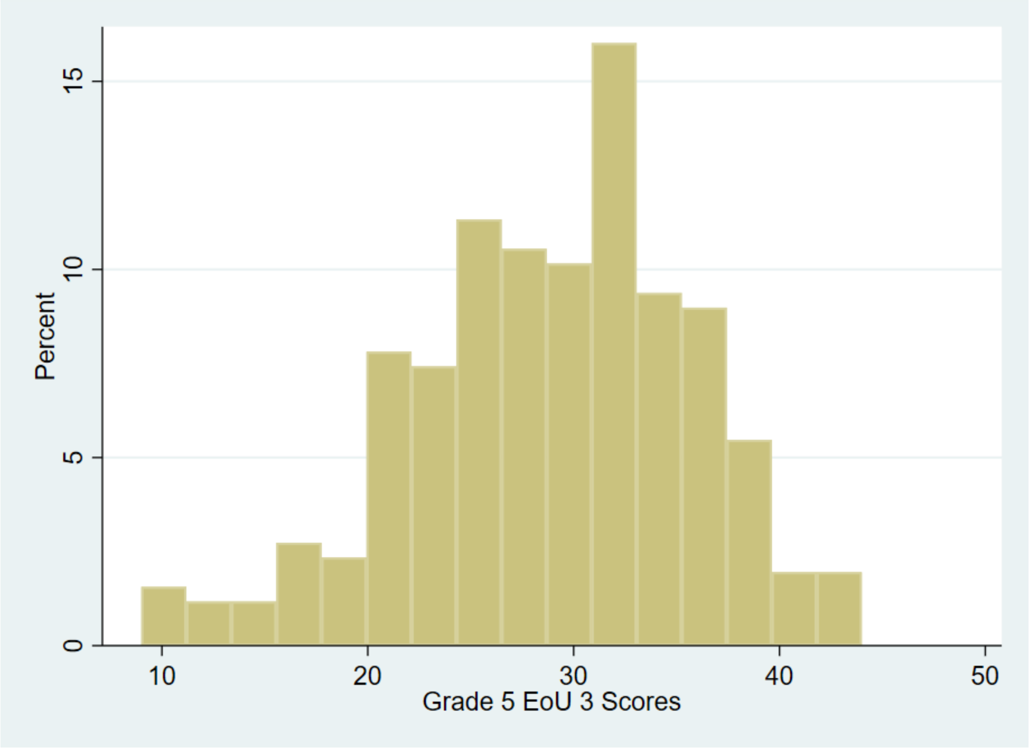


Figure 6.4.2.4: Score Distribution for Grade 5 EOU 4 (max possible score = 40)

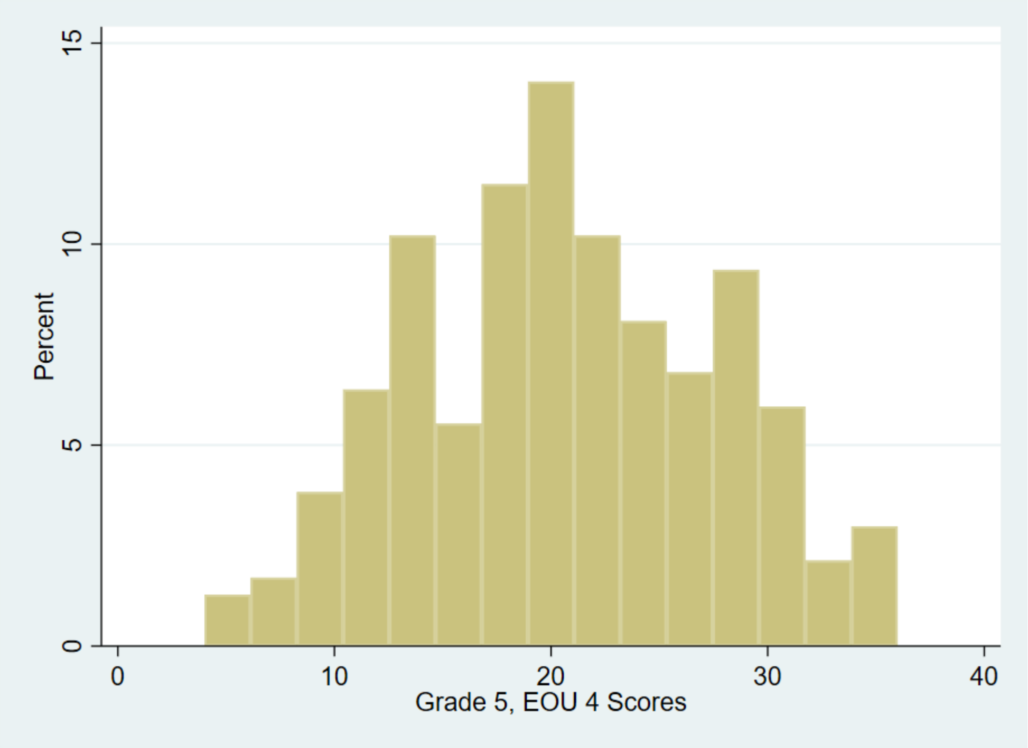


Figure 6.4.2.5: Score Distribution for Grade 8 EOU 1 (max possible score = 45)

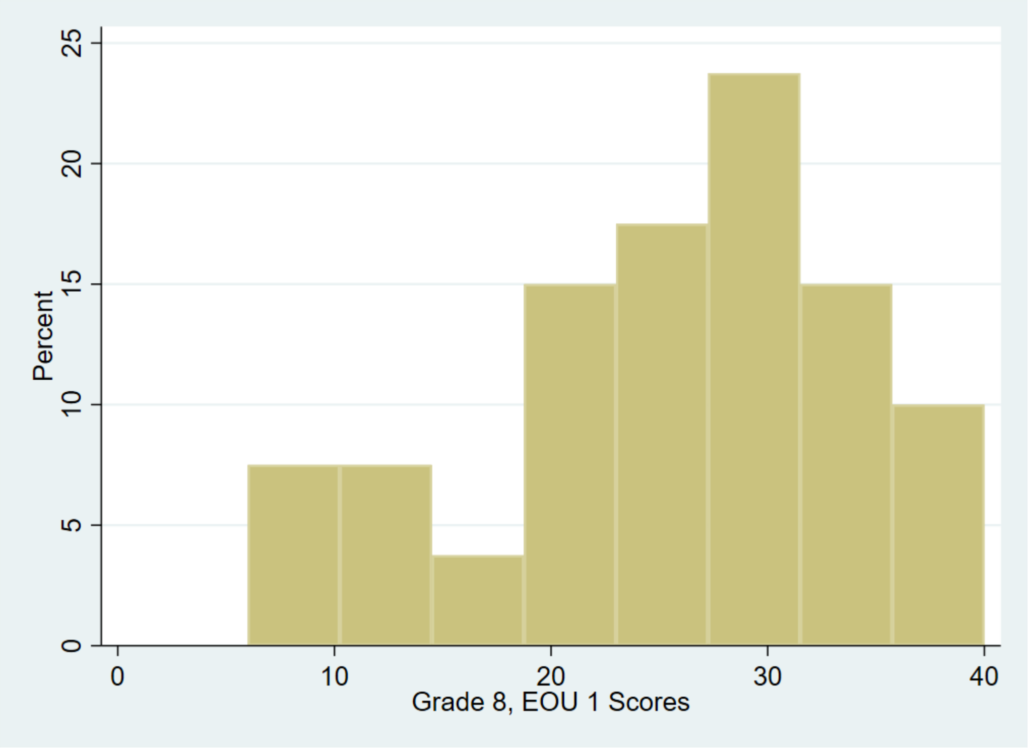


Figure 6.4.2.6: Score Distribution for Grade 8 EOU 2 (max possible score = 58)

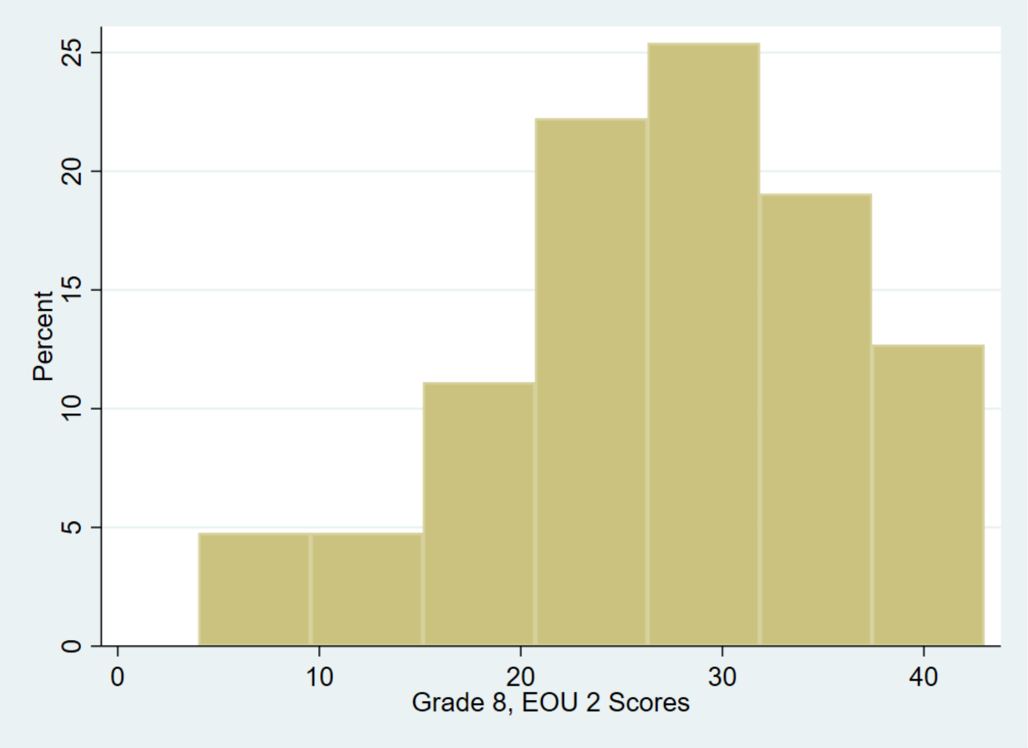


Figure 6.4.2.7: Score Distribution for Grade 8 EOU 3 (max possible score = 30)

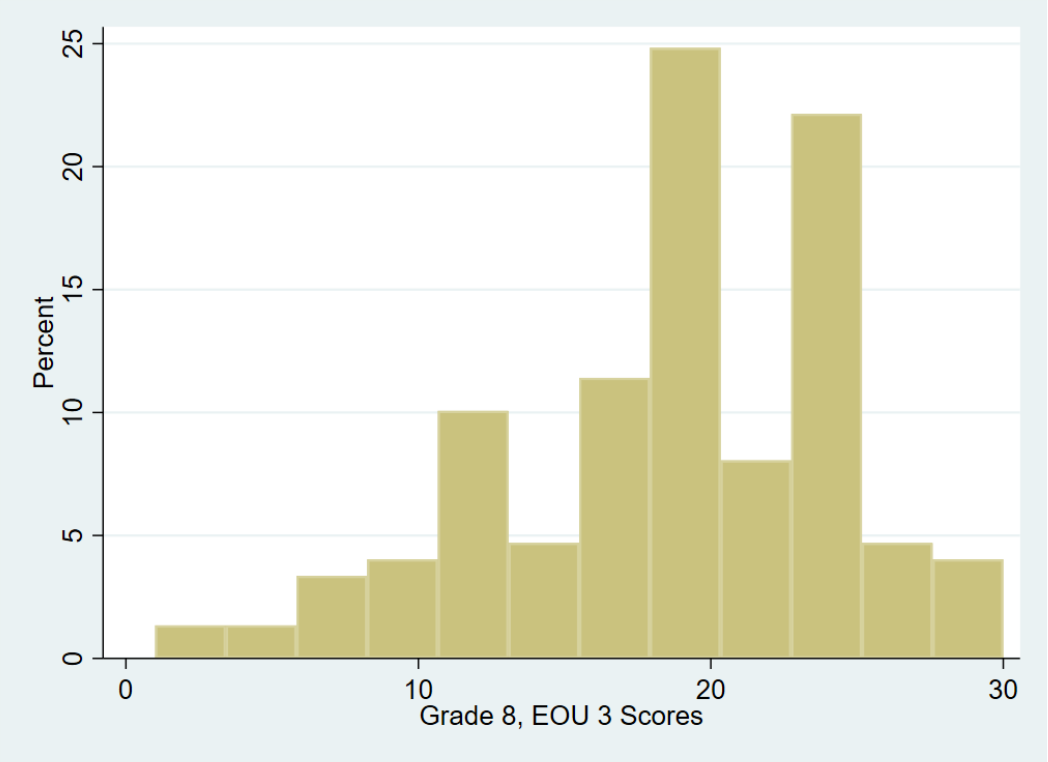
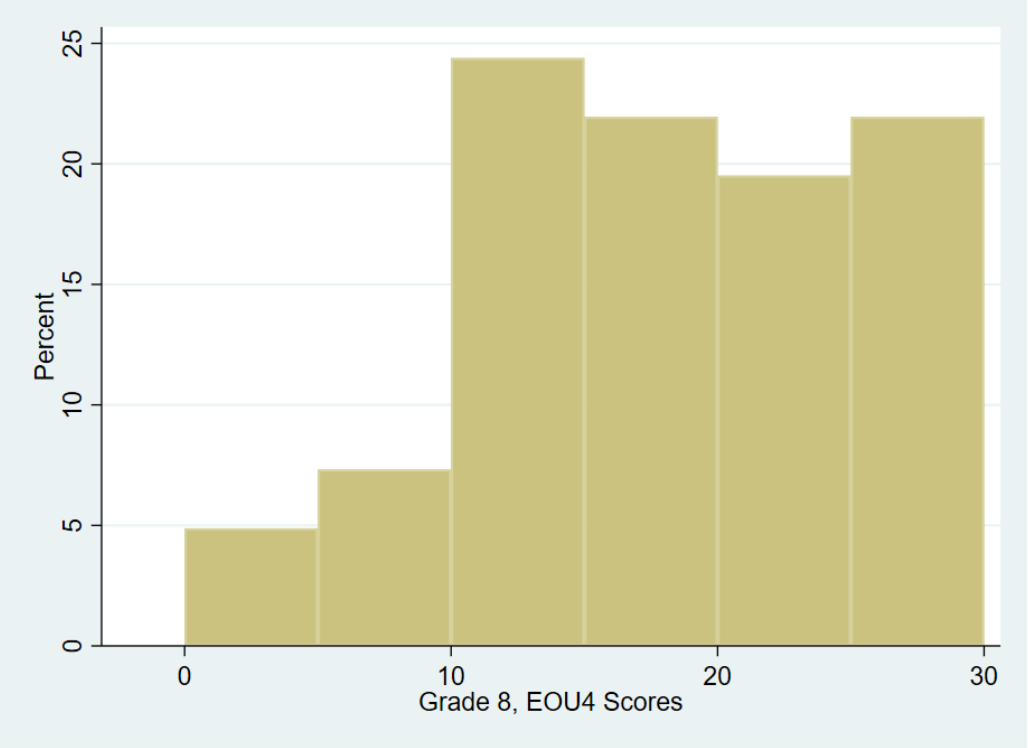


Figure: 6.4.2.8 Score distribution for Grade 8 EOU 4 (max possible score = 41)



Appendix E. Data Tables for RQ1 (Section 6.2.5 Analyses)

RQ1: To what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning?

6.2.5: Is performance on the EOU assessments associated statistically with other indicators of student learning (e.g., opportunity to learn (OTL), curriculum, or student performance on subsequent end-of-year (EOY) science assessments)?

Table 6.2.5.1: Frequency in which Educators Indicated they Included a Concept for Grade 5, EOU 1

Concept	Number Included	Number not Included
Describe that matter of any type is made of particles too small to be seen, and even then, can be detected by other means.	339	0
Understand that weight of matter is conserved when it changes form and no matter what reaction or change in properties occurs, the total weight of the substance does not change.	288	51
Describe how measurements of a variety of properties can be used to identify materials.	315	24
Describe that when two or more different substance are mixed, a new substance with different properties may be formed.	327	12
Develop and use models to demonstrate understanding of the structure and properties of matter.	316	23
Measure and graph quantities such as weight to demonstrate understanding of the structure and properties of matter and chemical reactions.	126	213
Use data based on observations and measurements from a given investigation to identify materials based on their properties.	339	0
Plan and use data from investigations to determine whether the mixing of two or more substances results in new substances.	297	42
Apply scale, proportion, and quantity using standard units to investigate natural objects from the very small to the immensely large.	137	202
Apply scale, proportion, and quantity using standard units to measure and describe physical quantities such as weight, time, temperature, and volume.	232	107
Apply cause and effect relationships as the organizing concept for explaining change.	298	41

Table 6.2.5.2: T-test results for educator identification if they have taught students to: Develop and use models to demonstrate understanding of the structure and properties of matter.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	181	116	7.11	6.78	0.33	1.01	295	0.16
Task 2	196	84	4.07	4.93	-0.86	-3.02	278	1.00
Task 3	172	104	4.38	4.71	-0.33	-1.22	274	0.89
EOU 1	154	74	16.10	17.15	-1.05	-1.39	226	0.92

Table 6.2.5.3: T-test results for educator identification if they have taught students to: Apply scale, proportion, and quantity using standard units to investigate natural objects from the very small to the immensely large.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	177	120	6.93	7.07	-0.14	-0.43	295	0.67
Task 2	182	98	4.10	4.76	-0.66	-2.39	278	0.99
Task 3	157	119	4.43	4.61	-0.17	-0.65	274	0.74
EOU 1	144	84	15.97	17.25	-1.28	-1.75	226	0.96

Table 6.2.5.4: T-test results for educator identification if they have taught students to: Apply scale, proportion, and quantity using standard units to measure and describe physical quantities such as weight, time, temperature, and volume.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	86	211	6.33	7.25	-0.93	-2.68	295	1.00
Task 2	85	195	3.60	4.65	-1.05	-3.72	278	1.00
Task 3	66	210	3.94	4.69	-0.75	-2.46	274	1.00
EOU 1	54	174	14.63	17.00	-2.37	-2.88	226	1.00

Table 6.2.5.5: Frequency in which Educators Indicated they Included a Concept for Grade 5, EOU2

Concept	Number Included	Number not Included
Describe how energy in animals’ food was once energy from the sun.	443	0
Understand the idea that plants get the materials they need for growth chiefly from air and water.	443	0

Describe the movement of matter among plants, animals, decomposers, and the environment.	443	0
Understand that food provides animals with the materials they need for body repair and growth and the energy they need for motion and to maintain body warmth.	283	160
Use models to describe that energy in animals’ food (used for body repair, growth, motion, and to maintain body warmth) was once energy from the sun.	355	88
Develop a model to describe the movement of matter among plants, animals, decomposers, and the environment.	422	21
Develop and/or use a model with provided information (i.e., a specific mammal, insect, set of living things, sun) to show that energy from the sun is transferred to animals through a chain of events that begins with plants producing food then being eaten by animals.	443	0
Use the evidence/data based on observations to construct a claim about the effects of a newly introduced species to an ecosystem	363	80
Support an argument with evidence that plants get the materials they need for growth chiefly from air and water	389	54
Use diagrams or flowcharts to describe the flow of energy within an ecosystem, tracing the energy in animals' food back to the energy from the sun that was captured by plants.	443	0

Table 6.2.5.6: T-test results for educator identification if they have taught students to: Understand that food provides animals with the materials they need for body repair and growth and the energy they need for motion and to maintain body warmth.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	143	261	5.61	5.92	-0.31	-1.63	402	0.95
Task 2	147	253	9.78	9.66	0.11	0.39	398	0.35
Task 3	145	262	8.46	8.12	0.34	1.33	405	0.10
EOU 2	126	235	24.32	24.03	0.29	0.46	359	0.32

Table 6.2.5.7: Frequency in which Educators Indicated they Included a Concept for Grade 5, EOU 3

Concept	Number Included	Number not Included
Understand the ways the geosphere, biosphere, hydrosphere and/or atmosphere interact.	184	157
Understand that nearly all of Earth’s available water is in the ocean.	341	0

Understand that most freshwater is in glaciers or underground and a tiny fraction is in streams, lakes, wetlands, and the atmosphere.	341	0
Understand that human activities have had major effects on the land, vegetation, streams, ocean, air and even outer space.	219	122
Understand that individuals and communities do things to help protect Earth’s resources and environments.	305	36
Develop and use models to demonstrate understanding of the interactions between Earth’s spheres.	190	151
Use mathematics and computational thinking to describe and represent quantities to address scientific questions and to demonstrate understanding of sources of Earth’s water.	297	44
Obtain and combine information to explain phenomena or solutions to a design problem about the effect of human activities on Earth’s resources and environments.	300	41
Apply systems and system models in terms of their components and interactions to describe ways Earth’s spheres interact.	220	121
Apply scale, proportion, and quantity to describe Earth’s water sources.	305	36
Apply systems and system models to describe the effects of human activity.	203	138

Table 6.2.5.8: T-test results for educator identification if they have taught students to: Understand the ways the geosphere, biosphere, hydrosphere and/or atmosphere interact.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	137	166	14.15	13.93	0.21	0.53	301	0.30
Task 2	147	167	8.70	9.20	-0.50	-1.36	312	0.91
Task 3	133	155	5.72	5.39	0.33	1.36	286	0.09
EOU 3	116	140	29.10	28.84	0.23	0.27	254	0.39

Table 6.2.5.9: T-test results for educator identification if they have taught students to: Understand that human activities have had major effects on the land, vegetation, streams, ocean, air and even outer space.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	96	207	13.30	14.37	-1.07	-2.53	301	0.99
Task 2	110	204	9.36	8.75	0.61	1.60	312	0.05
Task 3	104	184	5.93	5.32	0.61	2.41	286	0.01
EOU 3	88	168	28.93	28.96	-0.03	-0.03	254	0.51

Table 6.2.5.10: T-test results for educator identification if they have taught students to: Develop and use models to demonstrate understanding of the interactions between Earth's spheres.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	135	168	14.79	13.43	1.36	3.49	301	0.00
Task 2	141	173	9.27	8.72	0.55	1.51	312	0.07
Task 3	133	155	6.07	5.09	0.98	4.07	286	0.00
EOU 3	120	136	30.18	27.86	2.32	2.72	254	0.00

Table 6.2.5.11: T-test results for educator identification if they have taught students to: Apply systems and system models in terms of their components and interactions to describe ways Earth's spheres interact.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	105	198	14.31	13.88	0.44	1.05	301	0.15
Task 2	112	202	9.31	8.77	0.54	1.42	312	0.08
Task 3	107	181	5.90	5.33	0.57	2.24	286	0.01
EOU 3	95	161	29.29	28.75	0.55	0.62	254	0.27

Table 6.2.5.12: T-test results for educator identification if they have taught students to: Apply systems and system models to describe the effects of human activity.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	111	192	14.05	14.02	0.02	0.06	301	0.48
Task 2	121	193	9.17	8.84	0.33	0.87	312	0.19
Task 3	111	177	6.09	5.20	0.89	3.61	286	0.00
EOU 3	95	161	29.52	28.62	0.90	1.01	254	0.16

Table 6.2.5.13: Frequency in which Educators Indicated they Included a Concept for Grade 5, EOU 4

Concept	Number Included	Number not Included
Understand that the gravitational force of earth acts on objects, pulling them towards the planet's center.	244	104
Understand that the Sun is a star that appears larger and brighter than other stars because it is closer to Earth.	329	19
Understand that stars range greatly in their distance from Earth.	274	74

Develop an argument that the apparent brightness of the Sun and stars is due to their relative distance from the Earth.	248	100
Develop an argument that the gravitational force exerted by the Earth on objects is directed downward.	227	121
Develop a model that provides evidence to support a claim about how gravity affects objects on Earth.	163	185
Develop a model showing the relationship between distances and the apparent brightness of stars and use this to support their claim.	166	182
Understand that the orbits of Earth around the Sun and the orbits of the moon around the Earth's axis causes observable patterns.	298	50
Understand that there are observable patterns related to daily changes in length and direction of shadows, day and night, and the seasonal appearance of some stars in the night sky.	334	14
Organize data to reveal patterns related to the orbits of Earth around the Sun, of the moon around Earth, and the rotation of the Earth around its axis.	223	125
Apply cause and effect relationships as the organizing concept for explaining change.	181	167

Table 6.2.5.14: T-test results for educator identification if they have taught students to: Understand that the gravitational force of earth acts on objects, pulling them towards the planet's center.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	99	178	9.09	7.91	1.18	2.99	275	0.00
Task 2	97	192	5.24	4.60	0.63	1.94	287	0.03
Task 3	96	181	7.58	6.12	1.47	3.95	275	0.00
EOU 4	90	136	22.60	19.74	2.86	3.07	224	0.00

Table 6.2.5.15: T-test results for educator identification if they have taught students to: Develop an argument that the gravitational force exerted by the Earth on objects is directed downward.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	113	164	8.95	7.91	1.04	2.69	275	0.00
Task 2	113	176	5.07	4.65	0.42	1.32	287	0.09
Task 3	112	165	7.36	6.13	1.23	3.39	275	0.00
EOU 4	103	123	21.97	19.97	2.00	2.17	224	0.02

Table 6.2.5.16: T-test results for educator identification if they have taught students to: Develop a model that provides evidence to support a claim about how gravity affects objects on Earth.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	164	113	8.34	8.33	0.01	0.02	275	0.49
Task 2	164	125	4.70	4.98	-0.28	-0.90	287	0.82
Task 3	160	117	7.06	6.03	1.04	2.86	275	0.00
EOU 4	136	90	20.98	20.73	0.24	0.26	224	0.40

Table 6.2.5.17: T-test results for educator identification if they have taught students to: Develop a model showing the relationship between distances and the apparent brightness of stars and use this to support their claim.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	160	117	7.76	9.12	-1.36	-3.58	275	1.00
Task 2	154	135	4.47	5.21	-0.75	-2.43	287	0.99
Task 3	152	125	6.89	6.30	0.60	1.65	275	0.05
EOU 4	126	100	20.06	21.91	-1.85	-1.99	224	0.98

Table 6.2.5.18: T-test results for educator identification if they have taught students to: Organize data to reveal patterns related to the orbits of Earth around the Sun, of the moon around Earth, and the rotation of the Earth around its axis.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	105	172	6.57	9.41	-2.84	-7.93	275	1.00
Task 2	98	191	3.72	5.38	-1.65	-5.30	287	1.00
Task 3	93	184	5.86	7.01	-1.15	-3.04	275	1.00
EOU 4	71	155	16.28	22.99	-6.71	-7.49	224	1.00

Table 6.2.5.19: T-test results for educator identification if they have taught students to: Apply cause and effect relationships as the organizing concept for explaining change.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	130	147	7.18	9.35	-2.16	-5.96	275	1.00
Task 2	133	156	4.39	5.18	-0.79	-2.57	287	0.99
Task 3	125	152	5.59	7.47	-1.88	-5.42	275	1.00

EOU 4	91	135	17.92	22.87	-4.95	-5.58	224	1.00

Table 6.2.5.20 Frequency in which Educators Indicated they Included a Concept for Grade 8, EOU 1

Concept	Number Included	Number not Included
Apply Newton’s Third Law to design a solution to a problem involving the motion of two colliding objects.	136	15
Explain how the change in an object's motion depends on balanced (Newton's First Law) and unbalanced forces in a system.	151	0
Describe and explain the relationships of kinetic energy to the mass of an object and to the speed of an object.	151	0
Understand that gravitational interactions are always attractive and depend on the masses of interacting objects.	36	115
Evaluate competing design solutions involving the motion of two colliding objects based on jointly developed and agreed upon design criteria.	78	73
Develop a model to represent the motion of objects in colliding systems and their interactions	150	1
Plan an investigation to provide evidence that the change in an object's motion depends on the sum of the forces on the object and the mass of the object.	110	41
Examine the changes over time and forces at different scales to explain the stability and change in designed systems.	79	72
Use proportional relationships among quantities to analyze data related to the effect on kinetic energy when the mass of an object or its speed changes.	134	17
Use reasoning to connect the appropriate evidence about the forces on objects to construct an argument that gravitational forces are attractive and mass dependent.	80	71
Use models to represent the gravitational interactions between two masses.	51	100

Table 6.2.5.21: T-test results for educator identification if they have taught students to: Evaluate competing design solutions involving the motion of two colliding objects based on jointly developed and agreed upon design criteria.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	46	74	7.61	10.66	-3.05	-4.86	118	1.00
Task 2	35	56	4.97	8.18	-3.21	-5.21	89	1.00
Task 3	45	63	6.51	8.22	-1.71	-3.35	106	1.00

EOU 1	28	52	20.18	28.40	-8.23	-4.62	78	1.00

Table 6.2.5.22: T-test results for educator identification if they have taught students to: Examine the changes over time and forces at different scales to explain the stability and change in designed systems.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	52	68	8.23	10.46	-2.23	-3.46	118	1.00
Task 2	35	56	5.03	8.14	-3.11	-5.01	89	1.00
Task 3	46	62	6.89	7.97	-1.08	-2.05	106	0.98
EOU 1	28	52	20.18	28.40	-8.23	-4.62	78	1.00

Table 6.2.5.23: T-test results for educator identification if they have taught students to: Use reasoning to connect the appropriate evidence about the forces on objects to construct an argument that gravitational forces are attractive and mass dependent.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	57	63	10.39	8.68	1.70	2.61	118	0.01
Task 2	48	43	7.85	5.93	1.92	2.94	89	0.00
Task 3	43	65	8.16	7.08	1.09	2.05	106	0.02
EOU 1	41	39	28.29	22.62	5.68	3.14	78	0.00

Table 6.2.5.24: t-test results for educator identification if they have taught students to: Use models to represent the gravitational interactions between two masses.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	85	35	10.25	7.66	2.59	3.72	118	0.00
Task 2	61	30	7.90	5.00	2.90	4.40	89	0.00
Task 3	67	41	8.07	6.59	1.49	2.83	106	0.00
EOU 1	54	26	28.00	20.38	7.62	4.10	78	0.00

Table 6.2.5.25: Number of Students who were Taught Specific Concepts (as Reported by the Educator) for Grade 8, EOU 2

Concept	Number Included	Number not Included
Understand that patterns of motion of the sun, the moon, and stars in the sky can be observed, described, and explained and predicted with models.	189	0
Understand that seasons are a result of Earth's tilt of its axis of rotation and are caused by differential intensity of sunlight on different areas of Earth across the year.	177	0
Develop and use a model of the Earth-sun-moon system to describe the cyclic patterns of lunar phases, eclipses of the sun and moon and seasons.	177	0
Understand that Earth and its solar systems are part of the Milky Way galaxy.	167	10
Understand that the Milky Way galaxy is one of many galaxies in the universe.	147	30
Understand that the solar system consists of the sun and a collection of objects held in orbit around the sun by the sun's gravitational pull on them.	177	0
Develop and use a model to describe the role of gravity in the motions within galaxies and the solar system.	153	24
Understand that gravitational interactions are always attractive and depend on the masses of interacting objects.	156	21
Use the practices of engaging in argument from evidence to make sense of relationships between energy and forces.	121	56
Use evidence to support the claim that gravitational interactions are attractive and depend on the masses of interacting objects.	141	36
Explain how some effects of gravitational interactions, which apply universally, may only be observable in interactions between very massive objects.	131	46

Table 6.2.5.26: T-test results for educator identification if they have taught students to: Use the practices of engaging in argument from evidence to make sense of relationships between energy and forces.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	25	81	6.16	8.70	-2.54	-3.50	104	1.00
Task 2	25	90	6.32	8.28	-1.96	-2.53	113	0.99

Task 3	29	84	10.03	9.71	0.32	0.39	111	0.35
EOU 2	15	36	23.33	28.64	-5.31	-2.02	49	0.98

Table 6.2.5.27: T-test results for educator identification if they have taught students to: Explain how some effects of gravitational interactions, which apply universally, may only be observable in interactions between very massive objects.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	20	86	7.95	8.14	-0.19	-0.23	104	0.59
Task 2	25	90	7.24	8.02	-0.78	-0.99	113	0.84
Task 3	14	99	10.43	9.71	0.72	0.66	111	0.25
EOU 2	6	45	28.00	26.96	1.04	0.27	49	0.39

Table 6.2.5.28: Number of Students who were Taught Specific Concepts (as Reported by the Educator) for Grade 8, EOU 3

Concept	Number Included	Number not Included
Understand that geologic time can be interpreted from rock strata and provides a way to organize Earth's history.	220	38
Understand that the analysis of rock strata and the fossil record provides only relative dates and not an absolute scale.	220	38
Understand that the fossil record is a collection of fossils and their placement in chronological order.	220	38
Understand that the fossil record documents the existence, diversity, extinction, and change of many life forms throughout the history of life on Earth.	220	38
Understand that anatomical similarities and differences between various organisms living today and the fossil record enable the reconstruction of evolutionary history.	220	38
Understand that these similarities and differences can provide information to make inferences of lines of evolutionary descent.	220	38
Understand that natural selection leads to the predominance of certain traits in a population, and the suppression of others.	248	10
Understand that adaptation by natural selection is a process by which species change over time in response to changes in environmental conditions.	258	0
Understand that traits that support successful survival and reproduction in the new environment become more common; those that do not become less common.	248	10

Understand that the distribution of traits in a population changes.	155	103
Understand that genes are located in the chromosomes of cells and each distinct gene chiefly controls the production of specific proteins, which in turn affects the traits of the individual.	172	86
Understand that changes (mutations) to genes can result in changes to proteins which can affect the structure and functions of the organism and thereby change traits.	156	102
Understand that genetic mutations can result in changes to the structure and function of proteins which may be beneficial or harmful or neutral to the organism.	177	81
Construct scientific explanations based on valid and reliable evidence obtained from sources.	258	0
Analyze and interpret data to determine similarities and differences.	258	0
Construct explanations about real-world phenomena.	242	16
Construct explanations that include relationships between variables that describe phenomena.	212	46
Use mathematical representations to support scientific conclusions related to measurable changes in selected traits in a population over time.	92	166
Develop and use models to describe phenomena related to inheritance and variation of traits.	193	65
Apply scale, proportion, and quantity with respect to time, space, and energy observed at various scales using models.	107	151
Identify patterns in data using graphs, charts, and images.	210	48
Apply patterns to identify cause and effect relationships.	247	11
Identify that phenomena may have more than one cause and some cause and effect relationships in systems can only be described using probability.	162	96

Table 6.2.5.29: T-test results for educator identification if they have taught students to: Understand that the distribution of traits in a population changes.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	96	116	6.24	7.31	-1.07	-3.25	210	1.00
Task 2	80	149	5.54	6.91	-1.37	-4.82	227	1.00
Task 3	60	131	3.88	5.79	-1.90	-5.23	189	1.00

EOU 3	47	102	15.21	20.15	-4.93	-5.19	147	1.00

Table 6.2.5.30: T-test results for educator identification if they have taught students to: Understand that genes are located in the chromosomes of cells and each distinct gene chiefly controls the production of specific proteins, which in turn affects the traits of the individual.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	73	139	5.92	7.30	-1.38	-4.06	210	1.00
Task 2	78	151	4.95	7.19	-2.24	-8.62	227	1.00
Task 3	66	125	3.94	5.85	-1.91	-5.39	189	1.00
EOU 3	58	91	15.24	20.73	-5.48	-6.26	147	1.00

Table 6.2.5.31: T-test results for educator identification if they have taught students to: Understand that changes (mutations) to genes can result in changes to proteins which can affect the structure and functions of the organism and thereby change traits.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	91	121	5.77	7.62	-1.85	-5.87	210	1.00
Task 2	78	151	5.12	7.11	-1.99	-7.39	227	1.00
Task 3	69	122	3.90	5.92	-2.02	-5.83	189	1.00
EOU 3	55	94	15.29	20.52	-5.23	-5.83	147	1.00

Table 6.2.5.32: T-test results for educator identification if they have taught students to: Understand that genetic mutations can result in changes to the structure and function of proteins which may be beneficial or harmful or neutral to the organism.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	75	137	6.21	7.16	-0.95	-2.74	210	1.00
Task 2	60	169	5.50	6.76	-1.26	-4.03	227	1.00
Task 3	51	140	4.61	5.40	-0.79	-1.96	189	0.97
EOU 3	43	106	17.00	19.24	-2.24	-2.14	147	0.98

Table 6.2.5.33: T-test results for educator identification if they have taught students to: Use mathematical representations to support scientific conclusions related to measurable changes in selected traits in a population over time.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	153	59	6.88	6.69	0.18	0.48	210	0.32
Task 2	156	73	6.00	7.34	-1.34	-4.60	227	1.00
Task 3	132	59	5.13	5.32	-0.19	-0.49	189	0.69
EOU 3	123	26	18.48	19.12	-0.64	-0.50	147	0.69

Table 6.2.5.34: T-test results for educator identification if they have taught students to: Develop and use models to describe phenomena related to inheritance and variation of traits.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	57	155	6.54	6.93	-0.39	-1.02	210	0.84
Task 2	60	169	5.28	6.83	-1.56	-5.06	227	1.00
Task 3	48	143	4.71	5.35	-0.64	-1.55	189	0.94
EOU 3	46	103	16.83	19.38	-2.55	-2.51	147	0.99

Table 6.2.5.35: T-test results for educator identification if they have taught students to: Apply scale, proportion, and quantity with respect to time, space, and energy observed at various scales using models.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	142	70	7.13	6.20	0.93	2.65	210	0.00
Task 2	142	87	6.15	6.87	-0.72	-2.49	227	0.99
Task 3	110	81	5.31	5.02	0.28	0.78	189	0.22
EOU 3	102	47	18.78	18.17	0.61	0.59	147	0.28

Table 6.2.5.36: T-test results for educator identification if they have taught students to: Identify that phenomena may have more than one cause and some cause-and-effect relationships in systems can only be described using probability.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	88	124	6.27	7.22	-0.95	-2.82	210	1.00
Task 2	89	140	5.38	7.09	-1.71	-6.37	227	1.00
Task 3	70	121	4.01	5.87	-1.85	-5.30	189	1.00

EOU 3	63	86	15.70	20.71	-5.01	-5.69	147	1.00

Table 6.2.5.37: Number of Students who were Taught Specific Concepts (as Reported by the Educator) for Grade 8, EOU4

Concept	Number Included	Number not Included
Understand that a simple wave has a repeating pattern with a specific wavelength, frequency, and amplitude.	51	0
Understand that a sound wave needs a medium through which it is transmitted.	51	0
Understand that when light shines on an object, it is reflected, absorbed, or transmitted through the object, depending on the object's material and the frequency (color) of the light.	51	0
Understand that the path a light travels can be traced as straight lines, except at surfaces between different transparent materials where the light path bends.	39	12
Understand that a wave model of light is useful for explaining brightness, color, and the frequency-dependent bending of light at a surface between media.	39	12
Understand that because light can travel through space, it cannot be a matter wave, like sound or water waves.	51	0
Use mathematical representations to describe and/or support scientific conclusions and design solutions.	41	10
Apply logical and conceptual connections between evidence and explanations.	51	0
Develop and use models to describe phenomena related to waves.	51	0
Apply graphs and charts to identify patterns in data.	51	0
Apply the knowledge that structures can be designed to serve particular functions related to the transmission of waves.	51	0

Table 6.2.5.38: T-test results for educator identification if they have taught students to: Understand that the path light travels can be traced as straight lines, except at surfaces between different transparent materials where the light path bends.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	12	38	6.33	6.26	0.07	0.08	48	0.47
Task 2	12	38	4.83	6.68	-1.85	-1.44	48	0.92
Task 3	12	30	5.25	5.83	-0.58	-0.56	40	0.71

EOU 4	12	29	16.42	18.41	-2.00	-0.72	39	0.76

Table 6.2.5.3: T-test results for educator identification if they have taught students to: Understand that a wave model of light is useful for explaining brightness, color, and the frequency-dependent bending of light at a surface between media.

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	12	38	6.33	6.26	0.07	0.08	48	0.47
Task 2	12	38	4.83	6.68	-1.85	-1.44	48	0.92
Task 3	12	30	5.25	5.83	-0.58	-0.56	40	0.71
EOU 4	12	29	16.42	18.41	-2.00	-0.72	39	0.76

Appendix F. Data Tables for RQ2 (Section 6.3 Analyses)

RQ2. How well do latent variable measurement models fit the empirical EOU assessment data?

Table 6.3.1. Grade 5 Scale Reliability Estimates

EOU1 Total Score Scale	Score
Average interitem covariance	.21
Number of items in the scale	10
Scale reliability coefficient	.69
EOU2 Total Score Scale	
Average interitem covariance	.19
Number of items in the scale	.12
Scale reliability coefficient	.77
EOU3 Total Score Scale	
Average interitem covariance	.14
Number of items in the scale	20
Scale reliability coefficient	.82
EOU4 Total Score Scale	
Average interitem covariance	.21
Number of items in the scale	14
Scale reliability coefficient	.80

Table 6.3.1. Grade 8 Scale Reliability Estimates

EOU1 Total Score Scale	Score
Average interitem covariance	.21
Number of items in the scale	17
Scale reliability coefficient	.85
EOU2 Total Score Scale	
Average interitem covariance	.17
Number of items in the scale	23
Scale reliability coefficient	.86
EOU3 Total Score Scale	
Average interitem covariance	.33
Number of items in the scale	10
Scale reliability coefficient	.83
EOU4 Total Score Scale	
Average interitem covariance	.20
Number of items in the scale	17
Scale reliability coefficient	.87

Appendix G. Data Tables for RQ3 (Section 6.4.1 Analyses)

RQ3. Overall, what do the EOU assessment results tell us about students' science learning?

6.4.1 What do the EOU assessment results tell us about student learning in terms of variation across student groups?

Table 6.4.1.1: T-test results for comparison of scores by gender for Grade 5, EOU 1 (male = 0, female = 1)

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	139	158	6.67	7.26	0.59	1.87	295	0.97
Task 2	133	147	4.17	4.48	0.31	1.18	278	0.88
Task 3	133	143	4.34	4.66	0.33	1.25	274	0.89
EOU 1	109	119	15.63	17.18	1.54	2.18	226	0.99

Table 6.4.1.2: Task 1 scores by ELA achievement level for Grade 5, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	94	6.53	2.77	0	14
Level 1	82	7.48	2.68	0	12
Level 2	39	8.28	2.08	3	14
Level 3	67	6.13	2.74	1	13

Table 6.4.1.3: Task 1 ANOVA test results comparing ELA achievement level for Grade 5, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	157.64	4	39.41	5.65	0.00
Within Groups	2037.27	292	6.98		
Total	2194.92	296	7.42		

Table 6.4.1.4: Task 2 scores by ELA achievement level for Grade 5, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	87	4.15	2.35	0	9
Level 1	80	4.61	1.92	0	9
Level 2	35	5.29	1.99	0	9
Level 3	64	3.95	2.35	0	9

Table 6.4.1.5: Task 2 ANOVA test results comparing ELA achievement level for Grade 5, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	70.01	4	17.50	3.72	0.01
Within Groups	1293.76	275	4.70		
Total	1363.77	279	4.89		

Table 6.4.1.6: Task 3 scores by ELA achievement level for Grade 5, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	89	4.12	2.12	0	9
Level 1	80	4.71	2.38	0	10
Level 2	34	5.88	1.30	3	10
Level 3	60	4.00	2.12	0	8

Table 6.4.1.7: Task 3 ANOVA test results comparing ELA achievement level for Grade 5, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	96.35	4	24.09	5.43	0.00
Within Groups	1202.63	271	4.44		
Total	1298.99	275	4.72		

Table 6.4.1.8: EOU1 scores by ELA achievement level for Grade 5, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	69	15.55	5.48	5	32
Level 1	66	17.47	5.18	3	26
Level 2	28	20.11	3.10	14	26
Level 3	53	14.57	5.53	3	26

Table 6.4.1.9: EOU1 ANOVA test results comparing ELA achievement level for Grade 5, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	696.01	4	174.00	6.62	0
Within Groups	5864.13	223	26.30		
Total	6560.14	227	28.90		

Table 6.4.1.10: Task 1 scores by mathematics achievement level for Grade 5, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	96	6.36	2.82	0	14
Level 1	91	7.51	2.36	0	12
Level 2	32	8.47	2.68	3	14
Level 3	63	6.29	2.74	1	13

Table 6.4.1.11: Task 1 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	167.37	4	41.84	6.03	0.00
Within Groups	2027.55	292	6.94		
Total	2194.92	296	7.42		

Table 6.4.1.12: Task 2 scores by mathematics achievement level for Grade 5, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	92	3.72	2.26	0	9
Level 1	87	5.14	1.94	0	9
Level 2	27	4.85	1.88	0	9
Level 3	60	4.13	2.32	0	9

Table 6.4.1.13: Task 2 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	120.72	4	30.18	6.68	0
Within Groups	1243.05	275	4.52		
Total	1363.77	279	4.89		

Table 6.4.1.14: Task 3 scores by mathematics achievement level for Grade 5, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	97	4.24	2.25	0	9
Level 1	83	4.80	2.22	0	10

Level 2	27	5.30	1.81	2	10
Level 3	56	4.14	2.10	0	8

Table 6.4.1.15: Task 3 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	38.36	4	9.59	2.06	0.0861
Within Groups	1260.63	271	4.65		
Total	1298.99	275	4.72		

Table 6.4.1.16: EOU1 scores by ELA achievement level for Grade 5, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	74	15.03	5.98	3	32
Level 1	72	17.96	4.58	4	26
Level 2	21	19.62	3.67	13	26
Level 3	49	15.18	5.31	4	26

Table 6.4.1.17: EOU1 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	612.10	4	153.03	5.74	0.00
Within Groups	5948.04	223	26.67		
Total	6560.14	227	28.90		

Table 6.4.1.18: t-test results for comparison of scores by gender for Grade 5, EOU 2

	# of Males	# of Females	Mean Males	Mean Females	diff	t	df	pr(T>t)
Task 1	214	217	5.64	5.92	-0.28	-1.59	429	0.94
Task 2	211	216	9.73	9.73	0.01	0.03	425	0.49
Task 3	216	218	8.32	8.19	0.13	0.54	432	0.30
EOU 2	192	195	24.27	23.98	0.29	0.50	385	0.31

Table 6.4.1.19: Task 1 scores by ELA achievement level for Grade 5, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	109	5.36	1.82	1	9
Level 1	119	5.79	1.84	0	9
Level 2	78	6.36	1.53	2	9
Level 3	125	5.78	1.86	0	9

Table 6.4.1.20: Task 1 ANOVA test results comparing ELA achievement level for Grade 5, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	51.48	4	12.87	4.04	0.00
Within Groups	1362.97	428	3.18		
Total	1414.45	432	3.27		

Table 6.4.1.21: Task 2 scores by ELA achievement level for Grade 5, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	104	8.73	3.01	1	15
Level 1	118	10.26	2.35	4	15
Level 2	76	10.88	2.01	3	15
Level 3	129	9.37	2.87	0	15

Table 6.4.1.22: Task 2 ANOVA test results comparing ELA achievement level for Grade 5, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	255.82	4	63.95	9.24	0
Within Groups	2935.89	424	6.92		
Total	3191.71	428	7.46		

Table 6.4.1.23: Task 3 scores by ELA achievement level for Grade 5, EOU 2

	Obs	Mean	Std. Dev.	Min	Max
Level 0	109	7.08	2.61	1	13
Level 1	123	8.63	2.02	3	13
Level 2	75	9.52	1.96	4	13
Level 3	127	8.15	2.53	1	13

Table 6.4.1.24: Task 3 ANOVA test results comparing ELA achievement level for Grade 5, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	290.08	4	72.52	13.44	0
Within Groups	2326.17	431	5.40		
Total	2616.25	435	6.01		

Table 6.4.1.25: EOU2 scores by ELA achievement level for Grade 5, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	98	21.36	6.12	7	37
Level 1	109	25.14	4.72	13	36
Level 2	73	26.75	4.28	16	36
Level 3	107	23.82	5.72	5	34

Table 6.4.1.26: EOU2 ANOVA test results comparing ELA achievement level for Grade 5, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	1380.45	4	345.11	12.22	0
Within Groups	10845.12	384	28.24		
Total	12225.57	388	31.51		

Table 6.4.1.27: Task 1 scores by mathematics achievement level for Grade 5, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	107	5.16	1.91	0	9
Level 1	120	5.93	1.62	2	9
Level 2	83	6.29	1.63	2	9
Level 3	121	5.83	1.89	0	9

Table 6.4.1.28: Task 1 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	71.57	4	17.89	5.7	0.0002
Within Groups	1342.88	428	3.14		
Total	1414.45	432	3.27		

Table 6.4.1.29: Task 2 scores by mathematics achievement level for Grade 5, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	81	10.77	2.13	2	15
Level 1	124	9.47	2.83	0	15
Level 2	27	4.85	1.88	0	9
Level 3	60	4.13	2.32	0	9

Table 6.4.1.30: Task 2 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	229.93	4	57.48	8.23	0
Within Groups	2961.78	424	6.99		
Total	3191.71	428	7.46		

Table 6.4.1.31: Task 3 scores by mathematics achievement level for Grade 5, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	111	6.98	2.72	1	13
Level 1	118	8.62	2.16	3	13
Level 2	82	9.33	1.97	3	13
Level 3	123	8.34	2.30	3	13

Table 6.4.1.32: Task 3 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	292.18	4	73.04	13.55	0
Within Groups	2324.07	431	5.39		
Total	2616.25	435	6.01		

Table 6.4.1.33: EOU2 scores by mathematics achievement level for Grade 5, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	101	21.29	6.21	5	37
Level 1	105	25.28	4.48	14	35
Level 2	78	26.35	4.65	9	36
Level 3	103	24.04	5.66	6	34

Table 6.4.1.34: EOU2 ANOVA test results comparing mathematics achievement level For Grade 5, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	1341.91	4	335.48	11.84	0
Within Groups	10883.66	384	28.34		
Total	12225.57	388	31.51		

Table 6.4.1.35: t-test results for comparison of scores by gender for Grade 5, EOU 3

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	108	127	14.15	14.04	0.11	0.23	233	0.41
Task 2	107	127	8.88	8.79	0.09	0.22	232	0.41
Task 3	97	114	5.22	5.51	-0.29	-1.01	209	0.84
EOU 3	90	103	28.58	29.11	-0.53	-0.53	191	0.71

Table 6.4.1.36: Task 1 scores by ELA achievement level for Grade 5, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	80	12.63	3.57	5	21
Level 1	76	15.11	2.71	7	21
Level 2	27	15.44	2.99	10	21
Level 3	38	14.32	4.63	5	21

Table 6.4.1.37: Task 1 ANOVA test results comparing ELA achievement level for Grade 5, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	301.39	3	100.46	8.45	0
Within Groups	2578.79	217	11.88		
Total	2880.17	220	13.09		

Table 6.4.1.38: Task 2 scores by ELA achievement level for Grade 5, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	80	7.90	2.97	1	14
Level 1	77	9.30	3.21	1	15
Level 2	28	10.00	2.83	5	15
Level 3	36	8.44	3.42	2	14

Table 6.4.1.39: Task 2 ANOVA test results comparing ELA achievement level for Grade 5, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	128.08	3	42.69	4.4	0.01
Within Groups	2104.22	217	9.70		
Total	2232.30	220	10.15		

Table 6.4.1.40: Task 3 scores by ELA achievement level for Grade 5, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	70	4.67	1.75	1	9
Level 1	67	5.40	1.92	2	9
Level 2	29	6.31	1.95	3	10
Level 3	32	5.50	2.78	0	10

Table 6.4.1.41: Task 3 ANOVA test results comparing ELA achievement level for Grade 5, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	59.24	3	19.75	4.78	0.00
Within Groups	801.77	194	4.13		
Total	861.01	197	4.37		

Table 6.4.1.42: EOU3 scores by ELA achievement level for Grade 5, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	61	25.56	6.11	9	37
Level 1	63	30.10	5.51	16	40
Level 2	27	32.04	6.00	19	44
Level 3	29	29.41	9.28	9	43

Table 6.4.1.43: EOU3 ANOVA test results comparing ELA achievement level for Grade 5, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	1040.25	3	346.75	8.17	0
Within Groups	7466.48	176	42.42		
Total	8506.73	179	47.52		

Table 6.4.1.44: Task 1 scores by mathematics achievement level for Grade 5, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	76	12.49	3.96	5	21
Level 1	73	14.92	2.52	9	21
Level 2	33	15.15	3.09	8	21
Level 3	39	14.90	4.08	5	21

Table 6.4.1.45: Task 1 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	307.85	3	102.62	8.66	0
Within Groups	2572.33	217	11.85		
Total	2880.17	220	13.10		

Table 6.4.1.46: Task 2 scores by mathematics achievement level for Grade 5, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	81	7.75	3.34	1	14
Level 1	73	9.40	2.64	3	14
Level 2	32	9.66	3.38	2	15
Level 3	35	8.83	3.18	2	14

Table 6.4.1.47: Task 2 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	137.57	3	45.86	4.75	0.00
Within Groups	2094.73	217	9.65		
Total	2232.30	220	10.15		

Table 6.4.1.48: Task 3 scores by mathematics achievement level for Grade 5, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	73	4.49	1.76	1	9
Level 1	63	5.59	1.96	2	9
Level 2	31	6.16	1.92	3	10
Level 3	31	5.71	2.64	0	10

Table 6.4.1.49: Task 3 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	80.91	3	26.97	6.71	0.00
Within Groups	780.10	194	4.02		
Total	861.01	197	4.37		

Table 6.4.1.50: EOU3 scores by mathematics achievement level for Grade 5, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	63	25.13	6.78	9	39
Level 1	60	30.55	5.13	19	42
Level 2	29	30.93	6.36	16	44
Level 3	28	30.71	7.95	10	43

Table 6.4.1.51: EOU3 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	1267.32	3	422.44	10.27	0
Within Groups	7239.41	176	41.13		
Total	8506.73	179	47.52		

Table 6.4.1.52: t-test results for comparison of scores by gender for Grade 5, EOU 4

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	154	164	8.05	8.84	-0.78	-2.18	316	0.99
Task 2	155	173	4.77	4.70	0.07	0.26	326	0.70
Task 3	154	168	6.54	6.43	0.11	0.33	320	0.37
EOU 4	109	126	20.28	21.04	-0.76	-0.83	233	0.80

Table 6.4.1.53: Task 1 scores by ELA achievement level for Grade 5, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	79	7.86	3.66	0	15
Level 1	100	9.05	3.01	3	15
Level 2	57	7.37	3.16	0	14

Level 3	16	8.06	2.35	4	12

Table 6.4.1.54: Task 1 ANOVA test results comparing ELA achievement level for Grade 5, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	177.92	3	59.31	5.85	0.00
Within Groups	3023.42	298	10.15		
Total	3201.34	301	10.64		

Table 6.4.1.55: Task 2 scores by ELA achievement level for Grade 5, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	83	4.11	2.44	0	11
Level 1	101	5.43	2.79	0	11
Level 2	67	5.52	2.49	1	11
Level 3	63	3.89	2.27	0	10

Table 6.4.1.56: Task 2 ANOVA test results comparing ELA achievement level for Grade 5, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	166.47	3	55.49	8.61	0
Within Groups	1997.66	310	6.44		
Total	2164.13	313	6.91		

Table 6.4.1.57: Task 3 scores by ELA achievement level for Grade 5, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	86	5.73	3.01	0	13
Level 1	102	7.28	2.93	1	13
Level 2	62	7.69	2.63	2	13
Level 3	58	5.43	2.67	0	11

Table 6.4.1.58: Task 3 ANOVA test results comparing ELA achievement level for Grade 5, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	265.80	3	88.60	10.92	0
Within Groups	2467.01	304	8.12		
Total	2732.81	307	8.90		

Table 6.4.1.59: EOU4 scores by ELA achievement level for Grade 5, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	59	18.76	7.345	5	36
Level 1	71	22.72	6.91	8	35
Level 2	50	23.52	5.38	12	34
Level 3	41	18.00	6.74	4	33

Table 6.4.1.60: EOU4 ANOVA test results comparing ELA achievement level for Grade 5, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	1191.25	3	397.08	8.88	0
Within Groups	9703.52	217	44.72		
Total	10894.78	220	49.52		

Table 6.4.1.61: Task 1 scores by mathematics achievement level for Grade 5, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	77	7.87	3.08	0	14
Level 1	102	9.31	3.21	1	15
Level 2	53	7.36	3.23	0	14
Level 3	16	8.06	2.35	4	12

Table 6.4.1.62: Task 1 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	151.38	3	50.46	7.77	0.00
Within Groups	2012.74	310	6.49		
Total	2164.13	313	6.91		

Table 6.4.1.63: Task 2 scores by mathematics achievement level for Grade 5, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	83	4.22	2.71	0	11
Level 1	102	5.68	2.58	0	11
Level 2	68	4.91	2.57	1	11
Level 3	61	3.95	2.22	0	10

Table 6.4.1.64: Task 2 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	151.38	3	50.46	7.77	0.00
Within Groups	2012.74	310	6.49		
Total	2164.13	313	6.91		

Table 6.4.1.65: Task 3 scores by mathematics achievement level for Grade 5, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	84	5.92	2.83	0	12
Level 1	101	7.36	3.00	0	13
Level 2	68	7.24	2.89	2	13
Level 3	55	5.38	2.73	0	11

Table 6.4.1.66: Task 3 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	206.00	3	68.67	8.26	0
Within Groups	2526.80	304	8.31		
Total	2732.81	307	8.90		

Table 6.4.1.67: EOU4 scores by mathematics achievement level for Grade 5, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	55	19.60	6.28	7	34
Level 1	80	23.08	7.18	6	36
Level 2	47	21.47	6.74	9	35
Level 3	39	17.97	6.87	4	33

Table 6.4.1.68: EOU4 ANOVA test results comparing mathematics achievement level for Grade 5, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	819.35	3	273.12	5.88	0.00
Within Groups	10075.43	217	46.43		
Total	10894.78	220	49.52		

Table 6.4.1.69: t-test results for comparison of scores by gender for Grade 8, EOU 1

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	68	52	9.68	9.25	0.43	0.63	118	0.26
Task 2	54	37	6.67	7.35	0.69	0.99	89	0.84
Task 3	62	46	7.48	7.54	0.06	0.11	106	0.54
EOU 1	48	32	25.35	25.78	0.43	0.22	78	0.59

Table 6.4.1.70: Task 1 scores by ELA achievement level for Grade 8, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	33	7.12	3.78	1	15
Level 1	31	9.35	2.82	3	16
Level 2	6	9.67	3.14	6	13
Level 3	13	8.85	3.26	3	14

Table 6.4.1.71: Task 1 ANOVA test results comparing ELA achievement level for Grade 8, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	94.43	3	31.48	2.85	0.04
Within Groups	873.64	79	11.06		
Total	968.07	82	11.81		

Table 6.4.1.72: Task 2 scores by ELA achievement level for Grade 8, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	18	4.89	3.55	0	12
Level 1	23	6.70	1.94	4	10
Level 2	3	7.00	1.00	6	8
Level 3	10	4.60	3.03	0	10

Table 6.4.1.73: Task 2 ANOVA test results comparing ELA achievement level for Grade 8, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	51.79	3	17.26	2.27	0.09
Within Groups	381.05	50	7.62		
Total	432.83	53	8.17		

Table 6.4.1.74: Task 3 scores by ELA achievement level for Grade 8, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	25	6.08	2.55	1	11
Level 1	29	8.31	1.93	4	12
Level 2	4	6.00	0.82	5	7
Level 3	13	6.08	3.23	1	11

Table 6.4.1.75: Task 3 ANOVA test results comparing ELA achievement level for Grade 8, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	86.02	3	28.67	4.96	0.00
Within Groups	386.97	67	5.78		
Total	472.99	70	6.76		

Table 6.4.1.76: EOU1 scores by ELA achievement level for Grade 8, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	16	18.00	8.93	6	33
Level 1	21	25.29	5.77	12	35
Level 2	2	26.50	0.71	26	27
Level 3	4	22.75	12.34	7	34

Table 6.4.1.77: EOU1 ANOVA test results comparing ELA achievement level for Grade 8, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	518.74	3	172.91	2.91	0.05
Within Groups	2319.54	39	59.48		
Total	2838.28	42	67.58		

Table 6.4.1.78: Task 1 scores by mathematics achievement level for Grade 8, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	42	7.60	3.43	1	15
Level 1	24	9.79	3.05	2	16
Level 2	3	9.33	2.31	8	12
Level 3	14	8.29	3.77	1	14

Table 6.4.1.79: Task 1 ANOVA test results comparing mathematics achievement level for Grade 8, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	76.47	3	25.49	2.26	0.09
Within Groups	891.60	79	11.29		
Total	968.07	82	11.81		

Table 6.4.1.80: Task 2 scores by mathematics achievement level for Grade 8, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	22	4.73	3.09	0	12
Level 1	19	7.32	1.73	4	10
Level 2	2	8.00	0.00	8	8
Level 3	11	4.55	2.88	0	10

Table 6.4.1.81: Task 2 ANOVA test results comparing mathematics achievement level for Grade 8, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	95.64	3	31.88	4.73	0.01
Within Groups	337.20	50	6.74		
Total	432.83	53	8.17		

Table 6.4.1.82: Task 3 scores by mathematics achievement level for Grade 8, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	33	6.52	2.12	1	10
Level 1	22	8.36	2.46	3	12
Level 2	3	6.67	1.53	5	8
Level 3	13	5.92	3.33	1	11

Table 6.4.1.83: Task 3 ANOVA test results comparing mathematics achievement level for Grade 8, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	64.06	3	21.35	3.5	0.02
Within Groups	408.92	67	6.10		
Total	472.99	70	6.76		

Table 6.4.1.84: EOU1 scores by mathematics achievement level for Grade 8, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	20	18.20	7.39	6	33
Level 1	16	28.06	4.06	19	35
Level 2	2	25.50	2.12	24	27
Level 3	5	19.80	12.56	7	34

Table 6.4.1.85: EOU1 ANOVA test results comparing mathematics achievement level for Grade 8, EOU 1

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	918.84	3	306.28	6.22	0.00
Within Groups	1919.44	39	49.22		
Total	2838.28	42	67.58		

Table 6.4.1.86: t-test results for comparison of scores by gender for Grade 8, EOU 2

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	66	52	7.92	8.56	-0.63	-1.03	116	0.85
Task 2	59	68	7.49	8.24	-0.74	-1.24	125	0.89
Task 3	62	63	9.10	10.63	-1.54	-2.30	123	0.99
EOU 2	33	30	25.18	29.50	-4.32	-2.08	61	0.98

Table 6.4.1.87: Task 1 scores by ELA achievement level for Grade 8, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	35	7.17	3.39	0	14
Level 1	51	9.49	2.77	1	16
Level 2	6	8.67	3.72	2	12
Level 3	21	7.05	3.26	1	13

Table 6.4.1.88: Task 1 ANOVA test results comparing ELA achievement level for Grade 8, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	150.49	3	50.16	5.18	0.00
Within Groups	1056.00	109	9.69		
Total	1206.50	112	10.77		

Table 6.4.1.89: Task 2 scores by ELA achievement level for Grade 8, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	35	6.69	2.53	2	13
Level 1	55	9.38	2.73	4	13
Level 2	8	9.88	3.80	1	13
Level 3	23	5.74	3.53	0	13

Table 6.4.1.90: Task 2 ANOVA test results comparing ELA achievement level for Grade 8, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	310.76	3	103.59	12.17	0
Within Groups	995.83	117	8.51		
Total	1306.60	120	10.89		

Table 6.4.1.91: Task 3 scores by ELA achievement level for Grade 8, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	50	8.36	3.53	0	16
Level 1	42	12.21	3.10	4	17
Level 2	3	11.00	6.93	3	15
Level 3	27	8.89	3.43	0	15

Table 6.4.1.92: Task 3 ANOVA test results comparing ELA achievement level for Grade 8, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	374.64	3	124.88	10.47	0
Within Groups	1407.26	118	11.93		
Total	1781.90	121	14.73		

Table 6.4.1.93: EOU2 scores by ELA achievement level for Grade 8, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	19	25.11	7.35	9	39
Level 1	24	32.13	6.69	18	43
Level 2	2	19.50	19.09	6	33

Level 3	15	22.87	8.41	4	38

Table 6.4.1.94: EOU2 ANOVA test results comparing ELA achievement level for Grade 8, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	1065.69	3	355.23	5.93	0.00
Within Groups	3354.65	56	59.90		
Total	4420.33	59	74.92		

Table 6.4.1.95: Task 1 scores by mathematics achievement level Grade 8, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	42	7.57	3.24	0	14
Level 1	42	9.55	2.51	2	14
Level 2	7	9.43	4.76	1	16
Level 3	22	6.82	3.36	1	13

Table 6.4.1.96: Task 1 ANOVA test results comparing mathematics achievement level Grade 8, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	144.82	3	48.27	4.96	0.00
Within Groups	1061.68	109	9.74		
Total	1206.50	112	10.77		

Table 6.4.1.97: Task 2 scores by mathematics achievement level Grade 8, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	33	7.24	2.97	2	13
Level 1	51	8.94	2.70	1	13
Level 2	13	10.00	3.367	3	13
Level 3	24	5.67	3.47	0	13

Table 6.4.1.98: Task 2 ANOVA test results comparing mathematics achievement level Grade 8, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	246.38	3	82.13	9.06	0
Within Groups	1060.22	117	9.06		
Total	1306.60	120	10.89		

Table 6.4.1.99: Task 3 scores by mathematics achievement level Grade 8, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	50	8.86	3.74	0	16
Level 1	36	11.83	3.50	3	17
Level 2	8	11.50	3.34	6	15
Level 3	28	8.68	3.55	0	15

Table 6.4.1.100: Task 3 ANOVA test results comparing mathematics achievement level Grade 8, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	250.77	3	83.59	6.44	0.00
Within Groups	1531.13	118	12.98		
Total	1781.90	121	14.73		

Table 6.4.1.101: EOU2 scores by mathematics achievement level Grade 8, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	24	27.54	6.75	14	39
Level 1	18	30.89	9.36	6	43
Level 2	2	30.50	3.54	28	33
Level 3	16	22.00	8.83	4	38

Table 6.4.1.102: EOU2 ANOVA test results comparing mathematics achievement level Grade 8, EOU 2

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	702.10	3	234.03	3.52	0.02
Within Groups	3718.24	56	66.40		
Total	4420.33	59	74.92		

Table 6.4.1.103: t-test results for comparison of scores by gender for Grade 8, EOU 3

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	107	101	6.42	7.17	-0.75	-2.23	206	0.99
Task 2	103	122	6.12	6.61	-0.50	-1.75	223	0.96
Task 3	92	95	4.58	5.75	-1.17	-3.27	185	1.00
EOU 3	72	73	17.22	19.66	-2.44	-2.55	143	0.99

Table 6.4.1.104: Task 1 scores by ELA achievement level for Grade 8, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	57	5.09	2.52	0	10
Level 1	81	7.40	2.10	3	11
Level 2	59	7.61	2.02	0	11
Level 3	15	7.27	2.40	2	11

Table 6.4.1.105: Task 1 ANOVA test results comparing ELA achievement level for Grade 8, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	237.66	3	79.22	16.11	0
Within Groups	1022.89	208	4.92		
Total	1260.54	211	5.97		

Table 6.4.1.106: Task 2 scores by ELA achievement level for Grade 8, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	59	5.58	2.21	0	9
Level 1	93	6.80	2.01	0	9
Level 2	59	6.95	2.00	2	9
Level 3	18	5.61	2.17	0	9

Table 6.4.1.107: Task 2 ANOVA test results comparing ELA achievement level for Grade 8, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	83.41	3	27.80	6.47	0.00
Within Groups	966.65	225	4.30		
Total	1050.06	228	4.61		

Table 6.4.1.108: Task 3 scores by ELA achievement level for Grade 8, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	55	3.93	2.30	0	9
Level 1	70	5.30	2.17	1	9
Level 2	52	6.48	2.40	0	10
Level 3	14	4.79	2.75	0	9

Table 6.4.1.109: Task 3 ANOVA test results comparing ELA achievement level for Grade 8, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	177.47	3	59.16	11.04	0
Within Groups	1001.75	187	5.36		
Total	1179.22	190	6.21		

Table 6.4.1.110: EOU3 scores by ELA achievement level for Grade 8, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	31	13.94	5.71	1	22
Level 1	57	19.30	5.02	7	28
Level 2	50	20.72	5.25	3	30
Level 3	11	18.36	6.30	5	28

Table 6.4.1.111: EOU3 ANOVA test results comparing ELA achievement level for Grade 8, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	927.60	3	309.20	10.85	0
Within Groups	4132.43	145	28.50		
Total	5060.03	148	34.19		

Table 6.4.1.112: Task 1 scores by mathematics achievement level for Grade 8, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	80	5.75	2.38	0	10
Level 1	71	7.37	2.37	0	11
Level 2	41	7.78	2.07	2	11
Level 3	20	7.25	2.22	2	11

Table 6.4.1.113: Task 1 ANOVA test results comparing mathematics achievement level for Grade 8, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	154.29	3	51.43	9.67	0
Within Groups	1106.25	208	5.32		
Total	1260.54	211	5.97		

Table 6.4.1.114: Task 2 scores by mathematics achievement level for Grade 8, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	72	5.74	2.30	0	9
Level 1	85	6.60	2.03	0	9
Level 2	49	7.43	1.61	2	9
Level 3	23	5.83	2.23	0	9

Table 6.4.1.115: Task 2 ANOVA test results comparing mathematics achievement level for Grade 8, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	94.37	3	31.46	7.41	0.00
Within Groups	955.69	225	4.25		
Total	1050.06	228	4.61		

Table 6.4.1.116: Task 3 scores by mathematics achievement level for Grade 8, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	70	4.10	2.19	0	9
Level 1	62	5.21	2.28	0	9
Level 2	40	7.18	2.02	3	10
Level 3	19	4.95	2.68	0	9

Table 6.4.1.117: Task 3 ANOVA test results comparing mathematics achievement level for Grade 8, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	241.92	3	80.64	16.09	0
Within Groups	937.30	187	5.01		
Total	1179.22	190	6.21		

Table 6.4.1.118: EOU3 scores by mathematics achievement level for Grade 8, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	54	16.04	5.62	1	26
Level 1	46	18.89	5.72	3	28
Level 2	34	22.06	4.71	10	30
Level 3	15	19.00	5.45	5	28

Table 6.4.1.119: EOU3 ANOVA test results comparing mathematics achievement level for Grade 8, EOU 3

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	767.76	3	255.92	8.65	0
Within Groups	4292.27	145	29.60		
Total	5060.03	148	34.19		

Table 6.4.1.120: t-test results for comparison of scores by gender for Grade 8, EOU 4

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	12	11	7.00	5.82	1.18	0.96	21	0.17
Task 2	12	11	4.08	4.27	-0.19	-0.13	21	0.55
Task 3	12	10	4.92	5.70	-0.78	-0.60	20	0.72
EOU 4	12	10	16.00	16.50	-0.50	-0.15	20	0.56

Table 6.4.1.121: Task 1 scores by ELA achievement level for Grade 8, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	26	5.58	2.66	0	9
Level 1	16	7.25	2.14	2	10
Level 2	1	9.00	0.00	9	9
Level 3	6	6.17	2.40	2	9

Table 6.4.1.122: Task 1 ANOVA test results comparing ELA achievement level for Grade 8, EOU 4

Source	Sum of Squares (SS)	Df	MS	F	Prob>F
Between Groups	35.37	3	11.79	1.94	0.14
Within Groups	274.18	45	6.09		
Total	309.55	48	6.45		

Table 6.4.1.123: Task 2 scores by ELA achievement level for Grade 8, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	26	5.96	3.94	0	12
Level 1	16	8.44	3.33	2	14
Level 2	1	3.00	0.00	3	3

Level 3	6	3.00	2.10	1	6

Table 6.4.1.124: Task 2 ANOVA test results comparing ELA achievement level for Grade 8, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	152.20	3	50.73	3.96	0.01
Within Groups	576.90	45	12.82		
Total	729.10	48	15.19		

Table 6.4.1.125: Task 3 scores by ELA achievement level for Grade 8, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	23	4.83	2.93	0	10
Level 1	12	7.58	2.84	3	12
Level 2	1	9.00	0.00	9	9
Level 3	5	4.40	2.19	1	7

Table 6.4.1.126: Task 3 ANOVA test results comparing ELA achievement level for Grade 8, EOU 4

Source	Sum of Squares (SS)	Df	MS	F	Prob>F
Between Groups	79.46	3	26.49	3.29	0.03
Within Groups	297.42	37	8.04		
Total	376.88	40	9.42		

Table 6.4.1.127: EOU4 scores by ELA achievement level for Grade 8, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	22	16.32	9.03	0	30
Level 1	12	22.00	6.11	10	30
Level 2	1	21.00	0.00	21	21
Level 3	5	14.80	5.07	10	22

Table 6.4.1.128: EOU4 ANOVA test results comparing ELA achievement level for Grade 8, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	314.33	3	104.78	1.7	0.19
Within Groups	2223.57	36	61.77		
Total	2537.90	39	65.07		

Table 6.4.1.129: Task 1 scores by mathematics achievement level for Grade 8, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	24	5.42	2.47	1	9
Level 1	18	7.28	2.40	0	10
Level 2	2	8.00	1.41	7	9
Level 3	5	6.00	2.65	2	9

Table 6.4.1.130: Task 1 ANOVA test results comparing mathematics achievement level for Grade 8, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	42.11	3	14.04	2.36	0.08
Within Groups	267.44	45	5.94		
Total	309.55	48	6.45		

Table 6.4.1.131: Task 2 scores by mathematics achievement level for Grade 8, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	24	6.42	3.75	0	12
Level 1	18	7.56	4.05	0	14
Level 2	2	2.50	0.71	2	3
Level 3	5	3.20	2.28	1	6

Table 6.4.1.132: Task 2 ANOVA test results comparing mathematics achievement level for Grade 8, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	105.52	3	35.17	2.54	0.07
Within Groups	623.58	45	13.86		
Total	729.10	48	15.19		

Table 6.4.1.133: Task 3 scores by mathematics achievement level for Grade 8, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	21	5.10	2.86	0	10
Level 1	14	6.79	3.40	0	12
Level 2	2	5.00	5.66	1	9
Level 3	4	5.25	1.26	4	7

Table 6.4.1.134: Task 3 ANOVA test results comparing mathematics achievement level for Grade 8, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	25.96	3	8.65	0.91	0.44
Within Groups	350.92	37	9.48		
Total	376.88	40	9.42		

Table 6.4.1.135: EOU4 scores by math achievement level for Grade 8, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Level 0	20	16.95	8.80	2	30
Level 1	14	20.29	7.89	0	30
Level 2	2	15.50	7.78	10	21
Level 3	4	16.00	4.97	10	22

Table 6.4.1.136: EOU4 ANOVA test results comparing mathematics achievement level for Grade 8, EOU 4

Source	Sum of Squares (SS)	df	MS	F	Prob>F
Between Groups	123.59	3	41.20	0.61	0.61
Within Groups	2414.31	36	67.06		
Total	2537.90	39	65.07		

Appendix H. Data Tables for RQ3 (Section 6.4.2 Analyses)

RQ3. Overall, what do the EOU assessment results tell us about students' science learning?

6.4.2 What do the EOU assessment results tell us about student learning in terms of variation in performance across instructional programs, instructional units, and instructional unit sequences?

Table 6.4.2.1: Task 1 scores by curricular unit for Grade 5, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Amplify	40	5.56	2.47	0	10
Inspire	24	7.04	2.71	1	12
Mosa Mack	14	9.86	1.83	5	12
Sail: Garbage Unit	122	7.60	2.28	1	14
Matter curricular materials	73	6.03	2.86	0	12
Other/Unknown	24	7.38	3.05	2	14

Table 6.4.2.2: Task 2 scores by curricular unit for Grade 5, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Amplify	20	2.90	2.27	0	7
Inspire	24	3.96	1.78	0	7
Mosa Mack	15	5.13	1.55	1	7
Sail: Garbage Unit	123	4.79	2.35	0	9
Matter curricular materials	76	3.92	2.02	0	8
Other/Unknown	22	4.32	2.01	1	9

Table 6.4.2.3: Task 3 scores by curricular unit for Grade 5, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Amplify	27	4.07	2.43	0	10
Inspire	24	4.08	2.54	0	10
Mosa Mack	15	6.60	1.88	2	9
Sail: Garbage Unit	123	5.05	1.94	0	9
Matter curricular materials	68	3.75	2.02	0	8

Other/Unknown	19	3.21	1.40	2	7

Table 6.4.2.4: EOU 1 scores by curricular unit for Grade 5, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Amplify	0	0.00	0.00		
Inspire	24	15.08	5.59	5	26
Mosa Mack	14	21.57	4.80	10	26
Sail: Garbage Unit	119	17.50	4.78	3	32
Matter curricular materials	58	13.97	5.30	3	26
Other/Unknown	13	14.77	5.05	6	22

Table 6.4.2.5: t-test results for students taking SAIL with students taking something else for Grade 5, EOU 1

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	175	122	6.55	7.60	-1.04	-3.30	295	1.00
Task 2	157	123	3.97	4.79	-0.82	-3.13	278	1.00
Task 3	153	123	4.07	5.05	-0.98	-3.80	274	1.00
EOU 1	109	119	15.28	17.50	-2.21	-3.16	226	1.00

Table 6.4.2.6: Task 1 scores by curricular unit for Grade 5, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
BOCES: Deer, Deer, Everywhere	179	5.90	1.90	0	9
Inspire	22	4.86	2.14	1	8
Mosa Mack	10	6.10	1.52	4	8
NGSS	85	5.91	1.59	2	9
Other/Unknown	137	5.69	1.76	0	9

Table 6.4.2.7: Task 2 scores by curricular unit for Grade 5, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
BOCES: Deer, Deer, Everywhere	174	10.18	2.70	2	15
Inspire	24	8.58	3.43	2	14

Mosa Mack	7	10.57	1.72	8	13
NGSS	85	9.52	2.49	1	14
Other/Unknown	139	9.47	2.74	0	15

Table 6.4.2.8: Task 3 scores by curricular unit for Grade 5, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
BOCES: Deer, Deer, Everywhere	181	8.85	2.54	1	13
Inspire	24	7.58	3.01	1	11
Mosa Mack	10	7.70	1.25	6	10
NGSS	84	7.65	2.37	2	13
Other/Unknown	137	7.99	2.18	2	13

Table 6.4.2.9: EOU 2 scores by curricular unit for Grade 5, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
BOCES: Deer, Deer, Everywhere	156	25.46	5.93	9	37
Inspire	22	21.68	7.03	5	30
Mosa Mack	7	24.14	4.10	18	31
NGSS	84	23.02	4.84	7	33
Other/Unknown	120	23.63	5.15	6	34

Table 6.4.2.10: t-test results for students taking BOCES with students taking something else for Grade 5, EOU 2

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	254	179	5.71	5.90	-0.19	-1.10	431	0.86
Task 2	255	174	9.43	10.18	-0.75	-2.80	427	1.00
Task 3	255	181	7.83	8.85	-1.02	-4.38	434	1.00
EOU 2	233	156	23.24	25.46	-2.21	-3.88	387	1.00

Table 6.4.2.11: Task 1 scores by curricular unit for Grade 5, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
BOCES	37	13.78	3.02	7	18
Inspire	24	12.13	4.27	5	21

Mosa Mack	19	14.32	4.15	7	19
Mystery Science	83	13.72	2.78	5	19
NGSS Wastewater	70	14.29	3.51	5	20
Other/Unknown	70	14.84	3.56	5	21

Table 6.4.2.12: Task 2 scores by curricular unit for Grade 5, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
BOCES	37	7.35	2.75	2	12
Inspire	24	10.08	3.15	2	14
Mosa Mack	21	11.90	2.57	6	15
Mystery Science	93	9.89	3.16	2	15
NGSS Wastewater	70	8.27	3.22	1	14
Other/Unknown	69	8.00	2.82	2	15

Table 6.4.2.13: Task 3 scores by curricular unit for Grade 5, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
BOCES	36	4.08	1.36	0	7
Inspire	24	5.79	2.30	1	10
Mosa Mack	21	5.10	1.48	3	7
Mystery Science	91	6.11	1.83	0	10
NGSS Wastewater	57	5.84	2.27	1	10
Other/Unknown	59	5.32	2.31	1	10

Table 6.4.2.14: EOU 3 scores by curricular unit for Grade 5, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
BOCES	30	24.67	5.77	10	33
Inspire	24	28.00	8.27	9	44
Mosa Mack	19	31.37	6.57	17	40
Mystery Science	78	29.65	6.16	10	41
NGSS Wastewater	51	29.75	7.96	9	43
Other/Unknown	54	29.13	6.08	13	43

Table 6.4.2.15: Anova results for students using different curricular materials for Grade 5, EOU 3

	Within Group			Between Group				
	Sum of Squares (SS)	df	MS	Sum of Squares (SS)	df	MS	F	Prob>F
Task 1	3415.18	297	11.50	149.55	5	29.91	2.60	0.03
Task 2	2806.84	308	9.11	485.77	5	97.15	10.66	0.00
Task 3	1129.88	282	4.01	119.62	5	23.92	5.97	0.00
EOU 3	11364.52	250	45.46	755.82	5	151.16	3.33	0.01

Table 6.4.2.16: Task 1 scores by curricular unit for Grade 5, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Ambitious Science	49	9.53	2.15	2	13
Mosa Mack	19	9.37	3.21	0	13
Mystery Science	94	6.81	3.06	1	13
NGSS	45	9.44	2.68	3	15
Other/Unknown	111	8.82	3.45	0	15

Table 6.4.2.17: Task 2 scores by curricular unit for Grade 5, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Ambitious Science	50	7.34	1.52	4	10
Mosa Mack	19	8.79	2.53	4	11
Mystery Science	96	3.77	1.99	0	10
NGSS	44	4.59	2.25	1	9
Other/Unknown	119	3.82	2.21	0	9

Table 6.4.2.18: Task 3 scores by curricular unit for Grade 5, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Ambitious Science	50	7.92	2.29	3	12
Mosa Mack	21	8.52	3.43	2	12
Mystery Science	92	5.90	2.94	0	13
NGSS	44	5.75	2.57	1	10
Other/Unknown	115	6.23	3.01	0	13

Table 6.4.2.19: EOU 4 scores by curricular unit for Grade 5, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Ambitious Science	49	24.80	4.42	13	33
Mosa Mack	17	26.59	7.97	13	36
Mystery Science	67	16.82	6.44	4	31
NGSS	43	20.00	6.38	10	34
Other/Unknown	59	20.47	6.67	7	35

Table 6.4.2.20: Anova results for students taking different curriculum for Grade 5, EOU 4

	Within Group			Between Group				
	Sum of Squares (SS)	df	MS	Sum of Squares (SS)	df	MS	F	Prob>F
Task 1	2900.69	313	9.27	386.20	4	96.55	10.42	0
Task 2	1397.27	323	4.33	840.66	4	210.16	48.58	0
Task 3	2601.41	317	8.21	252.98	4	63.24	7.71	0
EOU 4	8982.64	230	39.05	2443.68	4	610.92	15.64	0

Table 6.4.2.21: Task 1 scores by curricular unit for Grade 8, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Amplify	46	8.15	3.29	1	14
Open Sci-Ed	70	10.37	3.60	3	16
STEMscopes	4	9.50	4.80	3	14

Table 6.4.2.22: Task 2 scores by curricular unit for Grade 8, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Amplify	15	5.47	2.00	2	8
Open Sci-Ed	69	7.52	3.25	0	13
STEMscopes	7	4.43	3.51	0	10

Table 6.4.2.23: Task 3 scores by curricular unit for Grade 8, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Amplify	29	7.17	2.22	2	11
Open Sci-Ed	70	7.99	2.75	0	12
STEMscopes	9	4.89	2.71	1	10

Table 6.4.2.24: EOU 1 scores by curricular unit for Grade 8, EOU 1

	Obs.	Mean	Std. Dev.	Min	Max
Amplify	8	23.88	7.74	8	33
Open Sci-Ed	69	25.96	8.43	6	40
STEMscopes	3	20.00	13.53	7	34

Table 6.4.2.25: t-test results for students taking Open Sci-Ed with students taking something else for Grade 8, EOU 1

	# of Group 0	# of Group 1	Mean Group 0	Mean Group 1	diff	t	df	pr(T>t)
Task 1	50	70	8.26	10.37	-2.11	-3.24	118	1.00
Task 2	22	69	5.14	7.52	-2.39	-3.15	89	1.00
Task 3	38	70	6.63	7.99	-1.35	-2.52	106	0.99
EOU 1	11	69	22.82	25.96	-3.14	-1.14	78	0.87

Table 6.4.2.26: Task 1 scores by curricular unit for Grade 8, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Combination LPS and Open Sci-Ed	0				
Open Sci-Ed	65	8.49	3.43	0	14
Other	53	7.85	3.14	1	16

Table 6.4.2.27: Task 2 scores by curricular unit for Grade 8, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Combination LPS and Open Sci-Ed	27	8.22	3.34	2	13
Open Sci-Ed	48	8.48	2.81	2	13
Other	52	7.17	3.80	0	13

Table 6.4.2.28: Task 3 scores by curricular unit for Grade 8, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Combination LPS and Open Sci-Ed	24	10.13	3.03	4	15
Open Sci-Ed	60	10.15	4.22	0	17
Other	41	9.32	3.57	0	16

Table 6.4.2.29: EOU 2 scores by curricular unit for Grade 8, EOU 2

	Obs.	Mean	Std. Dev.	Min	Max
Combination LPS and Open Sci-Ed	0				
Open Sci-Ed	38	29.71	7.51	14	43
Other	25	23.48	8.55	4	38

Table 6.4.2.30: t-test results for students taking Open Sci-Ed with students taking something else for Grade 8, EOU 2

	# of group 0	# of group 1	Mean group 0	Mean group 1	diff	t	df	pr(T>t)
Task 1	53	65	7.85	8.49	-0.64	-1.05	116	0.85
Task 2	79	48	7.53	8.48	-0.95	-1.54	125	0.94
Task 3	65	60	9.62	10.15	-0.53	-0.79	123	0.78
EOU 2	25	38	23.48	29.71	-6.23	-3.05	61	1.00

Table 6.4.2.31: Task 1 scores by curricular unit for Grade 8, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
LPS	0				
Open Sci-Ed	81	6.56	2.43	0	11
Other	131	7.00	2.44	0	11

Table 6.4.2.32: Task 2 scores by curricular unit for Grade 8, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
LPS	29	7.69	1.04	6	9
Open Sci-Ed	64	6.72	1.86	2	9

Other	136	6.02	2.32	0	9

Table 6.4.2.33: Task 3 scores by curricular unit for Grade 8, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
LPS	25	5.88	1.79	2	9
Open Sci-Ed	58	4.86	2.27	0	9
Other	108	5.20	2.72	0	10

Table 6.4.2.34: EOU 3 scores by curricular unit for Grade 8, EOU 3

	Obs.	Mean	Std. Dev.	Min	Max
LPS	0				
Open Sci-Ed	49	18.80	5.00	7	28
Other	100	18.49	6.24	1	30

Table 6.4.2.35: t-test results for students taking Open Sci-Ed or LPS with students taking something else for Grade 8, EOU 3

	# of group 0	# of group 1	Mean group 0	Mean group 1	diff	t	df	pr(T>t)
Task 1	131	81	7.00	6.56	-0.44	1.27	210	0.10
Task 2	165	64	6.32	6.72	-0.40	-1.28	227	0.90
Task 3	133	58	5.33	4.86	0.47	1.20	189	0.12
EOU 3	100	49	18.49	18.80	-0.31	-0.30	147	0.62

Table 6.4.2.36: Task 1 scores by curricular unit for Grade 8, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Open Sci-Ed	40	6.23	2.56	0	10
Other	10	6.50	2.46	2	9

Table 6.4.2.37: Task 2 scores by curricular unit for Grade 8, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Open Sci-Ed	40	7.15	3.80	0	14
Other	10	2.60	1.78	1	6

Table 6.4.2.38: Task 3 scores by curricular unit for Grade 8, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Open Sci-Ed	33	5.88	3.22	0	12
Other	9	4.89	2.20	1	9

Table 6.4.2.39: EOU 4 scores by curricular unit for Grade 8, EOU 4

	Obs.	Mean	Std. Dev.	Min	Max
Open Sci-Ed	32	18.72	8.57	0	30
Other	9	14.67	4.58	10	22

Table 6.4.2.40: t-test results for students taking Open Sci-Ed with students taking something else for Grade 8, EOU 4

	# of group 0	# of group 1	Mean group 0	Mean group 1	diff	t	df	pr(T>t)
Task 1	10	40	6.50	6.23	0.28	0.31	48	0.38
Task 2	10	40	2.60	7.15	-4.55	-3.67	48	1.00
Task 3	9	33	4.89	5.88	-0.99	-0.87	40	0.80
EOU 4	9	32	14.67	18.72	-4.05	-1.36	39	0.91

Appendix I. Data Tables for Section 6.5 (Using Data for Revisions)

Table 6.5.1: flags for prompts based on p-values and correlations for Grade 5 EOU 1

ItemID	N	Observed Max	Possible Max	pvalue	item total correlation	item total correlation adjusted	flag based on p-value	flag based on correlation
EOU1_T1_P1	325	4	4	0.56	0.50	0.29	0	0
EOU1_T1_P2	319	4	4	0.24	0.37	0.23	1	0
EOU1_T1_P3	315	3	3	0.73	0.55	0.40	0	0
EOU1_T1_P4	313	3	3	0.47	0.62	0.48	0	0
EOU1_T2_P1_AB	299	4	4	0.50	0.51	0.33	0	0
EOU1_T2_P1_C	291	3	3	0.14	0.35	0.25	1	0
EOU1_T2_P2	289	2	2	0.35	0.49	0.35	0	0
EOU1_T2_P3	282	3	3	0.51	0.53	0.39	0	0
EOU1_T3_P1	310	4	4	0.47	0.59	0.37	0	0
EOU1_T3_P2_AB	294	3	3	0.56	0.50	0.36	0	0
EOU1_T3_P3	276	4	4	0.19	0.48	0.38	1	0

Table 6.5.2: Flags for prompts based on IRT data for Grade 5 EOU 1

ItemID	N	b1	b2	b3	b4	outfit	z.outfit	infit	z.infit	Flag (irt data)
EOU1_T1_P1	325	-0.83	-1.12	1.46	-0.60	1.04	0.52	1.05	0.64	0
EOU1_T1_P2	319	-0.68	1.59	1.30	1.66	0.95	-0.37	0.93	-0.60	0
EOU1_T1_P3	315	0.78	-1.18	-1.25		0.62	-3.16	0.76	-2.67	1
EOU1_T1_P4	313	-0.05	-0.28	0.81		0.78	-3.37	0.79	-3.23	1
EOU1_T2_P1_AB	299	-0.81	-0.43	0.54	0.70	0.90	-1.27	0.91	-1.21	0
EOU1_T2_P1_C	291	0.77	2.32	2.32		0.88	-1.05	0.86	-1.25	0
EOU1_T2_P2	289	1.25	-0.28			0.86	-1.90	0.89	-1.97	0
EOU1_T2_P3	282	-0.37	-0.42	1.01		0.84	-2.23	0.85	-2.17	0
EOU1_T3_P1	310	0.23	-0.14	0.12	0.28	0.88	-1.59	0.88	-1.69	0
EOU1_T3_P2_AB	294	-0.83	-0.45	0.87		0.90	-1.30	0.90	-1.36	0

EOU1_T3_P3	276	-0.35	2.00	1.82	2.45	0.74	-2.49	0.74	-2.45	1

Table 6.5.3: flags for prompts based on p-values and correlations for Grade 5 EOU 2

ItemID	N	Observed Max	Possible Max	pvalue	item total correlation	item total correlation adjusted	flag based on p-value	flag based on correlation
EOU2_T1_P1	459	4	4	0.52	0.59	0.44	0	0
EOU2_T1_P2	456	2	2	0.87	0.45	0.36	0	0
EOU2_T1_P3	451	3	3	0.61	0.46	0.34	0	0
EOU2_T2_P1_A	464	4	4	0.80	0.62	0.47	0	0
EOU2_T2_P1_B	458	3	3	0.51	0.59	0.47	0	0
EOU2_T2_P2	456	3	3	0.41	0.52	0.41	0	0
EOU2_T2_P3_A	443	2	2	0.90	0.42	0.34	0	0
EOU2_T2_P3_B	443	3	3	0.59	0.57	0.45	0	0
EOU2_T3_P1	456	3	3	0.52	0.58	0.47	0	0
EOU2_T3_P2_A	453	4	4	0.72	0.56	0.41	0	0
EOU2_T3_P2_B	452	2	2	0.83	0.49	0.43	0	0
EOU2_T3_P3	444	4	4	0.51	0.67	0.54	0	0

Table 6.5.4: Flags for prompts based on IRT data for Grade 5 EOU 2

ItemID	N	b1	b2	b3	b4	outfit	z.outfit	infit	z.infit	Flag (irt data)
EOU2_T1_P1	459	-1.92	0.12	0.10	0.91	0.92	-1.32	0.92	-1.29	0
EOU2_T1_P2	456	-0.41	-2.46			0.78	-1.26	0.98	-0.16	1
EOU2_T1_P3	451	-1.64	-0.79	1.09		0.98	-0.28	0.99	-0.09	0
EOU2_T2_P1_A	464	-0.59	-0.64	-1.21	-1.24	0.88	-0.86	0.92	-0.88	0

EOU2_T2_P1_B	458	-1.19	0.10	0.86		0.86	-2.53	0.86	-2.49	0
EOU2_T2_P2	456	-1.47	0.87	1.68		0.91	-1.28	0.91	-1.38	0
EOU2_T2_P3_A	443	-0.73	-2.60			0.70	-1.52	1.00	0.00	1
EOU2_T2_P3_B	443	-1.46	-0.24	0.58		0.87	-2.17	0.88	-2.10	0
EOU2_T3_P1	456	-1.52	-0.09	1.34		0.84	-2.65	0.84	-2.67	0
EOU2_T3_P2_A	453	-1.73	-0.59	-1.23	0.21	0.94	-0.67	0.95	-0.63	0
EOU2_T3_P2_B	452	-4.15	-0.77			0.82	-2.79	0.85	-2.78	0
EOU2_T3_P3	444	-1.18	-0.40	0.19	1.61	0.84	-2.76	0.83	-2.94	0

Table 6.5.5: Flags for prompts based on p-values and correlations for Grade 5 EOU 3

ItemID	N	Observed Max	Possible Max	pvalue	item total correlation	item total correlation adjusted	flag based on p-value	flag based on correlation
EOU3_T1_P1_AD	334	4	4	0.77	0.48	0.39	0	0
EOU3_T1_P1_E	333	3	3	0.52	0.59	0.52	0	0
EOU3_T1_P2_A	332	3	3	0.85	0.34	0.24	0	0
EOU3_T1_P2_B	325	1	1	0.68	0.38	0.33	0	0
EOU3_T1_P2_C	325	3	3	0.36	0.59	0.53	0	0
EOU3_T1_P3_A	319	3	3	0.80	0.53	0.45	0	0
EOU3_T1_P3_B	317	2	2	0.67	0.37	0.30	0	0
EOU3_T1_P3_C	309	2	2	0.51	0.40	0.32	0	0
EOU3_T2_P1_A	332	3	3	0.90	0.41	0.36	0	0
EOU3_T2_P1_BC	330	2	2	0.71	0.50	0.44	0	0
EOU3_T2_P2	330	4	4	0.56	0.63	0.50	0	0
EOU3_T2_P3_A	325	3	3	0.40	0.51	0.39	0	0
EOU3_T2_P3_B	324	3	3	0.42	0.49	0.39	0	0

EOU3_T2_P4_A	313	3	3	0.70	0.45	0.36	0	0
EOU3_T2_P4_BC	308	3	3	0.44	0.62	0.54	0	0
EOU3_T3_P1_A	320	1	1	0.61	0.26	0.21	0	0
EOU3_T3_P1_B	308	3	3	0.47	0.46	0.38	0	0
EOU3_T3_P2_AB	308	2	2	0.70	0.61	0.56	0	0
EOU3_T3_P3_AB	304	2	2	0.66	0.56	0.50	0	0
EOU3_T3_P3_C	304	2	2	0.66	0.35	0.27	0	0
EOU3_T3_P4_AB	300	2	2	0.33	0.49	0.41	0	0

Table 6.5.6: Flags for prompts based on IRT data for Grade 5 EOU 3

ItemID	N	b1	b2	b3	b4	outfit	z.outfit	infit	z.infit	Flag (irt data)
EOU3_T1_P1_AD	334	-1.86	-1.82	-0.93	0.048	1.00	0.06	1.01	0.10	0
EOU3_T1_P1_E	333	-1.40	-0.11	1.26		0.79	-3.05	0.78	-3.10	1
EOU3_T1_P2_A	332	0.40	-1.38	-2.14		1.57	2.10	1.25	1.76	1
EOU3_T1_P2_B	325					0.94	-0.85	0.96	-0.78	0
EOU3_T1_P2_C	325	-0.95	0.98	2.25		0.76	-3.12	0.76	-3.25	1
EOU3_T1_P3_A	319	-1.55	-1.24	-0.72		0.83	-1.51	0.87	-1.33	0
EOU3_T1_P3_B	317	-1.93	0.17			0.99	-0.12	1.01	0.13	0
EOU3_T1_P3_C	309	-0.70	0.67			0.96	-0.65	0.97	-0.46	0
EOU3_T2_P1_A	332	-3.46	-2.16	-1.37		0.74	-1.89	0.84	-1.25	1
EOU3_T2_P1_BC	330	-0.86	-0.74			0.94	-0.65	0.93	-0.86	0
EOU3_T2_P2	330	-0.04	-0.23	-0.30	-0.01	0.96	-0.45	0.98	-0.27	0
EOU3_T2_P3_A	325	1.52	-0.54	-0.02		1.06	0.64	1.06	0.89	0
EOU3_T2_P3_B	324	-0.35	0.23	1.22		0.95	-0.62	0.98	-0.30	0
EOU3_T2_P4_A	313	-1.30	-0.98	0.02		1.02	0.22	1.04	0.51	0
EOU3_T2_P4_BC	308	-0.22	0.15	0.80		0.80	-2.83	0.83	-2.65	0

EOU3_T3_P1_A	320					0.96	-0.67	0.98	-0.37	0
EOU3_T3_P1_B	308	-1.52	0.18	2.01		0.98	-0.18	0.99	-0.15	0
EOU3_T3_P2_AB	308	-1.14	-0.44			0.79	-2.55	0.83	-2.37	1
EOU3_T3_P3_AB	304	-0.67	-0.50			0.78	-2.83	0.82	-2.69	1
EOU3_T3_P3_C	304	-0.77	-0.40			1.05	0.65	1.03	0.47	0
EOU3_T3_P4_AB	300	0.65	0.66			0.90	-1.25	0.94	-0.83	0

Table 6.5.7: Flags for prompts based on p-values and correlations for Grade 5 EOU 4

ItemID	N	Observed Max	Possible Max	pvalue	item total correlation	item total correlation adjusted	flag based on p-value	flag based on correlation
EOU4_T1_P1_A	374	3	3	0.73	0.39	0.26	0	0
EOU4_T1_P1_B	366	4	4	0.62	0.70	0.58	0	0
EOU4_T1_P1_C	358	2	2	0.52	0.45	0.36	0	0
EOU4_T1_P1_D	352	3	3	0.40	0.52	0.43	0	0
EOU4_T1_P2	346	4	4	0.27	0.55	0.44	0	0
EOU4_T2_P1	366	2	2	0.50	0.50	0.42	0	0
EOU4_T2_P2_A	354	2	2	0.49	0.54	0.43	0	0
EOU4_T2_P2_BC	345	4	4	0.39	0.59	0.47	0	0
EOU4_T2_P3	335	3	3	0.34	0.51	0.40	0	0
EOU4_T3_P1_A	376	2	2	0.40	0.60	0.52	0	0
EOU4_T3_P1_B	364	3	3	0.33	0.59	0.50	0	0
EOU4_T3_P2_AB	357	4	4	0.57	0.61	0.48	0	0
EOU4_T3_P2_C	345	2	2	0.66	0.43	0.35	0	0
EOU4_T3_P3	331	2	2	0.38	0.52	0.44	0	0

Table 6.5.8: Flags for prompts based on IRT data for Grade 5 EOU 4

ItemID	N	b1	b2	b3	b4	outfit	z.outfit	infit	z.infit	Flag (irt data)
EOU4_T1_P1_A	374	-0.37	-0.93	-0.85		1.18	1.32	1.14	1.46	0
EOU4_T1_P1_B	366	0.36	-0.94	-0.44	0.02	0.69	-3.33	0.69	-4.08	1
EOU4_T1_P1_C	358	-0.44	0.41			0.91	-1.29	0.93	-1.16	0
EOU4_T1_P1_D	352	-0.85	0.71	1.77		0.78	-2.96	0.78	-2.98	1
EOU4_T1_P2	346	0.30	0.21	1.60	3.66	0.89	-1.39	0.91	-1.24	0
EOU4_T2_P1	366	-0.80	0.80			0.88	-1.71	0.89	-1.63	0
EOU4_T2_P2_A	354	2.44	-2.35			0.87	-1.58	0.90	-1.67	0
EOU4_T2_P2_BC	345	-0.68	0.14	1.41	0.63	0.84	-1.93	0.84	-2.06	0
EOU4_T2_P3	335	-0.48	1.14	1.31		1.02	0.20	0.97	-0.33	0
EOU4_T3_P1_A	376	0.37	0.43			0.81	-2.82	0.82	-3.10	0
EOU4_T3_P1_B	364	-0.25	1.32	0.86		0.78	-2.65	0.77	-2.97	1
EOU4_T3_P2_AB	357	-1.27	-0.04	0.12	0.16	0.88	-1.52	0.86	-1.83	0
EOU4_T3_P2_C	345	-1.17	-0.08			0.96	-0.51	0.96	-0.49	0
EOU4_T3_P3	331	0.21	1.00			0.93	-1.01	0.94	-0.91	0

Table 6.5.9: Flags for prompts based on p-values and correlations for Grade 8 EOU 1

ItemID	N	Observed Max	Possible Max	pvalue	item total correlation	item total correlation adjusted	flag based on p-value	flag based on correlation
EOU1_T1_P1_A	141	4	4	0.54	0.72	0.63	0	0
EOU1_T1_P1_B	140	3	3	5.00	0.43	0.34	0	0
EOU1_T1_P1_C	140	2	2	0.48	0.66	0.61	0	0
EOU1_T1_P2_A	139	2	2	0.76	0.32	0.23	0	0
EOU1_T1_P2_B	134	2	2	0.33	0.55	0.49	0	0
EOU1_T1_P3_AB	134	4	4	0.52	0.65	0.55	0	0
EOU1_T2_P1	136	2	3	0.18	0.60	0.54	1	0

EOU1_T2_P2	129	3	3	0.66	0.73	0.67	0	0
EOU1_T2_P3_A	119	2	2	0.55	0.73	0.68	0	0
EOU1_T2_P3_B	113	3	3	0.26	0.63	0.55	0	0
EOU1_T2_P4_A	106	2	2	0.44	0.29	0.23	0	0
EOU1_T2_P4_B	106	2	2	0.53	0.4	0.41	0	0
EOU1_T3_P1_AB	136	2	2	0.65	0.51	0.44	0	0
EOU1_T3_P1_C	127	2	2	0.67	0.38	0.31	0	0
EOU1_T3_P2	119	3	3	0.31	0.62	0.55	0	0
EOU1_T3_P3	114	4	4	0.47	0.62	0.54	0	0
EOU1_T3_P4	114	3	3	0.56	0.57	0.49	0	0

Table 6.5.10: Flags for prompts based on IRT data for Grade 8 EOU 1

ItemID	N	b1	b2	b3	b4	outfit	z.outfit	infit	z.infit	Flag (irt data)
EOU1_T1_P1_A	141	0.35	-0.60	-0.07	-0.05	0.91	-0.42	0.82	-1.27	0
EOU1_T1_P1_B	140	-1.42	0.31	0.98		1.07	0.58	1.10	0.81	0
EOU1_T1_P1_C	140	-0.32	0.54			0.72	-2.57	0.74	-2.51	1
EOU1_T1_P2_A	139	0.22	-1.95			**2.34**	3.36	**1.51**	2.70	1
EOU1_T1_P2_B	134	0.05	1.78			0.87	-0.98	0.89	-0.90	0
EOU1_T1_P3_AB	134	0.50	-0.38	-0.89	1.16	0.92	-0.39	1.01	0.10	0
EOU1_T2_P1	136	0.51	1.75			0.81	-1.35	0.89	-0.88	0
EOU1_T2_P2	129	-0.44	-1.02	-0.03		0.63	-2.38	0.66	-2.57	1
EOU1_T2_P3_A	119	1.96	-2.08			0.58	-2.69	0.70	-2.85	1
EOU1_T2_P3_B	113	0.90	0.81	1.32		0.82	-0.95	0.84	-1.15	0
EOU1_T2_P4_A	106	-1.02	1.98			1.12	0.84	1.11	0.83	0
EOU1_T2_P4_B	106	-1.92	1.64			0.88	-0.76	0.89	-0.72	0
EOU1_T3_P1_AB	136	-0.05	-0.95			0.93	-0.29	0.99	-0.07	0
EOU1_T3_P1_C	127	-1.52	0.06			1.13	0.92	1.13	1.01	0
EOU1_T3_P2	119	0.16	0.77	1.88		0.79	-1.56	0.74	-2.10	1
EOU1_T3_P3	114	-0.63	-0.82	0.47	4.37	0.87	-0.85	0.92	-0.52	0

EOU1_T3_P4	114	-1.17	-0.17	0.96		1.00	0.06	1.02	0.16	0

Table 6.5.11: Flags for prompts based on p-values and correlations for Grade 8 EOU 2

ItemID	N	Observed Max	Possible Max	pvalue	item total correlation	item total correlation adjusted	flag based on p-value	flag based on correlation
EOU2_T1_P1_A	154	2	2	0.73	0.50	0.46	0	0
EOU2_T1_P1_BC	152	3	3	0.42	0.44	0.35	0	0
EOU2_T1_P2_AB	148	3	3	0.43	0.44	0.37	0	0
EOU2_T1_P3_AB	137	3	3	0.53	0.65	0.59	0	0
EOU2_T1_P3_C	128	2	2	0.31	0.35	0.30	0	0
EOU2_T1_P4	124	4	4	0.37	0.13	0.02	0	1
EOU2_T2_P1_A	157	2	2	0.54	0.50	0.44	0	0
EOU2_T2_P1_B	153	2	2	0.45	0.63	0.58	0	0
EOU2_T2_P1_C	151	2	2	0.45	0.49	0.43	0	0
EOU2_T2_P1_D	150	3	3	0.55	0.57	0.48	0	0
EOU2_T2_P2_A	148	1	1	0.39	0.33	0.28	0	0
EOU2_T2_P2_B	146	3	3	0.68	0.56	0.49	0	0
EOU2_T2_P2_C	145	2	2	0.43	0.53	0.48	0	0
EOU2_T2_P3_A	141	2	2	0.69	0.62	0.57	0	0
EOU2_T2_P3_B	133	4	4	0.47	0.54	0.42	0	0
EOU2_T2_P3_C	127	2	2	0.41	0.48	0.43	0	0
EOU2_T3_P1_A	151	2	2	0.54	0.52	0.47	0	0
EOU2_T3_P1_B	149	3	3	0.46	0.42	0.35	0	0
EOU2_T3_P1_C	146	2	2	0.40	0.43	0.37	0	0
EOU2_T3_P2_AB	143	3	3	0.71	0.47	0.41	0	0
EOU2_T3_P3_A	137	3	4	0.26	0.72	0.66	0	0
EOU2_T3_P3_B	132	2	2	0.42	0.63	0.57	0	0
EOU2_T3_P4_ABC	131	3	3	0.34	0.51	0.44	0	0

Table 6.5.12: Flags for prompts based on IRT data for Grade 8 EOU 2

ItemID	N	b1	b2	b3	b4	outfit	z.outfit	infit	z.infit	Flag (irt data)
EOU2_T1_P1_A	154	-0.56	-1.20			0.47	-2.35	0.56	-2.53	1
EOU2_T1_P1_BC	152	-0.59	0.53	0.88		1.15	0.99	1.14	0.92	0
EOU2_T1_P2_AB	148	-0.65	0.12	1.73		0.94	-0.35	0.93	-0.39	0
EOU2_T1_P3_AB	137	-0.68	-0.25	0.80		0.64	-2.52	0.65	-2.58	1
EOU2_T1_P3_C	128	0.04	2.11			0.95	-0.25	0.94	-0.38	0
EOU2_T1_P4	124	-0.78	0.16	3.11	-0.65	1.99	4.17	1.62	2.93	1
EOU2_T2_P1_A	157	0.02	-0.08			0.94	-0.40	0.93	-0.55	0
EOU2_T2_P1_B	153	0.67	-0.12			0.74	-1.83	0.79	-1.79	1
EOU2_T2_P1_C	151	-1.29	-0.12			0.93	-0.36	0.99	-0.00	0
EOU2_T2_P1_D	150	-0.08	-0.25	0.15		1.04	0.28	1.06	0.42	0
EOU2_T2_P2_A	148					0.92	-0.81	0.95	-0.62	0
EOU2_T2_P2_B	146	-1.31	0.02	-0.54		0.77	-1.20	0.81	-1.24	1
EOU2_T2_P2_C	145	-0.26	1.08			0.89	-0.74	0.89	-0.81	0
EOU2_T2_P3_A	141	0.04	-1.09			0.63	-1.85	0.70	-2.06	1
EOU2_T2_P3_B	133	-0.31	0.77	-0.24	0.76	1.13	0.74	1.17	1.12	0
EOU2_T2_P3_C	127	0.10	1.03			0.82	-1.26	0.85	-1.12	0
EOU2_T3_P1_A	151	-0.75	0.28			0.81	-1.32	0.84	-1.20	0
EOU2_T3_P1_B	149	-0.27	-1.43	0.69		0.99	0.00	1.01	0.12	0
EOU2_T3_P1_C	146	-0.66	-0.22			1.05	0.37	1.07	0.51	0
EOU2_T3_P2_AB	143	-0.71	-0.74	-0.71		0.70	-1.19	0.77	-1.20	1
EOU2_T3_P3_A	137	0.06	0.80	0.75		0.93	-0.34	0.93	-0.44	0
EOU2_T3_P3_B	132	0.36	0.23			0.83	-1.22	0.88	-0.98	0
EOU2_T3_P4_ABC	131	-1.35	0.28	1.98		0.81	-1.17	0.82	-1.12	0

Table 6.5.13: Flags for prompts based on p-values and correlations for Grade 8 EOU 3

ItemID	N	Observed Max	Possible Max	pvalue	item total correlation	item total correlation	flag based on p-value	flag based on correlation

						adjusted		
EOU3_T1_P1_A	230	2	2	0.83	0.50	0.42	0	0
EOU3_T1_P1_B	228	3	3	0.44	0.63	0.51	0	0
EOU3_T1_P2_AB	228	3	3	0.67	0.61	0.51	0	0
EOU3_T1_P2_C	224	3	3	0.54	0.60	0.48	0	0
EOU3_T2_P1_AB	248	4	3	0.99	0.61	0.47	1	0
EOU3_T2_P2	247	2	2	0.79	0.45	0.36	0	0
EOU3_T2_P3	245	3	3	0.57	0.80	0.72	0	0
EOU3_T3_P1_AB	224	4	4	0.65	0.68	0.55	0	0
EOU3_T3_P2_AB	216	3	3	0.40	0.64	0.53	0	0
EOU3_T3_P3_AB	208	3	3	0.36	0.69	0.57	0	0

Table 6.5.14: Flags for prompts based on IRT data for Grade 8 EOU 3

ItemID	N	b1	b2	b3	b4	outfit	z.outfit	infit	z.infit	Flag (irt data)
EOU3_T1_P1_A	230	-1.11	-1.82			0.92	-0.27	0.83	-1.11	0
EOU3_T1_P1_B	228	-1.20	1.06	0.53		0.91	-0.91	0.95	-0.51	0
EOU3_T1_P2_AB	228	-1.01	-1.26	0.18		0.91	-0.67	0.84	-1.43	0
EOU3_T1_P2_C	224	-1.05	-0.12	0.71		0.89	-1.05	0.89	-1.13	0
EOU3_T2_P1_AB	248	-1.55	-1.60	-0.32	-0.38	0.99	-0.00	1.04	0.34	0
EOU3_T2_P2	247	-1.56	-0.97			1.07	0.48	1.06	0.52	0
EOU3_T2_P3	245	-0.80	-0.24	0.43		0.65	-3.68	0.68	-3.67	1
EOU3_T3_P1_AB	224	-1.20	-0.77	-0.54	0.24	0.85	-1.20	0.91	-0.85	0
EOU3_T3_P2_AB	216	-0.91	0.90	1.24		0.87	-1.30	0.86	-1.45	0
EOU3_T3_P3_AB	208	-0.06	1.00	0.70		0.78	-1.99	0.80	-2.14	1

Table 6.5.15: Flags for prompts based on p-values and correlations for Grade 8 EOU 4

ItemID	N	Observed Max	Possible Max	pvalue	item total	item total	flag based	flag based

					correlation	correlation adjusted	on p-value	on correlation
EOU4_T1_P1	50	2	2	0.39	0.38	0.31	0	0
EOU4_T1_P2_A	50	3	3	0.75	0.60	0.51	0	0
EOU4_T1_P2_B	50	2	2	0.38	0.53	0.47	0	0
EOU4_T1_P3_A	50	2	2	0.82	0.45	0.38	0	0
EOU4_T1_P3_B	50	3	3	0.28	0.48	0.39	0	0
EOU4_T2_P1_A	50	2	2	0.42	0.34	0.24	0	0
EOU4_T2_P1_B	50	3	3	0.40	0.54	0.48	0	0
EOU4_T2_P1_C	50	2	2	0.46	0.57	0.50	0	0
EOU4_T2_P2_A	50	3	3	0.50	0.77	0.70	0	0
EOU4_T2_P2_B	50	2	2	0.51	0.73	0.67	0	0
EOU4_T2_P2_C	50	2	3	0.25	0.68	0.61	0	0
EOU4_T3_P1	48	2	2	0.25	0.45	0.37	0	0
EOU4_T3_P2_AB	48	2	2	0.53	0.47	0.39	0	0
EOU4_T3_P2_C	46	2	2	0.48	0.30	0.21	0	0
EOU4_T3_P3_A	43	2	2	0.35	0.64	0.59	0	0
EOU4_T3_P3_B	42	3	3	0.51	0.80	0.73	0	0
EOU4_T3_P3_C	42	2	3	0.32	0.56	0.48	0	0

Table 6.5.16: Flags for prompts based on IRT data for Grade 8, EOU 4

ItemID	N	b1	b2	b3	outfit	z.outfit	infit	z.infit	Flag (irt data)
EOU4_T1_P1	50	-0.47	1.75		1.14	0.75	1.13	0.72	0
EOU4_T1_P2_A	50	-2.20	-0.15	-1.12	0.78	-0.68	0.88	-0.51	1
EOU4_T1_P2_B	50	-0.77	2.52		0.98	-0.02	1.01	0.10	0
EOU4_T1_P3_A	50	-0.16	-2.33		2.27	1.91	0.90	-0.28	1
EOU4_T1_P3_B	50	-0.30	1.58	2.14	1.17	0.75	1.24	1.04	1
EOU4_T2_P1_A	50	0.31	0.43		1.43	1.72	1.24	1.35	1
EOU4_T2_P1_B	50	-1.69	1.53	1.21	0.67	-1.47	0.73	-1.23	1

EOU4_T2_P1_C	50	0.16	0.25		0.82	-0.79	0.86	-0.80	0
EOU4_T2_P2_A	50	-0.21	0.07	0.28	0.64	-1.69	0.70	-1.67	1
EOU4_T2_P2_B	50	0.30	-0.28		0.63	-1.75	0.68	-2.04	1
EOU4_T2_P2_C	50	0.87	0.13		0.66	-1.22	0.75	-1.47	1
EOU4_T3_P1	48	0.98	1.33		1.08	0.33	1.19	0.89	0
EOU4_T3_P2_AB	48	-1.08	0.82		1.01	0.13	1.05	0.31	0
EOU4_T3_P2_C	46	-0.47	0.68		1.25	1.27	1.29	1.57	1
EOU4_T3_P3_A	43	-0.46	2.13		0.75	-1.33	0.77	-1.22	1
EOU4_T3_P3_B	42	-0.43	0.99	-0.91	0.60	-1.74	0.62	-2.15	1
EOU4_T3_P3_C	42	-0.23	0.35		0.94	-0.23	0.97	-0.14	0