Innovations in Science Map, Assessment & Report Technologies

I-SMART

Evaluating Innovative Science Assessments: Evidence from the I-SMART Pilot Study

The authors wish to thank the ATLAS staff members who made significant writing contributions to this report, listed below.

Jennifer Burnes
Jeffrey C. Hoover
Meagan Karvonen
Elizabeth Kavitsky
Jennifer Kobrin
Brooke Nash
Noelle Pablo
W. Jake Thompson

The authors wish to acknowledge Mitch McCann, Cory Henson, and Michelle Shipman for their contributions to the item review that is included in this report. The authors wish to acknowledge Gail Tiemann and Russell Swinburne Romine of the I-SMART leadership team for helping formulate the research questions and specific analyses used in this report.

Preferred Citation: Accessible, Teaching, Learning, and Assessment Systems (2021). *Evaluating Innovative Science Assessments: Evidence from the I-SMART Pilot Study*.

The project described in this report was developed under a grant from the U.S. Department of Education (S368A170009). However, the content does not necessarily represent the policy of the U.S. Department of Education, and you should not assume endorsement by the Federal government.

Project Director: Michael Plummer, Maryland State Department of Education
Principal Investigator: Meagan Karvonen, ATLAS, University of Kansas

Table of Contents

List of Tables

List of Figures

Overview

The purpose of the Innovations in Science Map, Assessment, and Reporting Technologies (I-SMART) project is to improve science achievement and progress across grades for students with significant cognitive disabilities who participate in alternate assessments and for students with or without disabilities who are not meeting grade-level standards in science. The project builds from the Dynamic Learning Maps® (DLM®) Science Alternate Assessment System but extends assessments to align to learning map models in science. This report describes selected project activities associated with Goal 2: Design, develop, and evaluate assessments that incorporate science disciplinary content and science and engineering practices in highly engaging, universally designed, technology-delivered formats. This report focuses on pilot studies conducted to evaluate the new assessments and teacher and student experiences with the assessments. For information on the I-SMART learning map models (Goal 1) or the assessment purpose, framework, and design and development process (earlier Goal 2 activities), please see *Developing and Evaluating Learning Map Models in Science: Evidence from the I-SMART Project* and *Designing, Developing, and Evaluating Innovative Science Assessments: Evidence from the I-SMART project*, respectively.

Introduction

Science content has traditionally been taught as a collection of facts, without deep attention to connecting concepts, constructing knowledge from past experiences, or integration of science practices (DeBoer, 2014; NGSS Lead States, 2013). In 2012, *A Framework for K-12 Science Education* (National Research Council, 2012) introduced a new model of science with three dimensions: disciplinary core ideas, science and engineering practices, and crosscutting concepts. The Framework, used as the foundation for the Next Generation Science Standards (NGSS), changes the emphasis from presenting scientific inquiry as a separate topic to a routine application of science and engineering practices and allows students to explore and demonstrate understanding of concepts. The science and engineering practices overlap with and provide opportunities to strengthen literacy and mathematics knowledge (Stage et al., 2013). Students must learn to apply science concepts across multiple contexts to solve problems, demonstrate conceptual understanding, and learn science vocabulary in order to meet the NGSS expectations.

Organized learning models such as learning progressions and learning maps in science have promise for supporting NGSS-based instruction and assessment. These organized learning models are designed to represent targeted skills and prerequisite concepts or learning experiences that are important in developing conceptual understanding (Alonzo & Gotwals, 2012; Corcoran et al., 2009). Additional learning models are needed to address content in the new NGSS-based standards and to support a new generation of science assessments. I-SMART learning map models have been designed to fill this gap.

Students with significant cognitive disabilities, and students with or without disabilities who perform substantially below grade level, require innovative assessments to meet NGSS expectations. The I-SMART project developed science neighborhood maps (Goal 1) that represent knowledge, skills, and understandings that describe pathways students could take to learn the science performance expectations for students with significant cognitive disabilities (see Swinburne Romine et al., 2018). Those performance expectations are called **Essential Elements (EEs)**. The neighborhood maps provided a structure from which assessments were built (Goal 2 [see Karvonen et al., 2020]) that incorporated science disciplinary content and science and engineering practices into engaging, universally designed, technology-delivered formats. Assessments are comprised of **testlets**, short groups of items that share a context and

an engagement activity. Testlets measure nodes (observable and distinct knowledge or skills) on the maps. I-SMART testlets are designed to be suitable for a broad population of students. Furthermore, I-SMART testlets are designed to be of high interest and engaging, supportive of student decision-making and self-regulation, and aligned to instructional practices.

Pilot Study Goals and Research Questions

One of the culminating objectives of Goal 2 of the I-SMART project is to have a set of testlets that could be models for a future assessment system along with a scoring model that supports reporting of student mastery of nodes in I-SMART neighborhood maps associated with several EEs. As such, the I-SMART pilot study, conducted in the winter and fall of 2019, evaluated the I-SMART testlets, including item quality, item difficulty, and impact of item features on performance. The I-SMART pilot study provided data to select a final scoring model. The pilot study also gathered data on how students engaged with the testlets to inform response process patterns as well as teachers' perceptions of the assessment. The specific research questions included the following:

1. How do item and testlet features impact item and testlet performance?
2. Which type of psychometric scoring model is optimal for the student response data collected from the assessments?
3. Do empirical data support the structure of the learning map models?
4. What are students' and teachers' perceptions of students' experiences with the new testlets and how do teacher perceptions of testlet difficultly align with student performance?

I-SMART Assessment Design, Pilot Design & Administration

This section outlines the I-SMART assessment design, including a description of the learning map neighborhoods, testlet design, the pilot study design, pilot administration, and participant descriptions, including student and teacher demographics.

Assessment Design

The I-SMART project evaluated assessment designs intended to serve a **primary student population**: students with the most significant cognitive disabilities who are eligible to take the Dynamic Learning Maps® (DLM®) alternate assessment in their state, and a **secondary student population**: students who are performing significantly below grade level in science but do not qualify to take the alternate assessment in their state. The secondary population may include students with or without disabilities. While the two populations are expected to achieve different grade-level expectations based on grade-level general or extended content standards, the project evaluated assessment designs for both populations using items closely related to the content of students' science instruction.

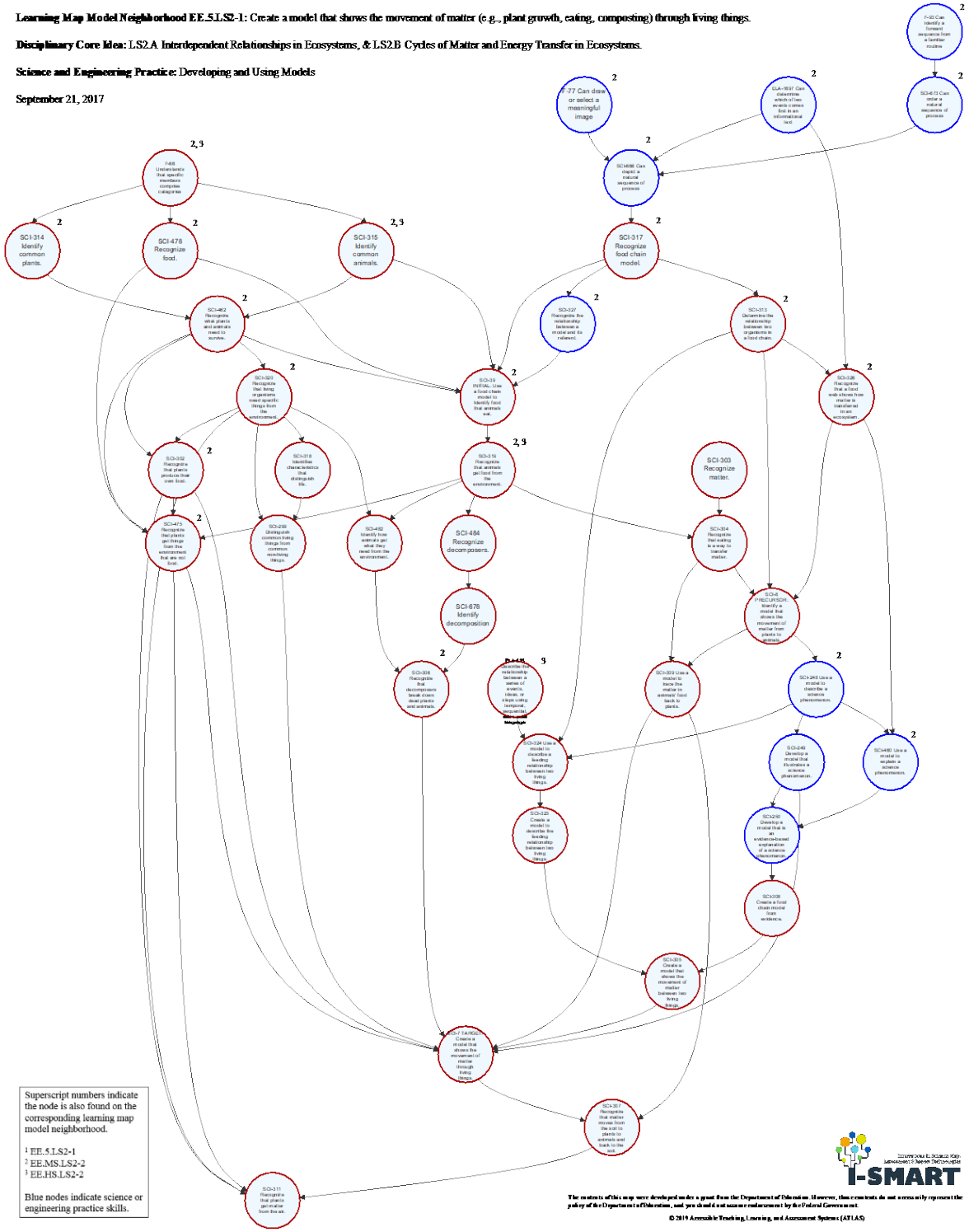
Learning Map Neighborhoods

At the foundation of the I-SMART assessment design are the learning map models that consist of a set of "map neighborhoods." Learning map neighborhoods show multiple pathways by which students develop multidimensional science knowledge, skills, and understandings (KSUs) on their way to, and beyond, mastery of extended content standards called **Essential Elements** (EEs). EEs align to the multidimensional performance expectations in the Next Generation Science Standards (NGSS) in three grades/grade bands: elementary (Grade 5), middle school (Grades 6–8) and high school (Grades 9–12). Neighborhood maps developed for the I-SMART project each represent one EE. I-SMART EEs encompass disciplinary core ideas and science and engineering practices. Two EEs were selected for each grade band to be included in the I-SMART pilot study. Depending on the grade band, EEs measure two of the three available science domains: physical science, life science, and earth and space science. Figure 1 displays an example learning map neighborhood for a single EE.

Within each EE's neighborhood map, groups of nodes are selected to represent three to four different levels of KSU complexity known as **linkage levels**: Initial, Distal (high school only), Precursor, and Target. See *Developing and Evaluating Learning Map Models in Science: Evidence From the I-SMART* project for more details on I-SMART map neighborhoods (Swinburne Romine et al., 2018). Linkage levels are used to assign students to test content that most closely aligns to their level of KSUs; additional information on linkage levels and test form assignment can be found in the Pilot Study Design section of this report.

Figure 1

The Learning Map Model for EE.5.LS2-1



Testlet Structure and Design

I-SMART testlets incorporate principles of Universal Design for Learning (UDL) to provide multiple means of engagement, representation, and action and expression (CAST, 2018). Options to promote UDL principles occur throughout the assessment based on the linkage level.

During the I-SMART pilot study, 34 testlets were administered. For students testing at the initial and precursor linkage levels, **choice items** were presented, which allowed students to pick a context for the academic items to be presented in (e.g., pick a location, animal, or character). The student's selection on the choice item dictated which set of science items the student received, referred to in this report as the "choice path." Subsequent science items were parallel for each choice path, differing only in the context.

A total of 14 testlets at the Initial, Distal, and Precursor linkage levels contained a choice item that branched students into one of two paths. These choice-based testlets were treated as two distinct testlets, yielding 28 testlets. There were two choices for two EEs at two linkage levels for the elementary and middle grade bands (16 testlets) and two choices for the two EEs at three linkage levels for the high school grade band (12 testlets). The six testlets at the Target level (two EEs and three grade bands) did not include a choice item and instead used an expanded science narrative (described below) to promote engagement.

The I-SMART testlets were based on nodes for one linkage level of one EE and included multiple-choice items for the student to complete. Testlets contained a **science narrative,** which provided a theoretical scenario that a student could use to interpret and explore concepts throughout the testlet. The narrative is intended to promote student engagement, activate prior knowledge, and provide appropriate background knowledge if needed for the assessed concepts. Narratives described a student investigating a science phenomenon and were grounded either in familiar context or in an experience found in a typical classroom. After the science narrative, testlets at the Precursor and Target linkage levels presented an unscored **wonder question**, which addressed a key concept in the science narrative. To determine whether the student's initial knowledge changed after completing the items in the testlet, the same wonder question was posed at the end of the testlet. Wonder questions used principles of UDL by providing an option for self-regulation. Precursor and Target linkage level testlets included unscored **think about it** questions to support students' planning and engagement. These questions asked students, "What should you do next?" or "How would you find this answer?"[1] Precursor, Distal, and Target level academic testlets included a **student self-evaluation** as the last item of the testlet. The self-evaluation item asked students if they felt happy, neutral, or sad regarding their performance on the testlet (see Figure 2). See *Designing, Developing, and Evaluating Innovative Science Assessments: Evidence From the I-SMART Project* for more details regarding I-SMART testlet design (Karvonen et al., 2020).

[1] These items had no answer options and are not evaluated in this report.

Figure 2

Student Self-Evaluation

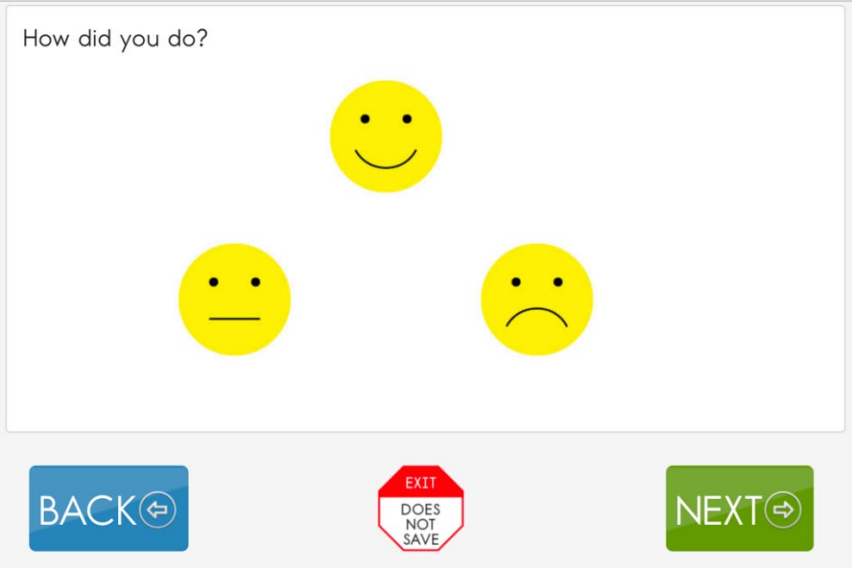
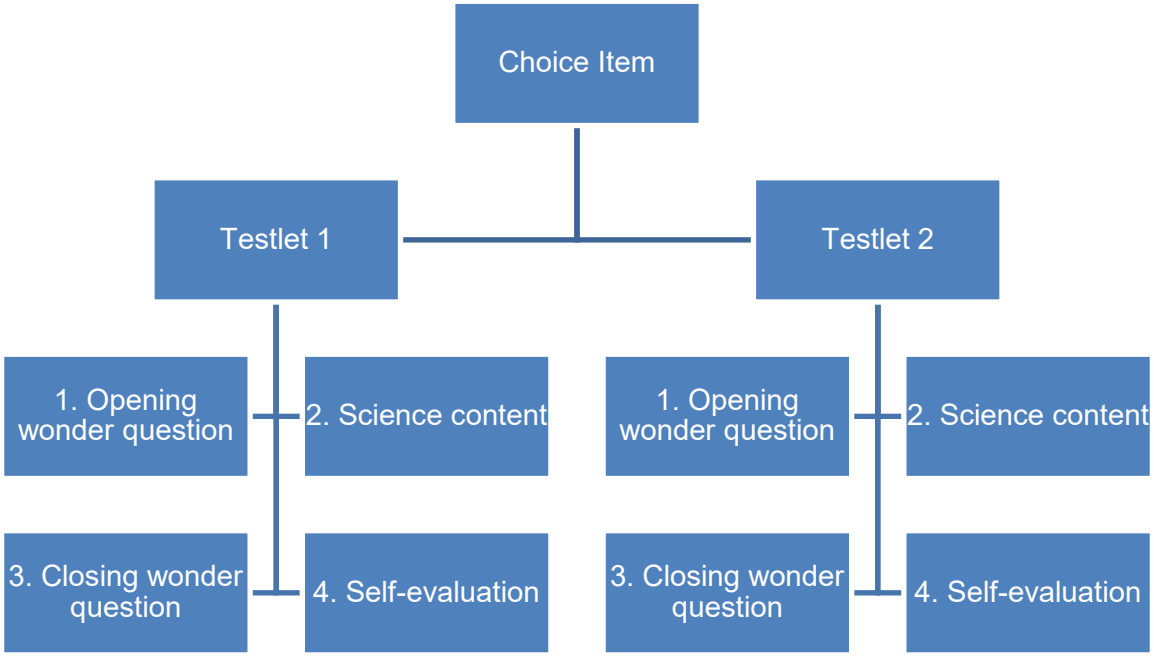


Figure 3 displays a chart of the I-SMART testlet structure beginning at the choice item, followed by the two possible choice paths.

Figure 3

I-SMART Testlet Structure for Initial, Precursor, and Distal Linkage Levels



Each testlet measured four nodes in a cluster or mini-progression. Adjacent linkage levels had one common (overlapping) node. For example, the least complex node in the Precursor linkage level was the same as the most complex node in the Distal linkage level. Each node was measured by three to six items. Specifically, nodes that did not overlap across adjacent linkage levels (i.e., non-overlapping nodes) were each measured by three items, whereas nodes that were shared across adjacent levels (i.e., overlapping nodes) were each measured by four to six items.[2] In total, each testlet comprised 13 to 18 items. For example, as shown in Figure 4, for the elementary grade band, two EEs (EE.5.LS2-1 and EE.5.PS1-3) contained three unique nodes (measured by three items each or total of nine items) and one overlapping node (measured by four to six items), for a total of 13 to 15 items for each testlet at the Initial and Target linkage levels. The Precursor testlets for the two elementary grade band EEs contained two unique nodes (measured by three items each or a total of six items) and two overlapping nodes (measured by eight to twelve items).

[2] The additional items for the overlapping nodes were needed to ensure adequate data were collected to estimate the relationships between linkage levels.

Figure 4

Example I-SMART Testlet Specifications for Elementary

Note. The figure represents the testlet design for the elementary grade band. The middle school grade band had the same design, but the high school grade band contained an additional linkage level (Distal) between the Initial and Precursor level.

Pilot Study Design

Student Populations

Due to differences in pilot design for the primary and secondary populations, the approaches are described separately for each group. Accessibility supports were available for both populations and are described below.

Primary Population

The primary population was students who are eligible for DLM alternate assessments. Students were assigned to pilot test forms by matching the testlet linkage levels with student complexity bands that were determined using responses to a teacher survey about students who take the DLM alternate assessments. The survey included questions about students' science and expressive communication skills and is administered or updated at least annually. The purpose of the complexity bands is to provide students with test content that most closely aligns to their skill level to provide them the greatest opportunity to demonstrate what they know and can do. Because the primary population was a subset of students who already take DLM assessments, the survey responses provided for their participation in the DLM assessment program also applied to the I-SMART pilot.

Students were categorized into one of four complexity bands based on their teacher's responses to the survey: Foundational, Band 1, Band 2, and Band 3. Once a student was assigned to a complexity band, the online assessment system assigned a test form corresponding to the band. Each form consisted of testlets at two adjacent linkage levels, one at the student's current skill level and one either higher or lower than their current skill level. Table 1 shows the correspondence between complexity bands and testlet linkage levels.

Table 1

Complexity Band and Testlet Level Assignments for Primary Population

Grade band	Science band (student)	Initial	Distal	Precursor	Target
EL	F or B1	X		X	
EL	B2 or B3			X	X
MS	F or B1	X		X	
MS	B2 or B3			X	X
HS	F	X	X		
HS	B1		X	X	
HS	B2 or B3			X	X

Note. EL = elementary; MS = middle school; HS = high school; F = Foundational; B1 = Band 1; B2 = Band 2; B3 = Band 3. Distal level only available at the HS level.

Two test forms were available for each complexity band, so students in a complexity band were randomly assigned to take one of the two forms. Table 2 displays the possible form assignments.

Table 2

Test Form Assignments for Primary Population Students

Grade band	Science band	Linkage levels	Form*	Testlet**
EL	F & B1	Initial Precursor	1	001
EL	F & B1	Initial Precursor	1	002
EL	F & B1	Initial Precursor	2	004
EL	F & B1	Initial Precursor	2	005
EL	B2 & B3	Precursor Target	3	002
EL	B2 & B3	Precursor Target	3	003
EL	B2 & B3	Precursor Target	4	005
EL	B2 & B3	Precursor Target	4	006
MS	F & B1	Initial Precursor	5	007
MS	F & B1	Initial Precursor	5	008
MS	F & B1	Initial Precursor	6	010
MS	F & B1	Initial Precursor	6	011
MS	B2 & B3	Precursor Target	7	008
MS	B2 & B3	Precursor Target	7	009
MS	B2 & B3	Precursor Target	8	011
MS	B2 & B3	Precursor Target	8	012
HS	F	Initial Distal	9	013
HS	F	Initial Distal	9	014
HS	F	Initial Distal	10	017
HS	F	Initial Distal	10	018
HS	B1	Distal Precursor	11	014
HS	B1	Distal Precursor	11	015
HS	B1	Distal	12	018

Grade band	Science band	Linkage levels	Form*	Testlet**
HS	B1	Precursor Distal Precursor	12	019
HS	B2 & B3	Precursor Target	13	015
HS	B2 & B3	Precursor Target	13	016
HS	B2 & B3	Precursor Target	14	019
HS	B2 & B3	Precursor Target	14	020

Note. EL = elementary; MS = middle school; HS = high school; F = Foundational; B1 = Band 1; B2 = Band 2; B3 = Band 3.
* Available forms within each grade and complexity band were randomly assigned to students.
** Each Initial- and Precursor-level testlet consisted of two choice paths, for a total of 34 unique testlets across all linkage levels.

In addition to the student self-evaluation question described above, two **teacher survey questions** were administered after each testlet. One question asked teachers about the amount of instructional time spent on the tested EE and the other asked about the teacher's perception of the student's mastery of the EE.

After the student completed both testlets and all embedded survey questions were answered, teachers completed a slightly longer survey about the student's overall experiences with the assessment. The survey was administered in the assessment platform and included questions about student engagement, interaction with the tests as intended, ease of use, familiarity with devices and accessibility features, testlet difficulty, and social-emotional factors that may have impacted the student's responses to test items. Appendix A shows an example pilot test experience for a student from the primary population in fifth grade. The Evaluation of Student Experiences section of the report provides results from the teacher and student surveys.

Secondary Population

The secondary population was students with or without disabilities who do not meet grade-level expectations in science. The testlets for the secondary population were set up to be administered through the same online platform but with a user interface designed for students who do not take alternate assessments.

The available test content for students in this population was limited to the Target level testlets at each grade band. The testlet content varied somewhat from the versions designed for the primary population although the items were nearly identical. The pilot test consisted of two Target level testlets, one per EE.

Students in the secondary population were assigned to pilot test forms using information from a short teacher survey. This survey, which was separate from the DLM teacher survey described above for the primary population, contained questions about students' science academic skills and mastery level of available test content. Depending on their academic skills and mastery levels, students could have been assigned to test content at a different grade level than their current enrolled grade. Secondary population students were to be assigned two Target level

testlets, one for each of the two EEs, at a grade level that was most closely aligned to their current skill level as reported on this survey.

After students had completed the testlets, teachers were to complete a second survey within the assessment platform about the student's experiences with the assessment. This survey was the same as the one that teachers completed for the primary population. Appendix A shows an example pilot test experience for a student from the secondary population in eighth grade.

Available Accessibility Supports

For the primary population, accessibility supports for the I-SMART pilot administration included supports provided in the Kite® technology platform (used to administer the DLM assessments), supports that required additional materials, and supports provided by the test administrator. Test administrators trained on the appropriate selection and use of the supports selected from the various settings options in the Personal Needs and Preferences (PNP) profile (which houses student-specific information that adjusts settings in the testing platform). Table 3 displays the available accessibility supports.

Table 3

Available Accessibility Supports

Supports provided in Kite	Supports requiring additional materials	Supports provided by the test administrator
Spoken audio	Individualized manipulatives	Human read aloud
Magnification	Calculator	Sign interpretation of text
Color contrast	Single-switch system	Language translation of text
Overlay color	Two-switch system	Test administrator entering response for student
Invert color choice		Partner-assisted scanning

In addition to the supports provided in Kite listed in Table 3, the secondary population also had access to additional tools and features in the Kite system that the student could turn on/off during test administration. These tools and features included: highlighter/striker/eraser, guide line, mark for review, notes, search, tags, sketch pad, and masking.

Pilot Administration

The first testing window took place in the winter 2019 (January 22 through March 1, 2019), and the second window took place in the fall of 2019 (November 5 through December 20, 2019).

Recruitment and Inclusion Criteria

Primary Population

State education agency staff from I-SMART state partners, as well as DLM state partners, were asked to recruit teachers to participate in the pilot test. Participation included administering the assigned testlets to one or more eligible students and completing the embedded survey questions and a survey after administering the assessment. Student eligibility criteria included (1) enrollment in DLM assessments for the current school year and (2) receiving instruction in the two selected science EEs for that grade band during the year.

Secondary Population

Districts that agreed to participate for the primary population were also asked if they had teachers willing to participate for the secondary population (in some cases, the same teachers were responsible for both populations). Teacher participation included administering the assigned testlets to one or more eligible students and completing the survey after administering the assessment. The only student participation criterion was that the student could *not* be eligible to take DLM assessments. This criterion was confirmed via a survey question administered to the student's teacher.

Test Administrator Training

Because the secondary population included students not eligible for alternate assessments, test administrators could have been unfamiliar with Kite Suite. Training and assistance were to be provided in the form of an administration manual, a website that served as a central hub of information about procedures, and service desk support.

Assessment Scoring

Latent class analysis is the basis for the scoring model of the I-SMART assessment. For every node in each of the assessed linkage levels within each EE, a probability of mastery was calculated. The nodes were treated as the latent variables in the latent class analysis. Students were then classified as masters and non-masters of the assessed nodes based on posterior probabilities. A posterior probability greater than or equal to .80 was required for node mastery. Students could also be classified as masters of a node if they responded correctly to at least 80% of the items assessing the node. Detailed information about the scoring model is provided in the Psychometric Model section of this report.

Participants

While recruitment efforts were made, unfortunately, no students from the secondary population participated in the pilot studies.[3] The following descriptions of participants and results only include students who participated from the primary population.

The I-SMART pilot study involved a total of 2,144 student participants and 995 teachers from five states. Table 4 summarizes the number of students who tested on each EE at each linkage level. Table 5 displays counts of students by grade band and Science complexity band for each testing window. Tables 6–8 show the number of students and teachers per testing window, by state and by grade. Note that some students and teachers participated in both testing windows. Students and teachers were only counted once in the overall totals.

[3] Teachers who were interested in the pilot study reported lack of time as the main reason for not participating. Some information related to RQ4 (student reactions) is reported elsewhere on cognitive labs with testlet prototypes for the secondary population. See ismart.works.

Table 4

Number of Students by Linkage Level and Essential Element

Linkage level	Essential Element	*n*
Initial	5.LS2-1.IP	150
	5.PS1-3.IP	389
	MS.LS2-2.IP	232
	MS.PS1-2.IP	244
	HS.ESS3-3.IP	129
	HS.LS2-2.IP	107
Distal (high school only)	HS.ESS3-3.DP	281
	HS.LS2-2.DP	350
Precursor	5.LS2-1.PP	190
	5.PS1-3.PP	484
	MS.LS2-2.PP	324
	MS.PS1-2.PP	344
	HS.ESS3-3.PP	232
	HS.LS2-2.PP	435
Target	5.LS2-1.T	40
	5.PS1-3.T	95
	MS.LS2-2.T	93
	MS.PS1-2.T	100
	HS.ESS3-3.T	79
	HS.LS2-2.T	191

Note. Students were able to test on multiple linkage levels and Essential Elements.

Table 5

Number of Students by Grade Band and Science Complexity Band in Each Testing Window

Grade band	Science complexity band	First window	Second window
Elementary school	Foundational	105	95
	Band 1	181	184
	Band 2	66	61
	Band 3	9	5
Middle school	Foundational	91	98
	Band 1	163	144
	Band 2	86	82
	Band 3	24	12
High school	Foundational	132	113
	Band 1	201	215
	Band 2	111	102
	Band 3	32	33
Total		1,201	1,144

Table 6

Number of Students and Teachers by Testing Window

Testing window	Students	Teachers
First	1,201	574
Second	1,144	600

Table 7

Number of Students and Teachers by State and Testing Window

State	Students in first window	Students in second window	Teachers in first window	Teachers in second window	Overall students	Overall teachers
Maryland	0	14	0	7	14	7
Missouri	1,118	1,121	516	589	2,038	926
New York	36	0	26	0	36	26
Oklahoma	47	1	32	1	48	33
West Virginia	0	8	0	3	8	3

Table 8

Number of Students by Grade and Testing Window

Grade	Students in first window	Students in second window	Teachers in first window	Teachers in second window	Overall students	Overall teachers
Grade 3	43	51	32	37	94	63
Grade 4	63	49	47	41	112	81
Grade 5	255	245	183	192	499	341
Grade 6	60	58	41	46	118	81
Grade 7	58	40	44	34	98	76
Grade 8	246	238	158	158	484	284
Grade 9	65	80	48	58	145	99
Grade 10	57	33	48	26	90	70
Grade 11	250	256	162	170	506	296
Grade 12	104	94	50	55	163	85

Student Demographics

The demographic characteristics of students who participated in the I-SMART pilot study are shown in Table 9. Nearly two thirds of the students were male (63.9%). While African American students made up 17.9% of the students, the majority of the students were white (74.6%). Most students were not of Hispanic ethnicity (94.3%) and were not English for Speakers of Other Languages (ESOL) eligible or monitored (97.8%).

Table 9

Student Demographics by Testing Window

Subgroup	First window	%	Second window	%	Overall	%
Gender						
Female	430	35.8	411	35.9	774	36.1
Male	771	64.2	733	64.1	1370	63.9
Race						
White	871	72.5	883	77.2	1600	74.6
African American	236	19.7	180	15.7	385	17.9
Asian	27	2.2	33	2.9	57	2.6
American Indian	15	1.2	7	0.6	21	1.0
Alaska Native	0	0	1	0.1	1	0.1
Two or more races	49	4.1	33	2.9	80	3.7
Native Hawaiian or Pacific Islander	3	0.3	7	0.6	10	0.5
Hispanic ethnicity						
No	1132	94.3	1082	94.6	2023	94.3
Yes	69	5.7	62	5.4	122	5.7
ESOL participation						
Not ESOL eligible or monitored	1175	97.8	1122	98.1	2096	97.8
Title III funded	16	1.3	12	1.1	28	1.3
Both Title III and state ESOL/bilingual funded	1	0.1	1	0.1	2	0.1
Monitored ESOL student	4	0.3	3	0.3	7	0.3
Eligible but not currently receiving services	1	0.1	1	0.1	2	0.1
Receives services but not Title III or state funded	4	.33	5	0.4	9	0.4

Note. ESOL = English for Speakers of Other Languages.

Teacher Experience

Teachers whose students participated in I-SMART had varying levels of experience with science and special education (see Table 10). Forty-five percent (*n* = 361) had between 0 and 5 years of experience teaching science, while almost 49% (*n* = 393) had 0–10 years of experience teaching special education. Over half (57.5%; *n* = 462) had a Master's degree.

Table 10

*Teacher Experience (*N = *803)*

Experience	*n*	%
Science		
0–5 years	361	45.0
6–10 years	162	20.2
11–15 years	101	12.6
16–20 years	69	8.6

21+ years	83	10.3
No answer	27	3.4
Special education		
0–5 years	200	24.9
6–10 years	193	24.0
11–15 years	137	17.1
16–20 years	115	14.3
21+ years	126	15.7
No answer	32	4.0
Highest degree earned		
Bachelor's	333	41.5
Master's	462	57.5
Doctorate	8	1.0

Use of Accessibility Supports

Table 11 shows the PNP profile selections for students who participated in the I-SMART pilot study. Teachers most often selected a human read aloud (89.5%) followed by test administrator enters responses for student (62.8%).

Table 11

I-SMART Personal Needs and Preferences Profile Selections

Selection	*n*	%
Human read aloud	1,918	89.5
Test administrator enters responses for student	1,347	62.8
Individualized manipulatives	876	40.9
Calculator	440	20.5
Color contrast	203	9.5
Magnification	197	9.2
Spoken audio	185	8.6
Partner assisted scanning	122	5.7
Alternate form–visual impairment	113	5.3
Overlay color	102	4.8
Single-switch system	75	3.5
Invert color choice	73	3.4
Two-switch system	52	2.4
Sign interpretation of text	32	1.5
Language translation of text	21	1.0
Uncontracted braille	1	<1.0
Total	5,757	

Note. Multiple selections were allowed, so totals add up to >100%.

Item and Testlet Performance

The pilot study examined student performance on the I-SMART testlets through data regarding node mastery, administration time, performance on choice paths and wonder questions, and item level performance.

Node Mastery

As described in the Testlet Structure and Design section, each testlet measured a total of four nodes within a single Essential Element (EE) and linkage level. Across all testlets, the average nodes mastered on each testlet was 1.1 nodes, with a standard deviation of 1.4 nodes. Appendix D contains bar graphs for node mastery by linkage level, EE, grade band, and domain. Table 12 displays the average nodes mastered on each testlet by linkage level. The Initial linkage level contained the highest average nodes mastered (1.8), followed by Distal (1.1), Precursor (0.9), and Target (0.9).

Table 12

Average Nodes Mastered on Each Testlet by Linkage Level

Linkage level	*n*	Average nodes mastered	Standard deviation
Initial	1,127	1.75	1.60
Distal	562	1.05	1.08
Precursor	1,871	0.87	1.22
Target	577	0.85	1.05

Table 13 displays the average nodes mastered on each testlet by EE. The MS.LS2-2 EE contained the highest average nodes mastered (1.7), while the HS.LS2-2 EE contained the lowest average nodes mastered (0.8).

Table 13

Average Nodes Mastered on Each Testlet by Essential Element

Essential Element	*n*	Average nodes mastered	Standard deviation
5.LS2-1	190	0.92	1.17
5.PS1-3	484	1.31	1.49
MS.LS2-2	325	1.29	1.60
MS.PS1-2	344	1.70	1.41
HS.ESS3-3	361	0.86	1.14
HS.LS2-2	542	0.79	1.04

Table 14 displays the average nodes mastered on each testlet by grade band. The highest average nodes mastered per testlet was in the middle school grade band (1.5), followed by the elementary grade band (1.2) and the high school grade band (0.8).

Table 14

Average Nodes Mastered on Each Testlet by Grade Band

Grade band	*n*	Average nodes mastered	Standard deviation
Elementary	648	1.20	1.42
Middle School	639	1.50	1.52
High School	824	0.82	1.08

Table 15 displays the average nodes mastered on each testlet by domain. Each testlet addressed one of three domains: physical science, life science, and earth and space science. The highest average nodes mastered per testlet was in the physical science domain (1.5), followed by life science (1.0) and earth and space science (0.9).

Table 15

Average Nodes Mastered on Each Testlet by Domain

Domain	*n*	Average nodes mastered	Standard deviation
Physical science	1,656	1.47	1.47
Life science	2,111	0.97	1.28
Earth and space science	721	0.86	1.14

Administration Time

The I-SMART testlets included many features aimed at increasing student engagement (e.g., choice options; see the Testlet Structure and Design section for a complete description). We examined total testlet administration time and time to complete only the scored portion of the testlet to evaluate the impact of the testlet features across the first and second administrations. The distribution of total administration time is displayed in Table 16. Although there were few extreme values greater than an hour, most testlets were completed in less than 6 minutes. This was slightly longer than the reported administration time for Dynamic Learning Maps® (DLM®) Science assessments (2–3 minutes; DLM Consortium, 2019). This discrepancy was likely due to the additional content (i.e., choice items, wonder questions, self-evaluation) and greater number of items that are included in the I-SMART testlets.

Table 16

Distribution of Total Response Times per Testlet in Minutes

Grade	*n*	Minimum	Median	Mean	Maximum	25Q	75Q	IQR
Elementary	1,347	0.18	3.12	4.04	152.27	1.68	5.05	3.37
Middle school	1,333	0.27	3.17	3.88	60.15	1.77	5.10	3.33
High school	1,806	0.20	4.22	4.90	90.00	2.14	6.38	4.24

Note. 25Q = lower quartile; 75Q = upper quartile; IQR = interquartile range.

The administration times for only the scored items measuring nodes were also collected. Table 17 summarizes the distribution of time spent on the scored items measuring nodes, which excludes choice items, wonder questions, and students' self-evaluations. The majority of students took around 4 minutes or less to complete the scored items measuring nodes. This was again in line with expectations based on DLM assessments, given that the I-SMART testlets contain more items than the DLM science testlets.

Table 17

Distribution of Response Times per Testlet in Minutes—Scored Items Measuring Nodes

Grade	*n*	Minimum	Median	Mean	Maximum	25Q	75Q	IQR
Elementary	1,347	0.07	2.53	3.18	133.23	1.36	3.98	2.62
Middle school	1,333	0.17	2.63	3.08	60.02	1.48	3.88	2.40
High school	1,804	0.10	3.58	4.15	67.05	1.78	5.45	3.67

Note. 25Q = lower quartile; 75Q = upper quartile; IQR = interquartile range.

The administration times for just wonder questions were also collected. Table 18 summarizes the distribution of time spent on wonder questions only. Students who skipped one or both wonder questions accounted for the very short times of less than 1 second. The majority of students spent less than 1 minute on the wonder questions.

Table 18

Distribution of Response Times per Testlet in Minutes—Wonder Questions Only

Grade	*n*	Minimum	Median	Mean	Maximum	25Q	75Q	IQR
Elementary	808	0.02	0.52	0.70	33.15	0.25	0.82	0.57
Middle school	858	0.02	0.62	0.75	45.00	0.23	0.93	0.70
High school	937	0.02	0.52	0.68	90.00	0.27	0.75	0.48

Note. 25Q = lower quartile; 75Q = upper quartile; IQR = interquartile range.

For testlets where students completed both wonder questions, the difference in time spent on the first wonder question and the second wonder question was calculated. Table 19 summarizes the distribution of differences in time spent on the first wonder question compared to the second wonder question. On average, students spent more time on the first wonder question compared to the second wonder question.

Table 19

Distribution of Differences in Response Times Between First and Second Wonder Question

Grade	*n*	Minimum	Median	Mean	Maximum	25Q	75Q	IQR
Elementary	748	–31.15	0.03	0.04	12.60	–0.02	0.15	0.17
Middle school	774	–2.43	0.83	0.16	12.70	0.00	0.23	0.23

High school	879	–4.30	0.50	0.08	3.27	0.00	0.17	0.17

Note. 25Q = lower quartile; 75Q = upper quartile; IQR = interquartile range.

Overall, testlet administration times starting from the choice item to the students' self-evaluation took 4 minutes to complete on average, while scored items measuring nodes took 3 minutes on average, and wonder questions took close to 1 minute on average to complete. Most students spent more time answering the first wonder question than the second wonder question.

Choice Items

The choice items did not have a "correct option," nor were they designed to demonstrate one option that was more or less "preferred" than the other. We evaluated the frequency with which students chose the first option versus the second. Of the 14 choice items, six were at the Initial level, two were at the Distal level, and six were at the Precursor level. Table 20 shows the distribution of choice item selections (Option 1 or Option 2) at each linkage level. For all three linkage levels, the majority of students selected the second option.

Table 20

Distribution of Students' Choice Items Selections by Linkage Level

Linkage level	Choice	*n*	%
Initial	1	570	45.6
	2	681	54.4
Distal	1	279	44.3
	2	351	55.7
Precursor	1	823	41.0
	2	1,186	59.0

Distribution of Choice Items by Linkage Level and Choice Items

Table 21 shows the distribution of selections on choice items at each linkage level and preceding choice items. Distal Choice Item 8 had the smallest difference in percentage with a discrepancy of fewer than 2 points between beach and mountain. Eight choice items had a difference in percentage greater than 10 points. The largest difference in percentage was in Precursor Choice Item 11 (83% vs 17%). Some options were available across multiple testlets, such as park on Items 4 and 5. The park was slightly more preferred than house on Choice Item 5 (3 points), but much less preferred than farm on Choice Item 4 (36 points).

Table 21

Distribution of Students' Choice Items Selections by Linkage Level and Choice Item

Linkage level	Choice Item	Choice	Order	*n*	%
Initial	1	Ocean	1	50	46.7
		Forest	2	57	53.3
	2	Park	1	46	35.7
		Beach	2	83	64.3
	3	Bear	1	66	44.0
		Cat	2	84	56.0
	4	Park	1	74	31.9

Linkage level	Choice Item	Choice	Order	*n*	%
		Farm	2	158	68.1
	5	House	1	189	48.6
		Park	2	200	51.4
	6	School	1	145	59.4
		House	2	99	40.6
Distal	7	Whale	1	136	39.0
		Tiger	2	213	61.0
	8	Beach	1	143	50.9
		Mountain	2	138	49.1
Precursor	9	House	1	256	52.9
		Park	2	228	47.1
	10	Camping	1	120	51.7
		Picnic	2	112	48.3
	11	Bread	1	59	17.2
		Cookies	2	285	82.8
	12	Meadow	1	75	39.5
		Pond	2	115	60.5
	13	Pond	1	142	32.6
		Forest	2	293	67.4
	14	Lake	1	171	52.8
		Arctic	2	153	47.2

Student Performance on Choice Paths

When students made a selection on a choice item, they were directed to a testlet with the chosen context. However, the scored items measuring nodes on the two choice items were the same. In other words, the context was the only difference. The choice paths were designed to assess the same four nodes in the same order, so the difficulty of the choice paths should be approximately the same. Student performance was expected to be roughly the same regardless of the chosen choice path. To determine whether student performance differed between choice paths, the average number of nodes students mastered on each testlet was calculated and is displayed in Table 22.[4] Overall, students performed similarly on the different choice paths. The largest difference in average number of nodes mastered was for the Initial school testlet, where the average number of nodes mastered was 2.4 compared to the house testlet, which had an average of 1.1 nodes mastered per student (d = 0.9).

Table 22

Distribution of Students' Choice Items Selections by Linkage Level

Linkage level	Choice Item	Essential Element	Choice	Order	*n*	%	Average nodes mastered
Initial	1	HS.LS2-2.IP	Ocean	1	50	46.7	1.1
		HS.LS2-2.IP	Forest	2	57	53.3	0.7
	2	HS.ESS3-3.IP	Park	1	46	35.7	0.8
		HS.ESS3-3.IP	Beach	2	83	64.3	0.6

[4] For a description of how node mastery was determined, see the section on the psychometric model in this report.

Linkage level	Choice Item	Essential Element	Choice	Order	*n*	%	Average nodes mastered
	3	5.LS2-1.IP	Bear	1	66	44.0	1.5
		5.LS2-1.IP	Cat	2	84	56.0	1.3
	4	MS.LS2-2.IP	Park	1	74	31.9	2.3
		MS.LS2-2.IP	Farm	2	158	68.1	2.4
	5	5.PS1-3.IP	House	1	189	48.6	2.1
		5.PS1-3.IP	Park	2	200	51.4	2.0
	6	MS.PS1-2.IP	School	1	145	59.4	2.4
		MS.PS1-2.IP	House	2	99	40.6	1.1
Distal	7	HS.LS2-2.DP	Whale	1	136	39.0	1.0
		HS.LS2-2.DP	Tiger	2	213	61.0	1.2
	8	HS.ESS3-3.DP	Beach	1	143	50.9	1.1
		HS.ESS3-3.DP	Mountain	2	138	49.1	0.8
Precursor	9	5.PS1-3.PP	House	1	256	52.9	1.0
		5.PS1-3.PP	Park	2	228	47.1	0.6
	10	HS.ESS3-3.PP	Camping	1	120	51.7	0.9
		HS.ESS3-3.PP	Picnic	2	112	48.3	0.7
	11	MS.PS1-2.PP	Bread	1	59	17.2	1.6
		MS.PS1-2.PP	Cookies	2	285	82.8	1.9
	12	5.LS2-1.PP	Meadow	1	75	39.5	0.6
		5.LS2-1.PP	Pond	2	115	60.5	0.5
	13	HS.LS2-2.PP	Pond	1	142	32.6	0.3
		HS.LS2-2.PP	Forest	2	293	67.4	0.5
	14	MS.LS2-2.PP	Lake	1	171	52.8	0.4
		MS.LS2-2.PP	Arctic	2	153	47.2	0.9

The proportion of students who mastered each node was also calculated for each choice path. Appendix B displays the proportion of students mastering each node in a choice path. Each choice path assessed the same four nodes in the same order. For example, on Choice Item 1, the student could choose either a forest or ocean-themed testlet. Both the forest and ocean testlets were written to assess the same four nodes: F-66, SCI-315, SCI-501, and SCI-527. Because only the context changes between the choice paths, it would be expected that the proportion of students to master each node on each choice path is similar, that is, the difficulty should not have been different between the two testlets. Overall, most nodes had similar proportions of mastery between the choice paths; 39 of 56 (70%) nodes had a difference in proportion that was less than 0.1. The largest difference in proportion of mastery between choice paths for a node was for SCI-117 on the house and school testlets. The house testlet had a proportion of mastery of 0.2 on the SCI-117 node, while the school testlet had a higher proportion of mastery of 0.6 on the SCI-117 node (d = 0.9).

Correlation Between Administration Time and Performance

The correlation between students' administration time and performance on scored items was also examined. Table 23 displays the correlation between time spent on items measuring nodes only and nodes mastered per testlet by linkage level. Because nodes mastered is a limited range of interval data, the Spearman correlation was calculated. The overall Spearman correlation between administration time for items measuring nodes only and nodes mastered per testlet is 0.08, suggesting that there is no correlation between students' administration time and their performance on the items measuring nodes. Correlations at each linkage level ranged from 0.09 to 0.26. The overall correlation (ρ = 0.08) is lower than any of the correlations for specific linkage levels because the overall correlation is not weighted by sample size and there are considerable differences across linkage levels in variance of the administration time and node mastery variables. We do see a trend where the correlation increases as the linkage level increases.

Table 23

Correlation Between Administration Time (Scored Items Only) and Nodes Mastered by Linkage Level

Linkage level	*n*	Spearman correlation
Initial	1,127	0.09
Distal (high school only)	562	0.13
Precursor	1,871	0.22
Target	577	0.26

Wonder Questions

A total of 18 testlets at the Precursor and Target linkage levels began and ended with wonder questions. As described in the Testlet Structure and Design section, the unscored wonder question was asked at the beginning of the testlet to gauge students' current knowledge about the science content to be assessed and engage the student in the testlet. The exact same wonder question was asked at the end of the testlet to see if any change in knowledge occurred after completing the academic tasks. Overall, 2,607 testlets containing wonder questions were administered across both administrations. A total of 2,401 testlets (92%) contained answers to both wonder questions. Of the 206 testlets not containing answers to both wonder questions, 59 testlets (29%) were missing an answer to the first question, 45 testlets (22%) were missing an

answer to the second question, and 102 testlets (49%) were missing answers to both questions. Table 24 summarizes the distribution of answer-changing patterns from the first wonder question to the second wonder question on testlets where both wonder questions were answered. The most frequent answer pattern was right to right (41%), followed by wrong to wrong (23%) and wrong to right (20%). Thus, 64% of students did not change their response. However, of those who did, the most common change was the desirable wrong to right change. That is, students originally answered incorrectly, but after engaging with the scored items measuring nodes, they provided a correct response to the same wonder question.

Table 24

Distribution of Answer Changing Patterns From Wonder Question 1 to Wonder Question 2

Answer pattern	*n*	%
RR	991	41.3
WW	557	23.2
WR	488	20.3
RW	365	15.2

Note. RR = right to right; WW = wrong to wrong; WR = wrong to right; RW = right to wrong.

To evaluate patterns of answer changes on the wonder question based on complexity of the content assessed, Table 25 summarizes the distribution of answer changing patterns from the first wonder question to the second wonder question by linkage level. For both Precursor and Target levels, the most frequent answer pattern was right to right (41%; 44%), followed by right to wrong (23%; 24%). Students changed their answers from wrong to right more often at the Precursor level (22%) than the Target level (16%). At the Target level, the wrong to right and right to wrong answer patterns were almost equally likely.

Table 25

Distribution of Answer Changing Patterns From Wonder Question 1 to Wonder Question 2 by Linkage Level

Answer pattern	Precursor *n* (%)	Target *n* (%)
RR	740 (40.6)	251 (43.5)
WW	271 (14.9)	94 (16.3)
WR	396 (21.7)	92 (15.9)
RW	417 (22.9)	140 (24.3)

Note. RR = right to right; WW = wrong to wrong; WR = wrong to right; RW = right to wrong.

We further evaluated the wonder question answer patterns in relation to overall performance on the scored items. Table 26 shows the average number of nodes mastered by answer pattern on wonder questions. Overall, students who answered both wonder questions correctly mastered more nodes on average than other students (1.2 nodes). Students who answered the final wonder question correctly mastered the second highest number of nodes (0.9 nodes). Similarly, students who answered the first wonder question correctly mastered more nodes on average (0.8 nodes) than students who answered both wonder questions incorrectly (0.5). This is expected, given the relationship between wonder question and the items measuring nodes of

the testlets. It is possible that higher mastery among students with right-to-wrong answers than wrong-to-wrong answers is indicative of partial understanding of the content.

Table 26

Average Number of Nodes Mastered by Answer Pattern

Answer pattern	*n*	Average nodes mastered
RR	991	1.20
WW	557	0.48
WR	488	0.94
RW	365	0.76

Note. RR = right to right; WW = wrong to wrong; WR = wrong to right; RW = right to wrong. The highest possible number of nodes mastered is four.

The correlation between administration time for only wonder questions and nodes mastered was also analyzed. Table 27 displays the correlation between time spent on wonder questions only and nodes mastered per testlet, by linkage level. The overall Spearman correlation between total time for wonder questions and nodes mastered per testlet is 0.21, suggesting there is little to no correlation between the amount of time spent on wonder questions and a students' overall performance on the scored portion of the testlet.

Table 27

Correlation Between Time Spent on Wonder Questions and Nodes Mastered by Linkage Level

Linkage level	*n*	Spearman correlation
Precursor	2,009	0.23
Target	598	0.14

Item Flagging Review

In addition to evaluating test content in relation to student outcomes (i.e., node mastery), item-level performance was examined to further investigate the psychometric properties of the I-SMART items. Item statistics were calculated and reviewed by the test development team. The following item statistics were provided to the test development team:

- **Item *p*-value**: the proportion of students answering the items correctly.
- **Weighted standardized difference**: the z-score for the item's *p*-values compared to the average *p*-value for that EE, linkage level, and node.
- **Conditional *p*-value for masters**: the proportion of students who mastered the node who provided a correct response.
- **Conditional *p*-value for non-masters**: the proportion of students who did not master the node who provided a correct response.

Based on these statistics, items were flagged if they met any of the following four criteria:

- The item was too challenging or too easy, as indicated by a *p*-value of less than .35 or greater than .95, respectively. The .35 threshold was chosen as most items offer three response options, so a value of less than .35 may indicate less than chance selection of the

options. The .95 threshold ensures that the item is not so easy that nearly every student provides a correct response.
- The item was significantly easier or harder than other items assessing the same EE and linkage level, as indicated by an absolute standardized difference z-score of greater than 1.96.
- The item conditional *p*-value (master or non-master) was outside of the model expected range, as determined by a 95% credible interval around the parameter estimate.

Overall, 133 items (29%) were not flagged by any of the statistics, 210 (46%) were flagged by only one criterion, 101 (22%) were flagged by two criteria, and 16 items (3%) were flagged by three criteria. No items were flagged by all four criteria. That 75% of items were flagged by one or fewer of the statistics indicates that the items largely performed as expected. Table 28 and Table 29 display the number and percentage of items in each linkage level and grade band, by the number of flagged criteria. Across linkage levels, the Initial level had the highest percentage of items that were flagged by one or more criteria (80%), while the Target level had the lowest percentage of items that were flagged by one or more criteria (53%). Across grade bands, the high school grade band had the highest percentage of items flagged by one or more criteria (76%), while the middle school grade band had the lowest percentage of items flagged by one or more criteria (65%). The item statistics and flags for items that were flagged by two or more statistics were sent to the test development team for review. Based on these statistics, the test development team identified trends in the items that will inform future work in the development of science content. See the Conclusions section of this report for a description of the future directions.

Table 28

Number and Percentage of I-SMART Items by Number of Flagged Criteria and Linkage Level

Linkage level	No flags	One flag	Two flags	Three flags
Initial	32 (20%)	90 (57%)	30 (19%)	6 (4%)
Distal	13 (23%)	21 (38%)	18 (32%)	4 (7%)
Precursor	51 (30%)	70 (42%)	43 (26%)	4 (2%)
Target	37 (47%)	29 (37%)	10 (13%)	2 (3%)

Table 29

Number and Percentage of I-SMART Items by Number of Flagged Criteria and Grade Band

Grade band	No flags	One flag	Two flags	Three flags
Elementary	40 (30%)	61 (46%)	29 (22%)	4 (3%)
Middle school	48 (35%)	63 (46%)	20 (15%)	5 (4%)
High school	45 (24%)	86 (45%)	52 (27%)	7 (4%)

Psychometric Model

This section organizes the progression from the foundation of the psychometric models used in the I-SMART assessments to the implementation of the final models in the I-SMART assessments. A brief overview of diagnostic classification models (DCMs) is presented to establish the psychometric foundation for the models utilized in the I-SMART assessments. The rationale for the choice of the fungible models as the initially chosen models is then presented. Next, a summary of the evidence supporting the selection of the fungible models over other possible models is presented. Finally, the implementation of the chosen fungible models is presented by examining the calibrated parameters and the scoring rules used to determine mastery assignment.

The I-SMART assessment uses learning map models to specify distinct science knowledge, skills, and understandings (KSUs) and their relationships to each other through a network of sequenced nodes and connections, which are intended to guide the progression of learning. By specifying the connections between nodes, the learning map models theorize the nodes that are precursors to each node in the learning map models. Using psychometric terminology, learning map models define latent variables for students' mastery status on assessed nodes and define the hierarchical relationship between the latent variables. For an example of a map for a single content standard, see Figure 1 in the Assessment Design section of the report.

The I-SMART assessment estimates mastery for each assessed node to provide fine-grained feedback on student achievement. The node mastery estimations form a profile of mastery across multiple nodes, which is created using DCMs. DCMs are confirmatory latent class models that map observed student responses to categorical latent variables (e.g., Rupp et al., 2010; Bradshaw, 2016). This allows for the reporting of more actionable results that can be used to support future instruction (Clark & Karvonen, 2019; Feldberg & Bradshaw, 2019). DCMs have been primarily applied across a variety of subjects in educational settings to provide fine-grained feedback about what skills students have learned (e.g., Bradshaw et al., 2014; Templin & Bradshaw, 2014; Templin & Henson, 2008), although Sessoms and Henson (2018) noted in a review of DCM applications that about 85% of DCM-based assessments have examined either mathematics or reading. The process for estimating item and student parameters in the I-SMART assessment can be described as specifying the Q-matrix (Tatsuoka, 1983), estimating the conditional probabilities for each latent class, and using the latent class conditional probabilities and observed responses to estimate the posterior probability of latent class membership for each student. For the I-SMART assessment, the probability of mastery is calculated at the node level.

I-SMART Model Specification

To provide detailed profiles of student mastery of attributes, DCMs utilize an item by attribute Q-matrix to specify which attributes are assessed by each item. For each item, *i* in an assessment measuring *k* attributes, the vector $q_i = [q_{i1}, q_{i2}, \ldots, q_{ik}]$ contains a dichotomous entry for each of the *k* attributes, where $q_{ik} = 1$ indicates the *i*th item measures the *k*th attribute and $q_{ik} = 0$ indicates the *i*th item does not measure the *k*th attribute.

The I-SMART assessment items are written so that each item only assesses a single node, which means the Q-matrix for each item was a column of ones. Because each item only measures a single node, there were two parameters estimated for each item: (a) the probability of a student who has not mastered the node providing a correct response and (b) the probability of a student who has mastered the node providing a correct response. A structural parameter,

which reflects the proportion of students who have mastered the assessed node, was also estimated.

Model Calibration

Latent class analysis was conducted for each node within each linkage level for each Essential Element (EE). The general form of latent class analysis (and DCMs) is:

$$P(X_j = x_j) = \sum_{c=1}^{C} v_c \prod_{i=1}^{I} \pi_{ic}^{x_{ij}} (1 - \pi_{ic})^{1-x_{ij}} \qquad (1)$$

where π_{ic} is the probability of a correct response to item *i* from a student in latent class *c*, x_{ij} is the dichotomously scored observed response to item *i* from student *j*, and v_c is the base rate probability that a randomly selected student is a member of latent class *c*. Thus, the probability of a student in latent class *c* providing the observed response, x_{ij}, is estimated by the $\pi_{ic}^{x_{ij}}(1 - \pi_{ic})^{1-x_{ij}}$ term in equation (1). The probabilities of each observed response for the student are then multiplied across all items, which estimates the likelihood of a student in latent class *c* providing the observed response pattern. The probability of the observed response pattern for student *j* in latent class *c* is then multiplied by v_c. The product of v_c and the probability of the observed response pattern represents the probability that the student is in latent class *c* and provided the observed response pattern. This calculation is done for all latent classes, and then summed, giving the total likelihood of the student's observed data.

Mathematically, π_{ic} is defined using the log-linear cognitive diagnosis model (Henson et al., 2009), modified as described by Thompson (2019):

$$\pi_{ic} = P(X_{ic} = 1|\alpha_c) = \frac{exp(\lambda_0 + b_{i,0} + (\lambda_{1,1} + b_{i,1,1})\alpha_c)}{1 + exp(\lambda_0 + b_{i,0} + (\lambda_{1,1} + b_{i,1,1})\alpha_c)} \qquad (2)$$

where λ_0 is the node-level intercept and $b_{i,0}$ is the item-level deviation from the node-level intercept for item *i*. Thus, the intercept for item *i* can be calculated as $\lambda_0 + b_{i,0}$. Similarly, $\lambda_{1,1}$ is the node-level main effect, and $b_{i,1,1}$ is the item-level deviation from the node-level main effect for item *i*. Finally, α_c is a binary indicator for the mastery status (0 = non-master, 1 = master) of students in latent class *c*.

Items used in the I-SMART diagnostic assessment were written to be interchangeable in measuring the node. Consequently, a *fungible* model is assumed, which constrains the item-level parameters to zero. Operationally, the $b_{i,0}$ and $b_{i,1,1}$ parameters from Equation (2) are set equal to zero. In addition to λ_0 and $\lambda_{1,1}$, the base rate of node mastery (v_c) is estimated.

In addition to being the theoretically preferred model due to the item writing practices that were used, the fungible model is also the most parsimonious. By constraining the item-level parameters to zero, fewer parameters need to be estimated, which reduces the data requirements and computational complexity of the model.

A weakly informative prior was used to estimate each parameter. When an adequate amount of data has been obtained, weakly informed priors only constrain the parameter space to span the plausible range of parameter values without overly influencing the posterior distribution (Kruschke & Liddell, 2018). The weakly informative prior for the probability of a non-master providing a correct response was a normal distribution with a mean of 0 and a scale of 2. The

weakly informative prior for the probability of a master providing a correct response was a log-normal distribution with a mean of 0 and a scale of 1. The weakly informative prior for the structural parameter was a beta distribution with shape parameters of 1 and 1.

Because of the design of the I-SMART assessment, no students were assessed on all of the map neighborhood nodes for an EE, or even all of the nodes selected as assessment targets. In the Patterns of Mastery Profiles section in Map Validation, there was a significant amount of misfit in the models calibrated concurrently at the linkage level. Consequently, each node within each EE was calibrated using separate DCMs. In total, there were 80 assessed nodes, which means 80 DCMs were estimated.[5]

Model Estimation

The models were estimated in *R* version 3.6.3 (R Core Team, 2019) using the rstan package interface (Guo et al., 2019). The rstan package incorporates the *Stan* (Carpenter et al., 2017) programming language, which allows for the posterior distribution to be further explored. To do this, the rstan package and *Stan* use specific processes and algorithms (e.g., Markov chain Monte Carlo, Hamiltonian Monte Carlo, No-U-Turn sampler) for navigating the posterior distribution (Betancourt & Girolami, 2013; Hoffman & Gelman, 2014; Neal, 2011). A complete description of the model estimation procedure can be found in Thompson (2019).

The DCMs for the I-SMART assessment were estimated with four chains per model, where each chain had a total of 2,000 iterations and the first 1,000 iterations were warm-up iterations that were discarded. Thus, there were a total of 4,000 iterations after discarding the warm-up iterations. Additionally, the adaptive threshold was set to 0.99, and the maximum tree depth was set to 15. A full description for the rationale of these choices can be found in Thompson (2019).

Table 30 presents the data available for estimating the models. Because students tested at two adjacent linkage levels in each EE, the total number of students testing in each EE is not the sum of the number of students testing at each linkage level within an EE.

Table 30

Sample Sizes by Linkage Level

Essential Element	Initial	Distal*	Precursor	Target	Total
5.LS2-1	150	N/A	190	40	190
5.PS1-3	389	N/A	484	95	484
MS.LS2-2	232	N/A	324	93	325
MS.PS1-2	244	N/A	344	100	344
HS.ESS3-3	129	281	232	79	361
HS.LS2-2	107	349	435	191	542

*Available at the high school level only.

[5] The pilot test form design means that students never tested on all three linkage levels. The shared nodes across adjacent linkage levels provides some information on the untested linkage level. However, low sample sizes, especially at the higher linkage levels, prevented the fully concurrent models from converging with this amount of missing data (i.e., approximately 33% missing for each EE, as one linkage level was always missing).

Model Fit

Because fungible models were chosen based on their consistency with the item writing practices as well as parsimony, the model fit of these models must be evaluated to confirm that the evidence supporting the selection of fungible models is sufficient. The fungible models are first evaluated in the absolute sense to assess whether the models conform to the observed item responses. The assumption of fungibility is then evaluated by assessing the absolute and relative model fit of models with varying levels of fungibility.

Absolute Model Fit

For the DCMs used to estimate profiles of mastery in the I-SMART project, absolute model fit is measured using posterior predictive model checking (PPMC). PPMC assesses absolute model fit by simulating replications of the data using parameter values at each iteration of the posterior distribution (Gelman et al., 2013). Because 4,000 iterations of the posterior distribution were retained, there are 4,000 replicated data sets used in the PPMC for each model.

For the replicated data sets at each of the iterations of the posterior distribution, Gelman et al. (1996) suggested using a discrepancy function to evaluate how the observed data relate to the replicated data. Béguin and Glas (2001) suggested a discrepancy measure defined as follows:

$$\chi^2_{obs} = \sum_{s=0}^{S} \frac{[n_s - E(n_s)]^2}{E(n_s)} \qquad (3)$$

where *s* is the score point (i.e., specific total score on the assessment), n_s is the number of respondents at score point *s*, and $E(n_s)$ is the expected number of respondents at score point *s*.

A posterior predictive *p*-value (*ppp*) is calculated with the formula $ppp = P(\chi^2_{rep} \geq \chi^2_{obs}|n_s)$. Put simply, *ppp* is the proportion of the replicated data sets with a χ^2 greater than or equal to the observed χ^2. Generally, *ppp* near .50 indicates good model fit (Crawford, 2014; Sinharay & Almond, 2007), *ppp* near zero indicates poor model fit (Thompson, 2019), and *ppp* near one indicates potential overfitting (Thompson, 2019). While the cutoff for identifying poor model fit is somewhat arbitrary, a cutoff of .05 could be used similar to the approach taken in null hypothesis significance testing. The proportion of the nodes with a *ppp* less than .05 reflects the number of nodes with poor fitting models.

The fungible models with weakly informative priors demonstrated good absolute model fit (Table 31), as the majority of linkage levels had 0% of nodes with a *ppp* less than .05 for each model. There were 12% of nodes at the Distal linkage level with *ppp* less than .05, which is indicative of poor model fit in these nodes. The Distal linkage level was only available at the high school level. Thus, these findings indicate that it may have been difficult to write items to this additional linkage level falling between the Initial and Precursor levels. These findings do not appear to have been influenced by sample size, as the sample sizes for the Distal linkage levels in the high school EEs were as large or larger than the sample sizes for the majority of the linkage levels in any of the EEs (Table 30).

Table 31

Percentage of Nodes Exhibiting Misfit

Linkage level	With misfit (%)
Initial	0
Distal	12
Precursor	0
Target	0

Evaluating the Fungibility Assumption

Although the items are intended to be fungible, this is an assumption that should be tested. To evaluate the fungibility assumption, additional models with varying degrees of fungibility were fit to the data collected from the pilot administration. The *equivalent slopes* model allows the item-level deviations for the node-level intercept to vary freely, while constraining the item-level deviations for the node-level main effect to zero. The item-level main effects are not estimated, but the node-level main effects are held constant within the node. In contrast, the node-level intercept is not estimated, but the item-level intercept is allowed to vary across items. The *fixed partial equivalency* model estimates the node- and item-level parameters, but the variance of the item-level main effects about the node-level main effects are fixed. The *estimated partial equivalency* model is similar to the fixed partial equivalency model, except the estimated partial equivalency model estimates a parameter for the variance of the prior for the item-level main effects parameter. The *non-fungible* model allows the item-level deviation parameters to freely vary without constraint. Thus, the non-fungible model allows for between-item variation regarding the probability of a correct response by non-masters and masters.

Absolute Model Fit

To test whether the I-SMART assessment items were fungible, the four additional models (equivalent slopes, fixed partial equivalency, estimated partial equivalency, and non-fungible) were calibrated and compared to the previously calibrated fungible model. The absolute fit of each model was evaluated as previously described using PPMC with *ppp* values.

The absolute model fit of each model was examined using the percentage of nodes with *ppp* less than .05 that was described previously. As can be seen in Table 32, the majority of linkage levels had 0% of nodes with a *ppp* less than .05 for each model. There were 8% of the nodes at the Initial linkage level in the non-fungible model with *ppp* less than .05, and there were 12% of the nodes at the Distal linkage level in the fungible model with *ppp* less than .05, indicating poor model fit.

Table 32

Percentage of Nodes Exhibiting Misfit, by Model

Linkage level	Fungible (%)	Equal slopes (%)	Fixed partial equivalency (%)	Estimated partial equivalency (%)	Non-fungible (%)
Initial	0	0	0	0	8
Distal	12	0	0	0	0
Precursor	0	0	0	0	0
Target	0	0	0	0	0

Because multiple models demonstrated adequate absolute model fit, relative model fit indices can be evaluated to determine the preferred model for each node. As mentioned previously, the items were intended to be fungible. Based on the results from Table 30, there was preliminary support for the fungible model.

Relative Model Fit

Model comparisons were also conducted as a measure of relative model fit. Relative model fit is evaluated by comparing the model fit of two models, often through the use of information criterion. Because the two models are compared to one another rather than to the observed data, it is important that the models compared in relative model fit analyses have both demonstrated evidence of adequate absolute model fit (Sen & Bradshaw, 2017).

Two information criteria were applied to the resulting models: the Pareto smoothed importance sampling leave-one-out cross-validation (PSIS-LOO; Vehtari et al., 2017) and the Widely Applicable Information Criterion (WAIC; Watanabe, 2010). The WAIC assesses model fit by estimating the log transformation of the expected value of the posterior distribution, while penalizing for overfitting based on the number of parameters. The PSIS-LOO assesses model fit by omitting one data point from the posterior distribution and estimating a replacement data point, while controlling for deviations from normality stemming from the values in the tails of the posterior distribution. The PSIS-LOO has generally been found to be more robust than the WAIC, despite their mathematical similarities (Vehtari et al., 2017). The between-model differences for both the PSIS-LOO and the WAIC are considered statistically significant if the absolute value of the difference in the models' statistics is greater than 2.5 times the standard error of the difference (Bengio & Grandvalet, 2004).

As shown in Table 33, both the PSIS-LOO and WAIC fit statistics prefer the fungible model to the other four models, with the proportion of nodes favoring the fungible model ranging between 61% and 64%. Thus, the fungible model appears to be the best fitting model in comparison to the other models for the observed response data for the majority of nodes.

Table 33

Percentage of Models Preferring the Fungible Model

Model 1	Model 2	Proportion of fungible models preferred by PSIS-LOO	Proportion of fungible models preferred by WAIC
Fungible	Equal slopes	64.4	64.4
Fungible	Fixed partial equivalency	60.6	60.6
Fungible	Estimated partial equivalency	60.6	60.6
Fungible	Non-fungible	63.4	63.4

Note. PSISI-LOO = Pareto smoothed importance sampling leave-one-out cross-validation; WAIC = Widely Applicable Information Criterion.

Based on the absolute and relative model fit, the assumption of item fungibility appears to be supported. The absolute model fit results of this study indicate that there may be some localized instances of misfit for nodes at the Distal linkage level in the fungible model. Because the Distal linkage level is only available for nodes in the EEs for high school content, future work may attempt to identify any sources of misfit. Despite this localized misfit, the fungible model was preferred over the other four models based on the relative model fit results.

Calibrated Parameters

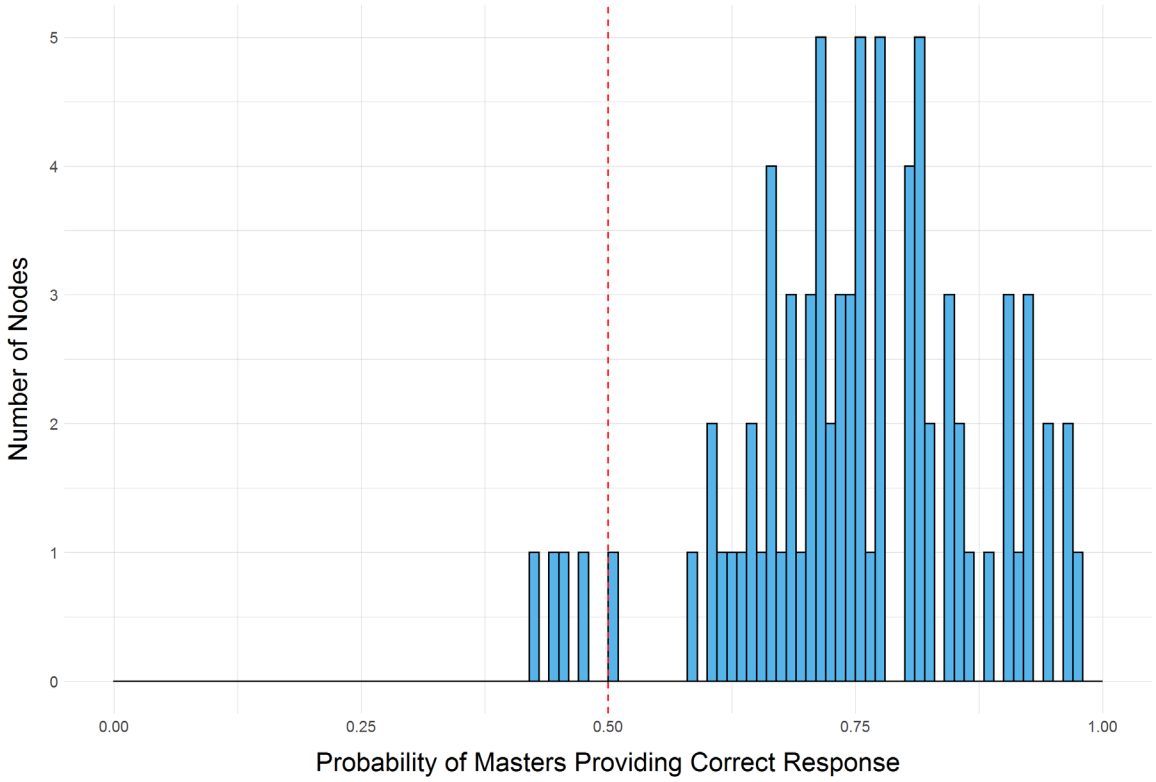
As described previously, the models were estimated with fungibility. That is, item-level parameters are shared across all items measuring the same node. With the fungibility assumption, only the node-level intercept (λ_0) and the node-level main effect (λ_1) from Equation (2) are estimated. In addition to the node-level parameters, the base rate of node mastery (v_c) is estimated. Having shown that the fungible models have adequate absolute model fit and that the assumption of fungibility is justified, the calibrated parameters need to be evaluated to demonstrate that the estimated parameters have expected values. A summary of the calibrated parameters used to classify students as masters or non-masters in the pilot administration of the I-SMART assessment is provided in the following sections.

Probability of Masters Providing Correct Response

Students who have mastered a node are expected to demonstrate a higher probability of responding correctly to items measuring the node. Using the pilot calibration, Figure 5 depicts the conditional probability of masters responding correctly to items for each of the 80 nodes. Because the conditional standard error of measurement is maximized at .50, masters should have at least a 50% chance of responding correctly. The results in Figure 5 demonstrate that most nodes (*n* = 76, 95%) performed as expected. Additionally, 93% of nodes (*n* = 74) had a conditional probability of masters responding correctly of .6 or greater. No nodes (0%) had a conditional probability of masters responding correctly of .40 or less. Thus, most nodes performed as expected. Appendix C lists nodes with a probability of masters providing a correct response that is outside of the ideal range by EE and linkage level.

Figure 5

Probability of Masters Responding Correctly to Items Measuring Each Node



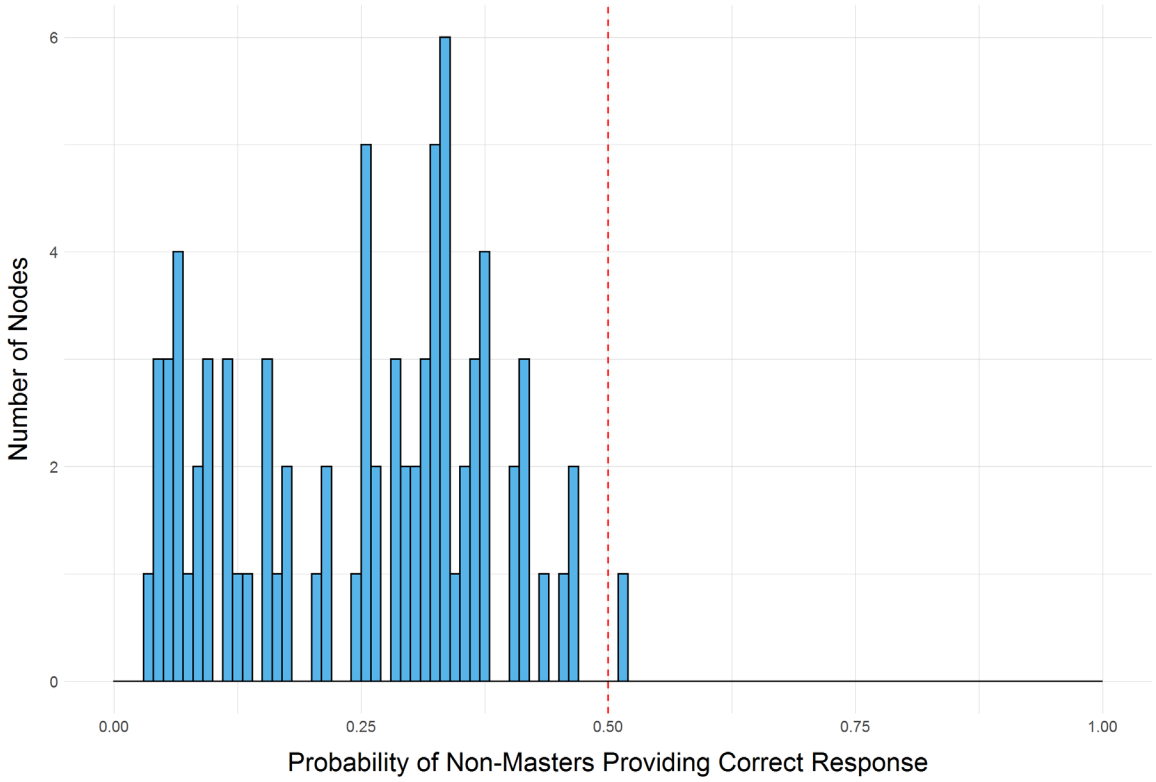
Note. Histogram bins are in increments of .01, and the reference line indicates .50.

Probability of Non-Masters Providing Correct Response

When items are functioning appropriately, non-masters should have a decreased probability of responding correctly to an item. High probabilities of a correct response by non-masters may indicate issues with the item. When non-masters are able to provide correct responses, there is a risk that non-masters will errantly be classified as masters, which undermines the validity of inferences made from the assessment. Figure 6 summarizes the probability of non-masters responding correctly to items measuring each of the 80 nodes. The majority of nodes (n = 79, 99%) performed as expected, with most nodes (n = 70, 88%) having a conditional probability of .40 or less that non-masters would respond correctly. Appendix C lists nodes with a probability of non-masters providing a correct response that is outside of the ideal range by EE and linkage level.

Figure 6

Probability of Non-Masters Responding Correctly to Items Measuring Each Node



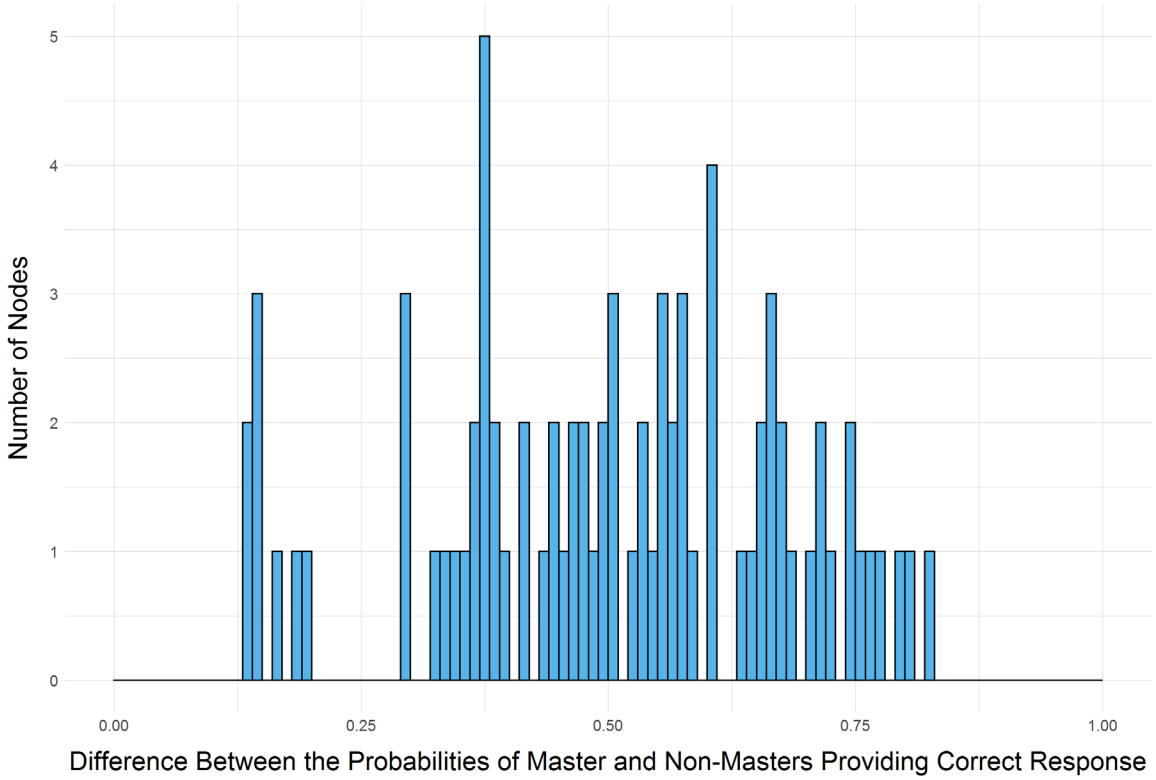
Note. Histogram bins are in increments of .01, and the reference line indicates .50.

Item Discrimination

The discrimination of a node reflects the ability to discern between masters and non-masters. This is quantified by subtracting the conditional probability of a non-master responding correctly from the conditional probability of a master responding correctly. Thus, the discrimination of a node is on a scale of 0 to 1, where 0 indicates no difference in conditional probabilities of responding correctly between masters and non-masters and 1 indicates perfect differentiation between masters and non-masters based on the conditional probability of responding correctly (e.g., masters have a 100% chance of providing a correct response and non-masters a 0% chance). Figure 7 displays the node discrimination values. Overall, 69% of nodes (n = 55) have a discrimination of .40 or greater, indicating a relatively large difference between the conditional probabilities of masters and non-masters. The .40 discrimination threshold was chosen as a reasonable difference between the conditional probabilities for the two mastery classes because the difference is large enough to be noticeably different yet not so large that masters always respond correctly and non-masters always respond incorrectly. Appendix C lists nodes with a discrimination that is outside of the ideal range by EE and linkage level.

Figure 7

Difference Between Masters' and Non-Masters' Probability of Responding Correctly to Items Measuring Each Node



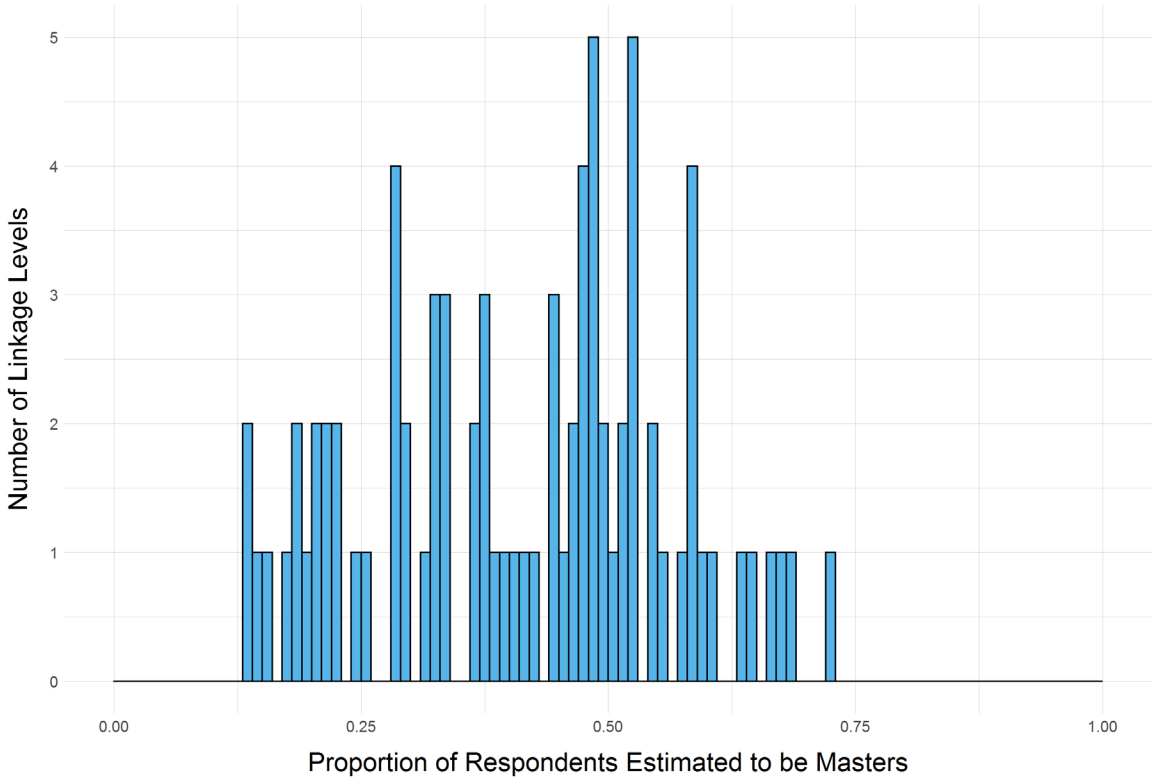
Note. Histogram bins are in increments of .01, and the reference line indicates .50.

Base Rate Probability of Mastery

The I-SMART assessment is intended to match student KSUs to appropriate content at the node level. The base rate of mastery reflects the proportion of students who are masters of the node. When the base rate of mastery is around .50, the students assessed on the node have a 50% likelihood of being a master and a 50% likelihood of being a non-master. Base rates closer to .50 indicate that the nodes match the KSUs. When the base rate of mastery approaches 0 or 1, nearly all students assessed on the node are classified as non-masters or masters, respectively. Figure 8 displays the base rate of mastery probabilities of the nodes. Overall, 81% of nodes (n = 65) had a base rate of mastery between .25 and .75, which suggests most nodes are performing as intended. Regarding base rate of mastery values outside of .25 and .75, 15 nodes (19%) had a base rate of mastery less than .25, and 0 nodes (0%) had a base rate of mastery greater than .75. This suggests students are more likely to be assessed on nodes they have not mastered than those they have mastered. Appendix C lists nodes with a base rate that is outside of the ideal range by EE and linkage level.

Figure 8

Base Rate of Node Mastery



Note. Histogram bins are shown in increments of .01.

Reliability

Because the fungible models were selected, the reliability of the fungible models should also be examined. Reliability of node-level mastery can be evaluated through classification consistency (Sinharay & Johnson, 2019). Classification consistency is defined as the probability that a student would receive the same mastery classification for a node on two parallel forms of the assessment. Wang et al. (2015) defined a classification consistency statistic, $\hat{\gamma}$, that approaches 1.0 as the probability of mastery for a node approaches 0 or 1. This makes sense intuitively. As we become more certain about a student's mastery classification, we also be more certain that they would achieve the same mastery status on a parallel form of the assessment.

The statistic proposed by Wang et al. (2015) assumes that posterior probability of mastery would be consistent across parallel forms, not just the classification. However, Johnson and Sinharay (2018) show that a consistent classification could result in different posterior probabilities of mastery. Thus, Johnson and Sinharay (2018) propose a new statistic, $\hat{P}_c$, that corrects for this assumption.

Table 34 and Figure 9 show the distribution of node reliability estimates for both the Wang et al. (2015) and Johnson and Sinharay (2018) indices. The Wang et al. and Johnson and Sinharay

indices are interpreted consistently with the recommendations from Landis and Koch (1977) that kappa estimates of .60 or above indicate acceptable agreement for categorical classifications. Overall, the reliability estimates from both Johnson and Sinharay (2018) and Wang et al. (2015) are consistent and fairly high, with only a small proportion of nodes showing a reliability below .60. Additionally, Figure 10 shows the distribution of reliability indices by linkage level. In general, the distributions are similar across most of the linkage levels. However, there does appear to be slightly lower reliability estimates for higher linkage levels. This finding is consistent with reliability findings for the Dynamic Learning Maps® (DLM®) assessment (DLM Consortium, 2019) and is likely due to smaller sample sizes at the higher linkage levels.

Table 34

Reliability Summaries Across All Nodes: Proportion of Nodes Within a Specified Index Range

Reliability index	<0.6	0.60–0.64	0.65–0.69	0.70–0.74	0.75–0.79	0.80–0.84	0.85–0.89	0.90–0.94	0.95–1.00
$\hat{P}_c$ (Johnson & Sinharay, 2018)	.125	.125	.025	.088	.088	.175	.138	.212	.025
$\hat{\gamma}$ (Wang et al., 2015)	.162	.100	.050	.025	.112	.150	.138	.200	.062

Figure 9

Summaries of Node Reliability

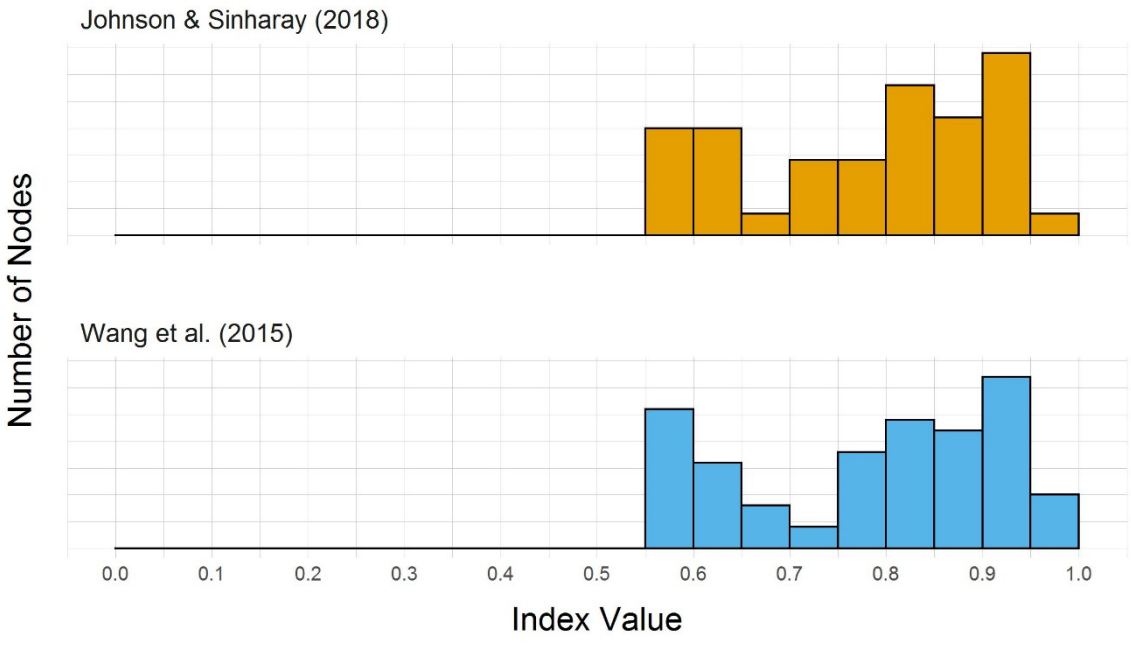
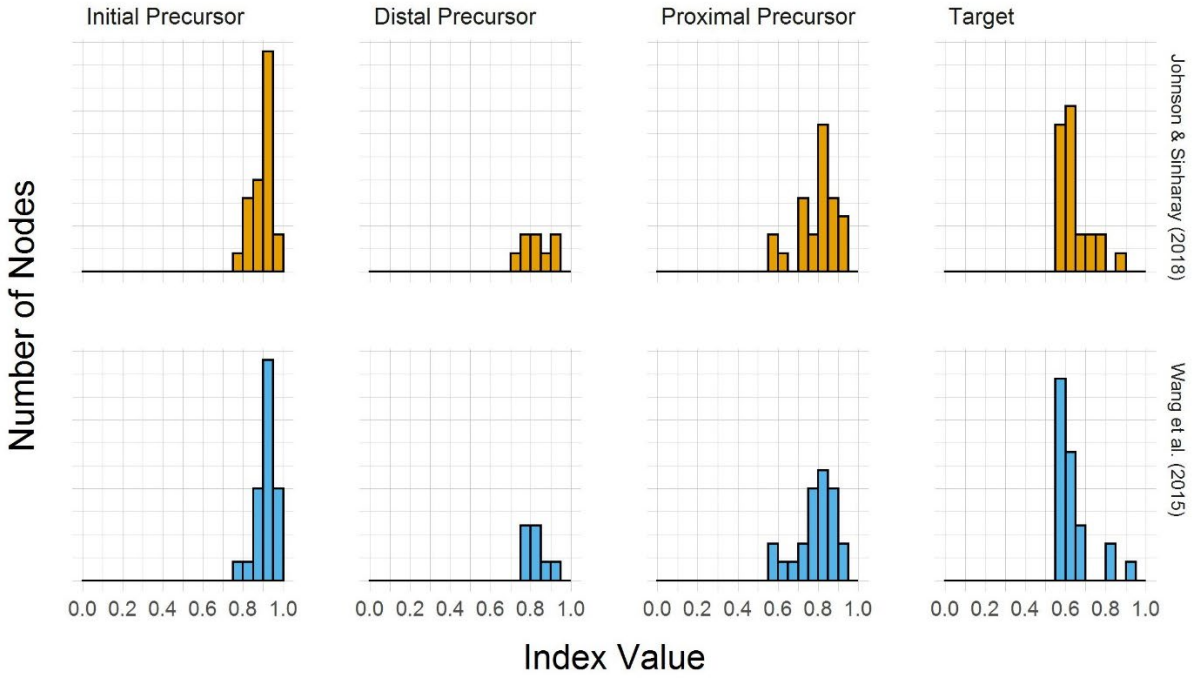


Figure 10

Conditional Reliability Evidence Summarized by Linkage Level



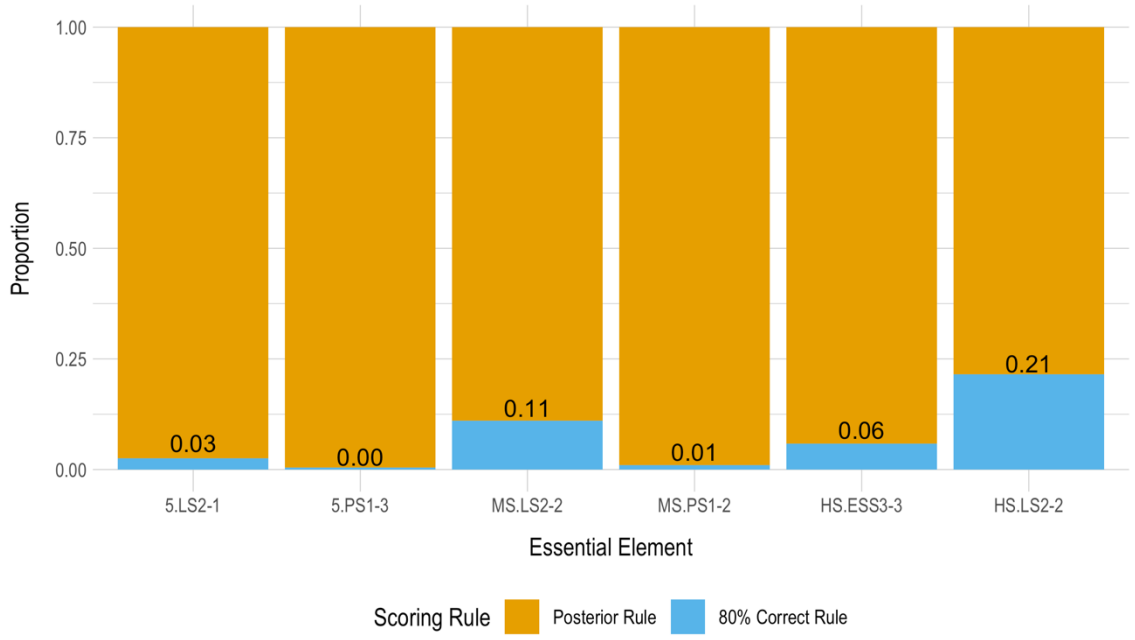
Mastery Assignment

Because the model fit evidence supports the selection of the fungible models, assignment of node mastery is based on the fungible models. There are two rules for assigning node mastery. Students can be classified as masters of the node when the estimated posterior probability is greater than .80 or when more than 80% of items were answered correctly. A posterior probability threshold of .80 provides more certainty than a lower threshold (e.g., .50) that students classified as masters are indeed masters. In other words, the Type I error rate is reduced. The 80% correct threshold allows students an alternative method to show mastery when they answer a high proportion of items correctly, but their posterior probability of mastery is still below .80 (e.g., because the overall rate of mastery for the node is extremely low). These mastery assignment rules are consistent with those used by the DLM alternate assessment (DLM Consortium, 2019).

To evaluate the frequency that each scoring rule was used to determine students' mastery status during the pilot administration of I-SMART assessment, the percentage of mastery statuses derived using each scoring rule was calculated. Posterior probability was given precedence, meaning students with a posterior probability greater than .80 and correct response rate of more than 80% were considered to have been classified as masters using the posterior probability estimates. For students with posterior probabilities less than .80, the 80% scoring rule was used. In comparing the usage of the two scoring rules (Figure 11), the majority of students in all EEs were assigned mastery status using the posterior probability scoring rule, although the proportion of students assigned mastery status using the 80% correct scoring rule was notably higher in EE.MS.LS2-2, EE.HS.ESS3-3, and EE.HS.LS2-2.

Figure 11

Mastery Assignment by Scoring Rule for Each Essential Element



To achieve a posterior probability greater than .80, students must often respond correctly to at least 80% of items; hence, the scoring rules were strongly correlated. Despite this relationship, there were instances where students had a correct response rate greater than 80%, yet their posterior probability was less than .80. The agreement between the scoring rules was estimated by comparing the mastery classifications using each of the scoring rules (e.g., a posterior probability of .60 and a correct response rate of 50% agree that the student has not mastered the node). For the pilot administration, the rate of agreement between the two scoring rules was 89%.

When the two scoring rules disagreed regarding the student's mastery status, the posterior probability scoring rule indicated a higher level of mastery in 82% of the 11% of cases where the scoring rules were inconsistent. Thus, in some instances the posterior probabilities allowed students to demonstrate mastery when the percentage correct was lower than 80%.

Empirical Map Validation

The I-SMART assessment is based on learning map models, which means validating the learning map models is central to supporting the inferences made from the I-SMART assessment. Prior to assessment development, preliminary map validation evidence is gathered through the initial map development and external review process (Swinburne Romine et al., 2018). After assessment administration, statistical evaluation of the learning map models has two components. First, validating the individual nodes serves to empirically demonstrate the uniqueness and consistency of the nodes. Second, validating the connections provides empirical support for the order of the nodes in the learning map models. Because the connections provide a sequence in which the nodes are likely to be mastered, mastery of later nodes should be contingent on mastering preceding nodes. Thus, inferences made from mastery classifications depend on accurate knowledge of which nodes are requisite for mastering a new node.

Node Validation

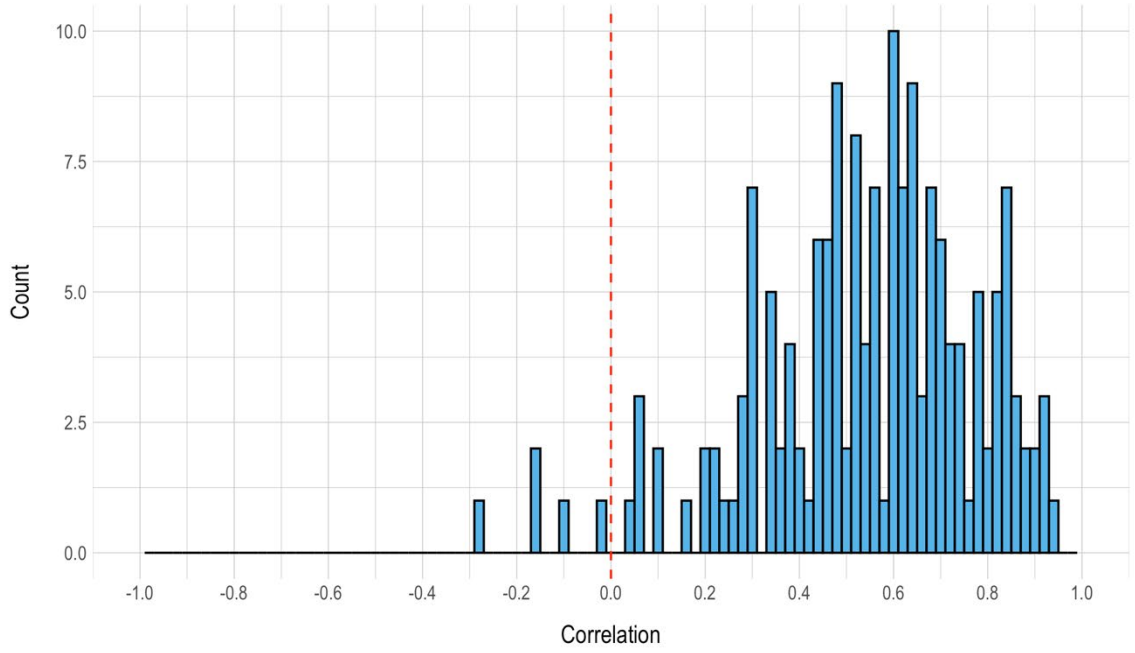
We took a three-step approach to validating the nodes of the I-SMART learning map models. First, we examined the uniqueness of the nodes by examining the between-node mastery correlations. Second, we examined the consistency of node-level weighted *p*-values across linkage levels for overlapping nodes. Finally, we evaluated the consistency of node mastery across linkage level for overlapping nodes.

Correlations of Node Mastery

To see if the nodes are as unique as intended, we assessed the pairwise node mastery classification correlations for all nodes within each Essential Element (EE) and within each EE and linkage level. Figure 12 displays the distribution of pairwise node mastery classification correlations within each EE. As can be seen, the distribution of correlation coefficients is centered around .60, with most of the correlation coefficients between .30 and .80. While the observed correlations between nodes were moderately positive, the correlations were not strong enough to assume mastery of a node given mastery of another node, which suggests uniqueness of nodes within EEs.

Figure 12

Pairwise Node Correlations, Within Essential Element

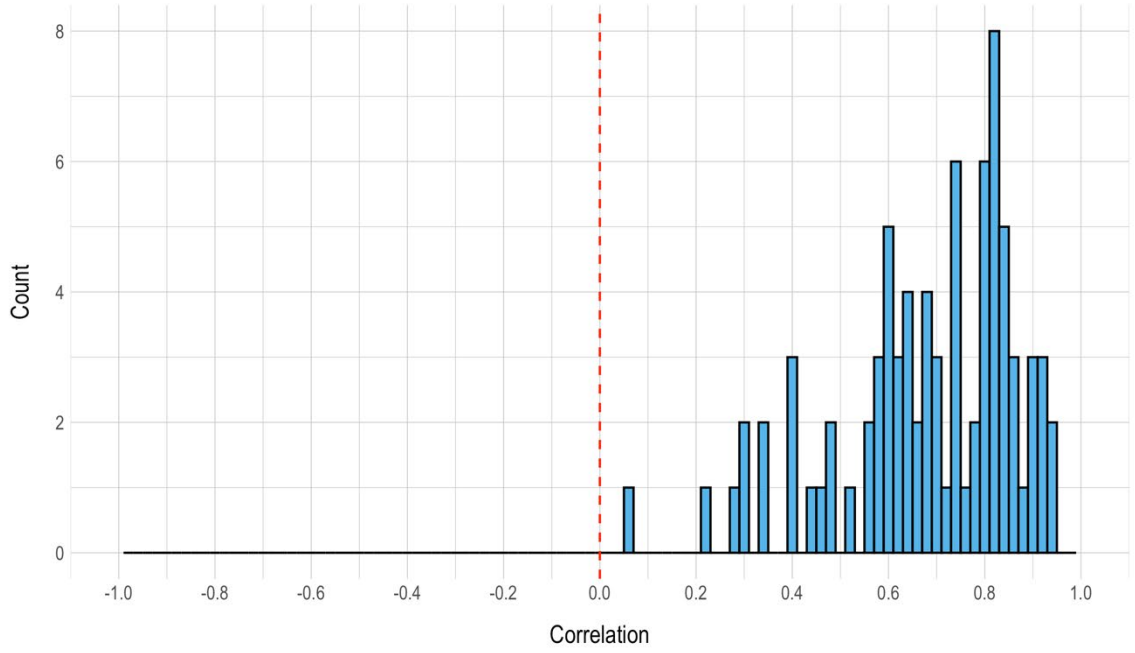


Within EE, all of the negative correlations were between nodes in EE.MS.PS1-2, and almost all of the correlations above .90 were between nodes in EE.MS.LS2-2. The remaining correlations with magnitudes above .80 tended to be spread evenly across the remaining EEs. Sample size appeared to be related to small mastery correlations. Within-EE correlations less than or equal to .20 tended to have at least one node with a smaller sample size than the sample sizes of within-EE correlations with larger magnitudes. More specifically, all of the within-EE correlations less than or equal to .20 involved at least one node with a sample size of 130 or less, while the majority of within-EE correlations with a magnitude greater than .20 involved both nodes having sample sizes greater than 130.

Figure 13 displays the distribution of pairwise node mastery classification correlations within each EE and linkage level. The distribution of the correlation coefficients is centered around .70. In Figure 13, most of the correlation coefficients are between .40 and .95. The higher correlation between nodes within linkage levels is expected, as these nodes are conceptually closer together than nodes within an EE. As nodes are compared that are further apart (i.e., in the same EE but not the same linkage level), lower correlations would be expected, as seen in Figure 12. The between-node mastery correlations within each linkage level and EE again suggest the uniqueness of nodes. The observed correlations are generally positive and of moderate strength, as would be expected for knowledge and skills related to a single content standard. However, the observed correlations are not so strong as to indicate that if one node is mastered, mastery of another node could be assumed.

Figure 13

Between-Node Correlations, Within Linkage Level and Essential Element



Within EE and linkage level, there were no negative correlations and there were 30 correlations above .80. Of the 30 correlations above .80, 25 of the 30 correlations (83%) were between nodes at the Initial linkage level. Almost all of the correlations above .90 were between nodes in EE.MS.LS2-2 at the Initial linkage level. The remaining correlations above .80 were relatively evenly spread across the other EEs. Of the correlations less than .40, eight of the 10 (80%) correlations were between nodes at the Target linkage level. Sample size again appeared to be related to small mastery correlations. The within–linkage level correlations with magnitudes less than or equal to .40 tended to have smaller node sample sizes than the within–linkage level correlations with larger magnitudes. For example, the lone within–linkage level correlation with a magnitude less than .20 only had 40 students testing on each node.

Overlapping Node Consistency

Overlapping nodes are defined as nodes that are common to testlets at different linkage levels. For example, node SCI-999 would be an overlapping node if it were included within both the Initial and Precursor linkage levels. I-SMART pilot assessment forms included testlets at two adjacent linkage levels, so a student would test on the same node in the context of two different linkage levels. When considering student performance at each node, students should be performing similarly on overlapping nodes when assessed on the node in the context of the testlet measuring each linkage level. That is, a student's performance on SCI-999 should be constant, regardless of whether the node is assessed on an Initial or Precursor testlet. Operationally, this means that overlapping nodes should have similar weighted *p*-values across linkage levels. More specifically, node SCI-999 would be expected to have a weighted *p*-value

near .6 at the Precursor linkage levels if node SCI-999 had a weighted *p*-value of .6 at the Initial linkage levels.

Based on the data obtained from the pilot administration, there are 12 assessed overlapping nodes (see Figure 4 for an example of overlapping nodes in EE.5.LS2-1 and EE.5.PS1-3). To evaluate whether students were performing similarly on overlapping nodes, we examined the weighted *p*-values for each node along with the 95% confidence interval for each weighted *p*-value. If the 95% confidence intervals overlapped for the overlapping nodes, the nodes were considered to have similar performance on the different linkage levels. Table 35 shows the weighted *p*-values for all of the overlapping nodes, including whether each node had similar performance at each of the measured linkage levels. Table 35 shows two weighted *p*-values for SCI-119 because it is a node in EE.5.PS1-3 and EE.MS.PS1-2.

Table 35

Percentage of Overlapping Nodes with Similar Performance, by Essential Element (EE)

Node	EE	LL	*N*	Weighted *p*-value (SE) of lower LL	Weighted *p*-value (SE) of higher LL	Similar performance
SCI-326	EE.5.LS2-1	I/P	150/190	.36 (.17)	.44 (.18)	Yes
SCI-309	EE.5.LS2-1	P/T	190/40	.35 (.17)	.60 (.25)	Yes
SCI-119	EE.5.PS1-3	I/P	389/484	.44 (.18)	.52 (.19)	Yes
SCI-121	EE.5.PS1-3	P/T	484/95	.39 (.19)	.61 (.25)	Yes
SCI-313	EE.MS.LS2-2	I/P	232/324	.53 (.19)	.54 (.18)	Yes
SCI-119	EE.MS.PS1-2	I/P	244/344	.39 (.17)	.61 (.21)	Yes
SCI-804	EE.MS.PS1-2	P/T	344/100	.54 (.21)	.61 (.25)	Yes
SCI-598	EE.HS.ESS3-3	I/D	129/281	.19 (.13)	.35 (.17)	Yes
SCI-601	EE.HS.ESS3-3	D/P	281/232	.44 (.20)	.50 (.18)	Yes
SCI-643	EE.HS.ESS3-3	P/T	232/79	.43 (.18)	.48 (.25)	Yes
SCI-501	EE.HS.LS2-2	I/D	107/349	.19 (.14)	.48 (.18)	Yes
SCI-80	EE.HS.LS2-2	D/P	349/435	.39 (.18)	.50 (.19)	Yes
SCI-528	EE.HS.LS2-2	P/T	435/191	.37 (.18)	.52 (.25)	Yes

Note. LL = linkage level; SE = standard errors; I/P = Initial/Precursor; P/T = Precursor/Target; I/D = Initial/Distal; D/P = Distal/Precursor. *N* reflects the sample size at the lower linkage level followed by the sample size at the upper linkage level.

As can be seen in Table 35, all of the overlapping nodes for each EE perform similarly across linkage levels. This result can be largely due to the relatively large standard error associated with the weighted *p*-values. The large standard errors may be due to variability in the student population, administration design where students were assessed at a linkage level above or below their skill level as defined by the First Contact survey, or small sample sizes. Future work may examine the impact of limited sample size on these findings. However, the current results provide preliminary evidence that the overlapping nodes are functioning similarly across linkage levels within each EE, as expected.

Overlapping Node Mastery

As was the case with examining the weighted *p*-values for overlapping nodes across linkage levels, we expected that students would perform similarly in terms of node mastery across

linkage levels. In this study, we examined the percentage of students with consistent node mastery statuses on overlapping nodes. Consistent node mastery statuses are defined as students having the same mastery status at both linkage levels of an overlapping node. If the overlapping node is performing as expected, students should be classified consistently for the node at both linkage levels. For example, a student mastering the node at the Initial and Precursor linkage levels of an overlapping node would have consistent node mastery statuses. We examined consistency using the percentage of consistent mastery classifications, Cohen's kappas, and tetrachoric correlations.

Table 36 presents the classification consistency for each overlapping node. Table 36 presents two sets of classification consistency estimates for SCI-119 because it is a node in EE.5.PS1-3 and EE.MS.PS1-2. The three consistency estimates were interpreted according to the recommendations from Landis and Koch (1977), who suggested kappa estimates between .40 and .60 indicated fair agreement for categorical classifications and kappa estimates greater than or equal to .60 indicated acceptable agreement for categorical classifications. The classifications are largely consistent, with a few exceptions. The three consistency estimates can be interpreted in both absolute and relative terms. For example, node SCI-804 on EE.MS.PS1-2 demonstrated suboptimal classification consistency (21%); additionally, the classification consistency for node SCI-804 (21%) was lower than the other overlapping node on this EE, indicating node SCI-804 was classified less consistently than the other overlapping node on this EE (SCI-119; 71%). Similarly, node SCI-601 on EE.HS.ESS3-3 demonstrated fair classification consistency (53%) but was also classified less consistently than the other overlapping nodes on this EE (SCI-598 and SCI-643; 85% and 75%, respectively).

Table 36

Percentage of Overlapping Nodes with Consistent Mastery Classifications, by Essential Element, Linkage Level, and Node

Essential Element	Linkage level	Node	Consistent mastery classifications (%)	Cohen's kappa	Tetrachoric correlation
EE.5.LS2-1	I/P	SCI-326	60.0	.12	.44
EE.5.LS2-1	P/T	SCI-309	77.5	.54	.87
EE.5.PS1-3	I/P	SCI-119	58.9	.23	.69
EE.5.PS1-3	P/T	SCI-121	71.6	.33	.60
EE.MS.LS2-2	I/P	SCI-313	51.7	.19	.72
EE.MS.PS1-2	I/P	SCI-119	70.9	.42	.65
EE.MS.PS1-2	P/T	SCI-804	21.0	.01	.09
EE.HS.ESS3-3	I/D	SCI-598	84.5	.43	.69
EE.HS.ESS3-3	D/P	SCI-601	53.3	.14	.54
EE.HS.ESS3-3	P/T	SCI-643	74.7	.46	.66
EE.HS.LS2-2	D/P	SCI-80	88.1	.29	.57
EE.HS.LS2-2	I/D	SCI-501	82.2	.38	.65
EE.HS.LS2-2	P/T	SCI-528	83.7	.28	.73

Note. I/P = Initial/Precursor; P/T = Precursor/Target; I/D = Initial/Distal; D/P = Distal/Precursor.

Map Structures

To validate the connections of the I-SMART learning map models, a three-step approach was used, each with different levels of model assumptions (see Thompson & Nash, 2019). First, within each linkage level, a log-linear cognitive diagnosis model (LCDM; Henson et al., 2009) and a hierarchical diagnostic classification model (HDCM; Templin & Bradshaw, 2014) were used to evaluate sections of the map neighborhoods. Specifically, we examined the absolute and relative model fit of the saturated LCDMs with all possible mastery profiles and the constrained HDCMs, which include only mastery profiles that are plausible given the connections between nodes in the underlying map structures. If the learning map models are structured correctly, the constrained model should demonstrate equivalent model fit to the saturated model. Second, we examined the node mastery patterns using the separate node calibrations for each of the 80 nodes. As previously mentioned, the learning map models specify the order in which knowledge, skills, and understandings (KSUs), as defined by the nodes, should be acquired. This order should be reflected in the node mastery patterns. Finally, we examined the weighted *p*-values for each node to determine whether the weighted *p*-values are consistent with the learning map models. That is, we expect nodes further along in the progression to be more difficult than preceding nodes because each node represents additional KSUs that must be mastered. This means that subsequent nodes should have lower weighted *p*-values than their preceding nodes.

Patterns of Mastery Profiles

To evaluate the connections of the I-SMART learning map models, the nodes in the learning map neighborhoods can be analyzed with and without the constraints imposed by the connections. In the *saturated model*, the learning map models' constraints are not imposed on mastery classifications. Take, for example, Figure 14. Here, SCI-599 and SCI-601 are two nodes in the Distal linkage level of EE.HS.ESS3.3. The learning map models specify that SCI-601 depends on SCI-599, meaning that the KSUs assessed in SCI-601 depend on the KSUs assessed in SCI-599. In the saturated model, SCI-599 and SCI-601 are analyzed without

considering the connection between these two nodes. That is, SCI-601 could be mastered without mastering SCI-599. In the *constrained model*, SCI-599 and SCI-601 are analyzed under the assumption that SCI-599 leads to SCI-601, which implies SCI-599 should be mastered in order for SCI-601 to be mastered.

Table 37 presents the possible mastery profiles for EE.HS.ESS3-3 at the Distal linkage level, where the highlighted mastery profiles indicate which mastery profiles are allowed in the constrained model. One example of an unexpected node mastery pattern was for students who mastered SCI-598 and did not master SCI-599.This profile is unexpected because SCI-599 is a precursor for SCI-598, which suggests students should master SCI-599 before mastering SCI-598.

Figure 14

Nodes in EE.HS.ESS3-3 at the Distal Linkage Level

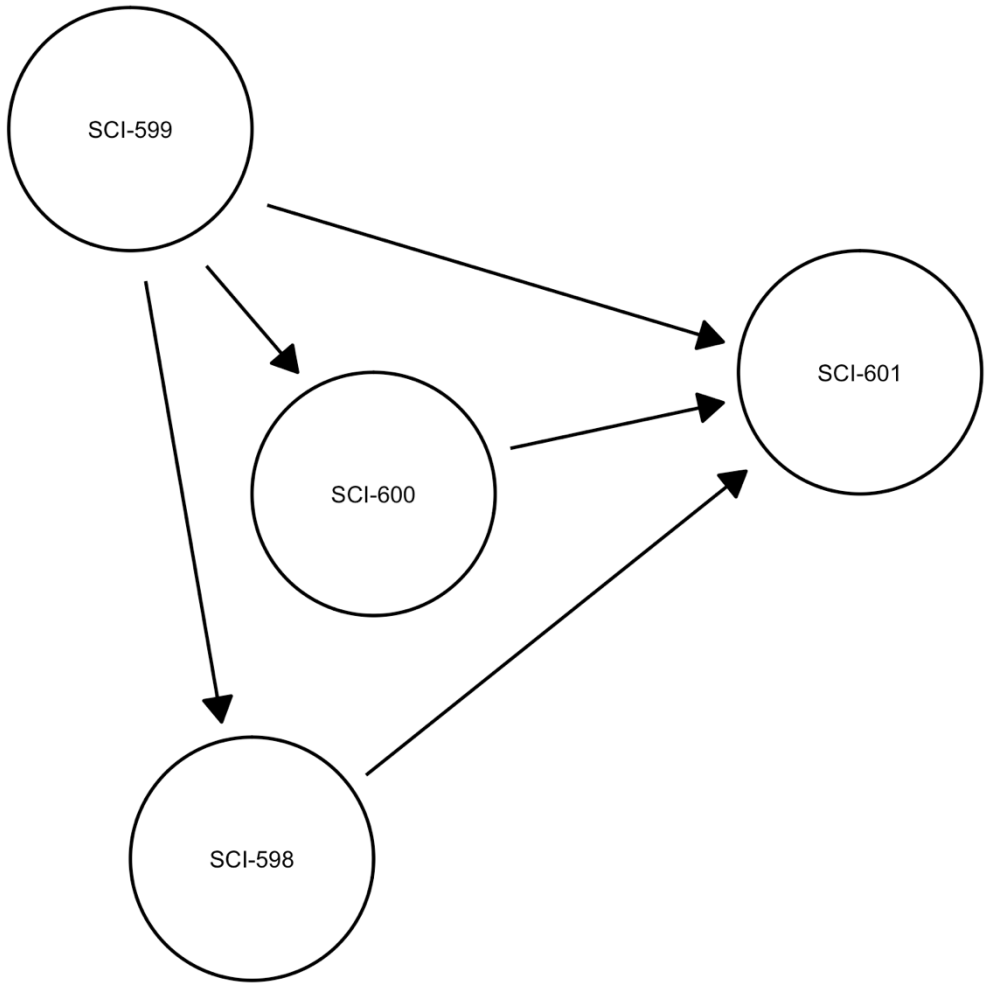


Table 37

All Possible Mastery Profiles for EE.HS.ESS3-3 at the Distal Linkage Level

SCI-599	SCI-598	SCI-600	SCI-601	Allowed in constrained model
0	0	0	0	Yes
1	0	0	0	Yes
0	1	0	0	No
0	0	1	0	No
0	0	0	1	No
1	1	0	0	Yes
1	0	0	1	No
0	1	0	1	No
1	0	1	0	Yes
0	1	1	0	No
0	0	1	1	No
1	1	0	1	No
1	1	1	0	Yes
1	0	1	1	No
0	1	1	1	No
1	1	1	1	Yes

Note. The mastery profiles highlighted in gray are the mastery profiles allowed in the constrained model.

This study examined whether the empirical data collected during the pilot administration support the proposed structure of the learning map models. Because of the underlying learning map models, it is expected that the constrained models will show equivalent absolute and relative model fit in comparison to the saturated model. As in the Psychometric Model section, absolute model fit was assessed using *ppp*, and relative model fit was assessed using the Pareto smoothed importance sampling leave-one-out cross-validation (PSIS-LOO; Vehtari et al., 2017) and the Widely Applicable Information Criterion (WAIC; Watanabe, 2010).

Of the 20 linkage levels, five (25%) showed adequate absolute model fit when all mastery profiles were included in the model. When the model was constrained to include only the map-implied mastery profiles, six (30%) show adequate absolute model fit. For the unconstrained LCDMs, all of the Target linkage level models demonstrated adequate absolute model fit, except for the Target linkage level of EE.HS.LS2-2. For the constrained HDCMs, all of the Target linkage level models demonstrated adequate absolute model fit. In terms of relative model fit, the constrained model fit the linkage level as well or better than the saturated model in 100% of the linkage levels according to the PSIS-LOO and the WAIC. This indicates that the model fit was not degraded by imposing the underlying map structures. Thus, there is preliminary support for the I-SMART learning map models. However, the results pertaining to the relative model fit comparisons should be interpreted with caution given the limited number of linkage levels showing adequate absolute model fit (Sen & Bradshaw, 2017).

Patterns of Mastery Assignment

Because many of the models showed insufficient model fit when nodes were estimated concurrently within linkage levels, patterns of mastery assignment across the separately

calibrated nodes was also evaluated. The separately calibrated nodes generally showed good model fit, as described in the Model Fit section of this report). Using the separate node calibrations, a mastery profile was created for each student across the assessed nodes. Student mastery profiles were then examined to determine if the profiles were consistent with the underlying learning map models. As previously mentioned when describing the constrained HDCM models, the ordering of the nodes in the learning map models should preclude certain mastery profiles from occurring if the proposed node structure is correct. This analysis compared the node mastery profiles achieved by students to the expected mastery profiles and identified the number of students with an unexpected mastery profile.

Individual node mastery classifications can be summarized as a profile of node mastery for each linkage level the student was assessed on, just like the profiles presented in Table 35. These profiles were then compared to our *a priori* expectations for existing mastery profiles. Using EE.HS.ESS3-3 at the Distal linkage level as an example, Table 35 presents the 16 total possible node mastery patterns and the six node mastery patterns that were expected, given the map structures.

To examine where the map structures may be misspecified, node mastery patterns and connections can be flagged based on the rate of unexpected mastery patterns. Linkage levels were flagged when more than 25% of students had an unexpected node mastery pattern or when a given connection between two nodes was flagged for an unexpected mastery pattern for more than 50% of students. The flagging thresholds were chosen in accordance with the work by Thompson and Nash (2019) so that the flagging method erred on the side of over-reviewing content for potential misspecifications.

Table 38 provides the percentages of students with at least one unexpected mastery pattern across connected nodes, by linkage level and EE. In Table 38, EE.HS.LS2-2 at the Target linkage level has around 50% of students with unexpected mastery patterns. The remaining linkage levels have 35% or less of students with unexpected mastery patterns, and most of these linkage levels have less than 20% of students with unexpected mastery patterns.

In the elementary grade band, four linkage levels were flagged for the 25% rule, and no linkage levels were flagged for the 50% rule. In the middle school grade band, no linkage levels were flagged for the 25% rule or the 50% rule. Finally, in the high school grade band, four linkage levels were flagged for the 25% rule, and no linkage levels were flagged for the 50% rule. Thus, in total, eight out of 20 linkage levels (40%) were flagged for unexpected patterns of mastery assignment, all on the basis of the 25% rule.

Table 38

Percentage of Students With an Unexpected Mastery Pattern, by Essential Element and Linkage Level

Grade band	Essential Element	Initial	Distal*	Precursor	Target
EL	EE.5.LS2-1	26.0	N/A	25.3	35.0
	EE.5.PS1-3	27.0	N/A	23.6	10.5
MS	EE.MS.LS2-2	12.5	N/A	9.9	0.0
	EE.MS.PS1-2	17.2	N/A	10.5	16.0
HS	EE.HS.ESS3-3	8.5	31.0	29.7	25.3
	EE.HS.LS2-2	13.1	0.0	4.8	52.4

Note. EL = elementary; MS = middle school; HS = high school.
*Available at the high school level only.

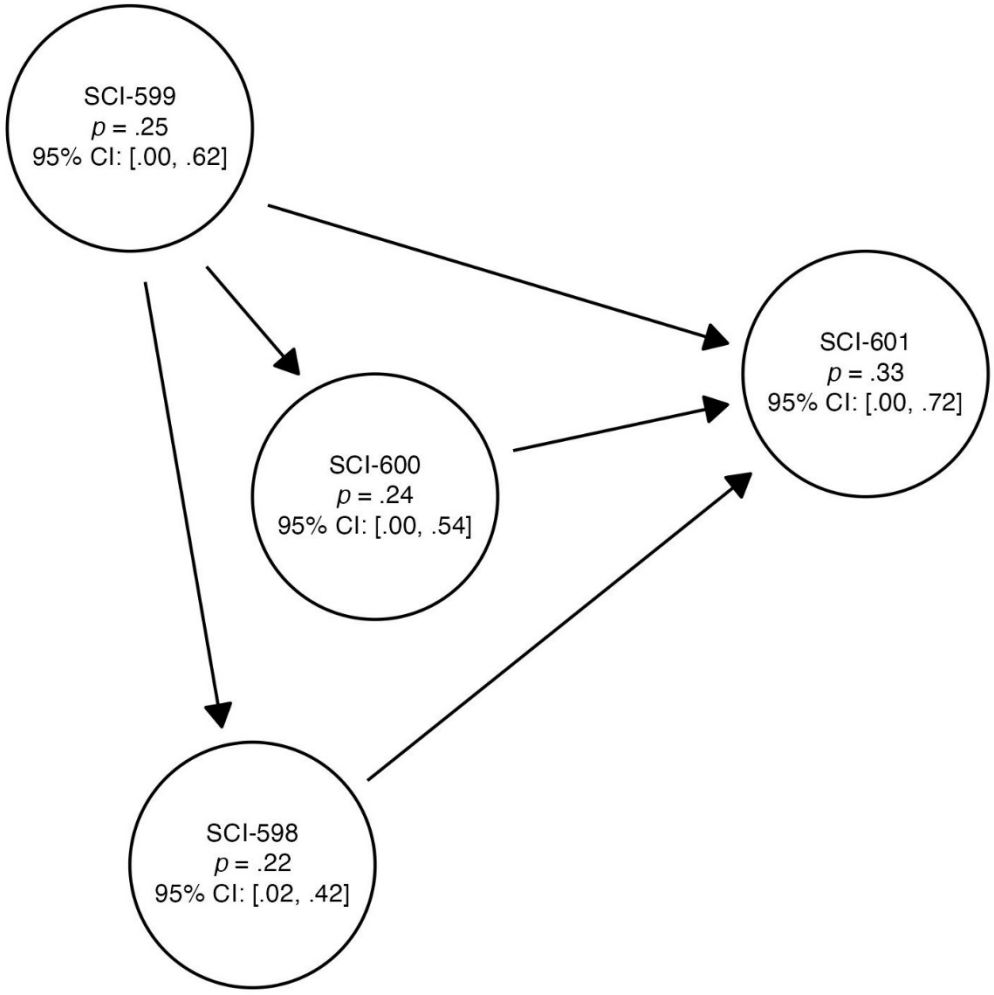
Patterns of Node Difficulty

Because the KSUs in the precursor node are necessary to learn the KSUs in the subsequent node, it is expected that students should perform better on the first node in the hierarchical sequence. That is, as students progress through the learning map, the content should become more difficult. In this study, we first placed similar students into cohorts based on the students' complexity band, which was determined from teacher responses to the First Contact survey. The First Contact survey contains items reflecting each student's skills in communication and science. After placing similar students into cohorts, students' node-level performance was measured using weighted *p*-values and their standard errors. We then identified unexpected *p*-value patterns within each cohort's performance by identifying places where the preceding node was significantly more difficult than the subsequent node, as defined by the 95% confidence intervals around each node's weighted *p*-value.

Figure 15 shows the nodes in EE.HS.ESS3.3 at the Distal linkage level. In Figure 15, the connections between nodes are presented, as well as the node code, weighted *p*-value, and 95% confidence interval for the weighted *p*-value. The directionality of the connections between nodes are given with the use of arrows. As an example, the connection between SCI-599 and SCI-598 indicates that the SCI-599 node precedes the SCI-598 node in the learning map models, which implies the expectation that students should have a higher weighted *p*-value on node SCI-599 than on node SCI-598. Note that SCI-601 has a higher weighted *p*-value than SCI-600, despite SCI-600 being a precursor. However, because the confidence intervals for the two nodes overlap, this discrepancy may be due to sampling variability, and is therefore not flagged.

Figure 15

Nodes and Weighted p*-values in EE.HS.ESS3-3 at the Distal Linkage Level*



The connections between these nodes were aggregated by EE, and we calculated the percentage of the node connections within each EE that had an unexpected *p*-value pattern. For the 90 assessed connections, all of the connections demonstrate *p*-value patterns consistent with the I-SMART learning map models. It should be noted, however, that many of the weighted *p*-values have wide 95% confidence intervals.

Evaluating Structure Misspecifications

The different types of evidence of map structure can be evaluated together to offer insights for potential improvements for the map structures. For example, unexpected findings from patterns of mastery profiles and patterns of mastery assignment can be compared. Figure 16 presents the nodes and connections of EE.5.LS2-1 at the Target linkage level, and Figure 17 presents the profiles of mastery for EE.5.LS2-1 at the Target linkage level. Figure 17 shows areas where the students demonstrated unexpected profiles of mastery in the saturated model. The profiles of attribute mastery in Figure 17 are presented as vectors where SCI-311 is the first entry, SCI-

307 is the second entry, SCI-7 is the third entry, and SCI-309 is the fourth entry (i.e., [SCI-311, SCI-307, SCI-7, SCI-309]). The patterns of mastery profiles study indicated that the most common unexpected profile was [0,1,0,1]. This pattern represents students who have mastered SCI-307 and SCI-309, but have not mastered SCI-7. This profile is unexpected because students should master SCI-7 prior to mastering SCI-307. The finding that this particular profile was flagged by the patterns of mastery profiles method is also consistent with the findings from the patterns of mastery assignment method. In that method, we examined unexpected patterns of mastery assignment for students across nodes. The most common unexpected node patterns for EE.5.LS2-1 at the Target linkage level are shown in Table 39, where the percentage variable reflects the number of students with the unexpected node mastery pattern relative to the total number of students completing testlets at that linkage level. The unexpected node mastery patterns identified in Table 39 are consistent with the unexpected patterns of mastery assignment identified in Figure 17.

Figure 16

Nodes in EE.5.LS2-1 at the Target Linkage Level

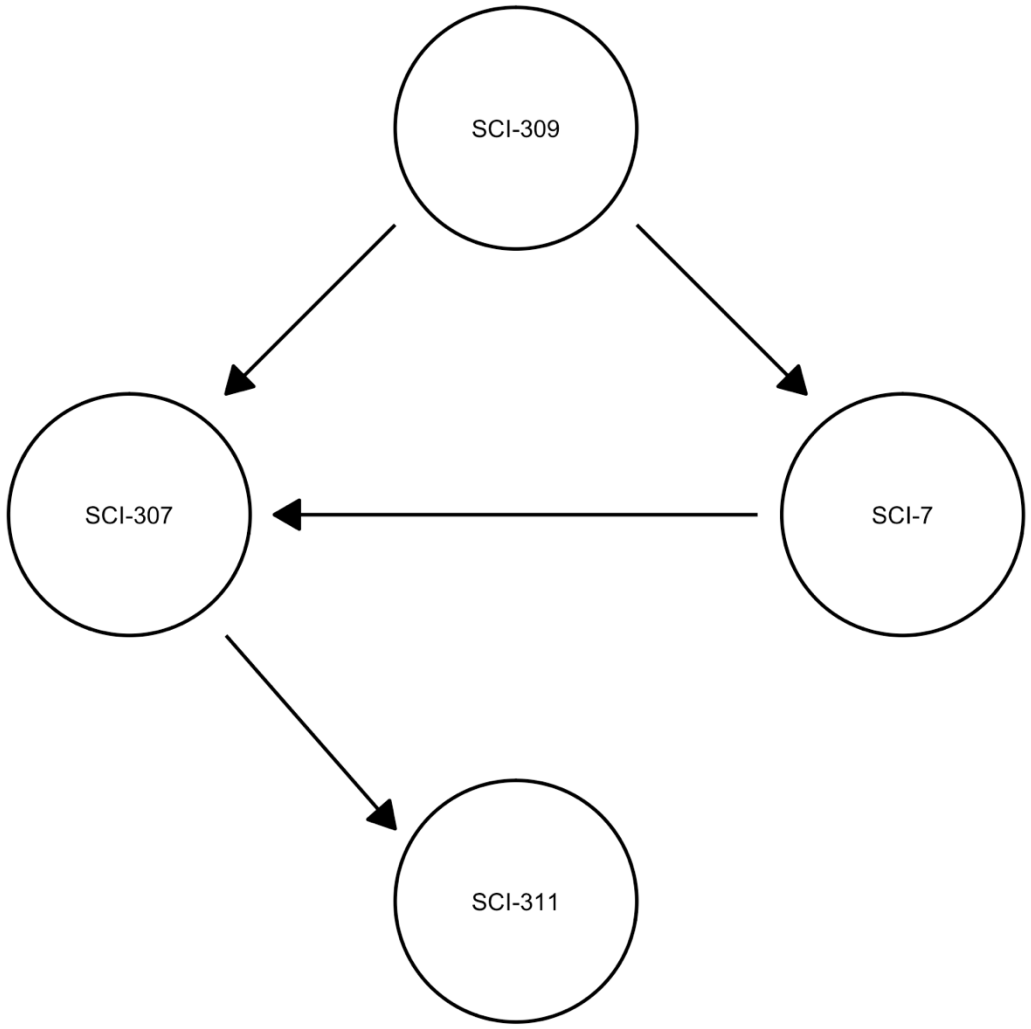


Figure 17

*The Proportion of Students With Each Mastery Profile in the Saturated and Constrained Models for EE.5.LS2-1 at the Target Linkage Level (*N = *40)*

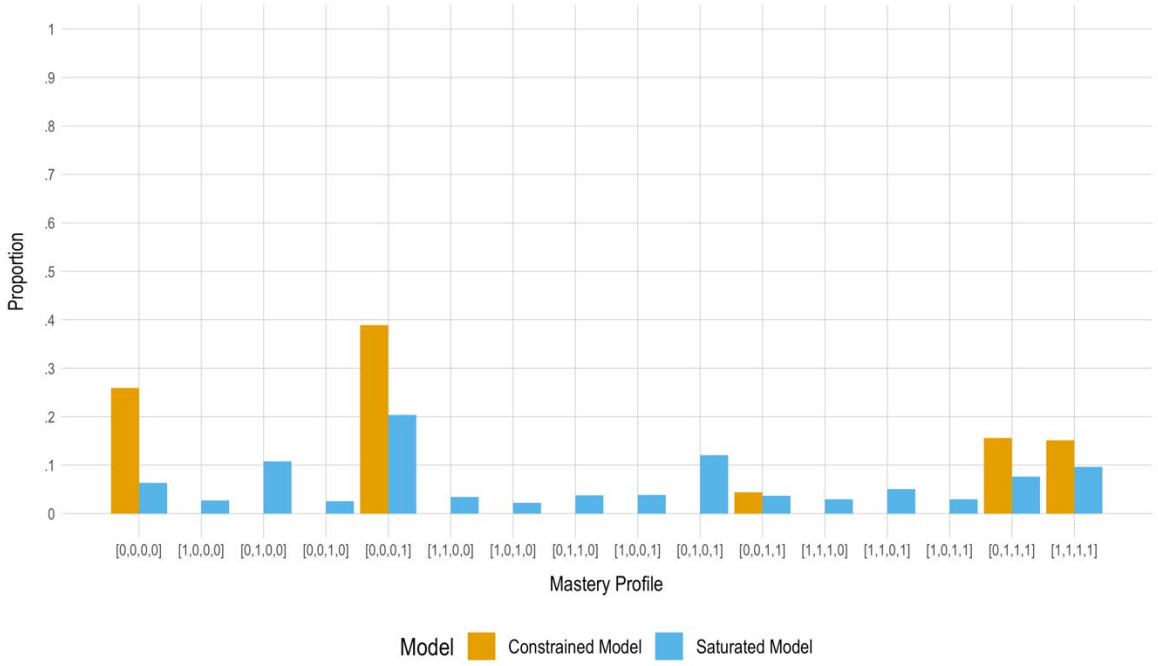


Table 39

Number of Students With an Unexpected Mastery Pattern on EE.5.LS2-1 at the Target Linkage Level

From	To	Count	%
SCI-309	SCI-7	8	20.0
SCI-309	SCI-307	6	15.0
SCI-7	SCI-307	6	15.0

Together, the first two methods indicate that there may be a misspecification in the defined map structure. However, results from the third method, patterns of node difficulty, did not indicate any reversals in the expected *p*-values for nodes in this linkage level. There are a couple of reasons for why the third method might provide results inconsistent with the previous two. The total sample size for this linkage level (and therefore each node) was only $n = 40$. Thus, the *p*-values from the third method have very large standard errors. The large amount of uncertainty in the *p*-values could be masking differences that would manifest with more data. Alternatively, the small sample sizes could be impacting the results of the first two methods. With small samples, there is also greater uncertainty in parameter estimates, which are used for mastery classifications. Thus, it is possible that additional data may resolve the unexpected mastery profiles and assignment patterns that were observed. Therefore, although the methods presented in this section provide general support for the defined map structures, it is important to consider the context of the data collection (i.e., small sample sizes, sparse data across linkage levels) when drawing inferences and interpreting the results.

Evaluation of Student Experience

As described in the Pilot Study Design section of this report, the I-SMART pilot included embedded and post-administration survey items to learn about teachers' perceptions of their students' experience with the testlets as well as students' perceptions of their own performance on each testlet. The relationship between survey items and student performance was also explored.

After each testlet, teachers were asked two embedded survey questions to indicate the amount of instructional time spent on the tested Essential Element (EE) during the school year and student mastery of the EE.[6] After the student completed both testlets, teachers completed a more comprehensive survey about the student's overall experiences with the assessment. A total of 2,144 students and 995 teachers participated in the I-SMART pilot study, and 2,056 students (95.9%) and 949 teachers (95.4%) completed at least one survey question over the two testing windows. Some students and teachers participated during both testing windows.

To evaluate the relationship between survey items and student performance, students' node mastery (described in the Psychometric Model section of this report) was aggregated for each linkage level and EE. Linkage level mastery is the mean performance (i.e., mastered = 1, did not master = 0) on the nodes at each linkage level of an EE. EE mastery is the mean performance (i.e., mastered = 1; did not master = 0) on the nodes that comprise the EE. Both linkage level and EE mastery range from 0 (no mastery) to 1 (complete mastery).

For analyses involving the student self-evaluation question, which was administered after each linkage level testlet, linkage level mastery is used as the outcome measure. For analyses that involve the teacher embedded survey questions or the final teacher survey, which were administered after the assessment, EE mastery is used as the outcome measure.

In analyses where a correlation coefficient is calculated, Cohen's (1988) conventions are used to interpret effect sizes. A correlation coefficient of .10 represents a weak or small association, a correlation coefficient of .30 represents a moderate correlation, and a correlation coefficient of .50 or larger represents a strong correlation.

Student Self-Evaluation and Testlet Performance

After completing each testlet at the Distal, Precursor, and Target linkage levels, students completed self-evaluation items. Students were asked to evaluate their own performance by selecting one of the three options (happy, neutral, or sad; see Figure 2). Overall, 68.1% of the evaluations were happy, 16.3% were neutral, 11.9% were sad, and 3.8% were missing (i.e., students did not provide a response). The percentage of student self-evaluations by grade level and linkage level are shown in Table 40 and Table 41, respectively.

[6] Since the pilot took place in fall 2019, teachers' reports of instructional time reflect only a small portion of the school year.

Table 40

*Student Self-Evaluation by Grade Level (*N = 3,237)

Grade level	Happy *n (%)*	Neutral *n (%)*	Sad *n (%)*	Missing *n (%)*
3–5	535 (66.1)	145 (17.9)	101 (12.5)	28 (3.5)
6–8	612 (71.3)	122 (14.2)	86 (10.0)	39 (4.5)
9–12	1,056 (67.3)	259 (16.5)	197 (12.6)	57 (3.6)
Total	2,203 (68.1)	526 (16.3)	384 (11.9)	124 (3.8)

Note. Students are represented in this table for each testlet that they completed.

Students across all linkage levels most often reported happy on their self-assessment, though students at the Distal level chose sad more often (17.6%) than those at the Precursor (12.0%) or Target (5.5%) levels.

Table 41

*Student Self-evaluation by Linkage Level (*N = 3,237)

Linkage level	Happy *n (%)*	Neutral *n (%)*	Sad *n (%)*	Missing *n (%)*
Distal	371 (58.9)	107 (17.0)	111 (17.6)	41 (6.5)
Precursor	1,358 (67.6)	335 (16.7)	240 (12.0)	76 (3.8)
Target	474 (79.0)	84 (14.1)	33 (5.5)	7 (1.2)
Total	2,203 (68.1)	526 (16.3)	384 (11.9)	124 (3.8)

Note. Students are represented in this table for each testlet that they completed.

The Spearman correlation coefficient between student self-assessment (1 = sad, 2 = neutral, 3 = happy) and linkage level mastery was 0.19, which was statistically significant ($p < .0001$), although it demonstrates a small effect size. Table 42 displays the median linkage level mastery status by student self-assessment.

Table 42

Median Student Self-Assessment by Linkage Level Mastery Status

Student self-assessment	*n*	Median
Happy	2,203	0.25
Neutral	526	0.00
Sad	384	0.00
Total	3,237	0.00

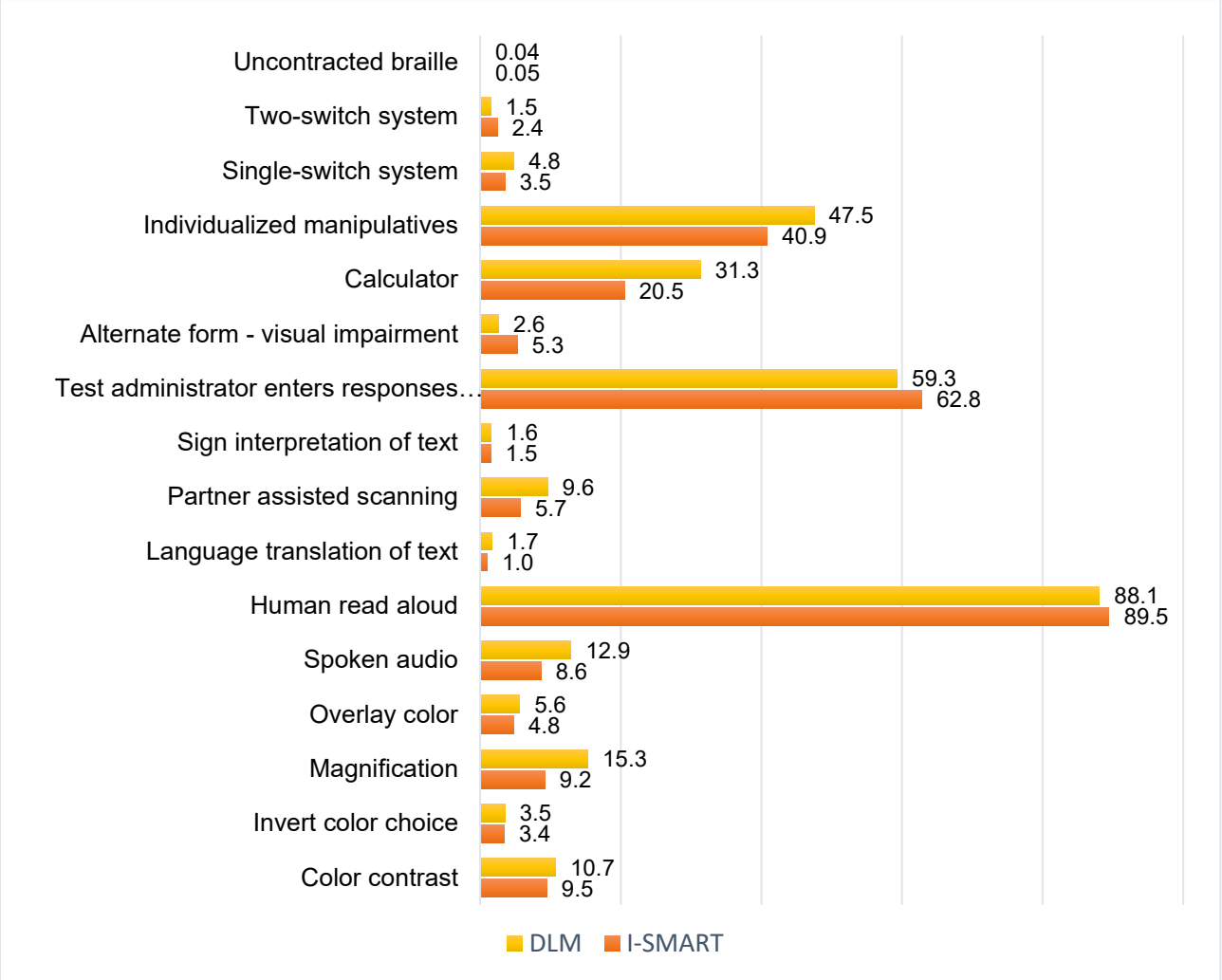
Note. Linkage level mastery is the mean performance (i.e., mastered = 1, did not master = 0) on the nodes at each linkage level of an Essential Element.

Students' Use of Accessibility Supports

Figure 18 shows the Personal Needs and Preferences (PNP) profile selection rates for I-SMART students compared to the full Dynamic Learning Maps® (DLM®) population. Overall, teachers selected accessibility supports at similar frequencies in the I-SMART and DLM populations. However, teachers tended to select individualized manipulatives and the calculator less often for students participating in I-SMART.

Figure 18

Personal Needs and Preferences Profile Selection Rates Comparison



After students completed the testlets, teachers were asked to indicate the accessibility supports students actually needed during test administration that required additional materials and those that were provided by a test administrator. Teachers reported that when completing the assessment, just over 33% of students (*n* = 578) used individualized manipulatives, while over 63% (*n* = 1,092) did not require any supports that required additional materials. Additionally, 88.1% of students (*n* = 1,524) required a human read aloud. Table 43 summarizes teacher reported accessibility supports for students.

There are some differences in the number of supports recorded in the PNP profile and those reported in Table 43 due to different teacher response rates (Ns) to the two surveys, as well as the fact that teachers sometimes selected supports on the PNP profile that were not ultimately used during testing. For example, on the PNP profile, teachers recorded that 40.9% of students required individualized manipulatives; however, teachers reported that 33.4% of students used individualized manipulatives during the I-SMART assessment. Similarly, the PNP profile indicated that 62.8% of students required the test administrator to enter responses for students, while teachers reported this support was actually used by only 36.2% of students. These discrepancies are not surprising because the PNP profile indicates which supports should be available to the student across assessments in all subjects, while the teacher has flexibility in deciding what to actually provide for each assessment.

Table 43

*Teacher-Reported Accessibility Supports Used During Administration (*N = *1,730)*

Support	*n*	%
Supports that required additional materials		
None	1,092	63.1
Two-switch system	54	3.1
Single-switch system/access profile enabled	92	5.3
Individualized manipulatives	578	33.4
Supports provided by the test administrator		
None	153	8.8
Test administrator entering responses for student	626	36.2
Sign interpretation of text	37	2.1
Partner-assisted scanning	23	1.3
Language translation of text	50	2.9
Human read aloud	1,524	88.1

Note. Missing responses = 516.

Teachers were asked about their students' access to accessibility supports (see Table 44). Of the approximately 70% of teachers who responded, most agreed (*n* = 789; 46.4%) or strongly agreed (*n* = 537; 31.6%) that students used similar accessibility supports on the I-SMART pilot assessment to what they use in instruction.

Table 44

Teachers' Level of Agreement to Questions about Accessibility Supports

Teacher survey item	Strongly disagree *n (%)*	Disagree *n (%)*	Agree *n (%)*	Strongly agree *n (%)*	Not applicable *n (%)*	Missing *n*	Total *n*
The student was able to effectively use available accessibility supports.	113 (6.6)	99 (5.8)	780 (45.5)	525 (30.6)	198 (11.6)	531	1,715
The accessibility supports this student used on the assessment were similar to the ones s/he uses for instruction.	76 (4.5)	107 (6.3)	789 (46.4)	537 (31.6)	190 (11.2)	547	1,699

Note. Percentages are based on total number of responses. Row totals do not include missing counts.

Teachers' Perceptions of Students' Experiences With the I-SMART Assessments

As shown in Table 45, across all judgments of student mastery, teachers were most likely to report providing a total of 1–10 hours of instruction on the tested EE (*n* = 1,510; 53.1%) during the 2018–2019 school year up until the time the I-SMART assessment was administered. Of note, approximately 17% reported mastery or partial mastery of an EE with no instruction, and approximately 24% of those reporting that the skill was not taught on the mastery question indicated at least 1 hour of instruction.

Table 45

*Teacher's Judgement of Student Mastery of the Essential Element by Hours of Instruction on the Essential Element (*N = *2,827)*

Mastery	None *n (%)*	1–10 hours of instruction *n (%)*	11–20 hours of instruction *n (%)*	21–30 hours of instruction *n (%)*	>30 hours of instruction *n (%)*
Mastered	11 (10.8)	59 (57.8)	13 (12.8)	8 (7.8)	11 (10.8)
Partially mastered	53 (6.0)	451 (50.7)	241 (27.1)	100 (11.3)	44 (5.0)
Not mastered	88 (6.9)	892 (69.6)	193 (15.1)	53 (4.1)	56 (4.4)
Skills not taught	421 (76.0)	108 (19.5)	14 (2.5)	3 (0.5)	8 (1.4)
Total	573 (20.3)	1,510 (53.1)	461 (16.3)	164 (5.8)	119 (4.2)

Note. Missing responses = 410.

The survey that teachers completed about the student's overall experiences with the assessment consisted of multiple-choice and Likert scale items focusing on teachers' perceptions of students' responses to assessment items, accessibility supports, and student engagement and interest in the assessment.

Table 46 displays the results regarding teachers' level of agreement to questions about their students' responses to the assessment items. For the majority of students, teachers agreed (*n* = 824; 49.0%) or strongly agreed (*n* = 620; 36.9%) that the student responded to the items on the assessment to the best of his or her knowledge. However, for 26% (*n* = 432) of students, teachers indicated that the student was not able to respond to items regardless of his or her disability, behavior, or health concerns. In comparison, teachers responding to the 2019 DLM Teacher Survey gave this same response for 15% (n = 7,992) of students.

Table 46

Teachers' Level of Agreement to Questions About Their Students' Responses to the Assessment

Teacher survey item	Strongly disagree *n (%)*	Disagree *n (%)*	Agree *n (%)*	Strongly agree *n (%)*	Missing *n*	Total responses *n*
The student has responded to the items on this assessment to the best of his or her knowledge or ability.	112 (6.7)	124 (7.4)	824 (49.0)	620 (36.9)	566	1,680
The student was able to respond to items regardless of his or her disability, behavior, or health concerns.	218 (13.1)	214 (12.9)	755 (45.5)	473 (28.5)	586	1,660
The student had access to all necessary supports in order to participate on the assessment.	56 (4.0)	65 (4.7)	663 (48.0)	598 (43.3)	864	1,382

Note. Percentages are based on total number of responses. Row totals do not include missing counts.

Teachers answered questions about student effort on the assessment and interest in science instructional activities (see Table 47). For most students, teachers agreed or strongly agreed (79.9%; *n* = 1,367) that the student tried to do his or her best on the assessment. Nearly 70% (*n* = 816) of students are typically interested in science instructional activities.

Table 47

Teachers' Level of Agreement to Questions About Student Effort and Interest

Teacher survey item	Strongly disagree *n (%)*	Disagree *n (%)*	Agree *n (%)*	Strongly agree *n (%)*	Not applicable *n (%)*	Missing *n*	Total responses *n*
This student tried to do his or her best on the assessment.	125 (7.3)	194 (11.3)	831 (48.6)	536 (31.3)	25 (1.5)	535	1,711
This student is typically interested in classroom instructional activities related to science.	137 (11.4)	225 (18.7)	593 (49.2)	223 (18.5)	27 (2.2)	1,041	1,205

Note. Percentages are based on total number of responses. Row totals do not include missing counts.

For nearly 55% students, teachers agreed or strongly agreed that the student was interested and engaged in the assessment. Teachers disagreed or strongly disagreed that students were interested and engaged at a slightly higher rate (47.8%; *n* = 322) for students in Grades 9–12 than for students in Grades 3–5 or 6–8. When examining differences by students' complexity band (see Pilot Study Design section), teachers judged students at the Foundational level (72.8%, *n* = 134) and Band 1 level (63.2%, *n* = 338) as being less engaged than students at higher complexity bands. Table 48 displays teachers' level of agreement by students' grade level and complexity band.

Table 48

Teachers' Level of Agreement to "This Student Was Interested and Engaged in the Assessment" by Grade Level and Complexity Band

Grade level	Strongly disagree *n (%)*	Disagree *n (%)*	Agree *n (%)*	Strongly agree *n (%)*	Not applicable *n (%)*	Missing *n*	Total responses *n*
3–5	108 (20.9)	117 (22.6)	220 (42.6)	70 (13.5)	2 (0.4)	157	517
6–8	88 (16.6)	138 (26.0)	213 (40.2)	88 (16.6)	3 (0.6)	137	530
9–12	142 (21.1)	180 (26.7)	241 (35.8)	103 (15.3)	8 (1.2)	231	674
Complexity band							
Foundational	78 (42.4)	56 (30.4)	38 (20.7)	9 (4.9)	3 (1.6)	53	184
Band 1	169 (31.6)	169 (31.6)	152 (28.4)	44 (8.2)	1 (0.2)	215	535
Band 2	53 (12.7)	101 (24.2)	203 (48.6)	57 (13.6)	4 (1.0)	99	418
Band 3	38 (6.5)	109 (18.7)	281 (48.1)	151 (25.9)	5 (0.9)	158	584
Total	338 (19.6)	435 (25.3)	674 (39.2)	261 (15.2)	13 (0.8)	525	1,721

Note. Percentages are based on total number of responses. Row totals do not include missing counts.

As described in the Testlet Structure and Design section, choice items provided a choice path for students at the Initial, Precursor, and Distal levels. Teachers were asked if their students understood the options. Overall, teachers reported nearly half (48.1%, *n* = 823) of the students did not understand the choice options (see Table 49), and students in Grades 9–12 (52.1%, n = 339) were less likely to understand the choice options than students in Grades 3–5 or 6–8. When examining differences by students' complexity band, teachers judged students placed at the Foundational level (73.4%, *n* = 135) and Band 1 level (67.8%, *n* = 361) as having more difficulty understanding choice options than students at the higher complexity bands.

Table 49

Teachers' Level of Agreement With "This Student Understood the Choice Options" by Grade Level and Complexity Band

Grade level	Strongly disagree *n (%)*	Disagree *n (%)*	Agree *n (%)*	Strongly agree *n (%)*	Not applicable *n (%)*	Missing *n*	Total responses *n*
3–5	110 (21.3)	144 (27.9)	188 (36.4)	72 (14.0)	2 (0.4)	158	516
6–8	79 (15.0)	141 (26.8)	220 (41.8)	84 (15.9)	3 (0.6)	140	527
9–12	136 (20.3)	213 (31.8)	223 (33.3)	88 (13.2)	9 (1.4)	236	669
Complexity band							
Foundational	73 (39.7)	62 (33.7)	32 (17.4)	12 (6.5)	5 (2.7)	53	184
Band 1	164 (30.8)	197 (37.0)	141 (26.5)	27 (5.1)	3 (0.6)	218	532
Band 2	48 (11.6)	132 (31.8)	164 (39.5)	68 (16.4)	3 (0.7)	102	415
Band 3	40 (6.9)	107 (18.4)	294 (50.6)	137 (23.6)	3 (0.5)	161	581
Total	325 (19.0)	498 (29.1)	631 (36.9)	244 (14.3)	14 (0.8)	534	1,712

Note. Percentages are based on total number of responses. Row totals do not include missing counts.

When asked if their students enjoyed the assessment experience, almost half of the responses (48.4%, *n* = 800) indicated teachers agreed or strongly agreed (Table 50). Teachers' responses indicated they agreed or strongly agreed at higher rates for students in Grades 3–5 and 6–8 than for students in Grades 9–12. When examining differences by students' complexity band, teachers judged students at the Foundational level (68.4%, *n* = 121) and Band 1 level (65.1%, *n* = 332) as enjoying the assessment experiences less than students at the higher complexity bands.

Table 50

Teachers' Level of Agreement With "This Student Enjoyed the Assessment Experience" by Grade Level and Complexity Band

Grade level	Strongly disagree *n (%)*	Disagree *n (%)*	Agree *n (%)*	Strongly agree *n (%)*	Not applicable *n (%)*	Missing *n*	Total responses *n*
3–5	109 (22.1)	120 (24.3)	196 (39.8)	59 (12.0)	9 (1.8)	181	493
6–8	86 (16.9)	156 (30.7)	190 (37.4)	64 (12.6)	12 (2.4)	159	508
9–12	145 (22.3)	202 (31.0)	225 (34.6)	66 (10.1)	13 (2.0)	254	651
Complexity band							
Foundational	75 (42.4)	46 (26.0)	43 (24.3)	5 (2.8)	8 (4.5)	60	177
Band 1	161 (31.6)	171 (33.5)	132 (25.9)	35 (6.9)	11 (2.2)	240	510
Band 2	56 (13.9)	113 (28.0)	183 (45.3)	43 (10.6)	9 (2.2)	113	404
Band 3	48 (8.6)	148 (26.4)	253 (45.1)	106 (18.9)	6 (1.1)	181	561
Total	340 (20.6)	478 (28.9)	611 (37.0)	189 (11.4)	34 (2.1)	594	1,652

Note. Percentages are based on total number of responses. Row totals do not include missing counts.

Table 51 shows the survey results regarding how many of the students' assessment tasks had content that matched instruction, overall, by grade level, and by students' complexity band. Teachers indicated that only 24.7% (*n* = 425) of students had assessment tasks with most or all content that matched instruction. This finding was surprising given that the eligibility criteria included students receiving instruction on the two grade-band specific science EEs. Teachers reported a smaller number of high school students (21.9%, *n* = 148) had assessment tasks that matched instruction compared to Grades 3–5 or 6–8. Additionally, for 88.1% of students (*n* = 163) in the Foundational complexity band, less than half of their assessment tasks matched their instruction.

Table 51

*Proportion of Assessment Tasks with Content that Matched Instruction (*N = *1,726)*

Grade level	None *n (%)*	Some (less than half) *n (%)*	Most (more than half) *n (%)*	All *n (%)*
3–5	102 (19.7)	287 (55.5)	106 (20.5)	22 (4.3)
6–8	100 (18.8)	282 (53.1)	128 (24.1)	21 (4.0)
9–12	138 (20.4)	392 (57.8)	128 (18.9)	20 (3.0)
Complexity Band				
Foundational	54 (29.2)	109 (58.9)	15 (8.1)	7 (3.8)
Band 1	143 (26.6)	303 (56.3)	82 (15.2)	10 (1.9)
Band 2	81 (19.3)	245 (58.5)	81 (19.3)	12 (2.9)
Band 3	62 (10.6)	304 (52.1)	184 (31.5)	34 (5.8)
Total	340 (19.7)	961 (55.7)	362 (21.0)	63 (3.7)

Note. Missing responses = 520.

The survey also asked teachers how well the difficulty level of the majority of the content in the testlets matched the students' skill level. For over 60% of students (*n* = 1,037), teachers felt that the assessment tasks were "too hard". Nearly 86% of students (*n* = 156) at the Foundational level and over 81% of Band 1 students (*n* = 437) had assessment content that was judged to be "too hard." Table 52 displays the distribution of responses by grade level and students' complexity band.

Table 52

*How Did the Difficulty Level of the Majority of the Content Match the Student's Skill Level? (*N = *1,726)*

Grade level	Too easy *n (%)*	About right *n (%)*	Too hard *n (%)*	I don't remember *n (%)*
3–5	11 (2.1)	191 (36.9)	310 (59.9)	6 (1.2)
6–8	10 (1.9)	216 (40.7)	296 (55.7)	9 (1.7)
9–12	6 (0.9)	225 (33.2)	431 (63.7)	15 (2.2)
Complexity band				
Foundational	1 (0.6)	22 (12.1)	156 (85.7)	3 (1.7)
Band 1	0 (0.0)	95 (17.7)	437 (81.2)	6 (1.1)
Band 2	6 (1.4)	181 (43.1)	224 (53.3)	9 (2.1)
Band 3	20 (3.4)	334 (57.0)	220 (37.5)	12 (2.1)
Total	27 (1.6)	632 (36.6)	1,037 (60.1)	30 (1.7)

Note. Missing responses = 520.

Teachers were asked to select factors related and not related to the assessment that may have impacted their students' responses to assessment items (Table 53). Teachers felt that the expectations were too high for nearly 62% of students (*n* = 1,070). For over 35% of students (*n* = 608), teachers believed that the student's overall temperament on the day of the assessment may have impacted their responses. For a large number of students, teachers indicated that "other" factors, both related (28.2%, *n* = 488) and unrelated (39.7%, *n* = 686) to the assessment impacted student performance, suggesting that there were many unspecified factors that took place in the pilot study.

Table 53

*Factors That May Have Impacted the Student's Response to Items (*N = *1,730)*

Factors	*N*	%
Factors related to the assessment		
It was too high an expectation for this student	1,070	61.9
Other	488	28.2
Graphics were ambiguous	225	13.0
Graphics were distracting	91	5.3
It was too low an expectation for this student	39	2.3
Factors *not* related to the assessment		
Other	686	39.7
Overall temperament that day	608	35.1
Environmental distractors (e.g., noisy room, temperature of room)	561	32.4
Physical or emotional behaviors that occurred that day	465	26.9

Note. Missing responses = 516. Teachers were allowed to select multiple responses, so percentages add up to more than 100.

One topic of particular interest in this part of the study was the relationship between effort and engagement. To gain preliminary understanding of student effort as it related to student interest and engagement, we calculated the Spearman correlation between teacher responses to two items, "the student was interested and engaged in the assessment," and "the student tried to do his/her best on the assessment." Those who chose *Not Applicable* were eliminated from the analysis. The correlation was .58 ($p < 0.0001$), which demonstrates a large effect size. Teachers who judged their students to be interested and engaged typically also believed that these students tried to do their best on the assessment.

Relationship of Teacher Perceptions With Student Performance

To evaluate the relationship between teacher perceptions and student performance, we examined the Spearman correlation between teachers' responses to five Likert items judging students' interest, engagement, effort, understanding, and enjoyment of the assessment with students' EE mastery results. We eliminated those who chose *Not Applicable* from the correlational analysis. Table 54 displays the correlations with each statement. Correlations range from 0.27 to 0.43. While all are statistically significant, the correlation of the statement "The student tried to do his or her best on the assessment" with students' EE mastery demonstrates a small effect size, while all other correlations demonstrate a moderate effect size. As teacher-reported student interest and engagement increased, EE mastery also increased.

Table 54

Spearman Correlation Coefficients Between Teacher's Responses to Selected Items and Students' Essential Element Mastery

Teacher survey item	Correlation coefficient
This student was interested and engaged in the assessment	0.41
This student tried to do his or her best on the assessment	0.27
The student understood the choice options	0.43
This student enjoyed the assessment experience	0.39
This student is typically interested in classroom instructional activities related to science	0.32

Note. $p < .0001$ for all correlation coefficients.

Table 55 displays median EE mastery across the agreement categories selected by teachers for each item. As students' interest and engagement as reported by teachers increased, EE mastery also increased. This pattern holds across all survey items.

Table 55

Median Essential Element Mastery by Teachers' Level of Agreement with Items

Teacher survey item	Strongly disagree *n*, median	Disagree *n*, median	Agree *n*, median	Strongly agree *n*, median
The student was interested and engaged in the assessment	338, 0.000	435, 0.125	674, 0.375	261, 0.500
This student tried to do his or her best on the assessment	125, 0.000	194, 0.125	831, 0.250	536, 0.375
The student understood the choice options	325, 0.000	498, 0.125	631, 0.375	244, 0.500
The student enjoyed the assessment experience	340, 0.000	478, 0.250	611, 0.375	189, 0.500
The student is typically interested in classroom instructional activities related to science	137, 0.000	225, 0.125	593, 0.250	223, 0.375

Note. Group sums do not add up to the overall total because of missing responses.

Teachers' perceptions of how well students mastered the content in relation to how they performed on the assessment was also considered. After each testlet, the teacher was asked how well the student had mastered the content. We correlated teachers' judgments (1 = Did not master skills, 2 = Partially mastered skills, 3 = Mastered skills) with students' EE mastery results. The Spearman correlation coefficient was 0.41 ($p < .0001$), which demonstrates a moderate effect size. Teachers' judgement of students' mastery was significantly correlated with their students' EE mastery score. Table 56 displays the median of students' EE mastery by teachers' perceptions of student mastery. Students whose teachers judged them to master or partially master skills had higher EE mastery results than those judged not to master the skills.

Table 56

Median Essential Element Mastery by Teachers' Perceptions of Student Mastery

Teachers' judgement of student mastery	Proportion of nodes mastered for the Essential Element *n*, median
Mastered skills	64, 0.750
Partially mastered skills	569, 0.500
Did not master skills	931, 0.125
Missing	293, 0.125
Overall total	1,857, 0.250

Note. Essential Element mastery is the mean performance (i.e., mastered = 1, did not master = 0) on the nodes of an Essential Element.

To evaluate the extent to which teacher perceptions of testlet difficulty were associated with student performance, we correlated teachers' perceptions of (a) the difficulty level of the testlet content matched to the student's skill level and (b) whether assessment items were "too high of an expectation" for the student with student's EE mastery.

The Spearman correlation between teachers' perception of the difficulty level of the testlet content (1 = too easy, 2 = about right, 3 = too hard) and EE mastery was –0.45 ($p < .0001$), which demonstrates a moderate effect size. Students whose teacher had perceptions of content being "too easy" had a higher median EE mastery than those whose teacher judged the content as "too hard." Table 57 displays the median of EE mastery by teachers' perceptions of content difficulty.

Table 57

Median Essential Element Mastery by Teachers' Perception of Content Difficulty

Content difficulty	Proportion of nodes mastered for the Essential Element *n*, median
Too easy	27, 0.750
About right	632, 0.500
Too hard	1,037, 0.125
Missing	520, 0.125
Total	2,216, 0.250

Note. Essential Element mastery is the mean performance (i.e., mastered = 1, did not master = 0) on the nodes of an Essential Element.

The point biserial correlation between teachers choosing "It was too high an expectation for this student" (0 = no; 1 = yes) as a factor impacting student response (see Table 53) and student EE mastery was –0.36. Students whose teachers indicated that items were too high of an expectation generally had lower EE mastery. Table 58 shows the median of EE mastery by instructional time. EE mastery increased as amount of instruction increased.

Table 58

Median Essential Element (EE) Mastery by Instructional Time

Instructional time	Proportion of nodes mastered for the Essential Element *n*, median
None	386, 0.125
1–10 hours	1,055, 0.250
11–20 hours	312, 0.313
21–30 hours	109, 0.375
>30 hours	93, 0.375
Missing	291, 0.125
Total	2,246, 0.250

Note. Essential Element mastery is the mean performance (i.e., mastered = 1, did not master = 0) on the nodes of an Essential Element .

Conclusions and Future Directions

One of the culminating objectives of Goal 2 of the I-SMART project is to have a set of testlets that could be models for a future assessment system along with a scoring model that supports reporting of student mastery of nodes in I-SMART neighborhood maps associated with several Essential Elements (EEs). As such, the I-SMART pilot study, conducted in the winter and fall of 2019, evaluated the I-SMART testlets, including item quality, item difficulty, and impact of item features on performance, and provided data to select a final scoring model. The pilot study also gathered data on teachers' and students' perceptions of the assessment. This section organizes the main findings by each research question addressed in the pilot study, followed by potential future directions.

Research Question 1: How do item and testlet features impact item and testlet performance?

The I-SMART pilot study examined student performance on the I-SMART testlets through data regarding node mastery, administration time, performance on choice paths and wonder questions, and item level performance. For node mastery, the number of nodes mastered was examined by linkage level, EE, grade band, and domain. Students at lower linkage levels tended to master more nodes than students at higher linkage levels. Students mastered more nodes in the physical science EEs than in the life science and earth and space science EEs. Students in the middle school grade band tended to master more nodes than students in the elementary and high school grade bands. The total administration time was less than 6 minutes for most testlets, and the administration time of scored items was less than 4 minutes for most testlets. For the choice items, students tended to choose the second choice option. Despite this tendency, the number of nodes mastered was similar across the choice paths, suggesting similar difficulty across the choice paths. Most students selected the same answer to the wonder question presented at the beginning of the testlet and again at the end. Correct responses, whether to one or both of the wonder questions, were positively related to the number of nodes mastered. For item level performance, items were flagged for unexpected psychometric properties using four statistics. Of the 460 items developed for the I-SMART pilot study, 133 items (29%) were not flagged by any statistics, and 210 items (46%) were only flagged by a single statistic. These results suggest the majority of items performed as expected.

The test development team reviewed the flagged items and noted two major themes. First, the test development team noted that there was a significant percentage of students who did not respond to items at the Initial level. This may suggest that these items were insufficiently engaging for some students at the Initial level. Alternatively, this may indicate that some students were working at a level lower than the Initial level and that these students require further reduced complexity to access the testlet content. Second, the test development team noted that distractors could be better targeted to specific misconceptions. There were several items where non-masters had a high probability of providing a correct response, indicating that the distractors were too obviously incorrect. Similarly, there were items where masters had a low probability of providing a correct answer, suggesting that the distractors may have been too challenging for these items. A consequence of the findings related to distractors is that linkage levels may not optimally discriminate between masters and non-masters. This was supported in our analysis of the data, as a few linkage levels did show a discrimination value of less than .25 (see Figure 7).

The findings pertaining to the impact of item and testlet features on item and testlet performance were positive. The item level performance results suggest the items performed as expected, the

difficulty across the choice paths was similar, and correct responses to the wonder questions were related to the number of nodes mastered. The total administration time and administration time of the scored items suggest the testlets developed for the I-SMART pilot study were appropriate for the intended population, in terms of the amount of time we would expect students with significant cognitive disabilities to devote to completing assessment items. The time spent by students on the I-SMART pilot study testlets was comparable to what has been observed for other assessment programs for this population of students [i.e., the Dynamic Learning Maps® (DLM®) assessments].

Research Question 2: Which type of scoring model is optimal for the student response data collected from the assessments?

The I-SMART modeling approach estimates student mastery of assessed knowledge, skills, and understandings using diagnostic classification modeling. A latent class analysis is applied to each learning map node to estimate the posterior probability of node mastery. After examining different item constraints, the fungible model appears to fit the student responses the best. Thus, the items are assumed to be equivalent within each node, meaning the probability that masters will respond correctly to an item is the same for all items within each node and the probability that non-masters will respond correctly to an item is the same for all items within each node. In determining node mastery, students were classified as masters if they had a posterior probability of node mastery greater than or equal to .80 or if they had a correct response rate of 80% or greater. Students who do not meet either of these criteria are classified as non-masters of the node. The majority of nodes had reliability statistics greater than .60, indicating acceptable levels of agreement for categorical classifications (Landis & Koch, 1977). The nodes with reliability statistics less than .60 occurred at higher linkage levels, which suggests the lower reliability statistics are likely due to smaller sample sizes at the higher linkage levels. Thus, reliability analyses indicate that, overall, student mastery classifications have a high degree of consistency.

Research Question 3: Do empirical data support the structure of the learning map models?

The map validation studies indicated general support for the proposed map structures. The node validation studies indicated that nodes tended to provide unique information about students' mastery of the assessed skills. The correlations of node mastery study demonstrated a stronger relationship within linkage level than within EE, but these relationships were not so strong that mastery of one node implied mastery of another node. The overlapping node consistency study showed that all of the overlapping nodes demonstrated consistent performance based on the node-level weighted *p*-values and standard errors. Further, the majority of overlapping nodes demonstrated adequate classification consistency. The studies evaluating map structure supported the proposed learning map models. While the patterns of mastery profiles study showed limited evidence of adequate absolute model fit for the saturated and the constrained models, the patterns of mastery assignment showed the majority of linkage levels and connections were accurately specified. Further, the patterns of node difficulty demonstrated weighted node *p*-values that were consistent with the underlying learning map models. Taken together, these results suggest the models largely conformed to the hypothesized learning map models. There were, however, a few notable discrepancies between student mastery patterns and the hypothesized learning map models. As such, future work in this area should continue gathering data to further evaluate the proposed structure of the learning map models in areas where these discrepancies were noted.

Research Question 4: What are students' and teachers' perceptions of students' experiences with the new testlets?

Analyses showed that teachers' perceptions of both testlet difficulty and student experiences were related to student performance on the I-SMART assessments. Teachers' perceptions regarding student engagement and interest, effort, understanding of choice options, enjoyment of the assessment experience, and interest in classroom science instructional activities were all significantly correlated with student mastery of EEs. The results also support the notion that students have better results when they receive more instruction on the EE. However, teachers reported a large number of students across grade levels as not having assessment tasks that matched content they received in instruction. This indicates a potential mismatch between the science instruction teachers are providing and the complexity of the I-SMART assessment content.

While the results for Research Question 1 showed that students at lower linkage levels tended to master more nodes than students at higher linkage levels, teacher survey findings suggest students at the lower complexity bands (Foundational and Band 1) are not engaged in science instruction or assessments. These contradictory results suggest the need for additional research to understand relationships between student complexity band, teacher expectations, opportunity to learn science content, and student performance.

Teachers generally agreed that their students used similar accessibility supports on the I-SMART pilot assessments to what they use in instruction. Comparisons of I-SMART Personal Needs and Preferences profile selections to the full DLM population also showed that teachers select accessibility supports at similar frequencies.

Significance and Future Directions

Results from the I-SMART pilot study provide many useful insights for developing innovative, Next Generation Science Standards–based science assessments for students with significant cognitive disabilities. From these findings, there are several areas for potential future research.

Test Development

First, the test development team noted that additional work is needed to evaluate how best to provide accessible content at an appropriate level for all students, but especially those testing at the Initial level. Specifically, they noted that for these Initial items, there was often a conditional *p*-value for masters above .8, with an overall *p*-value below .35, suggesting that apart from the masters, the items were too difficult for most of the students. Notably, many of the items show almost equal numbers of students answering correctly as students who "attended to other stimuli," indicating that these items may not have been engaging for some students.

Second, the test development team noted that many items, across all linkage levels, had a relatively high probability of non-masters providing a correct response, suggesting that students who have not yet mastered the content are able to guess the correct answer. Improved distractors that more purposefully target specific misconceptions may help better discriminate between masters and non-masters, parsing the difference between true masters and those students who were able to correctly guess the correct response to items.

Finally, two patterns were discerned regarding item types and content. For item types, the test development team noted that multiple-choice multiple-select items, where the student can select multiple responses, tend to perform worse than standard multiple-choice items, likely due to students being unclear on how many responses are necessary to respond satisfactorily. In terms of content, the test development team noted that items related to food chains tended to have mixed performance. Thus, future work will focus on providing resources and revisions to item writers to make the expectations for these items more explicit.

Psychometric Modeling

The psychometric modeling research conducted on the I-SMART pilot assessment data presents several avenues for future work. First, the current work used weakly informed priors for estimating the diagnostic models. It is possible that using empirical priors would improve the model estimation process. The research literature would benefit from work evaluating under what conditions empirical priors are the most helpful or appropriate, including sample size and model complexity.

Second, the I-SMART pilot data were used to provide evidence supporting the structure of the developed learning maps. This was accomplished using diagnostic classification model framework described by Thompson and Nash (2019). Although the totality of evidence supported the defined structure, certain diagnostic models showed insufficient absolute model data fit. There are many reasons this could be the case. Because there were relatively small sample sizes, it is possible that there simply were not enough data to get reliable parameter estimates, even though the models converged. Thus, it is possible that regularizing priors could stabilize the estimates, leading to better model fit outcomes. Additionally, the models described in this report were restricted to single linkage level models. That is, the models were limited to 3–5 nodes that were grouped together in the linkage level. However, the full learning map for each EE contains 9–20 tested nodes from three or four linkage levels. Future work may examine the feasibility of concurrently estimating all nodes within an EE. This is much more computationally intensive, but it could also provide node mastery probabilities for nodes from linkage levels on which the students were not assessed. That is, it may be possible to infer mastery status for untested nodes by propagating information from the tested nodes through the defined map structure.

Learning Maps Evaluation

Finally, the results from the evaluation of the map structure were used to examine possible misspecifications in the map structure for a single EE. Future work could take these empirical recommendations and apply them to the map structure to evaluate the impact of the changes on model fit and classification accuracy. Simulation studies may examine which type of modifications to the proposed map structure are most likely to result in improved model fit and student classification accuracy.

Student Experience

While overall student mastery on the I-SMART assessments was low and teachers indicated that the assessment tasks were too difficult for most students, the majority of students gave positive self-evaluations of their performance and were interested and engaged in the assessments. These findings may be due to the innovative features (wonder questions and choice options) included in the I-SMART testlets. Future research should further explore student experience taking tests that include specific features designed to increase engagement, even if these features do not improve test performance. These features have great potential to activate

students' prior knowledge and foster a greater sense of agency and self-efficacy while completing assessments.

The results also suggest that student choice is an important area for future research. Choice-making is a test-taking skill that students can be taught, and there is a need to help teachers understand the intent of choice-making in assessment so they can help students get beyond low-level binary choice-making. Students should be taught to make choices based on their prior knowledge so that the choices they make give them the greatest chances of successful performance.

In addition, in this pilot study, the choice options were always presented in the same positions (e.g., if the choice options were cat and bear, the cat was always presented first and the bear was presented second). Future research using choice options should counterbalance the position of the choice items so that it is possible to distinguish students' choice of content from the position (first or second) because some students in the population may have a tendency to always select the last choice option.

Finally, there were many factors impacting student engagement and performance on the I-SMART assessments, including students' complexity band, their opportunity to learn the content on the assessments, and their teachers' perceptions, experiences, and administration load. The pilot study findings revealed that students' opportunity to learn the content of the I-SMART assessments was relatively low. Future research should further investigate students' engagement, interest, and access at lower complexity bands and the impact of targeted teacher professional development focusing on how to teach science to the expectations specified by the Next Generation Science Standards.

References

Alonzo, A. C., & Gotwals, A. W. (Eds.). (2012). *Learning progressions in science: Current challenges and future directions*. Sense Publishers.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541–561. https://doi.org/10.1007/BF02296195

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, *5*, 1089–1105. http://www.jmlr.org/papers/v5/grandvalet04a.html

Betancourt, M., & Girolami, M. (2013). Hamiltonian Monte Carlo for hierarchical models. https://arxiv.org/abs/1312.0906

Bradshaw, L. (2016). Diagnostic classification models. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 297–327). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch13

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, *33*(1), 2–14. https://doi.org/10.1111/emip.12020

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76,* 1–32. https://doi.org/10.18637/jss.v076.i01

CAST (2018). *Universal Design for Learning Guidelines version 2.2.* http://udlguidelines.cast.org

Clark, A. K., & Karvonen, M. (2019, April). Teacher assessment literacy: Implications for diagnostic assessment systems. *Assessment as Feedback for Teachers and*

Students [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, Toronto, Ontario, Canada.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Academic Press.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). Learning progressions in science: An evidence based approach to reform (CPRE Research Report #RR-63). Consortium for Policy Research in Education.

Crawford, A. (2014). *Posterior predictive model checking in Bayesian networks* [Doctoral Dissertation, Arizona State University]. ASU Repository. https://repository.asu.edu/attachments/134809/content/Crawford_asu_0010E_13643.pdf

DeBoer, G. E. (2014). The history of science curriculum reform in the United States. In N. G. Lederman, & S. K. Abell (Eds.), *Handbook of research on science education* (Vol. II, pp. 559–578). Routledge.

Dynamic Learning Maps Consortium. (2019, December). *2018–2019 Technical Manual Update—Science*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/publication/2018-2019_Science_Technical_Manual_Update.pdf

Feldberg, Z., & Bradshaw, L. (2019, April). Reporting results from diagnostic classification models for teachers. *Assessment as Feedback for Teachers and Students* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, Toronto, Ontario, Canada.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman; Hall/CRC.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–760.

Guo, J., Gabry, J., & Goodrich, B. (2019). *RStan: R interface to Stan*. https://CRAN.R-project.org/package=rstan

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210. https://doi.org/10.1007/S11336-008-9089-5

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamilttonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623. http://jmlr.org/papers/v15/hoffman14a.html

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement, 55*, 635–664. https://doi.org/10.1111/jedm.12196

Karvonen, M., Ruhter, L., Shipman, M., Swinburne Romine, R., & Tiemann, G. (2020). Designing, developing, and evaluating innovative science assessments. University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://ismart.works/sites/default/files/documents/Publications/I-SMART_Goal2_TestDev_TechnicalReport_FINAL.pdf

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178-206.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

National Research Council (NRC). (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* National Academies Press. https://doi.org/10.17226/13165

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 113–162). Chapman; Hall/CRC.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. https://doi.org/10.17226/18290

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* Guilford.

Sen, S., & Bradshaw, L. (2017). Comparison of relative fit indices for diagnostic model selection. *Applied Psychological Measurement*, *41*(6), 422–438. https://doi.org/10.1177/0146621617695521

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research & Perspective*, *16*(1), 1–17. https://doi.org/10.1080/15366367.2018.1435104

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement*, *67*(2), 239–257. https://doi.org/10.1177/0013164406292025

Sinharay, S., & Johnson, M. S. (2019). Measures of agreement: Reliability, classification accuracy, and classification consistency. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models* (pp. 359–377). Springer Nature. https://doi.org/10.1007/978-3-03-05584-4_17

Stage, E. K., Asturias, H., Cheuk, T., Daro, P. A., & Hampton, S. B. (2013). Opportunities and challenges in next generation standards. *Science, 340(*6130*)*, 276–277. https://doi.org/10.1126/science.1234011

Swinburne Romine, R., Andersen, L., Schuster, J., & Karvonen, M. (2018). *Developing and evaluating learning map models in science: Evidence from the I-SMART project*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://ismart.works/sites/default/files/documents/Publications/I-SMART_Goal_1_Technical_Report_FINAL_0.pdf

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317–339. https://doi.org/10.1007/s11336-013-9362-0

Templin, J., & Henson, R. A. (2008, March). *Understanding the impact of skill acquisition: Relating diagnostic assessments to measurable outcomes* [Paper presentation]. Annual Meeting of the American Educational Research Association, New York, New York.

Thompson, W. J. (2019). *Bayesian psychometrics for diagnostic assessments: A proof of concept* (Research Report #19-01). University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Bayes_Proof_Concept_2019.pdf

Thompson, W. J., & Nash, B. (2019, April). *Empirical methods for evaluating maps: Illustrations and results* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, Toronto, Ontario, Canada.

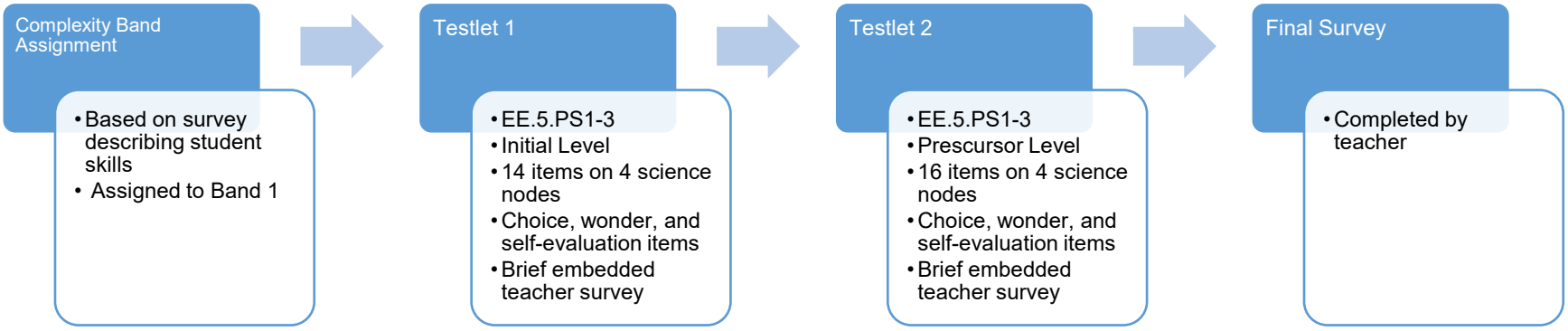
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement, 52*, 457–476. https://doi.org/10.1111/jedm.12096

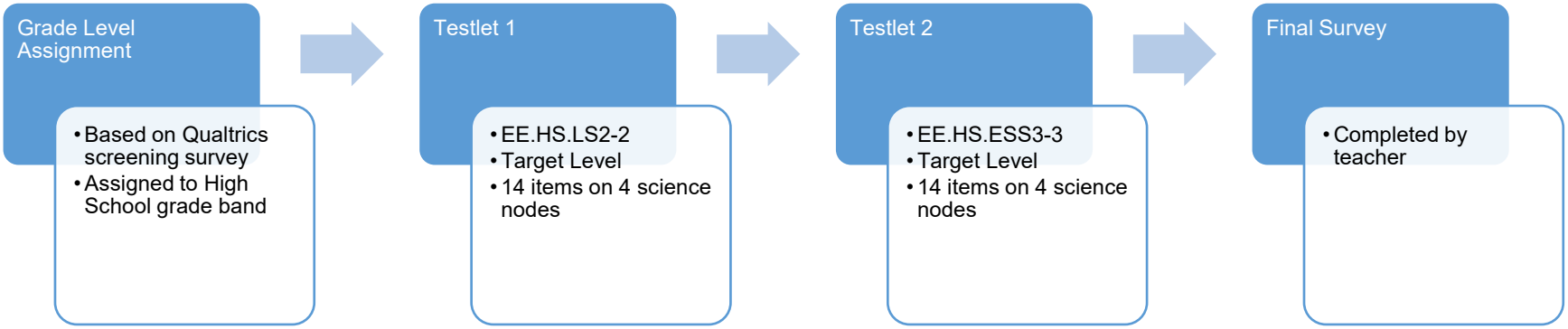
Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594. http://www.jmlr.org/papers/v11/watanabe10a.html

Appendix A

Example Pilot Test Administration



Example pilot test administration for a fifth grade primary population student



Example pilot test administration for an eighth grade secondary population student

Appendix B

Proportion of Mastery for Each Node by Linkage Level and Choice Item

Linkage level	Choice Item	Node	Choice	n	Proportion
Initial	1	F-66	Forest	57	0.12
Initial	1	F-66	Ocean	50	0.26
Initial	1	SCI-315	Forest	57	0.21
Initial	1	SCI-315	Ocean	50	0.36
Initial	1	SCI-501	Forest	57	0.11
Initial	1	SCI-501	Ocean	50	0.18
Initial	1	SCI-527	Forest	57	0.23
Initial	1	SCI-527	Ocean	50	0.26
Initial	2	F-121	Beach	83	0.23
Initial	2	F-121	Park	46	0.28
Initial	2	SCI-597	Beach	83	0.06
Initial	2	SCI-597	Park	46	0.15
Initial	2	SCI-598	Beach	83	0.17
Initial	2	SCI-598	Park	46	0.13
Initial	2	SCI-614	Beach	83	0.16
Initial	2	SCI-614	Park	46	0.24
Initial	3	F-77	Bear	66	0.62
Initial	3	F-77	Cat	84	0.52
Initial	3	SCI-317	Bear	66	0.33
Initial	3	SCI-317	Cat	84	0.27
Initial	3	SCI-326	Bear	66	0.45
Initial	3	SCI-326	Cat	84	0.44
Initial	3	SCI-666	Bear	66	0.09
Initial	3	SCI-666	Cat	84	0.10
Initial	4	F-66	Farm	158	0.54
Initial	4	F-66	Park	74	0.57
Initial	4	SCI-313	Farm	158	0.67
Initial	4	SCI-313	Park	74	0.57
Initial	4	SCI-314	Farm	158	0.60
Initial	4	SCI-314	Park	74	0.62
Initial	4	SCI-315	Farm	158	0.58
Initial	4	SCI-315	Park	74	0.51
Initial	5	F-105	House	189	0.74
Initial	5	F-105	Park	200	0.65
Initial	5	SCI-117	House	189	0.32
Initial	5	SCI-117	Park	200	0.36
Initial	5	SCI-119	House	189	0.58
Initial	5	SCI-119	Park	200	0.51
Initial	5	SCI-151	House	189	0.50
Initial	5	SCI-151	Park	200	0.44
Initial	6	F-75	House	99	0.10
Initial	6	F-75	School	145	0.33
Initial	6	SCI-117	House	99	0.19
Initial	6	SCI-117	School	145	0.59
Initial	6	SCI-119	House	99	0.28

Linkage level	Choice Item	Node	Choice	n	Proportion
Initial	6	SCI-119	School	145	0.66
Initial	6	SCI-151	House	99	0.52
Initial	6	SCI-151	School	145	0.78
Distal	7	SCI-497	Tiger	213	0.14
Distal	7	SCI-497	Whale	136	0.12
Distal	7	SCI-501	Tiger	213	0.34
Distal	7	SCI-501	Whale	136	0.37
Distal	7	SCI-661	Tiger	213	0.64
Distal	7	SCI-661	Whale	136	0.46
Distal	7	SCI-80	Tiger	213	0.08
Distal	7	SCI-80	Whale	136	0.04
Distal	8	SCI-598	Beach	143	0.29
Distal	8	SCI-598	Mountain	138	0.22
Distal	8	SCI-599	Beach	143	0.27
Distal	8	SCI-599	Mountain	138	0.11
Distal	8	SCI-600	Beach	143	0.04
Distal	8	SCI-600	Mountain	138	0.07
Distal	8	SCI-601	Beach	143	0.49
Distal	8	SCI-601	Mountain	138	0.46
Precursor	9	SCI-119	House	256	0.20
Precursor	9	SCI-119	Park	228	0.17
Precursor	9	SCI-121	House	256	0.28
Precursor	9	SCI-121	Park	228	0.07
Precursor	9	SCI-130	House	256	0.19
Precursor	9	SCI-130	Park	228	0.14
Precursor	9	SCI-27	House	256	0.34
Precursor	9	SCI-27	Park	228	0.24
Precursor	10	SCI-573	Camping	120	0.16
Precursor	10	SCI-573	Picnic	112	0.12
Precursor	10	SCI-574	Camping	120	0.27
Precursor	10	SCI-574	Picnic	112	0.23
Precursor	10	SCI-601	Camping	120	0.14
Precursor	10	SCI-601	Picnic	112	0.15
Precursor	10	SCI-643	Camping	120	0.38
Precursor	10	SCI-643	Picnic	112	0.21
Precursor	11	SCI-119	Bread	59	0.44
Precursor	11	SCI-119	Cookies	285	0.55
Precursor	11	SCI-121	Bread	59	0.32
Precursor	11	SCI-121	Cookies	285	0.54
Precursor	11	SCI-35	Bread	59	0.12
Precursor	11	SCI-35	Cookies	285	0.15
Precursor	11	SCI-804	Bread	59	0.75
Precursor	11	SCI-804	Cookies	285	0.68
Precursor	12	SCI-309	Meadow	75	0.27
Precursor	12	SCI-309	Pond	115	0.22
Precursor	12	SCI-313	Meadow	75	0.20
Precursor	12	SCI-313	Pond	115	0.14
Precursor	12	SCI-326	Meadow	75	0.12

Linkage level	Choice Item	Node	Choice	n	Proportion
Precursor	12	SCI-326	Pond	115	0.10
Precursor	12	SCI-8	Meadow	75	0.04
Precursor	12	SCI-8	Pond	115	0.04
Precursor	13	SCI-515	Forest	293	0.09
Precursor	13	SCI-515	Pond	142	0.02
Precursor	13	SCI-528	Forest	293	0.05
Precursor	13	SCI-528	Pond	142	0.04
Precursor	13	SCI-660	Forest	293	0.20
Precursor	13	SCI-660	Pond	142	0.15
Precursor	13	SCI-80	Forest	293	0.18
Precursor	13	SCI-80	Pond	142	0.11
Precursor	14	SCI-313	Arctic	153	0.38
Precursor	14	SCI-313	Lake	171	0.11
Precursor	14	SCI-324	Arctic	153	0.13
Precursor	14	SCI-324	Lake	171	0.07
Precursor	14	SCI-37	Arctic	153	0.22
Precursor	14	SCI-37	Lake	171	0.14
Precursor	14	SCI-474	Arctic	153	0.18
Precursor	14	SCI-474	Lake	171	0.05

Appendix C

Nodes Outside of the Ideal Range in the Probability of a Correct Response, Discrimination, and Base Rate, by Essential Element and Linkage Level

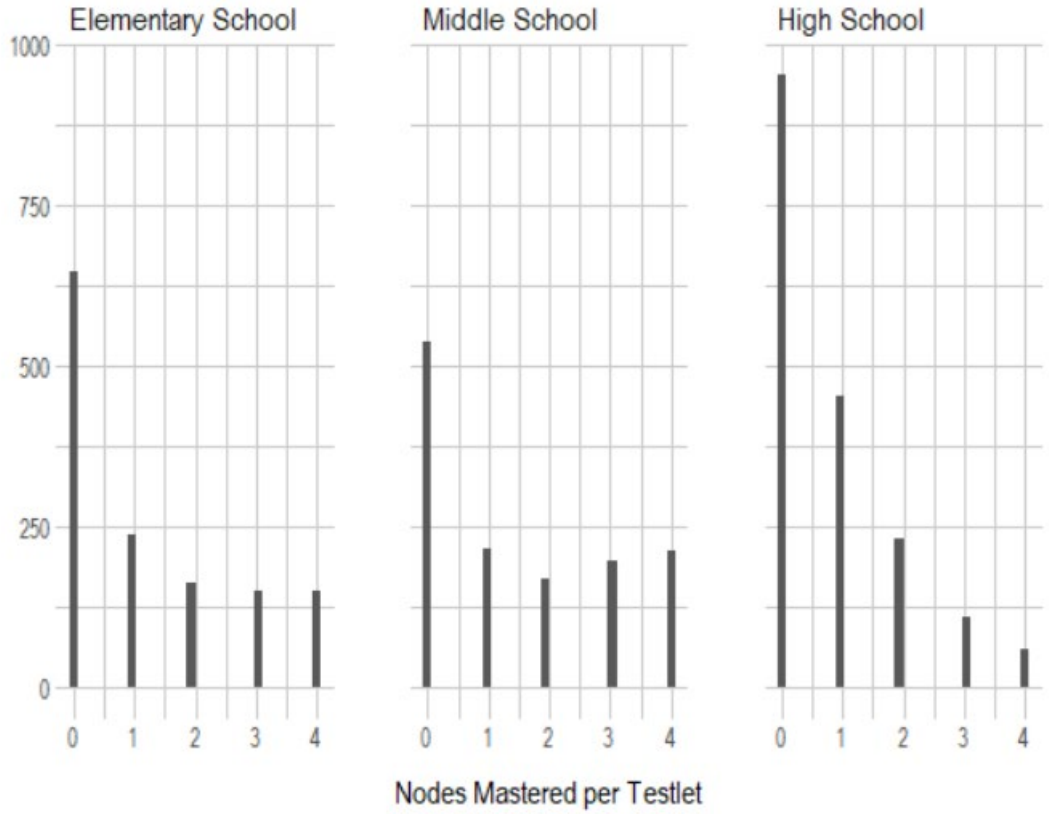
Essential Element - Linkage Level -node	Masters	Non-masters	Discrimination	Base rate
5.LS2-1				
Precursor				
SCI-326	.81	.35	.46	.22*
SCI-8	.43*	.29	.13*	.48
SCI-313	.86	.32	.54	.23*
Target				
SCI-311	.45*	.30	.15*	.46
SCI-309	.85	.47*	.38*	.37
SCI-307	.73	.36	.37*	.52
5.PS1-3				
Precursor				
SCI-121	.64	.25	.39*	.39
SCI-130	.93	.33	.60	.21*
Target				
SCI-155	.71	.34	.37*	.48
SCI-804	.50	.36	.14*	.50
SCI-121	.77	.40*	.37*	.51
MS.LS2-2				
Precursor				
SCI-324	.82	.34	.48	.15*
Target				
SCI-656	.81	.43*	.37*	.36
SCI-518	.77	.47*	.30*	.45
SCI-481	.59	.40*	.18*	.45
MS.PS1-2				
Precursor				
SCI-35	.92	.32	.60	.15*
Target				
SCI-142	.66	.32	.34*	.53
SCI-214	.70	.33	.37*	.59
SCI-804	.69	.52*	.17*	.49
SCI-189	.78	.46*	.32*	.48
HS.ESS3-3				
Initial				
SCI-614	.68	.04	.64	.24*
SCI-597	.73	.07	.67	.18*
SCI-598	.78	.06	.72	.19*
Distal				

Essential Element - Linkage Level -node	Masters	Non-masters	Discrimination	Base rate
SCI-600	.73	.25	.48	.18*
Precursor				
SCI-574	.76	.41*	.35*	.49
SCI-573	.83	.37	.46	.19*
Target				
SCI-96	.71	.41*	.30*	.46
SCI-95	.45*	.30	.15*	.47
SCI-94	.67	.37	.29*	.57
SCI-643	.66	.29	.38*	.51
HS.LS2-2				
Initial				
SCI-501	.81	.06	.76	.21*
Distal				
SCI-80	.72	.32	.40	.21*
SCI-497	.90	.35	.56	.14*
Precursor				
SCI-528	.74	.33	.41	.14*
SCI-660	.88	.38	.50	.22*
SCI-515	.48*	.34	.14*	.47
Target				
SCI-76	.61	.41*	.19*	.53
SCI-550	.63	.26	.37*	.59
SCI-554	.61	.26	.35*	.38

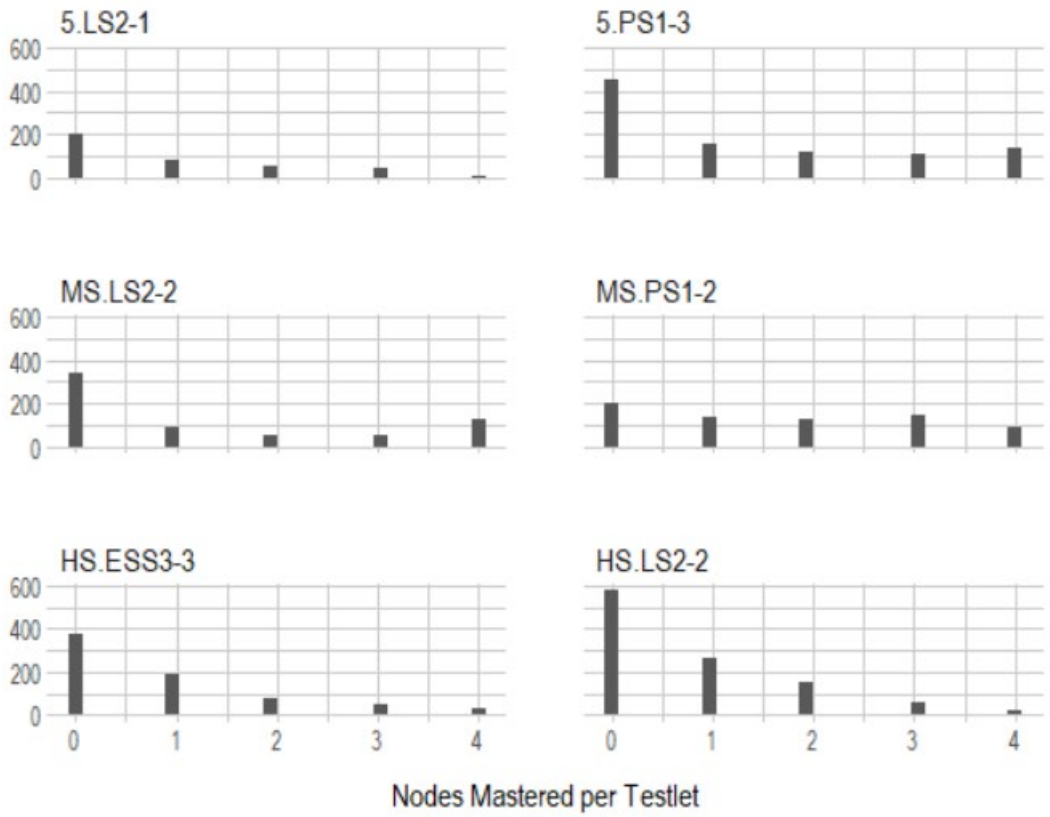
* Indicates the estimate is outside of the ideal range.

Appendix D

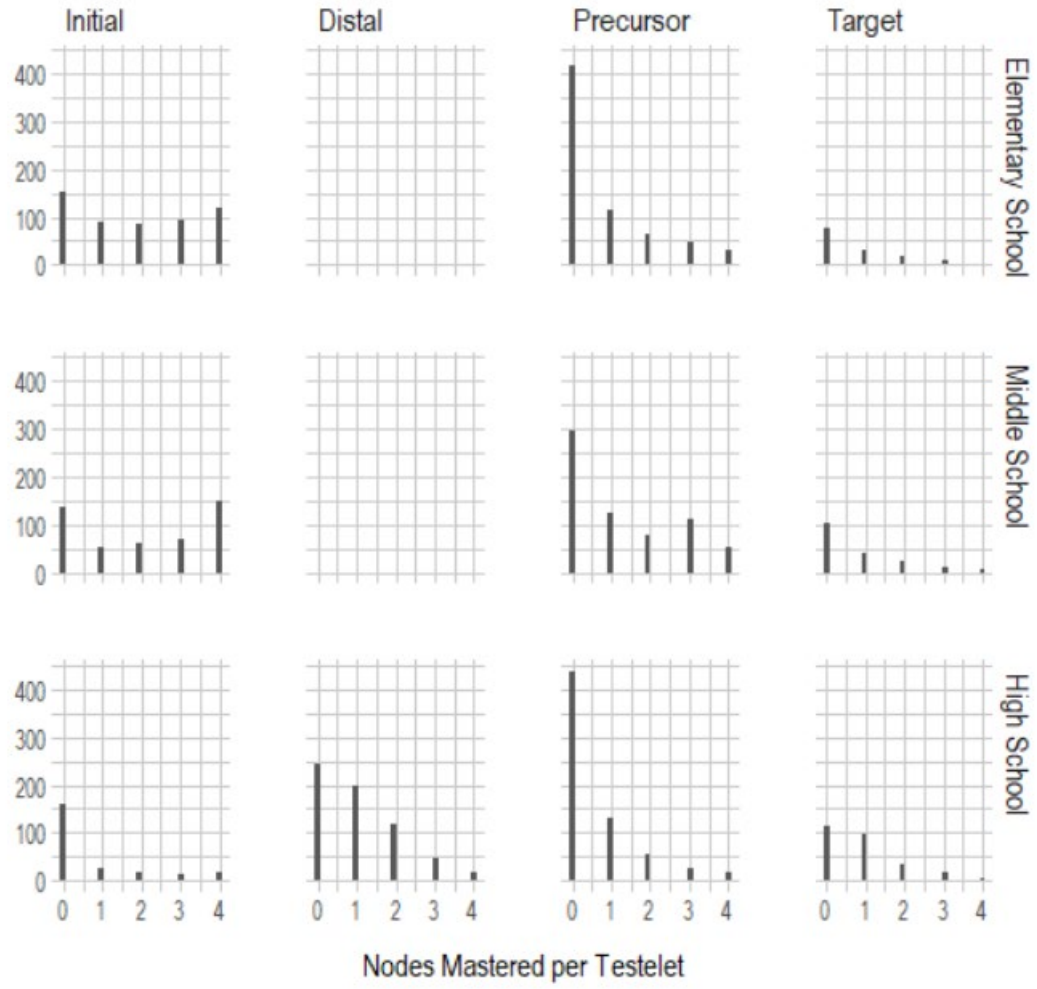
Node Mastery by Grade Band



Node Mastery by Essential Element



Node Mastery by Linkage Level and Grade Band



Node Mastery by Domain and Grade Band