Evaluating the Impact of Multidimensionality on Science Item Performance

Brooke Nash and Sue Bechard

University of Kansas

Nash, B. & Bechard, S. (2018, April). *Evaluating the impact of multidimensionality on science item performance*. Presented at the annual meeting of the American Educational Research Association, New York, NY.

Abstract

New science standards focus on an interwoven approach whereby students engage in science and engineering practices (SEPs) to demonstrate knowledge of disciplinary core ideas (DCIs) and cross-cutting concepts (CCCs). This brings significant challenges to assessment design when considering students with the most significant cognitive disabilities (SCD). A science consortium developed science alternate assessments that combined two dimensions - DCIs and SEPs. This paper presents findings on the relationship between item dimensionality and item performance and patterns of difficulty by SEP for students with SCD. Results suggest that for most grade bands and levels of cognitive complexity, item performance is not negatively impacted by the anticipated complexity of the multiple dimensions. However, there is evidence to suggest that some SEPs are easier than others particularly at the elementary and middle school grade bands.

Evaluating the Impact of Multidimensionality on Science Item Performance

The Framework for K-12 Science Education (Framework) and the Next Generation Science Standards (NGSS; National Research Council, 2012) shifts the focus of science assessment from separate measurement of content and inquiry practices to an interwoven 3-D approach that includes science and engineering practices (SEPs), disciplinary core ideas (DCIs) and cross-cutting concepts (CCCs). Students are expected to engage in science and engineering practices to demonstrate knowledge of science content and concepts. Due to the cognitive and communication characteristics of students with the most significant cognitive disabilities (SCD), alternate assessments (AAs) typically constrain reliance on prior knowledge, abstract thinking, and generalization abilities. Students with SCD who are eligible for AAs comprise about ten percent of the population of students with disabilities, or about one percent of the overall student population. The students in this highly heterogeneous population have a disability or multiple disabilities that significantly impact intellectual functioning and adaptive behaviors and require individualized instruction and substantial supports (Nash, Clark & Karvonen, 2015).

To meet the needs of these students and provide optimum accessibility, the Dynamic Learning Maps (DLM) science assessment system was created using principles of universal design for learning (UDL) at multiple steps of the test development process. UDL principles suggest ways to vary task features (e.g., adjusting difficulty and supports) to support students in communicating what they understand and are able to do. Test design features are carefully controlled to reduce construct-irrelevant variance for these students by employing item writing guidelines, such as using single syllable, decodable words and simple sentences, avoiding commas, negation and pronouns. Technical vocabulary is used only when necessary for the linkage level being assessed (Dynamic Learning Maps Consortium, 2017). In addition, science item writers use the DLM Core Vocabulary list, which is a comprehensive list of words, spanning grades K–12, that reflects the research in core vocabulary in Augmentative and

Alternative Communication (AAC) and words needed to successfully communicate in academic settings where the EEs are being taught (Dennis, Erickson & Hatch, 2013).

Purpose

AAs are challenged to present problem situations that accurately elicit evidence of student mastery on multiple dimensions simultaneously. The purpose of this paper is to describe how the new multidimensional science standards (based on the *Framework* and the NGSS) impact performance of items written to assess those standards. Specifically,

1. Is item dimensionality related to item difficulty?
2. For items that measure a SEP, are there patterns in item difficulty by SEP?

Theoretical Framework

The *Framework* and the NGSS defined significant changes in science education that differed from previous science standards in two ways. First, they presented a focus on progressions of skill development in eight SEPs rather than the more generic inquiry process evident in previous standards. Second, as mentioned, each NGSS performance expectation is expressed as a combination of multiple dimensions. In effect, these changes require students to develop a deep understanding of content knowledge through application of one or more of the practices rather than just knowing science facts (Dynamic Learning Maps Consortium, 2017).

The Dynamic Learning Maps (DLM) alternate assessment system is based on alternate content standards, called Essential Elements (EEs), that are linked to grade-band expectations identified in the *Framework* and the DCI Arrangement of Performance Expectations in the NGSS. The EEs address the three dimensions of the *Framework* (SEPs, DCIs, and CCCs), while the assessments specifically measure the SEPs and DCIs. The goal of the assessment is to provide a means for students with SCD to demonstrate what they know and can do in science in elementary (3rd – 5th), middle school (6th – 8th) and high school (9th – 12th) grade bands. Based on the theory of action that guided the design of the DLM system for science, ultimately the goal is to raise expectations of students with SCD and improve their educational experiences by utilizing valid assessments that are useful for making instructional decisions (see Figure 1).

[Figure 1]

The DLM Essential Elements (EEs) for science are the alternate content standards which are specific statements of knowledge and skills intended build a bridge from the content in the *Framework* to academic expectations for students with SCD. Each EE is expressed at three levels of cognitive complexity known as linkage levels. The Target linkage level is intended to represent the grade-band expectation while the Precursor and Initial linkage levels represent the standard at reduced levels of depth, breadth and complexity. The purpose of the linkage levels is to provide access to the content at varying levels of complexity in order to meet the diverse needs of students with SCD.

DLM science assessments are delivered as a series of testlets (i.e., set of related items), each of which contains three to five selected response items. Assessment items are written to align to one of the three linkage levels for each of the 27 EEs specified on the test blueprints. Figure 2 shows the connections between DLM content standards, EEs, linkage levels and test items. Students take one testlet for each EE at a linkage level that matches the students' skill level. Linkage levels are allowed to vary or adapt between testlets during assessment administration in order to best meet students' needs. Therefore, testlets are delivered using an adaptive algorithm based on students' performance and students do not take a "fixed form" with the same set of testlets.

[Figure 2]

As shown in table 1, EEs across the grade bands cover three science domains: life science, physical science and earth and space science, 10 DCIs and 7 SEPs. Nine EEs are assessed at each grade band, resulting in 27 linkage level testlets per grade band. Each testlet as a whole assesses both the DCI and SEP addressed by the EE linkage level, however every item in the testlet is not multidimensional. Therefore, students were administered an assessment comprised of items that were unidimensional (DCI or SEP only) and multidimensional (DCI and SEP).

[Table 1]

Science testlets all begin with a non-scored engagement activity to increase access for this population by setting the context, activating prior knowledge, and increasing student interest. Initial level testlets contain the least complex contexts, are administered offline by the test administrator, and involve students responding to picture response cards. The engagement activity involves introducing the materials to the student. In general, engagement activities at the Target linkage level provide contexts that are most conducive to including multidimensional items. Some engagement activities are science stories that describe a hypothetical student engaging in a science activity. Testlet developers write science stories such that the student can use the science knowledge that they have been taught to answer questions about concepts that have been broken down into more manageable “chunks.” One effect of this chunking is the inclusion of unidimensional items in testlets to build a logical order of test questions within a testlet.

We expected that multidimensional items would be more difficult. Several sources of evidence contributed to this hypothesis. First, in the 2015-2016 operational assessment, where approximately 20,000 students participated, overall performance was low and the majority of students were categorized as either Emerging or Approaching the Target performance levels (the lowest 2 of 4 performance levels). The percent of students who achieved at the Target or Advanced performance levels was slightly under 20% for all grade bands (Dynamic Learning Maps Consortium, 2017). This pattern was a departure from states’ results based on previous science alternate assessments. When interpreting the results, states hypothesized that multidimensionality was one potential cause of the change in achievement.

An additional source of evidence that supported the expectation that multidimensional items would be more difficult than unidimensional items was a teacher survey administered during the fall 2015 field test on opportunity to learn (OTL) science content and practices. Based on teacher responses considering 837 students, most teachers (ranging from 70% to 87%) spent (or planned to spend) 10 or fewer hours teaching their students various science topics over the course of the 2015-16 school year. Responses to opportunity to engage in science practices was more varied depending on the practice. While 81% of teachers engaged (or planned to engage) their student in the practice of asking question and defining problems, not quite 20% had their student(s) engage in argument from evidence and only 29% had their students constructing explanations and design solutions (Nash & Bechard, 2016). The OTL survey results suggest that there is a significant need for improvement with respect to providing students with SCD access to science curriculum that involves a multidimensional approach to instruction. Given the varied lack of instruction in the SEPs, we anticipated that items that included a SEP would also vary in difficulty.

Methods

It is important to understand the relationship of the dimensions assessed to guide assessment design and inform the interpretation of assessment results. To evaluate how the new multidimensional science standards impact performance of items written to assess those standards, item difficulty is evaluated across the unidimensional and multidimensional items as well as within the various SEPs represented by the items. This study investigates the relationship between item dimensionality, the SEPs and item difficulty across three grade-band Essential Elements at three levels of content complexity.

First, items were coded by two science content experts as either measuring the DCI *or* SEP only (0 = unidimensional) or the DCI *and* SEP (1 = multidimensional). The content experts rated a subset of items independently (10% in each grade band) and then convened to compare and discuss ratings. After this Initial calibration, they rated an additional subset of items independently for a total of approximately 25% of items in each grade band rated by both content experts. Interrater agreement was calculated using Cohen's Kappa to measure the degree to which the two raters agreed beyond what could be expected to occur by chance. As shown in Table 2 below, Kappa values for each grade band were equal to 1.0 with perfect agreement of ratings for the subset of items. The remainder of the item pool was subsequently rated by only one content expert. For each item that measured a SEP (either as a unidimensional or a multidimensional item), a code was also assigned to indicate which SEP it measured based on the SEP of the associated EE (which came directly from the grade-level achievement standard). Tables 3 and 4 display the total number of items rated by dimensionality code and SEP, respectively.

[Table 2]

[Table 3]

[Table 4]

For each research question, item p-values were used as the indicator of item difficulty. To address the first research question, average item p-values were compared across all unidimensional and multidimensional items within each linkage level and for each grade band. Furthermore, a regression analysis was conducted for each grade band and linkage level to determine if item dimensionality was a significant predictor of item difficulty. To address the second research question, a bivariate frequency distribution was used to compare the average p-value for all multidimensional items by each SEP and also by SEP and linkage level for each grade band.

Data Sources

Data for this study came from the 2016-2017 end-of-year spring operational testing window were collected which included students from all three grade bands (elementary, middle and high school) and the full range of students eligible for alternate assessment in their state. The number of students who participated in the 2016-2017 science assessment by grade band is shown in table 5. All states within the DLM science consortium that participated in the spring testing window were included. The proportion of students responding correctly to each item (p-value) was calculated and merged with the item dimensionality and SEP codes.

[Table 5]

Results

The average p-value (weighted by item sample size) by linkage level and item dimensionality for each grade band is provided in Table 6. The results were fairly similar across item dimensionality, linkage levels and grade bands with little variation in p-values. While unidimensional items at the Target level appeared to be slightly more difficult than multidimensional items, the opposite was true for items at the Precursor linkage level. At the Initial level, there was not a consistent pattern across grade bands. In general, items appeared to become easier as linkage levels increased with larger p-values at the Target level than at the Initial level. However, the assessment is adaptive based on student performance which means that students do not necessarily take the same testlets at the same linkage levels; students testing at the Target level may be higher performing than students testing at the Initial level. Therefore, within linkage level comparisons are the most meaningful.

[Table 6]

While the differences in p-values by item dimensionality were small within linkage levels, it was important to evaluate the statistical relationship between item dimensionality and item difficulty at linkage level and grade band. Tables 7-9 display the results of the regression analyses for each linkage level and grade band. In the elementary and middle school grade bands, the non-significant beta coefficients for each linkage level analysis indicates that item dimensionality is not a significant predictor of item difficulty. In high school, item dimensionality is also not a significant predictor of item difficulty at the Precursor or target linkage levels but *is* significant at the Initial level. At the Initial level in high school, the negative Beta coefficient means that the multidimensional items are significantly harder than unidimensional items.

[Table 7]

[Table 8]

[Table 9]

For the items that measured a SEP, again the weighted item p-value was used to compare item difficulty across SEPs at each grade band and for each combination of SEP and linkage level and grade band. As shown in Figure 3, not all SEPs were represented within each grade band. In the elementary grade band, items measuring the practice of *Developing and Using Models* tended to be substantially easier than items measuring the other measured practices. Similarly, in middle school, items measuring the practice of *Constructing Explanations and Designing Solutions* were substantially easier than items measuring the other measured practices. However, in high school, those same two practices also tended to be easier than the others but *Using Mathematics and Computational Thinking* also emerged as a practice where, on average, approximately half of students answer items correctly.

[Figure 3]

Average p-values were also calculated by SEP and linkage level as shown in Figures 4-6. Of the SEPs that were measured in the elementary grade band, items measuring the practice of *Obtaining, Evaluating, and Communicating Information* tended to be more difficult at the Initial and Precursor levels than items measuring other SEPs at those levels. Conversely, items measuring the practice of *Developing and Using Models* tended to be slightly easier than items measuring other SEPs at those levels. In middle school, items that measured *Planning and Carrying out Investigations* at the Initial level and *Analyzing and Interpreting Data* at the target level were the most difficult while items measuring *Developing and Using Models* were the easiest at each linkage level. Conversely, in high school, items that measured *Developing and Using Models* at the Initial level were the most difficult but items measuring this same SEP at the Precursor level were the most difficult compared to items measuring other SEPs at that level. Items measuring the practice of *Engaging in Argument from Evidence* at the target level tended to be the easiest at the high school grade band.

Overall, patterns of item difficulty across SEPs within linkage levels appear to vary within each grade band. Some patterns of SEP difficulty also emerged across linkage levels. For example, in elementary, several practices (e.g., *Analyzing and Interpreting Data* and *Obtaining, Evaluating, and Communicating Information*) appeared to become easier as linkage levels increased from Initial to Target. The same general pattern appeared to be true for most SEPs in middle school but with more variability at the Precursor level. Patterns were less evident in high school. In any case, comparisons of SEPs across linkage levels should be interpreted with caution until the interplay between the DCIs and SEPs is better understood, particularly at different levels of cognitive complexity and across grade bands.

[Figure 4]

[Figure 5]

[Figure 6]

Overall, with the exception of the Initial level in high school, the results of the relationship between item dimensionality and difficulty are somewhat unexpected given the complex nature of the multidimensional science standards that the items were written to. It was hypothesized that the multidimensional items introduced an additional level of cognitive complexity that would result in those items being more difficult for students. It may be the case that the additional context that the multidimensional items provide actually offsets any additional cognitive complexities introduced from measuring both dimensions. Further research is needed at the Initial level in high school to determine how these items are different from items at the same level in other grade bands. On the other hand, differential patterns of difficulty across the SEPs is not surprising given previous survey data that indicated teachers engage their students in certain SEPs more than others.

In any case, there are limitations to using item p-values as the only indicator of item difficulty in this study. While commonly used as an index of item difficulty, p-values are sample dependent and do not allow for an evaluation of item performance independent of the sample on which the statistic is based. Even though the sample sizes in this study were sufficiently large, the results could vary if repeated using other samples of students.

Significance

Results from this study can inform stakeholders, policy makers, and researchers about the how the SEPs and DCIs interact in the context of one alternate assessment system and possible next steps for assessment design for students with SCD. This information will be useful to other states and consortia that are developing alternate assessments based on the new science standards. It is also a topic of interest for states currently developing general education science assessments. As of May 2017, 18 states have adopted the new standards, and new science assessments are slowly being developed for all students (Loewus, 2017). Investigations are needed to help inform the interpretation of assessment results and the design of student performance reports to better support instruction.

While the focus of this paper is on how item dimensionality impacted item performance, specifically in terms of difficulty, it is also important to consider the role of dimensionality on the selection of a measurement model (Martineau, 2017). The current measurement model used to provide student results from the DLM assessments is a diagnostic classification model (DCM). DCMs are confirmatory latent class models that are designed to measure multiple latent variables or attributes (e.g., Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010). However, the latent attributes that are measured in the DLM science assessment are the multidimensional linkage levels. Thus, while the model accounts for multiple discrete skills, further investigation is needed to determine the degree to which the DCIs and SEPs function as multiple dimensions within each linkage level.

References

Dennis, A., Erickson, K., and Hatch, P. (2013). The Dynamic Learning Maps Core Vocabulary: Overview. Retrieved from http://www.med.unc.edu/ahs/clds/files/vocabulary-overview.

Dynamic Learning Maps® Consortium. (2017, June). *2015-2016 Technical Manual – Science*. Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Loewus, L., (May, 2017). Next Generation Science Slowly Takes Shape. *Education Week*. Retrieved from http://www.edweek.org/ew/articles.

Martineau, J. (2017). *The intersection of measurement model, equating, and the Next Generation Science Standards*. Center for Assessment. Retrieved from https://www.nciea.org/library/martineaurils-brief-ngss-measurement-models-and-equating-final.

Nash, B., & Bechard, S. (2016). *Summary of the Science Dynamic Learning Maps® Alternate Assessment Development Process* (Technical Report No. 16-02). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

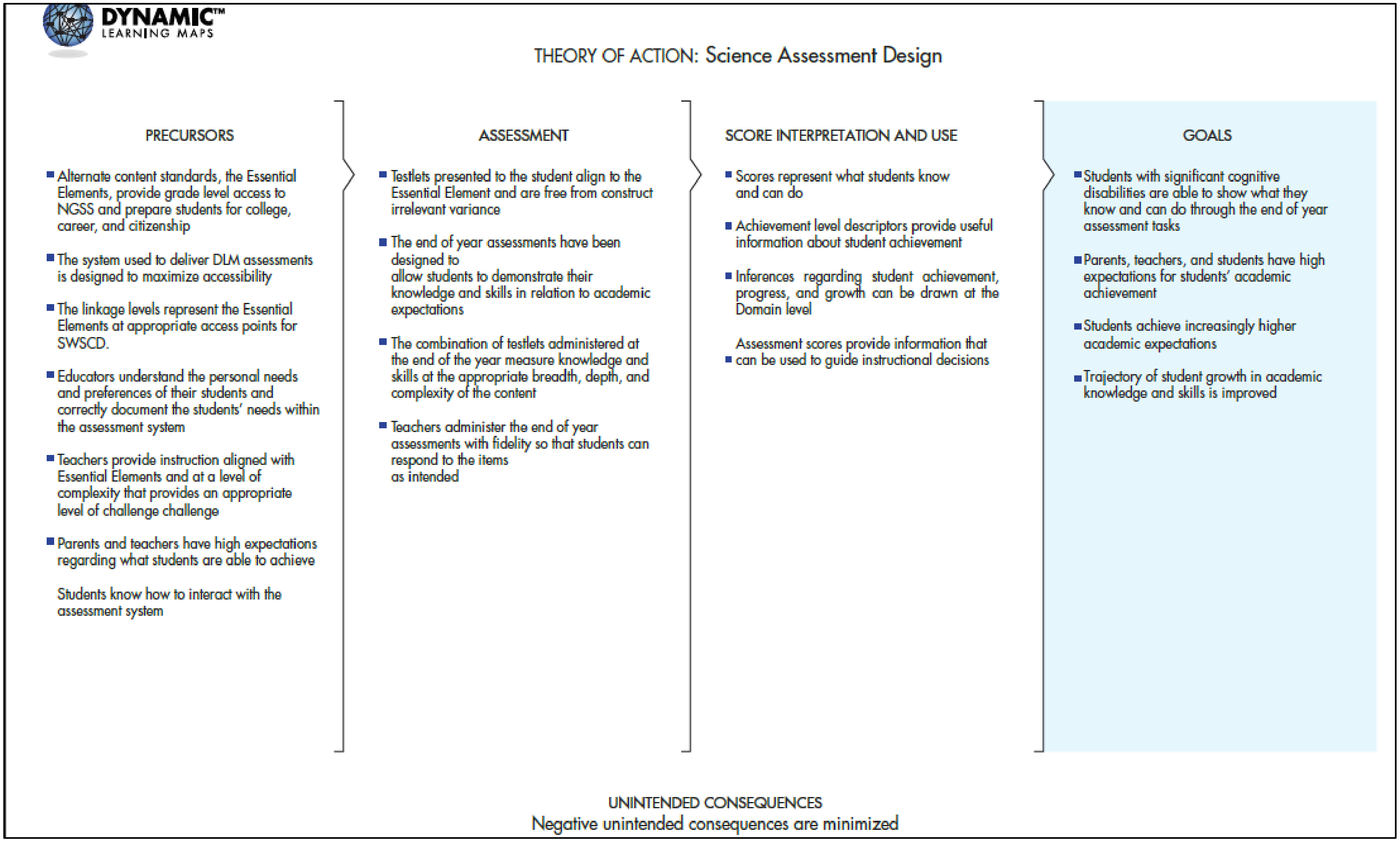
Nash, B., Clark, A. K., & Karvonen, M. (2015). *First contact: A census report on the characteristics of students eligible to take alternate assessments* (Technical Report No. 15-02). Lawrence: University of Kansas, Center for Educational Testing and Evaluation.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press.

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

Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic models: a review of the current state-of-the-art. *Measurement, 6*, 219-262.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York: Guilford.



DYNAMIC™ LEARNING MAPS

THEORY OF ACTION: Science Assessment Design

PRECURSORS

- Alternate content standards, the Essential Elements, provide grade level access to NGSS and prepare students for college, career, and citizenship
- The system used to deliver DLM assessments is designed to maximize accessibility
- The linkage levels represent the Essential Elements at appropriate access points for SWSCD.
- Educators understand the personal needs and preferences of their students and correctly document the students' needs within the assessment system
- Teachers provide instruction aligned with Essential Elements and at a level of complexity that provides an appropriate level of challenge challenge
- Parents and teachers have high expectations regarding what students are able to achieve

Students know how to interact with the assessment system

ASSESSMENT

- Testlets presented to the student align to the Essential Element and are free from construct irrelevant variance
- The end of year assessments have been designed to allow students to demonstrate their knowledge and skills in relation to academic expectations
- The combination of testlets administered at the end of the year measure knowledge and skills at the appropriate breadth, depth, and complexity of the content
- Teachers administer the end of year assessments with fidelity so that students can respond to the items as intended

SCORE INTERPRETATION AND USE

- Scores represent what students know and can do
- Achievement level descriptors provide useful information about student achievement
- Inferences regarding student achievement, progress, and growth can be drawn at the Domain level
- Assessment scores provide information that can be used to guide instructional decisions

GOALS

- Students with significant cognitive disabilities are able to show what they know and can do through the end of year assessment tasks
- Parents, teachers, and students have high expectations for students' academic achievement
- Students achieve increasingly higher academic expectations
- Trajectory of student growth in academic knowledge and skills is improved

UNINTENDED CONSEQUENCES

Negative unintended consequences are minimized

Figure 1. Dynamic Learning Maps theory of action for science.

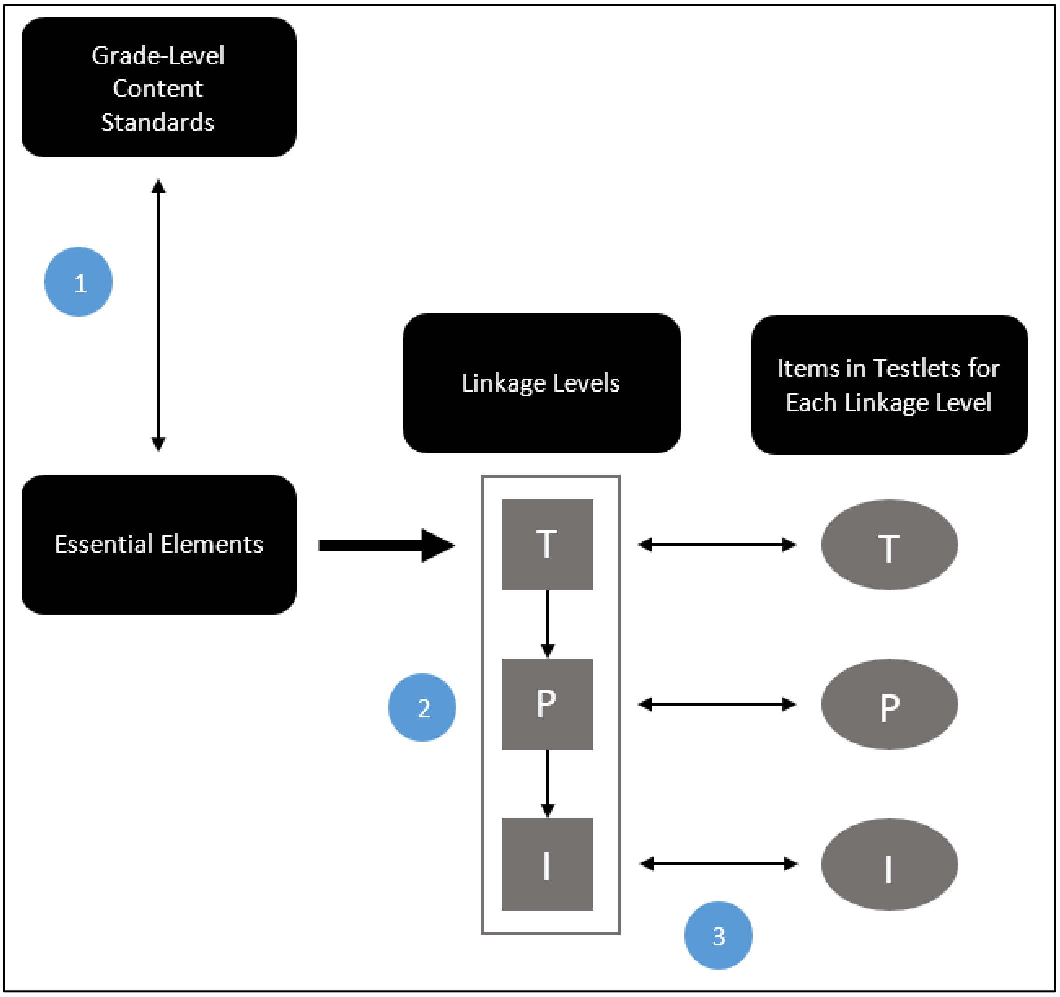


Figure 2. Design of the DLM science assessment. Linkage levels are Target (T), Precursor (P), and Initial (I).

Table 1.

Coverage of DLM EEs by Disciplinary Core Ideas and Science and Engineering Practices

Grade Band	**Disciplinary Core Idea (DCIs)**	**Science and Engineering Practices (SEPs)**
Elementary	Physical Science • PS 1 Matter and It’s Interactions • PS2 Motion and Stability: Forces and Interactions • PS3 Energy Life Science • LS1 From Molecules to Organisms: Structures and Processes • LS2 Ecosystems: Interactions, Energy, and Dynamics Earth and Space Science • ESS1 Earth's Place in the Universe • ESS2 Earth's Systems • ESS3 Earth and Human Activity	1. Developing and using models 2. Planning and carrying out investigations 3. Analyzing and interpreting data 4. Using mathematics and computational thinking 5. Engaging in argument from evidence 6. Obtaining, evaluating, and communicating information
Middle	Physical Science • PS 1 Matter and It’s Interactions • PS2 Motion and Stability: Forces and Interactions • PS3 Energy Life Science • LS1 From Molecules to Organisms: Structures and Processes • LS2 Ecosystems: Interactions, Energy, and Dynamics Earth and Space Science • ESS2 Earth's Systems • ESS3 Earth and Human Activity	1. Developing and using models 2. Planning and carrying out investigations 3. Analyzing and interpreting data 4. Constructing explanations (for science) and designing solutions (for engineering) 5. Engaging in argument from evidence

High	Physical Science • PS 1 Matter and It’s Interactions • PS2 Motion and Stability: Forces and Interactions • PS3 Energy Life Science • LS1 From Molecules to Organisms: Structures and Processes • LS2 Ecosystems: Interactions, Energy, and Dynamics • LS4 Biological Evolution: Unity and Diversity Earth and Space Science • ESS1 Earth's Place in the Universe • ESS3 Earth and Human Activity	1. Developing and using models 2. Planning and carrying out investigations 3. Using mathematics and computational thinking 4. Constructing explanations (for science) and designing solutions (for engineering) 5. Engaging in argument from evidence

Table 2

Interrater Agreement Results

Grade Band	N Items	n Items Rated Twice	Kappa
Elementary	185	48	1.00
Middle	200	54	1.00
High	204	54	1.00

Table 3

Number of Items Rated by Dimensionality

Grade Band	Unidimensional Items	Multidimensional Items
Elementary	48	137
Middle	41	159
High	76	128

Table 4

Number of Items Rated by SEP

Grade Band	1	2	3	4	5	6	7
Elementary	19	0	59	17	20	19	16
Middle	24	80	19	13	0	23	0
High	0	33	25	11	0	24	41

Note. 1 = Analyzing and Interpreting Data; 2 = Constructing Explanations and Designing Solutions; 3 = Developing and Using Models; 4 = Engaging in Argument from Evidence; 5 = Obtaining, Evaluating, and Communicating Information; 6 = Planning and Carrying out Investigations; 7 = Using Mathematics and Computational Thinking

Table 5

Student Participation for the DLM Science 2016-17 Spring Assessment (N=19,686)

Grade Band	n
Elementary	5,821
Middle	6,380
High	7,485

Table 6

Average Item P-value by Item Dimensionality and Linkage Level for each Grade Band

Linkage Level	Elementary	Middle	High
Initial	0.55	0.58	0.55
Unidimensional	0.53	0.58	0.56
Multidimensional	0.56	0.58	0.52
Precursor	0.59	0.61	0.59
Unidimensional	0.61	0.68	0.63
Multidimensional	0.58	0.59	0.58
Target	0.66	0.65	0.63
Unidimensional	0.64	0.43	0.62
Multidimensional	0.67	0.65	0.63

Table 7

Regression Coefficients for Elementary

Linkage Level	Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
Initial	(Constant)	0.576	0.025		23.201	0.000
	Item_Dimensionality	0.015	0.030	0.060	0.505	0.615
Precursor	(Constant)	0.668	0.040		16.544	0.000
	Item_Dimensionality	-0.006	0.045	-0.015	-0.130	0.897
Target	(Constant)	0.587	0.058		10.077	0.000
	Item_Dimensionality	0.034	0.069	0.084	0.490	0.627

Table 8

Regression Coefficients for Middle School

Linkage Level	Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
Initial	(Constant)	0.628	0.026		24.616	0.000
	Item_Dimensionality	0.009	0.033	0.033	0.277	0.782
Precursor	(Constant)	0.693	0.059		11.831	0.000
	Item_Dimensionality	-0.034	0.062	-0.058	-0.548	0.585
Target	(Constant)	0.428	0.178		2.411	0.022
	Item_Dimensionality	0.197	0.180	0.189	1.091	0.283

Table 9

Regression Coefficients for High School

Linkage Level	Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
Initial	(Constant)	0.617	0.016		39.560	0.000
	Item_Dimensionality	-0.102	0.027	-0.401	-3.786	0.000*
Precursor	(Constant)	0.708	0.041		17.352	0.000
	Item_Dimensionality	-0.050	0.046	-0.119	-1.092	0.278
Target	(Constant)	0.617	0.068		9.091	0.000
	Item_Dimensionality	-0.024	0.074	-0.052	-0.327	0.746

* p-value < 0.05

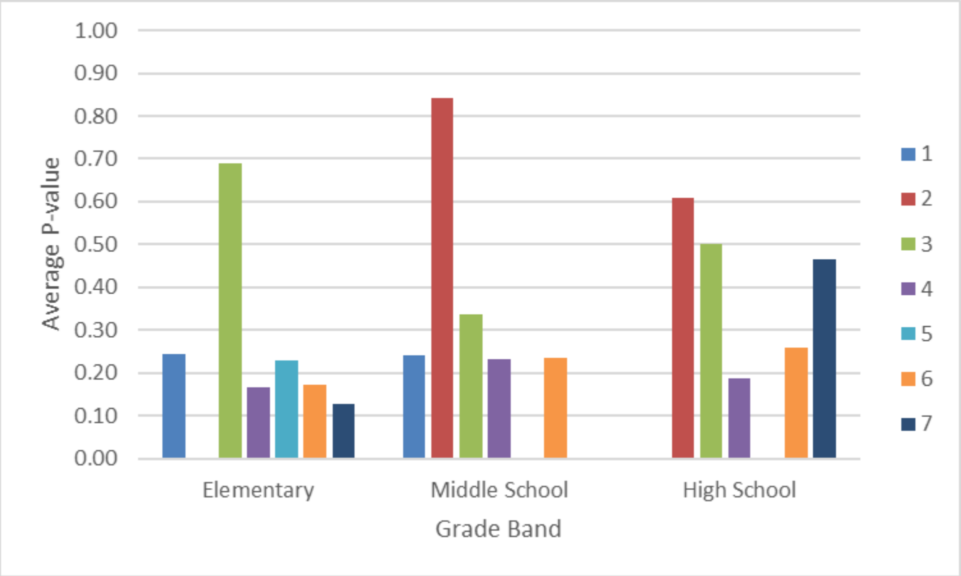


Figure 3. Average item p-value by SEP and grade band. 1 = Analyzing and Interpreting Data; 2 = Constructing Explanations and Designing Solutions; 3 = Developing and Using Models; 4 = Engaging in Argument from Evidence; 5 = Obtaining, Evaluating, and Communicating Information; 6 = Planning and Carrying out Investigations; 7 = Using Mathematics and Computational Thinking

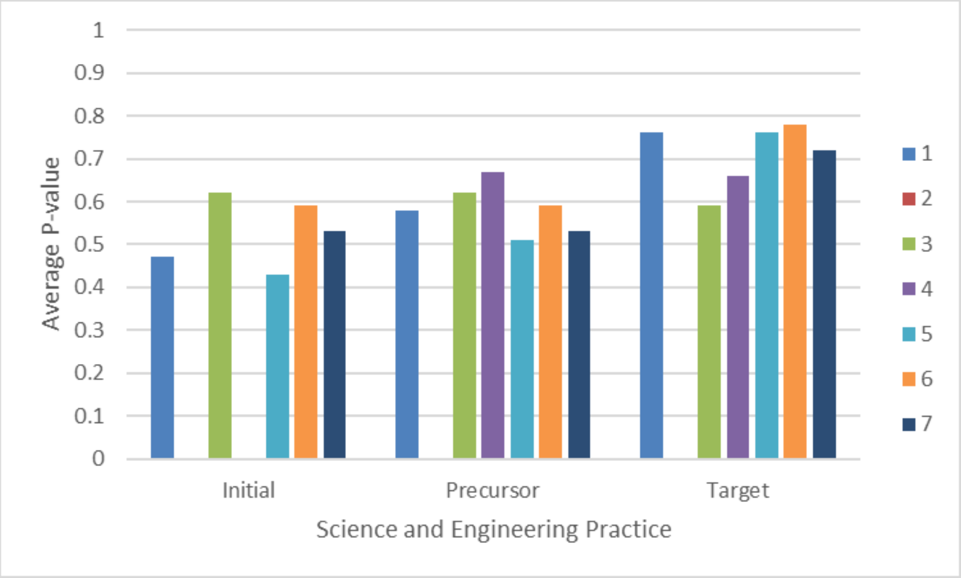


Figure 4. Average item p-value by SEP and linkage level for elementary. 1 = Analyzing and Interpreting Data; 2 = Constructing Explanations and Designing Solutions; 3 = Developing and Using Models; 4 = Engaging in Argument from Evidence; 5 = Obtaining, Evaluating, and Communicating Information; 6 = Planning and Carrying out Investigations; 7 = Using Mathematics and Computational Thinking

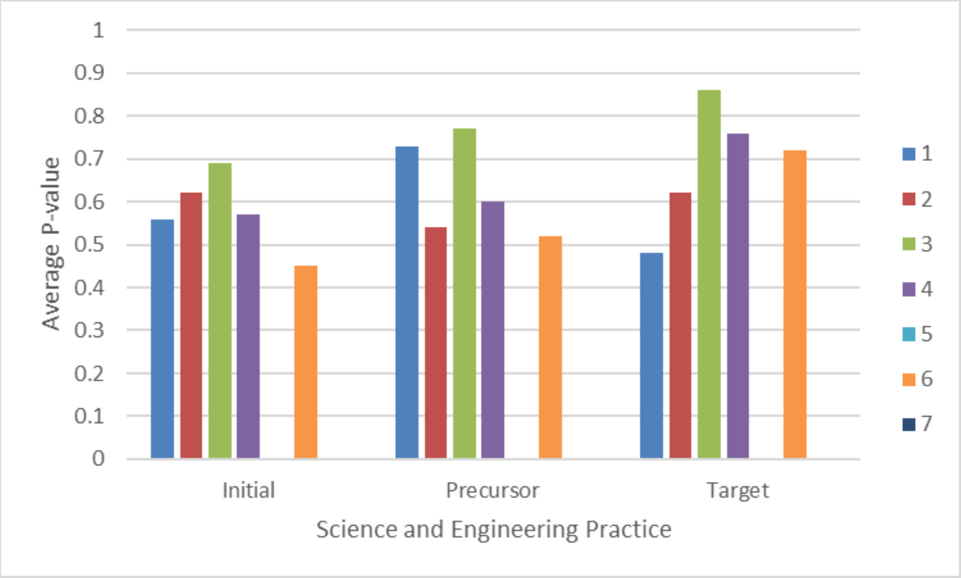


Figure 5. Average item p-value by SEP and linkage level for middle school. 1 = Analyzing and Interpreting Data; 2 = Constructing Explanations and Designing Solutions; 3 = Developing and Using Models; 4 = Engaging in Argument from Evidence; 5 = Obtaining, Evaluating, and Communicating Information; 6 = Planning and Carrying out Investigations; 7 = Using Mathematics and Computational Thinking

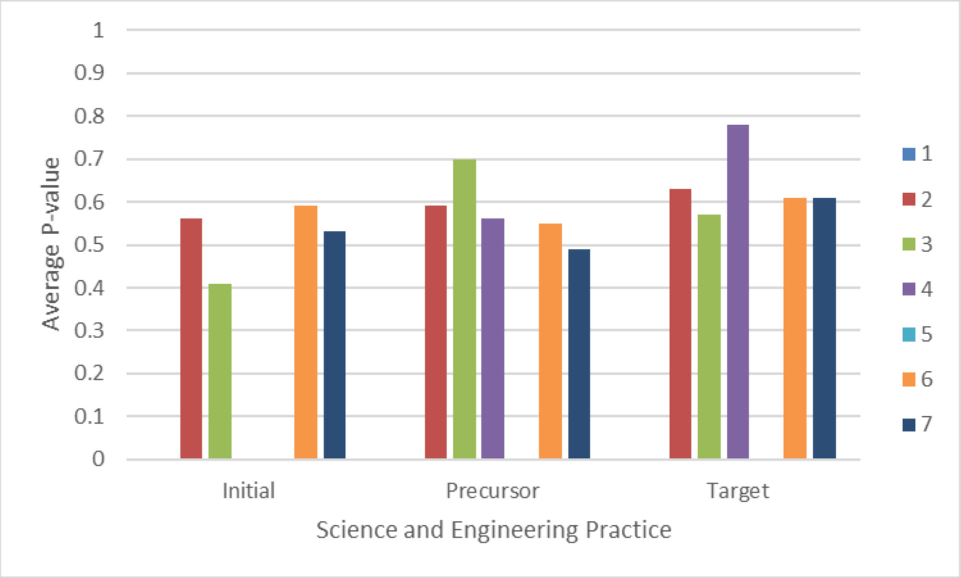


Figure 6. Average item p-value by SEP and linkage level for high school. 1 = Analyzing and Interpreting Data; 2 = Constructing Explanations and Designing Solutions; 3 = Developing and Using Models; 4 = Engaging in Argument from Evidence; 5 = Obtaining, Evaluating, and Communicating Information; 6 = Planning and Carrying out Investigations; 7 = Using Mathematics and Computational Thinking