Investigating the Comparability of Multiple-Choice and Constructed-Response Science Assessments

Cari F. Herrmann-Abell
BSCS Science Learning
Joseph Hardcastle and George E. DeBoer
AAAS Project 2061

Paper presented at the 2019 AERA Annual Conference
Division D: Measurement and Research Methodology
Section 1: Educational Measurement, Psychometrics, and Assessment
Toronto, Canada
April 5-9, 2019

Abstract

The *Next Generation Science Standards* calls for new assessments that measure students' integrated three-dimensional science learning. The National Research Council has suggested that these assessments utilize a combination of item formats including constructed-response and multiple-choice. In this study, students were randomly assigned three-dimensional assessments that contained either constructed-response or multiple-choice versions of items. Rasch analysis was used to compare the difficulty of these items on the same construct scale. We found that partial-credit constructed-response items were located at similar places on the scale as their multiple-choice counterpart but dichotomously-scored constructed-response items were considerably more difficult. This suggests that scoring method influences an item's difficulty, which has implications for what inferences are made from assessments.

1. Objectives or purposes

The *Next Generation Science Standards* (NGSS Lead States, 2013) calls for instruction that fosters an integrated understanding of science and engineering practices, crosscutting concepts, and disciplinary core ideas. Along with this new approach to instruction, new assessments are being called for to assess this vision of integrated, three-dimensional science learning. The National Research Council (NRC, 2014) recommends that assessments be designed to include multiple components to allow students to demonstrate their use of different practices in the context of disciplinary core ideas and crosscutting concepts, provide information that situates students' knowledge on learning progressions, and include tools to help teachers interpret and use students' responses to adapt instruction.

Our project aims to develop three-dimensional assessment tasks aligned to NGSS performance expectations (PEs) related to energy. The tasks consist of both constructed-response items and multiple-choice items. Because certain formats are more appropriate for different conditions, it is important to compare the inferences that can be made from the results of each format. The NRC (2104) speculates that the difference between high-quality multiple-choice items and constructed-response items may not be considerable if the multiple-choice items are used in coherent sets and are developed using a construct-centered approach. This paper reports our initial findings from a study to investigate the difference between using the two formats.

2. Perspective(s) or theoretical framework

Both constructed-response and multiple-choice items have features that warrant their use. The following describes the affordances and disadvantages of each item type.

Multiple-choice items can be written that require sophisticated mental processing and an understanding of complex ideas to answer them correctly. Although multiple-choice items cannot ask students to predict, explain, or design, they can be very effective at asking students to analyze complex relationships and to evaluate predictions, explanations, and designs. In terms of assessing science practices, multiple-choice assessments have been developed to measure students' abilities to identify and critique the evidence provided in scientific arguments (Knight, et al., 2014), and students' ability to read and interpret graphs and evaluate controlled experiments (Quellmalz, et al., 2012). Studies have also shown that when common misconceptions are used as distracters, the diagnostic power is increased (Hamilton et al., 1997; Sadler, 1998). Additionally, multiple-choice items can focus students' attention on a particular aspect of the targeted knowledge and practice, and thus control the response space. Multiple-choice items also require less time for students to answer, making it possible to include more items that can sample a more extensive portion of the targeted construct and result in a more comprehensive evaluation of students' understanding. If well-designed, multiple-choice items can also be more efficiently and reliably scored than constructed-response items. Special challenges include finding ways to reduce the chance that students will effectively guess or use various test-wiseness strategies.

Constructed-response items require students to form their own response, something that multiple-choice items cannot do. This allows for great flexibility in the range of practices these items can target. Through constructed-response items, students can design experiments, formulate their own explanations, and draw their own models. Supporters of constructed-response items argue that this is a more authentic way to assess students' content knowledge and ability to use practices.

As with multiple-choice questions, there are special challenges associated with using constructed-response items that need to be considered. First, because the outcome space is potentially so broad, it is important to be very clear about what the question is asking students to do. And along with that, it is critical that the scoring rubrics give students credit for what they know, even when the students' writing may be imprecise. In addition, the specific response type used may make an item more challenging to some students, for example a student may be able to construct a drawing of their mental model but struggle with writing a scientific explanation. It is also the case that constructed-response items tend to be more memorable to students and therefore make them less suited for being reused, which causes problems when measuring growth.

Regarding the comparability of these formats, a review of 67 studies by Rodriguez (2003) found that when both multiple-choice and constructed-response items used the same stem scores on the items are highly correlated. However, when the stems are different the correlation significantly drops. This points toward construct equivalence being a function of item design and the item writers intended purpose.

3. Methods

We took a construct-centered approach to assessment development, which is summarized below.

Construct Definition

We started by identifying related PEs that progress with increasing sophistication through the grade bands (see Table 1). Then, we wrote statements explicitly indicating what students should and should not know and be able to do. Next, we identified scenarios around which the tasks were designed. Scenarios were selected that are based upon students' everyday experiences and are engaging to a wide range of students. These included bowling, the game of pool, determining safe speed limits, Newton's cradle, among others.

Table 1:
Target Performance Expectations Grouped by Theme

Theme	Performance Expectation	
Transfer of energy by forces and conservation of energy	4-PS3-3	Ask questions and predict outcomes about the changes in energy that occur when objects collide.
	MS-PS3-5	Construct, use, and present arguments to support the claim that when the kinetic energy of an object changes, energy is transferred to or from the object.
	HS-PS3-1	Create a computational model to calculate the change in the energy of one component in a system when the change in energy of the other component(s) and energy flows in and out of the system are known.

Task Development

We developed tasks that are made up of sets of 3-11 items. Some of these items are aligned with one dimension, some with two, and some with three dimensions but when taken together the items provide a complete picture of students' 3D understanding.

To compare the inferences that can be made about students' understanding using the different formats, pairs of tasks were developed. One task includes a multiple-choice version of an item and the other task includes a constructed-response version. Both versions used identical or

similar stems. This paper discusses the results from four pairs. Two pairs compare short-response and multiple-choice versions (Tables 2 and 3) and the other pairs compare create-a-graph and select-a-graph versions (Tables 4 and 5).

Table 2:
Multiple-choice and constructed-response versions of an item from the bowling tasks

Stem

A person who likes to have fun while bowling wants to find out how to make the loudest sound when the bowling ball hits the pins. She also wants to use what she knows about energy to think about how energy is transferred when the bowling ball hits the pins. She and her friend decide to do a little experiment. The first person rolls the ball toward the pins. Her friend measures the speed of the ball and the loudness of the sound. They record the data in the table below. Then she rolls the ball again.

	Speed (miles per hour)	Loudness (decibels)
Try 1	10	80
Try 2	15	

Multiple-choice version

If the speed of the ball during Try 2 is 15 miles per hour, will the loudness of the sound on the second try be louder or softer than 80 decibels? Why?

A. The sound will be louder than 80 decibels. The faster the ball is rolling the more force it has and the more force it can transfer as sound to the pins.
B. The sound will be louder than 80 decibels. The faster the ball is rolling the more energy it has and the more energy it can transfer to the surroundings as sound when it hits the pins.*
C. The sound will be softer than 80 decibels. The faster the ball is rolling the less energy it can transfer to the surroundings as sound when it hits the pins.
D. D. The sound will be 80 decibels on the second try. The speed of the ball will not affect how much energy can be transferred to the surroundings as sound.

Constructed-response version

If the speed of the ball during Try 2 is 15 miles per hour, how loud do you think the sound will be on the second try?

A. The sound will be louder than 80 decibels.
B. The sound will be softer than 80 decibels.
C. The sound will be 80 decibels on the second try.

Use energy ideas to explain your prediction about how loud the sound will be on the second try. Include ideas about the transfer of energy.

Table 3:
Multiple-choice and constructed-response versions of an item from the pool tasks

Stem	Now that the player has thought about the kinds of questions that can be answered scientifically, the player wants to use what he knows about energy to think about what affects the energy of the cue ball and the energy of a ball that is hit. The player puts the cue ball and green ball back where they started and tries again. This time the player hits the cue ball harder and the cue ball rolls faster. When the cue ball hits the green ball, the green ball starts to roll.
Multiple-choice version	How far will the green ball roll when the player hits the cue ball harder? Why? A. The green ball will roll farther because the cue ball will transfer a stronger force to the green ball causing the green ball to roll farther. B. The green ball will roll farther because the cue ball will transfer more energy to the green ball causing the green ball to roll faster.* C. The green ball will not roll as far because the green ball will transfer more energy to the cue ball causing the green ball to roll slower. D. The green ball will roll the same distance because how hard you hit a ball is not related to how fast or far it moves.
Constructed-response version	How far do you think the green ball will roll when the player hits the cue ball harder? A. The green ball will roll farther. B. The green ball will not roll as far. C. The green ball will roll the same distance as before. D. More information is needed to know how far the green ball will roll. Use energy ideas to explain your prediction about how far the green ball will roll on the second try. Include ideas about the transfer of energy.

Table 4:
Multiple-choice and constructed-response versions of an item from the Newton's cradle tasks

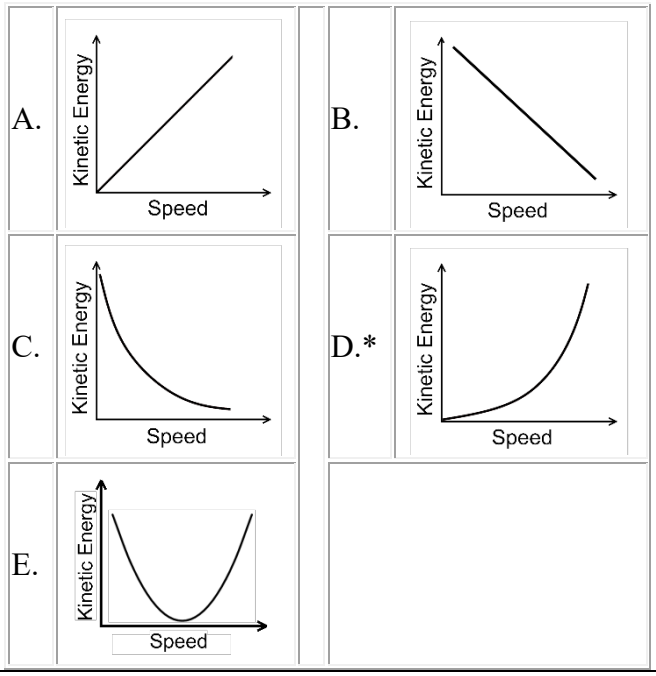
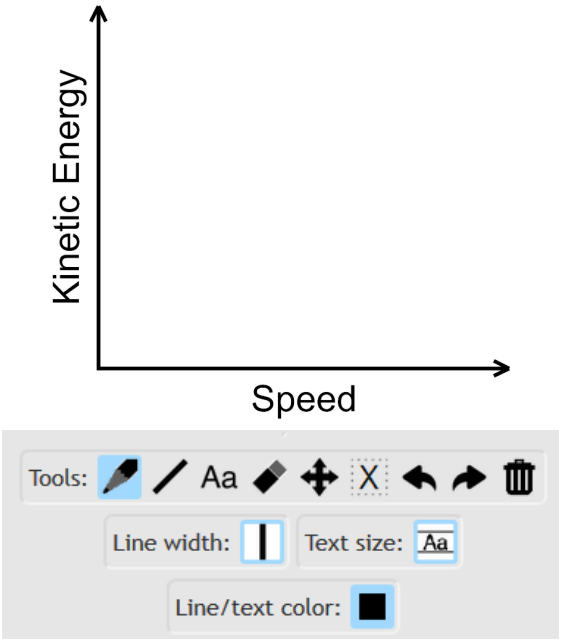
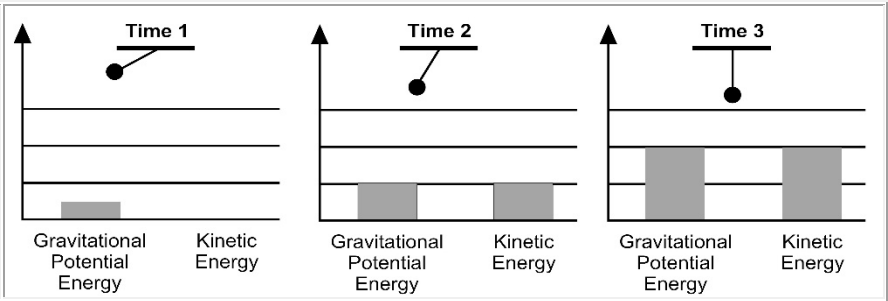
Stem	The student wants to know how the speed of the ball at the end of the pendulum is related to the kinetic energy of the ball.
Multiple-choice version	Which of the following graphs shows the relationship between the kinetic energy of the ball and the speed of the ball as the ball swings? 
Constructed-response version	To help think about that relationship, construct a graph that shows the relationship between the kinetic energy of the ball and the speed of the ball as the ball swings. 

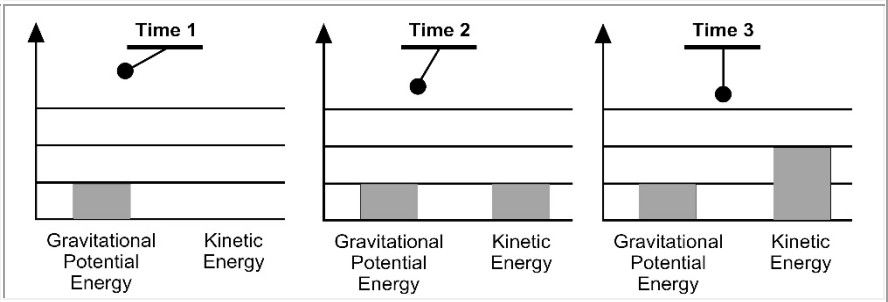
Table 5:
Multiple-choice and constructed-response versions of an item from the Newton's cradle tasks

Stem	The students start by pulling the ball to one side and letting go. They make some observations about the ball and use their observations to think about the energy changes that occur as the ball is swinging back and forth. Below are three graphs that can be used to represent the pendulum-Earth system at three points during the swing.
Multiple-choice version	Which set of bar graphs represents the energy in the pendulum-Earth system as the ball swings side to side?

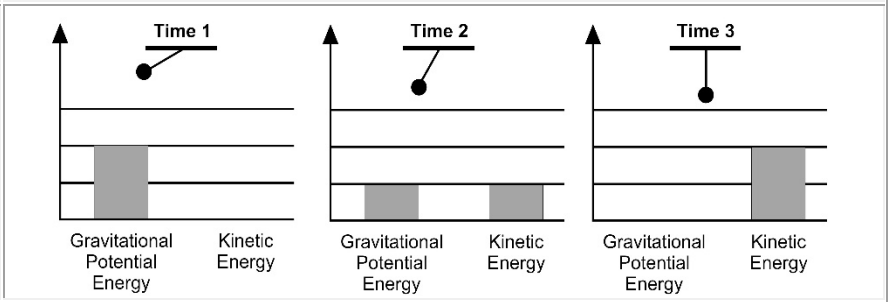
A.



B.



C.*



D.

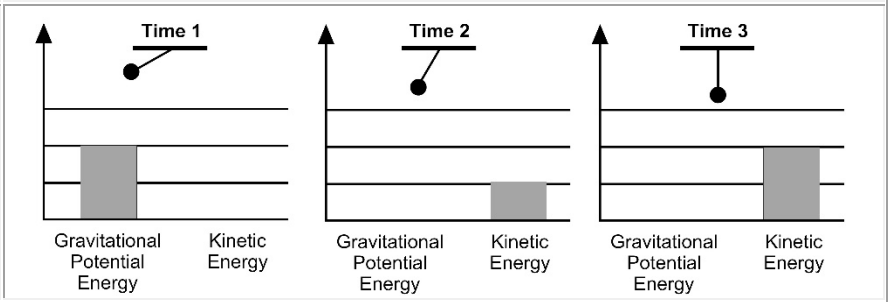
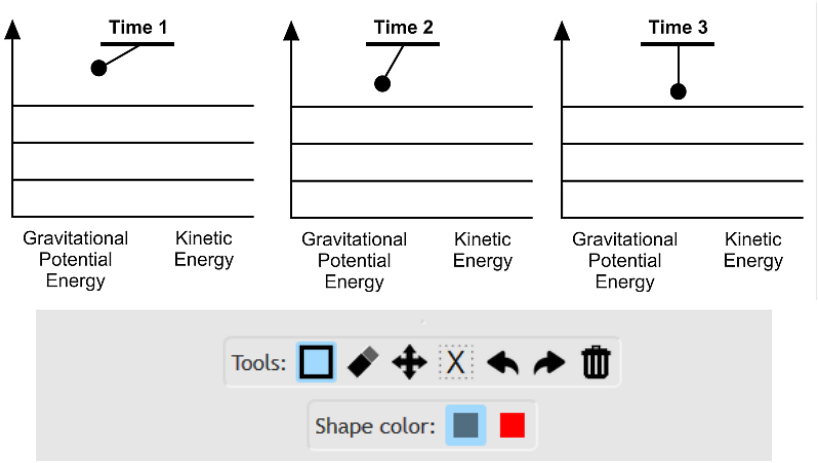


Table 5 continued:
Multiple-choice and constructed-response versions of an item from the Newton's cradle tasks

Constructed-response version	Complete the graphs by adding bars to represent the amount of energy at each time point. 

Defining the outcome space and scoring

In multiple-choice items, the outcome space is defined by the set of answer choices. Our guidelines for item construction ensure that the answer choices for an item are thematically related, and distractors target relevant student misconceptions and difficulties. Multiple-choice items within a task were scored dichotomously, meaning that there are two possible outcomes, right or wrong.

For constructed-response items, we attempted to control the outcome space by using clearly stated questions that target specific aspects of the construct and, when possible, elicit student misconceptions and difficulties. In developing rubrics, we began by creating an ideal response, which, when possible, was based on the correct answer to the multiple-choice version. We then deconstructed the ideal response to its essential elements, which serve as evidence of students' understanding of the content and ability to conduct the practice (see Tables 6 and 7). We also listed indicators of difficulty that provide evidence of misunderstanding. We calculated a score on these items in two ways: dichotomously and polytomously. In the polytomously scored case, the score on the item was determined by adding points for each element present in the response and subtracting points for indicators of difficulty with a minimum score of zero. In the dichotomously scored case, students who included all of the elements were given a score of one, and students who provided some or none of the elements and students who showed indicators of difficulty were given a score of zero.

Both construct-a-graph items were scored dichotomously. For the item in Table 4, students received a score of one if they drew a line graph that indicated that kinetic energy increases with the square of the speed. For the item in Table 5, students received a score of one if they drew bar graphs that indicated that gravitational potential energy decreased from Time 1 to Time 2 to Time 3 while kinetic energy increased.

Table 6:
Rubric for the constructed-response item in the bowling task shown in Table 2

Essential Elements	
1	Student selects A. The sound will be louder than 80 decibels.
2	Student's written response includes the idea that the ball is traveling at a faster speed so it has more energy.
3	Student's written response includes the idea that more energy the ball has, the more energy it can transfer to the surroundings as sound when the ball hits the pins
Indicators of difficulty	
1	Student selects either B or C indicating that they think the sound will be either softer than or the same as the sound in Try 1.
2	Student's response includes the misconception that energy can be created or destroyed.
3	Student's response includes the misconception that a force, not energy, is transferred when the ball hits the pins.
4	Student's response includes the incorrect idea that the ball's speed will not affect how much energy can be transferred as sound.

Table 7:
Rubric for the constructed-response item in the pool task shown in Table 3

Essential Elements	
1	Student selects A. The green ball will roll farther.
2	Student's written response included the idea that the cue ball will transfer more energy to the green ball.
3	Student's written response includes the idea that because more energy was transferred to the green ball, it will roll faster.
Indicators of difficulty	
1	Student selects B, C, or D indicating that they think the green ball will be either not roll as far or will roll the same distance.
2	Student's response includes the misconception that energy can be created or destroyed.
3	Student's response includes the misconception that a force, not energy, is transferred when the ball hits the pins.
4	Student's response includes the incorrect idea that the ball's speed will not affect how much energy can be transferred from the cue ball to the green ball.
5	Student's response included the incorrect idea that potential energy of the balls is changing as they roll across the table.

Comparability study

We investigated the extent to which the targeted construct can be measured using multiple-choice items versus constructed-response items. Students were randomly assigned to respond to either version of the items. Rasch analysis was used to determine if the two formats were assessing the same aspect of the construct. If this were the case, the multiple-choice and constructed-response versions would have the same Rasch difficulty. If the Rasch difficulties are different, we would conclude that the formats are assessing different aspects of the construct.

4. Data sources and Rasch fit

Pilot testing

The six tasks that contained the item pairs we were comparing were pilot tested with 1035 students from across the U.S. (see Table 8). Each pilot test form was made up of one task and 10-12 additional multiple-choice items. These multiple-choice items were selected from an existing item bank that assesses energy disciplinary core ideas (DCIs) and served as linking items. Table 9 summarizes the number of students who responded to each version. On average, 925 students responded to the linking items. The six tasks included seven other multiple-choice items, the data from which were also included in the analyses. These items were answered by an average of 330 students.

Table 8:
Demographic information for students included in the study.

	Percentage of students
Grade band	
Elementary	19%
Middle	40%
High	41%
Gender	
Male	51%
Female	49%
Ethnicity	
White	67%
Black	8%
Hispanic	8%
Asian	6%
Two or more ethnicities	9%
Primary Language	
English	96%
Other	4%

Table 9:
Description of the items being compared and the number of responses per item

Item Description	Multiple-choice Version	Constructed-response Version
Make and justify a prediction about the loudness of the sound made when a bowling ball hits the pins at a faster speed	AP09 N = 159	AP35 N = 165
Make and justify a prediction about how far a billiard ball will roll when the cue ball is hit harder.	AP10 N = 181	AP36 N = 159
Graph the relationship between kinetic energy and speed	AP37 N = 170	AP13 N = 144
Graph the changes in kinetic energy and gravitational potential energy as pendulum swings	AP13 N = 162	AP37 N = 138

Rasch analysis

WINSTEPS (Linacre, 2016) was used to estimate Rasch student and item measures. The measures for the linking items were anchored at their item bank values. Two analyses were performed; one with the polytomous scores and one with the dichotomous scores. The data's fit to the Rasch model was evaluated using the separation indices, infit and outfit mean-squares, standard errors, and point-measure correlations. Most of these values were in the acceptable ranges for both analyses except for the person separation/reliability, which was low (i.e. < 2).

5. Results

Multiple-choice vs. short-response formats

Table 10 compares the difficulties of the versions that asked students to make or identify a prediction and justification. When the constructed-response versions were scored polytomously, there is very little difference in the difficulties between the two formats. This would suggest that both formats are targeting the same aspect of the construct. However, when the constructed-response versions were scored dichotomously, their difficulty level increases an average of 2.6 logits. These results suggest that requiring students to provide their own complete justification is more difficult than asking them to identify a complete justification but allowing some flexibility in the students' justification is on the same level as asking them to identify the complete justification.

Table 10:
Item difficulties of the multiple-choice vs. short-response versions in logits

Scoring Method	Item Description	Multiple-choice Version	Constructed-response Version	Difference
Polytomous scoring	Make and justify a prediction about the loudness of the sound made when a bowling ball hits the pins at a faster speed	0.45	0.61	0.16
	Make and justify a prediction about how far a billiard ball will roll when the cue ball is hit harder.	0.31	0.24	-0.07
Dichotomous scoring	Make and justify a prediction about the loudness of the sound made when a bowling ball hits the pins at a faster speed	0.45	3.26	2.81
	Make and justify a prediction about how far a billiard ball will roll when the cue ball is hit harder.	0.31	2.71	2.40

Additionally, we found that some students didn't use energy ideas in their constructed-response response as requested by the stem. In the pool item, about 30% of the students used force ideas. In the bowling item, 14% used force ideas, and 23% used ideas about the speed of the ball. Another popular strategy for answering the bowling item was using proportional reasoning and ratios (9%). From these responses, we can conclude that energy is not these students' preferred model for thinking about these phenomena. However, if our goal was to determine what the students knew about energy, we did not accomplish this with almost a third of the students

because their responses did not include evidence of what they know about the targeted energy ideas.

Multiple-choice vs. construct-a-graph formats

Both items that required students to draw their own graphs were significantly more difficult than the items that asked students to select a correct graph (See Table 11). Because the constructed-response versions are located at different positions on the scale than the multiple-choice versions, we conclude that these items are measuring different aspects of the construct.

Table 11:
Item difficulties of the multiple-choice vs. construct-a-graph versions in logits

Item Description	Multiple-choice Version	Constructed-response Version	Difference
Graph the relationship between kinetic energy and speed	1.28	5.28	4.00
Graph the changes in kinetic energy and gravitational potential energy as pendulum swings	-0.09	1.36	1.45

6. Significance

This study provides some insights into the comparability of multiple-choice and constructed-response item formats. Our results show that asking students to provide their own answers to the questions may measure a different aspect of the construct than asking students to select from a list of possible answers. Furthermore, our results show that the location of an item on the difficulty scale is dependent on how the item is scored. In the case of the short-response items, the item becomes more difficult with a stricter dichotomous rubric. This finding points to the importance of clearly defining the construct before item and rubric development begins so that item writers are clear on what the expectations are.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A180512 to BSCS Science Learning. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Hamilton L. S., Nussbaum E. M. and Snow R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education*, 10, 169-207.

Knight, A.M., Alves, C.B, Cannady, M.A., McNeill, K.L. & Pearson, P.D. (2014). *Assessing middle school students' abilities to critique scientific evidence.* Paper presented at the annual meeting of NARST, Pittsburg, PA.

Linacre, J. M. (2016). WINSTEPS Rasch measurement computer program. Version 3.92.1. Beaverton, Oregon: Winsteps.com.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards.* Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, J.W. Pellegrino, M.R. Wilson, J.A. Koenig, and A.S. Beatty, *Editors.* Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: Integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching, 49*(3), 363–393.

Sadler, P.M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, 35(3), 265-296.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184.