

Exploring the Comparability of Multiple-Choice and Constructed-Response Versions of Scenario-Based Assessment Tasks

Cari F. Herrmann-Abell and Joseph Hardcastle
BSCS Science Learning

George E. DeBoer
Professor Emeritus, Colgate University

Paper presented at the 2022 NARST Annual International Conference
Vancouver, BC
March 27-30, 2022

Abstract

As implementation of the *Next Generation Science Standards* moves forward, there is a need for new assessments that can measure students' integrated three-dimensional science learning. The National Research Council has suggested that these assessments be multicomponent tasks that utilize a combination of item formats including constructed-response and multiple-choice. However, little guidance has been provided for determining the relative value or cost effectiveness of those two formats. In this study, students were randomly assigned assessment tasks that contained either a constructed-response or a multiple-choice version of an otherwise equivalent item. Rasch analysis was used to compare the difficulty of these items on the same construct scale. We found that the set of items formed a broad unidimensional scale, but the constructed-response versions were more difficult than their multiple-choice counterparts. This difficulty was found to be partially due to the increased writing demand and the reasoning element in the constructed-response rubric. Students were more likely to recognize a clearly reasoned argument in a multiple-choice item than they were to create that reasoning themselves and communicate it in writing. Our findings can help instrument developers select a set of items that balances the time and effort students must provide during testing and the time and effort scorers need to spend to evaluate and score students' responses. In cases where constructing a response is an essential part of the targeted understanding, as when the target learning goal is to be able to construct an argument or generate a model, CR items are needed, but in other cases, MC items may be more efficient.

The *Next Generation Science Standards* (NGSS Lead States, 2013) calls for instruction that fosters an integrated understanding of science and engineering practices (SEP), crosscutting concepts (CCC), and disciplinary core ideas (DCI). This approach to instruction established a need for new assessments that can measure students' integrated, three-dimensional science learning. The National Research Council (NRC, 2014) recommends that assessments be designed 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 that measure students' progress on developing a three-dimensional understanding of the energy concept. Each task consists of a series of items, both multiple choice (MC) and constructed response (CR), that share a common scenario or phenomenon for students to explore. These two item formats, even when testing the same construct, expect different mental processing and kinds of responses from students and thus it is important to formally compare how these item formats perform. The NRC (2014) speculates that the difference between high-quality MC and CR items may not be considerable if the MC items are used in coherent sets and are developed using a construct-centered approach. This paper reports the findings from a study to investigate the difference between using the two formats.

Item formats. MC and CR items both have features that warrant their use in assessment. The following describes the affordances and disadvantages of each item format.

Multiple-choice items. It is often said that MC items test only rote memorization of lower-level facts, but, in fact, MC items can be written that require sophisticated mental processing and an understanding of complex ideas to answer them correctly. While MC items cannot ask students to write, create, or synthesize they can be very effective at asking students to identify, evaluate, or rank. MC items can focus students' attention on a particular aspect of the knowledge and practice being targeted and, in that way, control the response space. They can also be designed to include common misconceptions as distractors, increasing their diagnostic power (Hamilton et al., 1997; Sadler, 1998). In addition, MC items tend to require less time for students to answer and can be more efficiently and reliably scored than CR items. This makes it possible to include more items that can reliably sample a more extensive portion of the targeted construct. These features have made MC items popular; however, they also present challenges including finding ways to reduce the chance that students will effectively guess or use various test-wiseness strategies to eliminate incorrect answer choices.

Constructed-response items. In contrast to MC items, CR items require students to form their own response. This allows for great flexibility in the range of practices these items can target. Through CR items, students can design experiments, formulate their own explanations, and draw their own models. Supporters of CR items argue that this is a more authentic way to assess students' content knowledge and ability to use practices. However, CR items also have challenges associated with using them. First, because the outcome space is potentially so broad, it is important to be clear about what the students are expected to do. That means that it is also important to decide how much emphasis to place on clear writing. If the learning goal is to understand a particular science idea, the scoring rubrics should give students credit for what they know, even when the students' writing may be imprecise. On the other hand, if the learning goal is for students to construct a well-reasoned argument, that is impossible to do without a certain

degree of clarity in their writing. In addition, the specific response type that is asked for may have an effect on the student's ability to respond correctly. For example, a student may be able to construct a drawing of their mental model but struggle with writing an explanation of that mental model. All these factors present challenges in the scoring and interpretation of CR items.

Comparability of multiple-choice and constructed-response formats. A review of 67 studies by Rodriguez (2003) found that when both MC and CR items used the same stem, scores on the items are highly correlated. However, when the stems are different, even when both items are meant to be testing the same idea, the correlation significantly drops. More recently, Morell, et al. (2019) investigated the comparability of forced-choice and constructed-response items that target students' ability to argue from evidence within the disciplinary core ideas of structure of matter and ecosystems and found that the difficulties of some of the item pairs were reasonably close but, on average, the constructed-response versions were more difficult.

Methodology

We took a construct-centered approach to assessment development as recommended by the NRC (2014). This approach draws from frameworks such as Evidence Centered Design (Mislevy, Almond, & Lukas, 2003) and Construct Modeling (Wilson, 2005) as well as our own previous work (DeBoer, et al., 2008). That approach is summarized below.

Construct Definition. We started by identifying thematically-related NGSS performance expectations for the topic of energy that progress through the grade bands. Then we clarified the component dimensions (SEPs, CCCs, and DCIs) by consulting the relevant sections of the NRC *Framework* (NRC, 2012) and the appendixes to NGSS to identify the appropriate level of understanding we could expect for each grade band. Next, we identified scenarios around which the energy tasks were designed. Scenarios were selected that are based upon students' everyday experiences and that should be engaging to a wide range of students.

Task Development. We developed tasks that are made up of sets of 3-11 discrete items. Some of these items are aligned with one NGSS dimension, some with two, and some with three dimensions. When taken together, the items are intended to provide a complete picture of students' three-dimensional understanding. To compare how the MC and CR formats performed, pairs of tasks were developed. One task includes an MC version of an item, and the other task includes a CR version of the same item. Both versions used identical or nearly identical stems. This study includes the results from 19 pairs of items in ten tasks.

MC and CR versions of an item from the bowling tasks. In Table 1, we give an example of one of these 19 pairs of MC and CR versions that use the same stem. The example demonstrates that even though two items may use the same stem, there are inevitably differences in the information students are given and what is expected of them. The MC item includes a claim, relevant evidence to support the claim, and the science idea that energy is transferred. The student is asked to recognize the correct statement. The CR item asks student to pay attention to what they observed and to use those observations along with a relevant science idea to construct an explanation for why the ball slowed down. Although the item is heavily scaffolded, the item does not tell students what the critical observations are (e.g., that the sound is a critical observation), and the word "evidence" does not appear in the CR stem, although it does appear in the MC answer choices.

Table 1: *MC and CR versions of an item from the bowling tasks & the rubric for the CR version*

Multiple-choice version	<p>The friends notice that the ball slows down after it hits the pin. Which of the following explains why the ball slows down after it hits the pins?</p> <p>A. The ball slows down because it has less energy after it hits the pin. Energy is moved from the ball to the pin and the air when the ball hits the pin. The increase in motion of the pin and the sound are evidence that energy was moved.</p> <p>B. The ball slows down because it has less energy after it hits the pin. Energy is moved only from the ball to the air when the ball hits the pin. The sound is evidence that energy was given to the air. The motion of the pin is not related to energy.</p> <p>C. The ball slows down because it has less energy after it hits the pin. Energy is moved only from the ball to the pin when the ball hits it. The motion of the pin is evidence that energy was given to the pin. The sound is not related to energy.</p> <p>D. The ball slows down because it has less force after it hits the pin. A force, not energy, is moved from the ball to the pin when the ball hits the pin. This force is changed into energy. The increase in motion of the pin and the sound are evidence that the force was changed into energy.</p>
Constructed-response version	<p>The friends notice that the ball slows down after it hits the pin. Use energy ideas to explain why the ball slows down after it hits the pin. Be sure to write about the observations and include ideas about how energy can move from place to place.</p>
Rubric for the Constructed-response version	
Ideal response	<p>The ball slows down because it has less energy after it hits the pin. Energy is moved from the ball to the pin and the air when the ball hits the pin. The increase in motion of the pin and the sound are evidence that energy was moved.</p>
Student makes a claim	<p>The ball slows down because it has less energy after hitting the pin <i>or</i> because it transfers energy to the pin and/or the air during the collision.</p>
Student cites evidence	<ul style="list-style-type: none"> • The pin starts moving (falls down) after it was hit. • A sound was heard when the ball hit the pin.
Student either states or uses a science idea (See bullet 1 for an example of using a science idea.)	<ul style="list-style-type: none"> • The faster/slower an object is moving, the more/less energy it has. (i.e. The ball is moving slower so it has less energy.) • When objects collide, energy can be transferred from one object to another. • Sound results from the transfer of energy to the surroundings during a collision.
Student uses reasoning	<ul style="list-style-type: none"> • The ball transferred energy to the pin and air as indicated by the increased speed of the pin and the sound heard during the collision, which means the ball has less energy and will therefore slow down.

Defining the outcome space and scoring. In MC items, the outcome space is restricted by a set of answer choices. Our guidelines for item construction ensure that the answer choices are thematically related, that the distractors are plausible, and that they target relevant student alternative mental models and preconceptions. MC items within a task were scored dichotomously, either right or wrong.

For CR items, we constrained the outcome space by using clearly stated questions that target specific aspects of the construct and by providing appropriate scaffolding. In developing rubrics, we created an ideal response, which was based on the correct answer to the MC version. Then we identified “elements” that we looked for in student responses including recognition of a pattern, inclusion of relevant components in a model, identification of evidence, use of a science idea, or use of reasoning to connect the evidence to the science idea. For tasks that involve the practice of scientific explanation, the elements typically cluster into the following categories: claim, evidence, science ideas, and reasoning. The score on the item is based on how many of these categories the student covers in their response. Table 1 above shows the scoring rubric for the CR version of an item from the bowling task.

Reliability of CR items. To evaluate the scoring reliability of CR rubric elements, a randomly selected set of thirty responses were scored by two researchers and the percentage match and Cohen’s kappa were calculated for each rubric element. The percentage match was uniformly high, with greater than 90% agreement between the two reviewers on most scoring elements. In addition, an acceptable kappa reliability (> 0.70) was achieved for most rubric elements. When the kappa reliability was found to be below 0.70, it was for elements on which very few students received points (Byrt, Bishop, & Carlin, 1993). All scoring mismatches were reviewed by the researchers so that a final decision on scoring could be made and scoring guidance could be written to ensure consistent scoring of the remaining responses.

Field test design. Data for this study were collected during field testing of the tasks. Field test forms were made up of three three-dimensional tasks and 15 content-focused, multiple-choice items. The content-focused, multiple-choice items were drawn from an existing item bank that assesses energy disciplinary core ideas (Herrmann-Abell & DeBoer, 2018) and served as linking items. Students were given one class period to complete the test. On average, 200 students responded to each version of the three-dimensional tasks. Only students who completed at least one of the three tasks and five of the 15 MC items were included.

Participants. Students in the classes of 49 teachers participated in the field test during the spring of 2021. The data used in this study includes the responses from the 1268 students who responded to at least five content-focused, multiple-choice items and one three-dimensional tasks. Table 2 shows a summary of the demographic information for the students in this data set.

Table 2: Summary of Demographic Information

	Percentage of Sample		Percentage of Sample
Grade Band		Race/Ethnicity	
Elementary	18%	American Indian	2%
Middle	43%	Asian	8%
High	38%	Black	9%
Gender		Hispanic	14%
Female	51%	White	58%
Male	49%	Other	5%
Primary Language		Two or more	5%
English	94%		
Other	6%		

Rasch analysis. Partial credit Rasch analysis was used to investigate the extent to which the construct can be measured using the set of MC items and CR items that were developed for this study: (1) Winsteps (Linacre, 2022) was used to estimate person and item measures, which are reported in logits. A logit of zero represents the average item difficulty, values greater than zero are more difficult, and values less than zero are less difficult. (2) A principal component analysis of the Rasch residuals was conducted to investigate whether the items were unidimensional, which would suggest the items were all measuring the same construct. (3) A Wright map was constructed to investigate the comparability of the MC and CR versions. We assumed that if the MC and CR versions were measuring the same aspect of the construct, the two versions would have the same Rasch difficulty and be located at the same place on the map. If the versions were separated on the map, we would conclude that the formats were assessing different aspects of the construct.

Results & Discussion

Fit statistics. The data was fit to the Rasch model with person and item reliabilities of .74 and .98, respectively. The average person measure was small and close to zero (-0.24), indicating that the items were well matched to the students. The average person measure increases from -.71 for elementary school students to -.25 for middle school students to -.01 for high school students. Table 3 summarizes the Rasch item fit statistics. Based on these fit statistics, we conclude that the data have an adequate fit to the Rasch model.

Table 3: Rasch Item Fit Statistics

	Minimum	Maximum	Median
Standard error	0.05	0.47	0.14
Infit mean-square	0.65	1.48	0.94
Outfit mean-square	0.44	1.80	0.91
Separation index		7.59 (0.98)	

Dimensionality. A principal component analysis of the Rasch residuals was conducted to examine the dimensionality of the data. Ideally, the first component should be less than 2, which would be considered at the random noise level. However, components less than 3 are generally considered small and indicate the set of items is largely unidimensional but measuring a “broad”

dimension (Wilson, 1994). We found the first component of the correlation matrix of the residuals to be 2.3. Therefore, we conclude that the items that make up the three-dimensional tasks and the content-focused items make up a unidimensional scale. Additionally, this suggests that the MC and CR versions are measuring the same general construct.

Wright map. Figure 1 is a Wright map that shows the items and students on a single scale ranging from easier items and lower performing students (persons) at the bottom, and harder items and higher performing students (persons) at the top. Items shaded dark gray are the MC versions and items shaded light gray are the CR versions. The map shows that for almost all the pairs there is a difference in difficulty, and, for the most part, the MC versions are located below the average item difficulty and the CR versions are located above.

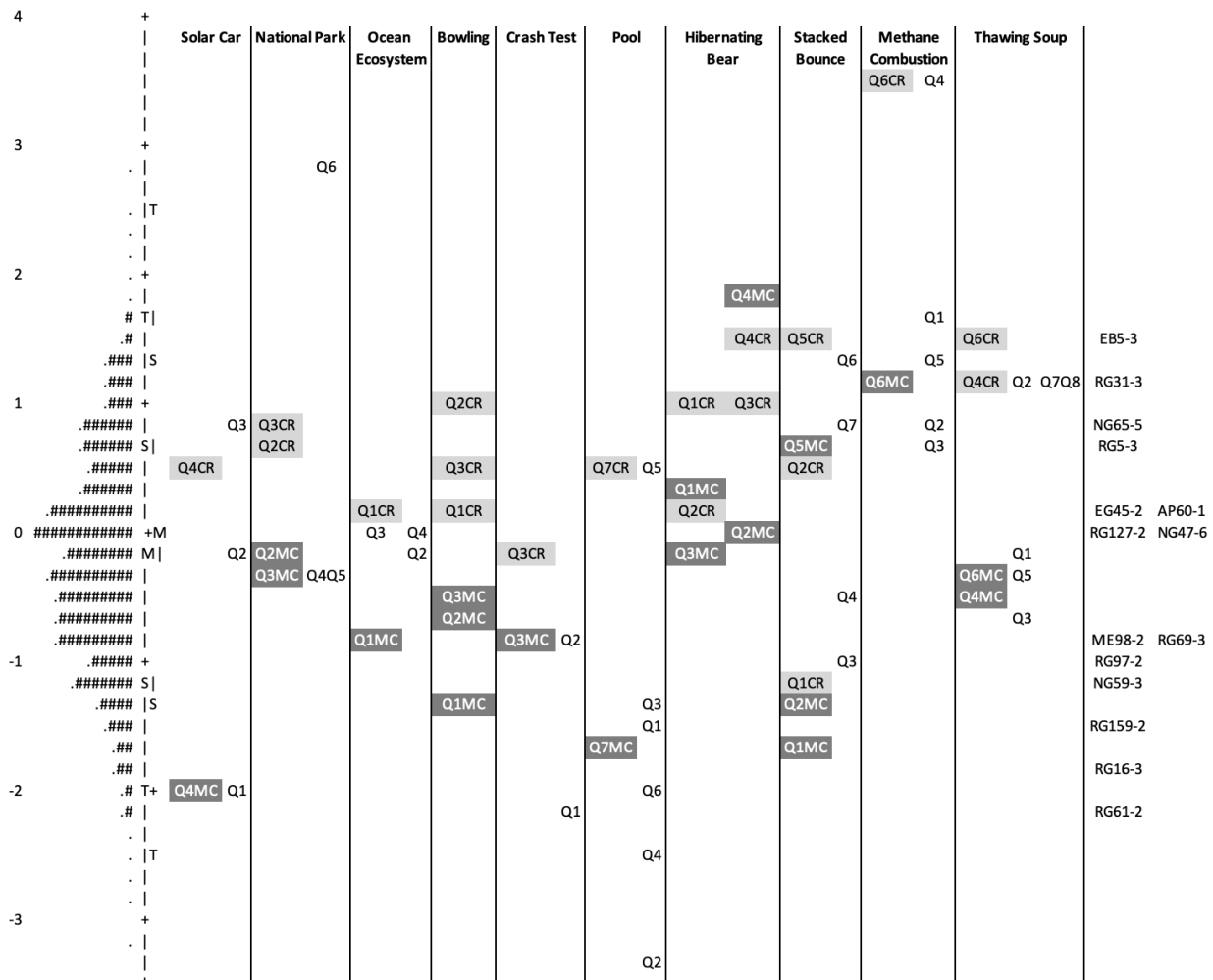


Figure 1. Wright map. Each # is 9 students and each “.” is 1 to 8 students.

Table 3 shows the difficulties of the MC and CR versions. The average difficulty for the MC versions was -0.44 logits and the average difficulty for the CR versions was 0.80. For all pairs except one, the MC version was easier than the CR version. The average difference in difficulty was 1.24 logits. This large difference in difficulty supports the idea that the MC and CR versions are assessing different aspects of the construct being measured by the items.

Table 3: *Item difficulties of the MC vs. CR versions in logits*

Context	Item Description	MC	CR	Difference
Bowling	Make and justify a prediction about the loudness of the sound made when a bowling ball hits the pins at a faster speed	-1.29	0.13	1.42
	Explain the pattern observed in data about the speeds of the ball and pin	-0.73	1.03	1.76
	Explain why a ball slows down after it hits a pin	-0.57	0.51	1.08
Stacked bounce	Analyze data to identify a pattern in the relative bounce heights of two balls	-1.71	-1.20	0.51
	Analyze data to identify a pattern in how the final height compares to the initial height	-1.29	0.53	1.82
	Identify data that support an explanation & describe how the data support the explanation	0.73	1.49	0.76
Ocean	Create a food web model for an ocean ecosystem	-0.82	0.20	1.02
Crash Test	Make a claim about how much energy a car has after crashing and cite evidence	-0.79	-0.11	0.68
Solar Car	Create a model of how energy is transferred between the sun, solar panel, and electric motor	-1.93	0.50	2.43
Pool	Make a prediction about how much energy a ball will have after hitting another ball and explain the prediction	-1.71	0.55	2.26
Methane Combustion	Explain why an input of energy was needed to start an exothermic reaction	1.19	3.54	2.35
National Park	Make a prediction about what will happen to an animal's sources of matter and energy when an organism is removed from the ecosystem	-0.18	0.74	0.92
	Make a prediction about what will happen to an animal's sources of matter and energy when an organism is removed from the ecosystem	-0.30	0.77	1.07
Hibernating Bear	Describe how oxygen a bear breathes in used by the bear to get energy	0.26	1.01	0.75
	Make a claim about where the carbon atoms in the carbon dioxide a bear breathes out comes from	-0.01	0.17	0.18
	Explain how a bear gets energy needed to stay alive during hibernation	-0.21	1.07	1.28
	Explain what caused a bear to decrease in mass during hibernation	1.79	1.58	-0.21
Thawing Soup	Make a claim about how energy is transferred among water, soup, and air in a metal pot	-0.51	1.18	1.69
	Make a claim about how energy is transferred among water, soup, and air in a Styrofoam pot	-0.35	1.42	1.77

Items involving the explanation practice. Table 4 shows some sample responses for the CR version of the bowling item presented in Table 1 and the category scores that the responses were assigned. Our past research on using the scoring rubric for the CR explanation items sheds some light on what makes the CR versions of explanation items more difficult (Hardcastle, Herrmann-Abell, & DeBoer, 2021). When we separated the categories of the rubrics and scored responses dichotomously based on whether the response included an element from the category, we saw a hierarchy of difficulty in the rubric categories. In this hierarchy, the science ideas and reasoning categories were more difficult than the claim category, which indicates that writing well-reasoned statements based on science ideas is a very challenging task for students. This increased writing demand is not required in a multiple-choice setting where students need only to recognize a well-reasoned statement.

Table 4: *Sample student responses to the explanation item in Table 1*

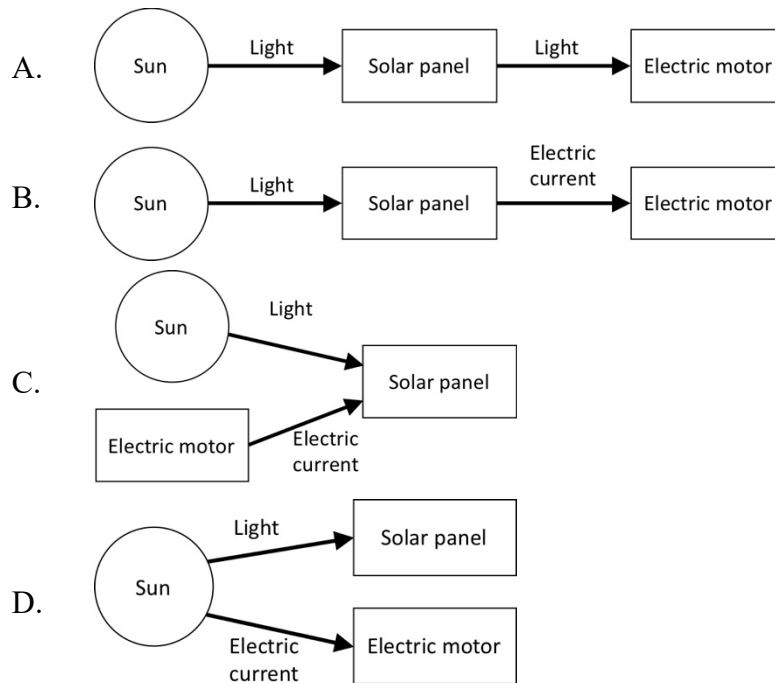
CR prompt	Claim	Evidence	Science Idea	Reasoning
The friends notice that the ball slows down after it hits the pin. Use energy ideas to explain why the ball slows down after it hits the pin. Be sure to write about the observations and include ideas about how energy can move from place to place.				
Student response	Claim	Evidence	Science Idea	Reasoning
Well the ball goes fast when its rolling and when the ball hits the pins I think they have a big impact and the ball goes slower.	0	0	0	0
Energy is transferred from the ball to the pin so the ball doesn't have as much energy as it did before so it will slow down.	1	0	1	1
The ball slowed down after it hit the pin because the energy that the ball had transferred to the pins. You can see this happen because after the pin got hit it was able to move.	1	1	1	1

The one exception to this pattern of difficulty is the Hibernating Bear item that asked students to explain what causes a bear to decrease in mass during hibernation. The MC version of this item was 0.21 logits more difficult than the CR version. One distractor in the MC version was selected by 50% of the students. This distractor stated that the matter the bear is made up of is turned into energy during weight loss. Many students likely chose this distractor because they don't understand where the energy the bear uses to stay alive comes from and thought it was plausible that the atoms inside the bear "turn into" energy. Fewer students (6%) wrote explanations including this alternative mental model on the CR version. Although we do not know for sure what "turned into" in the MC item means to students, we do know that it was a very popular answer choice.

Items involving the modeling practice. Two items required students to construct models (CR version) or identify models that describe phenomena (MC version). One asked students to construct a model of a food web in an ocean ecosystem, and the other asked students to construct a model that described how energy was transferred between the components of a solar car system. Table 5 shows the MC and CR versions of the solar car item and the rubric for the CR version. Table 6 includes sample student responses to the CR version.

Table 5: MC and CR versions of an item from the solar car tasks & the rubric and sample student responses for the CR version

Multiple-choice version Which of the following models best represents how energy was transferred between the sun, solar panel, and electric motor when light was allowed to shine on the car? Arrows in the models show the direction of energy transfer and the labels on the arrows show how energy moved between the electric motor, solar panel, and sun.



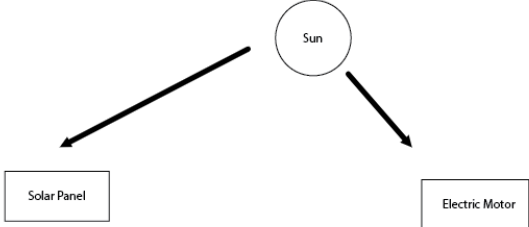
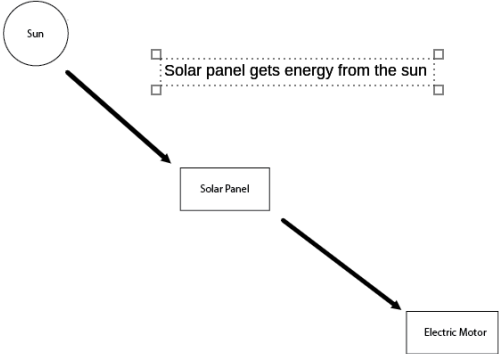
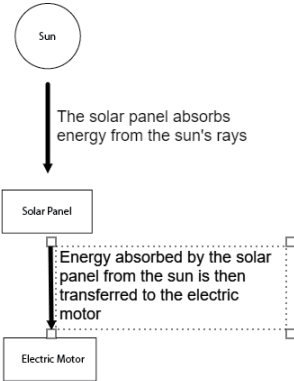
Constructed-response version Use the drawing tools below to create a model that shows how energy was transferred between the sun, solar panel, and electric motor when light was allowed to shine on the car. Your model should include all of the following:

1. The shapes in the drawing toolbox labeled Electric Motor, Solar Panel, and Sun
2. Arrows to show the flow of energy
3. Labels on the arrows to show how energy moved between the electric motor, solar panel, and sun.

Rubric for the Constructed-response version

Student includes the correct components	<ul style="list-style-type: none"> • Sun • Solar panel • Motor
Student includes the correct interactions	<ul style="list-style-type: none"> • Arrow from sun to solar panel • Arrow from solar panel to motor
Student labels the arrows using a general science idea	<ul style="list-style-type: none"> • Energy can be moved place to place by light (i.e. label on arrow between sun and solar panel says “light”). • Energy can be moved from place to place through electric currents (i.e. label on arrow between solar panel and motor says “current”).

Table 6: Sample responses to the CR version of the modeling item from the Solar Car tasks shown in Table 5

Sample student responses	Components	Interactions	Science ideas
	1	0	0
	1	1	0
	1	1	1

The CR versions of these items were over 2 logits more difficult than the MC items that required the students to select a correct model. The most common error that the students made when creating the food web diagram is that they reversed the direction of the arrows even though the stem told them to draw the arrows from the animal being eaten to the animal doing the eating. The most common error that the student made when creating the model of the solar car system is not indicating the mechanism of energy transfer even though the stem told them to include labels on the arrows to show how energy moved between the electric motor, solar panel, and sun. These errors were not included as distractor options in the MC versions of these items, and this factor is likely the primary cause of the difficulty difference between the MC and CR versions.

We found a similar phenomenon for other CR versions where some students many have not been reading the stem carefully and did not completely answer the question being asked. This may have been due to the fact that this was a “no-stakes” test given during a non-traditional, post-COVID school year. Students may have been rushing and not taking the time to read the whole question. For example, the item in Table 1 asks students to use energy ideas in their explanation but, despite the additional scaffolding reminding them, many students did not even mention energy in their responses. These students did not receive points because their response was not in the correct response space. This was not an issue with the MC items where the response space is defined by the answer choices. Therefore, the difficulty difference we observed between the two versions may be inflated because of differences between the MC and CR items even when the stems were the same.

Item involving making a claim about the source of carbon atoms in carbon dioxide exhaled by an organism. The item with the smallest difference in difficulty between the MC and CR versions was the item asking students where the carbon atoms in the carbon dioxide a bear breathes out comes from. The response space in both versions was fairly confined to naming the source molecule (i.e., the carbon-containing molecules from food the bear ate), and there was no requirement to communicate the reasoning behind the answer or to write a coherent sentence on the CR version. Therefore, it is not surprising that the difficulties of the CR and MC versions were very similar.

Conclusions

This study provides insights into the comparability of MC and CR item formats. Our results show that the MC and CR versions of items meant to test the same construct have different difficulties, suggesting that they are measuring different aspects of the construct being targeted. Based on this result, MC and CR items should not be used interchangeably, even when the same stem is used. The difference in difficulty may be due, in part, to the challenge for students of constructing well-reasoned statements versus the ease of recognizing one.

An analysis of students’ written responses revealed that some students might not have been reading the questions carefully, leading them to make incorrect assumptions about what was expected of them and write incorrect responses. CR items require students to pay close attention to the information given in the stem so that they include the correct information in their response (e.g., direction of arrows, required aspects of a model) whereas the answer choices in MC items tend to include all of the required information lowering the cognitive load.

Implications for instrument development. In addition to considering the difference in difficulty between the MC and CR items and the ways in which items that appear to be similar are actually measuring different aspects of a construct, item developers also need to consider the practical aspects of developing and using MC and CR items. When constructing instruments, assessment developers should weigh the importance of having students construct their own answer choice and the time-consuming task of scoring those constructed responses. This involves balancing the number of open-ended, human-scored items and multiple-choice, computer-scored items. Our findings can help inform efficient and effective instrument development by providing information about the comparability of MC and CR versions. We found that, overall, the different versions are measuring the same broad construct, but the CR versions are consistently more difficult than the MC versions. This indicates that they are

measuring different aspects of this construct. We attribute the additional difficulty to the CR version's requirement for students to communicate their reasoning in writing. This ability to communicate coherently in writing is an important part of several science and engineering practices, like constructing explanations and arguing from evidence. Therefore, items that require scientific explanations of phenomena may be best asked in constructed-response format to fully assess students' ability to construct an explanation.

In other instances, this skill may not be an essential part of the targeted understanding, and a MC item may be the more effective format to use. For example, the Crash Test task included an item asking students to cite evidence to support a claim about whether a car will have more, less, or the same amount of energy after crashing than when moving toward the wall. The goal of the item is to assess whether students can identify the relevant evidence, not whether they can write a full explanation. The difference between the difficulties of CR and MC versions of this item was 0.68 logits. Given the relatively small difference between the difficulties and the goal of the item to assess students' ability to identify evidence, the MC version is likely the most efficient version to use. Another example would be the question asking where the carbon atoms in carbon dioxide come from. This item may be more efficiently formatted as multiple-choice because there are a finite number of sources to choose from, and the target understanding being assessed with this question is whether the students know which is the correct source and not how well they can communicate their choice in writing.

Significance. These assessments and field test results will be informative to NARST members interested in assessing three-dimensional science understanding. As we navigate the context of NGSS, assessment plays a key role in helping to build the evidentiary basis for the efficacy of new instructional materials. Given that there are affordances and disadvantages of both CR and MC item formats, developing a better understanding of how the item formats compare will inform the development of more effective and efficient assessments.

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

- Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence, and kappa. *Journal of Clinical Epidemiology*, 46(5), 423–429.
- DeBoer, G. E., Herrmann-Abell, C. F., Gogos, A., Michiels, A., Regan, T., & Wilson, P. (2008a). Assessment linked to science learning goals: Probing student thinking through assessment. In J. Coffey, R. Douglas, & C. Stearns (Eds.), *Assessing student learning: Perspectives from research and practice* (pp. 231-252). Arlington, VA: NSTA Press.
- Hamilton, L. S., Nussbaum, E. M. and Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education*, 10, 169-207.
- Hardcastle, J.M., Herrmann-Abell, C.F., & DeBoer, G.E. (2021, April). Validating a Claim-Evidence-Science Idea-Reasoning (CESR) Framework for use in NGSS assessment Tasks.

- Paper presented at the NARST 2021 Annual Conference.* Online. Retrieved from <https://eric.ed.gov/?id=ED612227>.
- Herrmann Abell, C.F. & DeBoer, G.E. (2018). Investigating a Learning Progression for Energy Ideas from Upper Elementary Through High School. *Journal of Research in Science Teaching*, 55(1), 68-93.
- Linacre, J. M. (2022). WINSTEPS Rasch measurement computer program. Version 5.2.1. Beaverton, Oregon: Winsteps.com.
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report 03-16). Princeton, NJ: Educational Testing Service.
- Mislevy, R. J., Haertel, G., Riconscente, M., Rutstein, D.W. & Ziker, C. (2017). *Assessing Model-Based Reasoning using Evidence-Centered Design: A Suite of Research-Based Design Patterns*. Springer. 10.1007/978-3-319-52246-3.
- Morell, L., Suksiri, W., Dozier, S., Osborne, J., & Wilson, M. (2019). *Addressing the Next Generation Science Standards (NGSS) practice of arguing from evidence using forced-choice item formats: challenges & successes*. Paper presented at the Institute of Education Sciences: Annual Principal Investigators Meeting, Washington, DC.
- National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press.
- 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.
- 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.
- Wilson, M. (1994). *Objective Measurement: Theory into Practice*. Norwood NJ.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.