Developing NGSS-Aligned Tasks to Assess Elementary School Students' Ability to Explain Energy-Related Phenomena

Cari F. Herrmann-Abell

BSCS Science Learning

Joseph Hardcastle and George E. DeBoer

AAAS Project 2061

Presented at the 2020 AERA Annual Conference

Division H: Research, Evaluation, and Assessment in Schools; Section 3: Assessment in Schools

Abstract

We have developed assessment tasks aligned to NGSS that require students to use practices along with disciplinary core ideas to make sense of energy-related phenomena. In this paper, we present an analysis of field test data and feedback from expert reviewers on the validity and reliability of a set of elementary school tasks. These tasks focused on assessing students' ability to write explanations or arguments about energy-related phenomena. Field test data were scored using rubrics based on the claim, evidence, reasoning (CER) framework. Using Rasch modeling, we evaluated the reliability of the task's rubric categories. We found that rubric categories fit well to the Rasch model. Categories were found to cluster in a hierarchy of difficulty in which reasoning and applying science idea categories were more difficult than evidence, which were more difficult than claim. The observed hierarchy in difficulty of CER categories is consistent with other studies and validates the tasks as measures of the CER framework. In addition, a panel of experts agreed that the tasks were aligned to the targeted NGSS practices and ideas. Overall, our results show that our procedure for task development resulted in valid and reliable NGSS-aligned assessment tasks.

1. Objectives or purposes

The *Next Generation Science Standards* (NGSS Lead States, 2013) calls for instruction that fosters an integrated understanding of multiple dimensions of science, including what NGSS describes as: (1) science and engineering practices (SEPs), (2) crosscutting concepts (CCCs), and (3) disciplinary core ideas (DCIs). 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 allow students to demonstrate the use of science and engineering 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.

To address this need, we developed a set of assessment tasks aligned to NGSS that require students to use practices along with disciplinary core ideas and crosscutting concepts to make sense of energy-related phenomena. The tasks present a phenomenon or scenario followed by a series of constructed-response and multiple-choice items.

In this paper, we outline the procedure used for developing these tasks and report on field test results from a set of elementary school tasks in which students were presented a scenario that required them to use relevant energy concepts and the science practices of constructing explanations or writing arguments.

2. Perspective(s) or theoretical framework

As noted above, the NRC, in their report on developing assessments for NGSS, states that assessing students' three-dimensional science understanding will require "assessment tasks that examine students' performance of scientific and engineering practices in the context of crosscutting concepts and disciplinary core ideas" (NRC, 2014). Additionally, in order to sufficiently cover all three dimensions, the NRC recommends the use of sets of interrelated items where the individual items may target individual core ideas, practices, or crosscutting concepts but when taken as a whole provide a complete picture of students' three-dimensional science understanding.

In an effort to better conceptualize what these three-dimensional assessments would look like, Achieve, Inc. recently completed the Task Annotation Project in Science (Achieve, 2019b). In this project they collected tasks that were designed to be aligned to NGSS and convened a panel of experts to annotate the tasks. The project resulted in a framework that the panel thought all NGSS-aligned assessments should possess. The framework included: (1) a focus on real-world phenomena, (2) requiring students to engage in sense making, (3) requiring students to use both disciplinary core ideas and science practices, (4) being comprehensible to students, and (5) supporting the intended purpose and use of the assessment. This framework was used when constructing tasks for this project.

3. Method

The NRC (NRC, 2014) also suggests that three-dimensional assessments should be developed using a construct-centered approach, such as Evidence-Centered Design (Mislevy, Almond, & Lukas, 2003) or Construct Modeling (Wilson, 2004). Our development procedure followed a construct-centered approach and is briefly described below.

Construct Definition and Task Development

Assessment development started by selecting a set of thematically related NGSS performance expectations (PEs) that progress with increasing sophistication through the grade bands. The full set of tasks targets three energy themes: (1) transfer of energy by forces and conservation of energy, (2) thermal energy transfer and dissipation, and (3) energy and chemical reactions. Each of the PEs includes one of each of the three dimensions of science, i.e., a disciplinary core idea, a science practice, and a crosscutting concept. These three dimensions were further clarified 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, in this case, late elementary school. An example progression of PEs for the transfer of energy by forces and conservation of energy theme is shown in Table 1.

Table 1:
Example of targeted Performance Expectations for the Transfer of Energy by Forces 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.

After a progression of PEs was identified, we searched for phenomena and scenarios that required students to engage with the targeted set of DCIs, SEPs, and CCCs. Phenomena and scenarios were selected with the goal that they would be familiar and engaging to a wide range of students. These included bowling, the game of pool, and invasive species, among others.

Multiple-choice and constructed-response items were then created for each task that would move students through a process of phenomenon/problem introduction, sense-making, and final resolution. Some items were aligned with one dimension, some with two, and some with three dimensions. When taken together, the items were intended to provide a comprehensive picture of students' 3D understanding.

Defining the outcome space and scoring

Rubrics and scoring guidelines for tasks were done at the item level. For multiple-choice items, the outcome space was defined by the set of answer choices. Our guidelines for item construction ensured that all the answer choices for an item were thematically related to the question being

asked and that distractors targeted relevant student misconceptions and difficulties. Multiple-choice items within a task were scored dichotomously, either as right or wrong.

For constructed-response items, we controlled the outcome space by using clearly stated questions that target specific aspects of the construct and elicit student misconceptions and difficulties. We began rubric development for constructed-response items by first creating an ideal response. We then identified statements in the response that indicated use of the targeted DCIs, SEPs, and CCCs. These might include mentioning a trend in a data table, stating a critical science idea, or connecting a piece of evidence to a claim. These statements we called the rubric's "elements."

We then grouped the elements into "categories" that represented the types of features we were looking for. For example, a rubric for an item in which students were expected to draw a model might include rubric elements such as including appropriate components of the phenomenon, drawing arrows between those components, or grouping components into systems. These specific elements could then be grouped into two categories called "model components" and "interactions between components." The categories inform scorers about the types of features to look for in student responses, and the individual elements provide examples for each category.

In the argumentation and explanation tasks discussed in this paper, the categories followed a modified Claim, Evidence, Reasoning(CER) framework (McNeill & Krajcik, 2011). In addition to claims, evidence, and reasoning, we included a category called "States or Uses Science Ideas" in our rubrics. This category was included to identify when students stated or used a relevant science idea or principle in their explanation or argument. We considered "reasoning" to be the coherent linking of those science ideas to the claims and evidence statements. We differentiated "States or Uses Science Ideas" from reasoning because we found that students' explanations and arguments sometimes stated a science principle relevant to the question, but they did not use the principle to reason their way from evidence to claim in a coherent way. This approach is in contrast to other explanation and argumentation rubrics that focus on segmenting students' reasoning into weak and strong levels of reasoning (Hu Jin, Yan, Mehl, Llort, & Cui, 2020; Mcneill & Krajcik, 2008; Yao & Guo, 2018) or distinguishing logical versus scientific reasoning (Osborne et al., 2016). In summary, our explanation and argumentation rubrics include (1) a claim that answers the question, (2) evidence that supports the claim, (3) a statement or use of relevant science principles, and (4) reasoning based on science principles that coherently links the evidence to the claim. An example rubric using these categories is shown in Table 2.

To score the constructed-response items, scorers first identified which of the individual rubric elements were in a student's response. Dichotomous category-level scores were then generated from the elements present in the responses. Students were given a point for the category if their response included at least one of the elements from that category and zero if it did not include any elements from that category. For example, to get a point in the evidence category, a student's response would have to cite at least one of the evidence elements listed under the category (i.e. the ball hits the pin, the pin starts moving, or a sound was heard).

Table 2:
Rubric for a constructed-response item that is part of a task dealing with the game of bowling. The item asks students to explain why a bowling ball slows down after it hits a pin.

Prompt	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 data collected in both investigations and include ideas about how energy can move from place to place.
Ideal response	The ball slows down because when the ball hits the pin, energy is transferred from the ball to the pin and from the ball to the air, which means that the ball has less energy after hitting the pin. I know energy has been transferred from the ball to the pin and air because the pin starts to move when the ball hits it and because a sound is heard when the ball hits the pin. Both the increase in motion of the pins and the sound indicate an increase in energy, which had to come from somewhere else (i.e. the ball).
Category	**Individual Elements**
Student lists *evidence*	The ball hits the pin/there was a collision. The pin starts moving (falls down) after it was hit. A sound was heard when the ball hit the pin.
Students either state or use a general *science principle*	The slower an object moves, the less energy it has (i.e. the slower the ball moves the less energy it has). [*Student links speed and energy*.] Energy is transferred from one object to another when they collide [resulting in a change in motion] (i.e. the ball transfers energy to the pin when it hits the pin). [*Student links collisions and energy transfer*.] During a collision, some energy is transferred to the air and sound is produced (i.e. the sound heard when the ball hit the pin results from the transfer of energy from the ball to the pin). [*Student links sound and energy transfer*.]
Students use *reasoning* to link the evidence and science principle	If the ball transfers energy to the pin and air when it collides with the pin, it has less energy and will therefore slow down.

Expert Review

To evaluate the alignment of the tasks to the targeted DCIs, SEPs, and CCCs, tasks were sent to a panel consisting of science education and content knowledge experts with experience in the crafting and implementation of NGSS. Each reviewer was sent a set of ten tasks, and each task was evaluated by two reviewers. Reviewers were first asked to complete the task. Then they were asked to complete a survey based on criteria developed by Achieve as part of the Task Annotation Project in Science (Achieve, 2019b, 2019a). The survey asked reviewers to evaluate:

(1) the appropriateness of the task phenomenon/scenario, (2) the alignment of the task to the targeted SEPs, CCCs, and DCIs, and (3) the fairness and comprehensibility of the task. In addition, panel members were asked to evaluate the rubrics for the constructed-response items by considering the appropriateness of the ideal response for students who had mastered the relevant NGSS learning goals, the internal consistency of the rubric elements and categories, and whether enough detail was provided for raters to identify key elements in students' responses.

4. Data sources

Field testing

A random sample of 50 elementary, 100 middle school, and 100 high school students were drawn from participants who took part in a larger NGSS-aligned assessment development project. Their responses to four elementary school tasks in which students had to write explanations and arguments using ideas about energy were used in this study.

The larger set of tasks from which our sample was randomly drawn included responses from over 13,000 students in grades 4 through 12 from across the U.S in the spring of 2019. Table 3 shows a summary of the demographic information for the total data set.

Table 3: *Summary of Demographic Information*

	Percentage of Sample
Grade Band	
Elementary	6%
Middle	58%
High	27%
College	2%
Gender	
Female	47%
Male	45%
Race/Ethnicity	
American Indian	2%
Asian	7%
Black	10%
Hispanic	12%
Pacific Island	1%
White	51%
Other	10%
Primary Language	
English	86%
Other	7%

Each student responded to a test composed of two tasks and 12 additional DCI-focused multiple-choice items. These multiple-choice items were selected from a previously developed item bank that assesses energy DCIs across the grade bands and served as linking items.

5. Data Analysis

Scoring

As described earlier, students' written responses were scored based on the inclusion of any element under each category. In order to compare the difficulty of the categories from different tasks, we treated them as dichotomous items.

To evaluate the reliability of scoring, a subset of fifty student responses for each task were scored by two scorers. Scorers achieved an acceptable kappa reliability (>0.70) for most rubric elements. For some elements, only a few students received a point. A small number of rater disagreements for these elements produced an interrater reliability below the threshold. Scorers met to review the scoring of these rubric elements and found that their scoring matched a large percentage of the time (>90% matching). In the end, all scoring mismatches were reviewed and discussed by the scorers so a final decision on scoring could be made before moving onto Rasch analysis.

Rasch analysis

We used Rasch analysis to estimate item and student measures and investigate the relative difficulty of the claim, evidence, science idea, and reasoning categories. The Rasch analysis was conducted using the software WINSTEPS (Linacre, 2018). Each category from an item within a task was treated as a separate item for the Rasch analysis. The final data set for all the tasks included 7 claim items, 4 evidence items, 8 science idea items, 8 reasoning items, and 36 DCI-focused multiple-choice items that were used to link the tasks. The multiple-choice items were anchored at their previously determined item bank values. Difficulty values were reported in logits, with zero representing the average item difficulty, values greater than zero being more difficult, and values less than zero, less difficult.

Items that had poor fit to the Rasch model were examined to determine if there was anything in the item that could be responsible for the misfit. Based on that analysis, we eliminated one MC item. To decrease the influence of guessing on the MC item measures we used an approach outlined by Andrich *et al.* (Andrich, Marais, & Humphry, 2012) in which responses with large z-residual values are treated as missing data. For multiple choice items with a relatively large mean-square outfit statistic, we removed student responses with z-residuals greater than 3, which resulted in 32 responses being removed. These student responses were removed because they fell far outside the expected range for the student, for example a student with a low person measure responding correctly to a very difficult item.

Wright maps (Wilson & Draney, 2002) were used to compare student measures and item difficulties. On a Wright map, the distribution of student measures in logits appears vertically from lowest to highest, and next to it the distribution of item difficulties vertically from easiest to hardest.

6. Results

Rasch Analysis

Table 5 summarizes the fit statistics for all four tasks after anchoring the multiple-choice items. The item separation indexes, which indicate the number of levels into which items were high, indicating a wide range of item difficulties.

Table 5: *Summary of Rasch Fit Statistics*

	Item		
	Min	Max	Median
Standard error	0.1	0.25	0.14
Infit mean-square	0.68	1.36	0.97
Outfit mean-square	0.43	1.41	0.96
Point-measure correlation	0.06	0.62	0.41
Separation index (Reliability)	7.71 (.98)		

Dimensionality. Rasch analysis assumes unidimensionality in the construct being measured. To examine the extent to which our data showed multi-dimensionality, we conducted a principal component analysis on the item fit residuals. If the data was truly unidimensional, the first component of the correlation matrix of the residuals would be small. Simulation studies have shown that components less than 2 are at the random "noise" level (Wilson, 1994) with components less than 3 generally being considered small and indicates the test is largely unidimensional but measuring a "broad" dimension. We found the first component to be 2.6 in our data indicating the tasks and multiple-choice items were predominantly unidimensional, although the construct may be broad. This is to be expected given that items required both science practices and knowledge of the concept of energy.

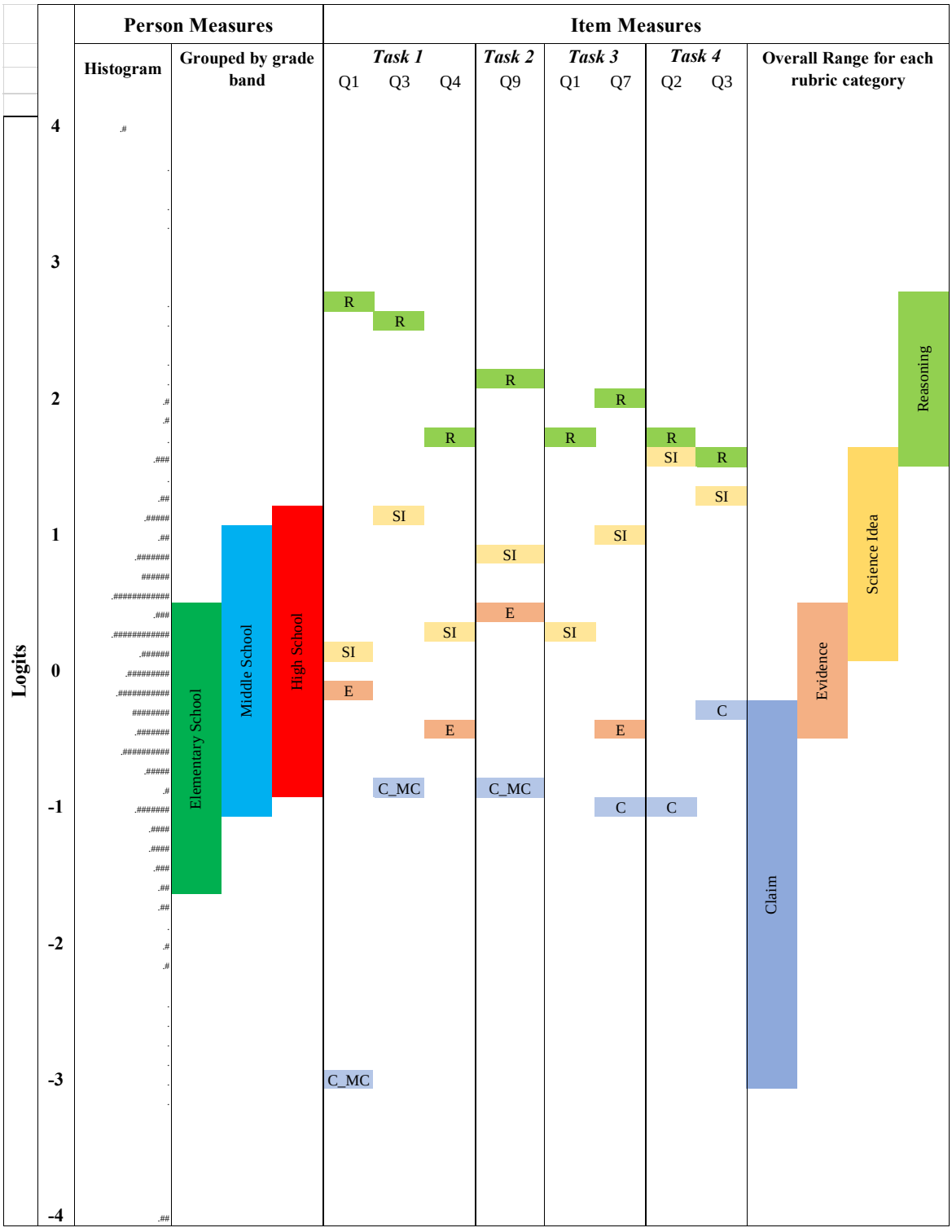
Wright map. Figure 1 shows a Wright map comparing the item measures. All item difficulties fell between -2.0 logits and 2.0 logits. Additionally, items within individual tasks also spanned the difficulty range. The varied difficulty of the items indicates the items were an appropriate difficulty for broad range of students.

As can be seen in the Wright map, the rubric categories differed in their item difficulties with rubric categories related to making a claim (C) having the lowest difficulty and elements related to using reasoning (R) being the most difficult. Table 6 summarizes the item difficulty for claim, evidence, science idea, and reasoning categories.

Table 6: *Summary of Item Difficulties per Rubric category*

Category	**Mean**	**Min**	**Max**	**Count**
Claims (C)	-1.0	-2.8	-0.1	6
Evidence (E)	0.0	-0.3	0.5	4
Science Ideas (SI)	1.0	0.3	1.7	8
Reasoning (R)	2.2	1.7	2.9	8

Figure 1: *Wright Map showing the difficulties of Claim (C), Evidence (E), Science Idea (SI), and Reasoning (R) elements of the argumentation and explanation rubric. (C_MC represent claim elements that were multiple-choice items).* The average person measures +/- a standard deviation for each grade band is also represented on the figure. In the histogram "#" represent eight students while "." represent 1 to 8 students.



Reviewer Feedback

Overall, reviewers agreed with our alignments of the tasks to the targeted NGSS DCIs, SEPs, and CCCs and thought that the phenomena and scenarios were appropriate and engaging. Feedback that they provided about the comprehensibility of the tasks was used to make minor modifications to the tasks. One question that was raised by some of the reviewers was whether current late elementary students would be able to include the science ideas and level of reasoning outlined in our rubrics.

7. Significance

We present an analysis of field test data and feedback from an expert review on the validity and reliability of several elementary school three-dimensional assessment tasks. Field testing data showed that rubric categories within the tasks fit well to a Rasch model with rubric categories spanning the difficulty range. This indicates the task's items were of appropriate difficulty for a wide range of students, although the reasoning category of the rubrics were found to be particularly difficult for most students. The difficulty of rubric categories was found to follow a progression with writing and identifying claims being relatively easy and including reasoning based on the targeted science idea being the most difficult. This progression in difficulty is supported by the findings of other researchers who used similar rubrics for scoring explanation and argument tasks (Gotwals & Songer, 2013; Hu Jin et al., 2020; Hui Jin, Mehl, & Lan, 2015; Osborne et al., 2016). Lastly, a review by science education and content knowledge experts indicated that the tasks were engaging and aligned to the targeted SEPs, CCCs, and DCIs. Together these results provide evidence for the validity of the tasks as NGSS assessments.

In addition to providing evidence of the validity of our tasks, the inclusion of a science idea element to our rubrics provides a unique insight into assessing students' explanations and arguments. It is noteworthy that the reasoning category of the rubrics was in general more difficult than the stating or applying a science idea category. Many student's explanations and arguments were descriptive, providing bits and pieces of information about their understanding of the relevant science concepts, while lacking clear reasoning from those concepts. This indicates that using relevant energy concepts to link evidence to a claim or to make clear deductions is more difficult then stating or applying those energy concepts outside the framework of an explanation or argument. The inclusion of science ideas in the rubrics allows for a task to identify students who may know the underlying science principles but struggle with reasoning with those principles in an argument or explanation. This is a helpful diagnostic tool in a three-dimensional assessment as it could tell whether a student requires instruction focused on the underlying DCI or SEP.

Our results also indicated that most students had difficulty writing coherent reasoning statements. Our panel of experts highlighted this in their comments, noting current students may not have had enough experience with NGSS to provide reasoning at the level expected in the tasks. The finding that writing coherent reasoning is particularly challenging for student has also be found in several others tasks using scientific explanation and argumentation (Gotwals & Songer, 2013; Hu Jin et al., 2020; Osborne et al., 2016). While students' ability to reason and write clearly were considered in our task development, tasks were designed to assess the standards as written. In the future, as NGSS instruction becomes more widespread students' performance on the reasoning elements of tasks such as these should improve.

Acknowledgements.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A180512 to the 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

Achieve. (2019a). NEXT GENERATION SCIENCE STANDARDS TASK SCREENER VERSION 1.0. Retrieved from https://www.nextgenscience.org/sites/default/files/resource/files/Achieve Task Screener_Final_9.21.18.pdf

Achieve. (2019b). Task annotation project in science. Retrieved from https://www.achieve.org/our-initiatives/equip/tools-subject/science/task-annotation-project-science

Andrich, D., Marais, I., & Humphry, S. (2012). Using a Theorem by Andersen and the Dichotomous Rasch Model to Assess the Presence of Random Guessing in Multiple Choice Items. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/1076998611411914

Gotwals, A. W., & Songer, N. B. (2013). Validity evidence for learning progression-based assessment items that fuse core disciplinary ideas and science practices. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21083

Jin, Hu, Yan, D., Mehl, C. E., Llort, K., & Cui, W. (2020). An Empirically Grounded Framework That Evaluates Argument Quality in Scientific and Social Context. *International Journal of Science and Mathematics Education*. https://doi.org/https://doi.org/10.1007/s10763-020-10075-9

Jin, Hui, Mehl, C. E., & Lan, D. H. (2015). Developing an analytical framework for argumentation on energy consumption issues. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21237

Linacre, J. M. (2018). Winsteps ® Rasch measurement computer program. Beaverton, Oregon. Retrieved from Winsteps.com

McNeill, K., & Krajcik, J. (2011). *Supporting grade 5-8 students in constructing explanations in science: The claim, evidence and reasoning framework for talk and writing*. Boston, MA: Pearson Education.

Mcneill, K. L., & Krajcik, J. (2008). Inquiry and Scientific Explanations: Helping Students Use Evidence and Reasoning. *Science as Inquiry in the Secondary Setting*.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A BRIEF INTRODUCTION TO EVIDENCE-CENTERED DESIGN. *ETS Research Report Series*. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

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

NRC. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and*

Core Ideas. (C. on a C. F. for N. K.-12 S. E. S. B. on S. E. D. of B. and S. S. and Education, Ed.). Washington DC: The National Academies Press.

NRC. (2014). *Developing Assessments for the Next Generation Science Standards*. Washington DC: The National Academies Press. https://doi.org/https://doi.org/10.17226/18409

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21316

Wilson, M. (1994). *Objective Measurement: Theory into Practice*. Norwood NJ.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. *Constructing Measures: An Item Response Modeling Approach*. https://doi.org/10.4324/9781410611697

Wilson, M., & Draney, K. (2002). A Technique for Setting Standards and Maintaining Them Over Time. In *Measurement and Multivariate Analysis* (pp. 325–332). Springer, Tokyo.

Yao, J. X., & Guo, Y. Y. (2018). Validity evidence for a learning progression of scientific explanation. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21420