Using Rasch Measurement to Develop 3D Assessment Tasks to Measure Students' Understanding of Energy

Cari F. Herrmann-Abell

BSCS Science Learning

George E. DeBoer

Professor Emeritus, Colgate University

Round table paper presented at the 2023 AERA Annual Meeting

Chicago, IL

April 13-17, 2023

Abstract

This study describes the role that Rasch measurement played in the development of assessments aligned to the *Next Generation Science Standards*, tasks that require students to use the three dimensions of science practices, disciplinary core ideas and cross-cutting concepts to make sense of energy-related phenomena. A set of 27 three-dimensional, multi-item tasks were developed and field tested along with a set of multiple-choice test items focused on disciplinary core ideas. Data were collected from elementary, middle, and high school students from across the U.S. Rasch modeling was used to investigate the fit of the data to the model, the dimensionality of the data, differential functioning of the tasks, the relative difficulties of the tasks, and the performance of students by grade band.

Objectives

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 in turn established a need for new assessments that can measure that integrated, three-dimensional (3D) science learning. The National Research Council (NRC, 2014) recommends that these 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 toward achieving a three-dimensional understanding of the *energy concept* from elementary school through high school. Each task consists of a series of items, both multiple choice and constructed response, that share a common scenario or phenomenon for students to explore. In this paper, we outline the procedure used for developing these tasks and report on the use of Rasch measurement to analyze field test data.

Theoretical Framework

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, p. 44). Additionally, to sufficiently cover all three dimensions, the NRC advocates for the use of sets of interrelated items that share a common scenario and that scaffold students to use their knowledge and skills to make sense of real-world phenomena or solve a problem. The discrete items within the set may target individual core ideas, practices, or crosscutting concepts, but, when taken as a whole, provide a picture of students' three-dimensional science understanding.

In an effort to better conceptualize what these three-dimensional assessments would look like, Achieve completed the Task Annotation Project in Science (Achieve, 2019a), which resulted in a list of features all NGSS-aligned assessments should possess. Those features include: (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.

Methods

The NRC (2014) suggests that these three-dimensional assessments should be developed using a construct-centered approach, such as Evidence-Centered Design (Mislevy *et al*., 2003) or Construct Modeling (Wilson, 2005). The development procedure for the three-dimensional tasks used this approach and builds on our previous assessment development work (DeBoer, et al., 2008).

Construct definition

Assessment development started by selecting a set of thematically related NGSS performance expectations (PEs) that progress with increasing sophistication through the grade bands (See Table 1 for example.). The 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. The three dimensions targeted by the PEs were further clarified by consulting the relevant sections of the NRC Framework (NRC, 2012) and the NGSS appendices to identify the grade band appropriate level of understanding.

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.

Choosing phenomena

After the PEs progressions were identified, we searched for phenomena that required students to engage with the targeted DCIs, SEPs, and CCCs. Phenomena were selected with the goal that they would be familiar and engaging to a wide range of students. A cluster of items were then created for each task that move students through a process of phenomenon introduction, sense-making, and final resolution of the problem or driving question that motived the task. We developed a set of 27 tasks.

Alternate versions of items

To provide options for users of these tasks that can help them balance the time and effort required of students during testing and the time and effort needed to score and evaluate students' responses, we developed alternative versions of some tasks that substituted multiple-choice versions of one to four constructed-response items within fourteen of the tasks. Both versions used identical or nearly identical stems. Our past research on the comparability of the multiple-choice and constructed-response versions of the items within the elementary tasks showed that the multiple-choice versions assessed the same unidimensional construct as the constructed-response versions, but the constructed-response versions were more difficult than their multiple-choice counterparts (Herrmann-Abell, Hardcastle, & DeBoer, 2022). This increased difficulty was found to be partially due to the added writing demand and the higher difficulty of the reasoning element in the constructed-response rubric. Students were more likely to recognize a clearly reasoned explanation in a multiple-choice setting than they were to create that reasoning themselves and communicate it in writing.

Developing the scoring rubrics

Rubrics for scoring were constructed at the item level. In multiple-choice items, the outcome space was defined by the set of answer choices. Our guidelines for item construction ensure that all the answer choices are thematically related to the question and that distractors target relevant student misconceptions. Multiple-choice items were scored dichotomously, either as right or wrong. For constructed-response items, we attempted to control the outcome space by using clearly stated questions and appropriate scaffolding, such as prompting students to use simpler precursor knowledge as a way to begin to make sense of the phenomenon being presented in the task.

We began the process of rubric development by first creating an ideal response. Then, we identified statements in the response that indicated use of the targeted dimensions. 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 were 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 modeling items includes the "Model Components" and "Interactions Between Components" categories. For explanation and argumentation items, the categories followed a modified Claim, Evidence, Reasoning (CER) framework (McNeill & Krajcik, 2011), with the addition of a category called "States or Uses Science Ideas" (Hardcastle, Herrmann-Abell, & DeBoer, 2021). An example rubric is shown in Table 2.

These categories inform scorers about the types of features to look for in student responses, and the individual elements provide examples for each category. Students' scores on the items were based on the number of categories included in their response. Task level scores were calculated by summing the number of points earned on the items within that task.

Before scoring the full set of data, a randomly selected subset of fifty student responses for each task were scored by two scorers. Scorers met to evaluate the interrater reliability and discuss disagreements. An acceptable kappa reliability (> 0.70) was achieved for most rubric elements. For some elements, only a few students received a point, and, therefore, a small number of rater disagreements produced kappa reliabilities below the 0.70 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 by the scorers so that a final decision on scoring could be made. The rubric elements were revised based on the scorers' discussions to clarify the meaning of the element, and then the full set of data was scored by one of the two scorers.

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 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.
Category	**Individual Elements**
Student makes a *claim*	• The ball slows down because it has less energy after hitting the pin *or* because it transfers energy to the pin and/or the air during the collision.
Student lists *evidence*	• 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 *science idea* (See bullet 1 for an example of using a science idea.)	• 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.
Students use *reasoning* to link the evidence and science idea	• 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.

Pilot testing

Draft versions of the tasks were pilot tested with elementary, middle, and high school students and crowdsourced adults to obtain feedback on the comprehensibility of the tasks and authentic responses to help in the development of scoring rubrics. Pilot testing with crowdsourced adults was conducted during the COVID-19 pandemic through Amazon's Mechanical Turk system as a way to continue assessment development and provide an opportunity to collect more correct responses at the upper end of the distribution, which might be difficult to obtain in a high school sample. During pilot testing, students responded to one task and a set of ten DCI-focused multiple-choice items, and adults responded to five or six tasks. Written responses from the pilot tests were analyzed, and these responses informed task and rubric revisions.

Student interviews

To evaluate the cognitive processes that students use while responding to the draft tasks and to obtain feedback on their length and difficulty, think-aloud interviews were conducted with elementary, middle, and high school students. During the interview, students were asked to think out loud as they responded to between one and four tasks. Overall, students used the intended

cognitive processes related to the targeted practices and concepts, and they found the task contexts to be familiar and engaging. Students also pointed out difficulties and confusion with the way some of the questions were worded and identified challenges with using a drawing tool implemented in some tasks. Our findings from these interviews were used to revise the tasks.

Expert review

We conducted two rounds of expert review to evaluate the tasks and rubrics. The advisory board review panel consists of science education and content knowledge experts with experience in the crafting and implementation of NGSS. Reviewers submitted their review using a survey based on criteria in the Achieve Task Screener (Achieve, 2019b). 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 scoring 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. Overall, reviewers agreed with our alignments of the tasks to the targeted NGSS DCIs, SEPs, and CCCs and thought that most of the phenomena and scenarios were appropriate and engaging. Feedback provided during each round of reviews was used to make modifications to the tasks.

Field testing

Each field test form was composed of three 3D tasks, one aligned to each theme, and a set of 15 additional DCI-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). The DCI-focused items were selected to cover the DCIs targeted by the themes and to span a wide range of difficulties, and they then served as linking items. Test forms for elementary students included the elementary and middle school tasks, but not the high school tasks. Test forms for middle and high school students included all the tasks, elementary through high school.

Rasch analysis

Because the items within the 3D tasks are polytomously scored, the Rasch partial credit model was applied (Masters, 1982). Fit to the Rasch model was evaluated using fit statistics including separation indices and reliabilities, infit and outfit mean-squares, and standard errors (Bond & Fox, 2007; Boone, Staver, & Yale, 2014). A principal component analysis of the Rasch residuals was conducted to examine the dimensionality of the data. Wright maps were used to compare the relative difficulties of the tasks and the relative performance levels of the students by grade band. Differential item functioning was used to investigate whether the tasks functioned differently for students who indicated their primary language is English versus students who indicated their primary language is not English, and to investigate whether tasks functioned differently for students who identified as male versus female.

Data sources

The sources of the data were field tests conducted during the spring of 2021 and 2022. In 2021, over 2,300 students in grades 4 through 12 in the classrooms of 50 teachers from 27 states across

the U.S. participated in the field testing. In 2022, over 1,500 students in grades 4 through 12 in the classrooms of 44 teachers from 16 states and Puerto Rico.

This study includes the data from the 3442 students who responded both to at least five DCI-focused items and all the items within at least one 3D task. Table 3 shows a summary of the demographic information for the students in this data set. Approximately 14% of the sample were elementary school students, 45% were middle school students, and 40% were high school students. Approximately 7% of the students indicated that English was not their primary language. On average, 192 students responded to each of the 3D tasks.

Table 3:
Summary of Demographic Information

	Percentage of Sample		Percentage of Sample
Grade Band		**Race/Ethnicity**	
Elementary	14%	American Indian	1%
Middle	45%	Asian	7%
High	40%	Black	10%
Gender		Hispanic	16%
Female	51%	White	57%
Male	49%	Other	4%
Primary Language		Two or more	5%
English	93%		
Other	7%		

Results & Discussion

Data fit

The data had a good fit to the Rasch model as shown in Table 4. The item separation index was very high (13.49) with a corresponding reliability of .99. The person separation index (1.61) was slightly lower than the desired minimum of 2, but still corresponded to a reliability of .72. The infit and outfit MNSQ values were within a range considered productive for measurement (0.5 – 1.5) except for one task that had MNSQ values above 1.5. The standard errors were low and the point measure correlations for the items were high.

Table 4:
Summary of Rasch Fit Statistics

	Item			Person		
	Min	Max	Median	Min	Max	Median
Standard error	0.03	0.10	0.05	0.29	1.92	0.38
Infit mean-square	0.64	1.72	0.88	0.06	5.33	0.78
Outfit mean-square	0.65	1.71	0.88	0.06	7.45	0.91
Point-measure correlation	0.04	0.90	0.74			
Separation index (Reliability)	13.49 (.99)			1.61 (.72)		

Dimensionality

To examine the extent to which our data showed multi-dimensionality, we conducted a principal component analysis of the Rasch residuals. If the data were truly unidimensional, the first component of the correlation matrix of the residuals would be less than 2 (Wilson, 1994). In our analysis, we found that approximately 68% of the raw variance in the data was explained by the measures, and the eigenvalue of the first contrast was less than two (1.67). This supports the idea that the 3D tasks and DCI-focused items form a unidimensional scale.

Task difficulties

The Wright map in Figure 1 includes the constructed-response versions of the 3D tasks and the DCI-focused, multiple-choice linking items. The map shows that the 3D tasks spread across the upper end of the difficulty scale and that there are several students with person measures lower than the easiest task. This indicates that the 3D tasks are challenging for the students in our sample.

Table 5 presents a summary of the task difficulties by theme and grade band. The overall trend by grade band was as expected with the elementary tasks being easier than the middle school tasks, which were easier than the high school tasks. When looking at the trend by theme, we see the expected trend for the Energy Transfer by Forces tasks and the Thermal Energy Transfer tasks. However, for the Chemical Reaction and Energy tasks, the middle school tasks were 0.16 logits more difficult than the high school tasks, possibly due to the students' familiarity with the chemical reactions involved in the scenarios in the tasks. The more difficult high school tasks (AP47-4 and AP43-6) involve the combustion of methane and butane. The easier high school tasks (AP33-6 and AP34-6) involve cellular respiration and photosynthesis, which are reactions that are central to middle and high school biology instruction. Overall, the tasks that were aligned to the Chemical Reactions and Energy theme were more difficult than the Energy Transfer by Forces tasks, which were more difficult than the Thermal Energy Transfer tasks.

Table 5:
Summary of average task difficulty by theme and grade band

	Elementary	Middle	High	Total
Chemical reactions and energy	0.18	0.86	0.70	0.64
Energy transfer by forces	-0.43	0.52	0.92	0.29
Thermal energy transfer	-0.35	-0.13	0.14	-0.06
Total	-0.23	0.43	0.52	0.29

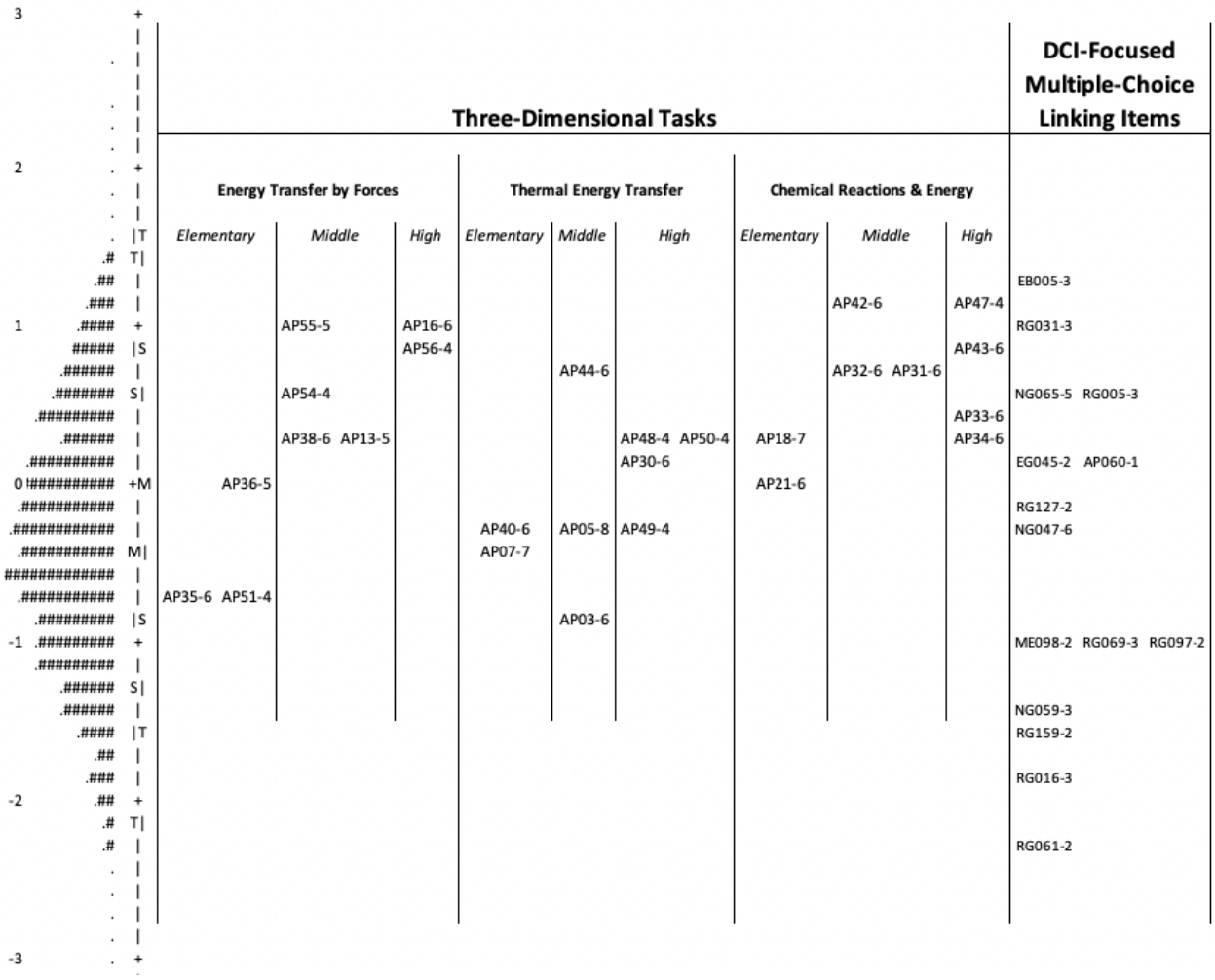

Figure 1. Wright map showing the task difficulties (right side) and person measures (left side). Each # is 18 people and each "." is one to 17 people.

Difficulties of alternate versions

The cross-plot in Figure 2 visually compares the task difficulties of the pairs of tasks for which multiple-choice and constructed-response versions of items were developed. In the cross-plot, dots above the dashed identity line represent tasks with constructed-response versions that are more difficult than multiple-choice versions, and dots below the line represent tasks with constructed-response versions that are less difficult. Dots that lie along the line represent tasks where the different versions are similar in difficulty. As we saw in our past study (Herrmann-Abell, Hardcastle, & DeBoer, 2022), the tasks with the constructed-response versions tend to be more difficult than the tasks with the multiple-choice versions. This is true for 10 of the 14 items. The four exceptions are three tasks (Catapult, Solar Car, and Thawing Soup) for which the constructed-response versions and multiple-choice versions have similar difficulties and one task (Creating Droplets) for which the constructed-response version is easier than the multiple-choice version. Below we describe the different versions of the items within these exceptional tasks and propose a possible explanation for the difference in difficulty or lack thereof.

Multiple-choice more difficult. The Creating Droplets task included two items for which there were multiple-choice and constructed-response versions. One item asked students to either describe the patterns they observe in the data for the temperature of different regions in a system or select the answer choice that described the correct pattern. The other item required students to either construct a graph that predicted the change in temperature over time for a system or select a graph that correctly illustrates the change in temperature. Both constructed-response versions were scored polytomously on a scale from zero to three, while the multiple-choice versions were scored dichotomously.

Equally difficult. The Thawing Soup tasks included two items for which there was multiple-choice and constructed-response versions. Both items asked students to use patterns that they observed in order to draw conclusions about how energy is transferred into and out of a system either by describing their conclusion in written form or by selecting the correct conclusion. The multiple-choice and constructed-response versions for both items were scored dichotomously.

The Catapult task included one item for which there was a multiple-choice and a constructed-response version. In this item, students either construct a graph of the amounts of potential energy and kinetic energy in a system or select a graph. The multiple-choice and constructed-response versions were scored dichotomously.

The Solar Car task included one item for which there was multiple-choice and constructed-response versions. In this item, students either construct a model that illustrates how energy moves between the sun, solar panel, and electric motor or select an answer choice containing the correct model. The constructed-response version was scored polytomously on a scale of one through three and the multiple-choice version was scored dichotomously.

Possible explanation. The way in which the items were scored could be one explanation for the pattern in difficulty difference. Most of the pairs of items in the tasks that had equal difficulty were scored the same. On the other hand, the pairs of items in the task with the less difficult constructed-response versions were scored differently, and students were able to get partial credit on the constructed-response version, which allowed them to earn more points. The exception to this is the item on the Solar Car task where the different versions had similar difficulties, but the constructed-response version was scored polytomously and the multiple-choice version was scored dichotmously. In this task, the model that the students construct or select is an elementary level model where students draw a flow chart with arrows from the sun to the solar panel to the electric motor and label the arrows with the mechanism of energy transfer. The simplicity of the model may have made it just as easy to construct the model as to select the correct model.

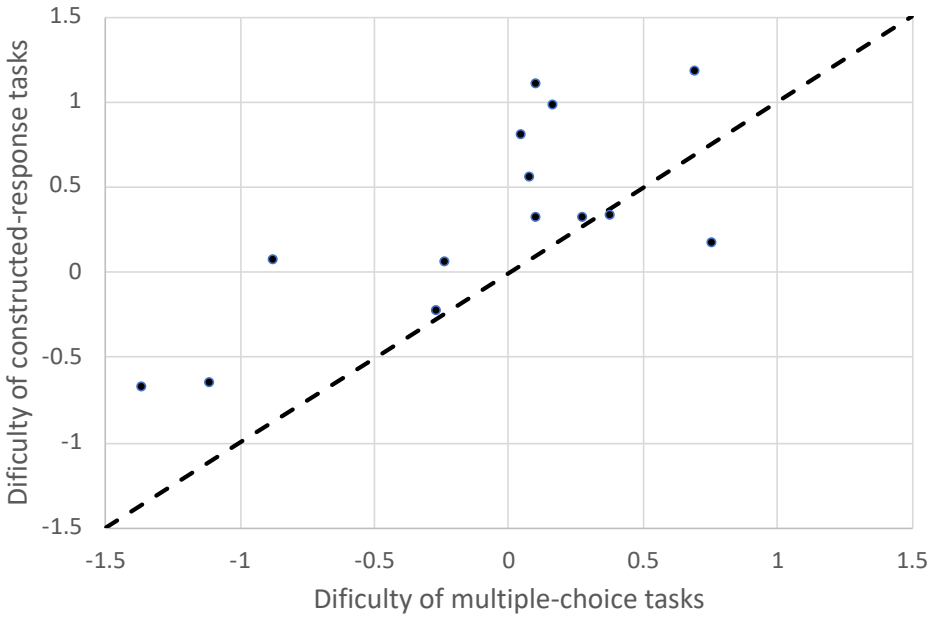

Figure 2. Cross-plot of the difficulties of the different versions of the paired tasks. The dashed line represents the identity line along which the difficulties of the different versions are equal.

Person measures

Table 6 shows the minimum, maximum, and mean of the person measures, and the Wright map in Figure 3 shows the distribution of person measures by grade band. The elementary students are represented by the letter "E," the middle school students are represented by the letter "M," and the high school students are represented by the letter "H." T-tests showed that the high school students performed better than the middle school students ($t = 10.96$, $p < .001$), and the middle school students performed better than the elementary school students ($t = 9.64$, $p < .001$). This is consistent with what we would expect from high school students who likely have had more instruction on energy than middle school students, who likely had more instruction on energy than elementary school students. According to the NGSS, fourth grade is the first grade in which energy is taught.

Differential item functioning

Differential item functioning (DIF) was used to investigate whether the tasks functioned differently for respondents based on whether the student identified English as their primary language and based on whether they identified themselves as male or female. In our DIF analyses, a task would be considered to have DIF if the DIF contrast value was greater than 0.43 logits and the p value was less than .05 (Linacre, n.d.). For the DIF analysis based on language, five of the tasks had DIF contrasts greater than 0.43, but the p values for these tasks were greater than .05. Therefore, we conclude that none of the tasks exhibited statistically significant DIF with respect to English as students' primary language. For the gender DIF analysis, no task had a DIF contrast greater than 0.43, indicating that the tasks also did not exhibit DIF with respect to gender.

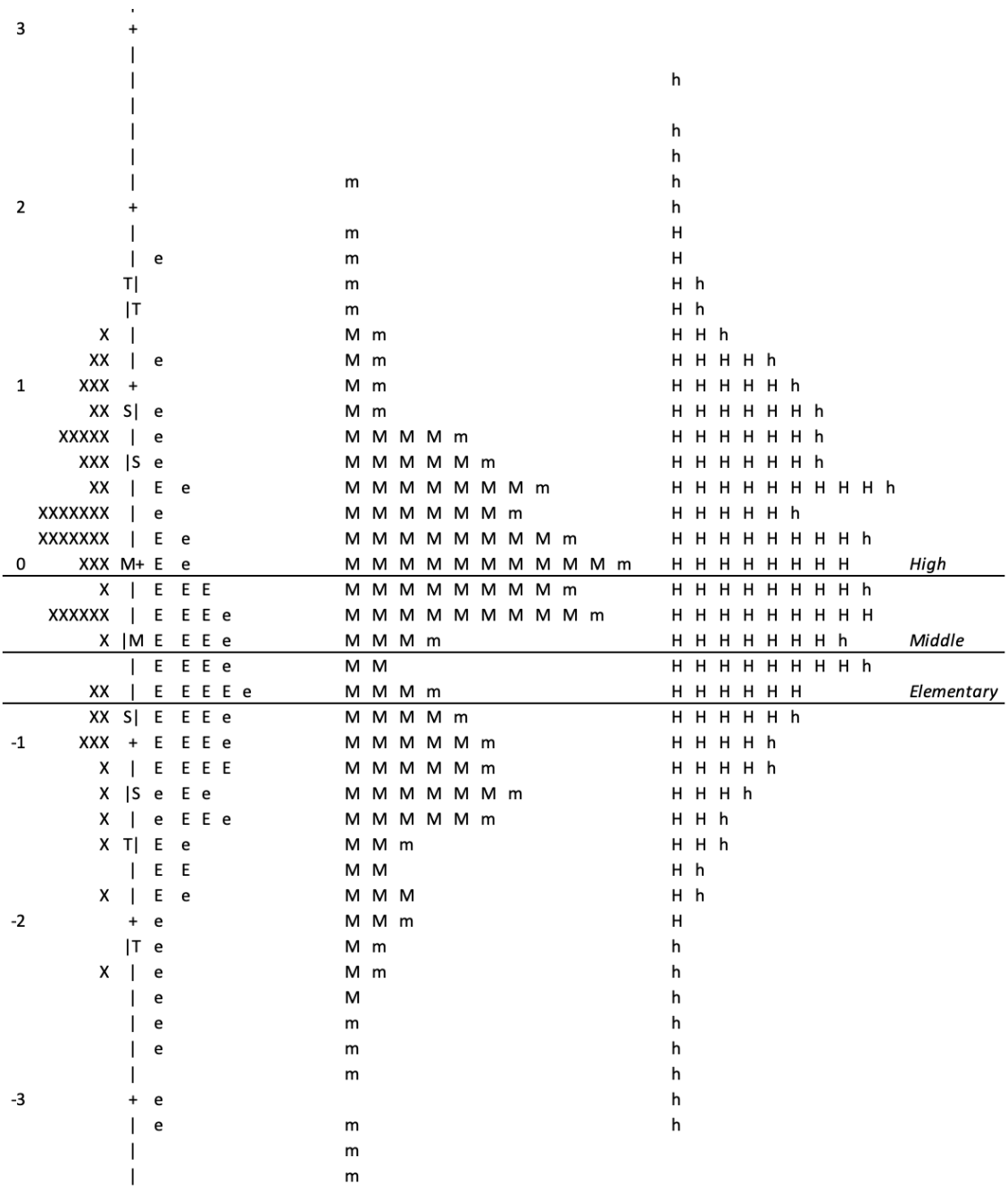

Figure 3. Person Wright map showing the spread of difficulties by grade band. Each capital letter is 10 students, and each lower-case letter is 1 to 9 students. The horizontal lines indicate the average measure for that grade band.

Table 6:
Summary of student performance by grade band

Grade Band	Minimum Measure	Maximum Measure	Mean Measure
Elementary	-3.08	1.69	-0.82
Middle	-3.56	3.13	-0.45
High	-3.13	3.31	-0.09

Conclusions & Significance

This study addresses the need for new assessments that measure students' multidimensional science understanding as described in the NGSS and provides insights into how Rasch measurement provides validity evidence for these assessments. The principal component analysis of the Rasch residuals showed that the 3D tasks and DCI-focused items form a unidimensional construct representing a 3D understanding of the energy concept. This means that tasks that ask students to make sense of phenomena using their science knowledge, practices, and crosscutting concepts can be assumed to be testing an integrated, unidimensional construct, most likely held together by the students' knowledge of the science ideas. This conclusion is also supported by the fact that the DCI-focused linking items were spread across the difficulty spectrum and did not fall out as a separate dimension.

The Wright map showed that the 3D tasks are located on the higher end of the difficulty scale, and there is a progression of difficulty for the tasks aligned to elementary, middle, and high school NGSS dimensions.

We also looked at the difficulty of alternative versions of tasks in which a selection of constructed-response items within the task were replaced with multiple-choice versions. Our findings supported past findings and showed that, overall, the constructed-response versions were more difficult than the multiple-choice versions. Our past item-level analysis concluded that this difference was most likely due to the increased writing demand and the higher difficulty of the reasoning element in the constructed-response rubric (Herrmann-Abell, Hardcastle, & DeBoer, 2022). The current analysis showed that task pairs that were close in difficult were ones in which both the constructed-response and multiple-choice versions were scored in the same way (dichotomously), whereas the task pair with the less difficult constructed-response version used polytomous scoring for the constructed-response version and dichotomous scoring for the multiple-choice version, making it possible for students to get partial credit for their response.

The analysis of the person measures showed the expected trend with high school students outperforming middle school students, and middle school students outperforming elementary school students. DIF analyses confirmed that the tasks were functioning the same for students whose primary language was English and students whose primary language was not English and for male and females.

As we navigate the context of the NGSS, assessment plays a key role in helping to understand how students build their three-dimensional understanding of science. This study focuses on the development of new assessments that can be used to measure students' use of disciplinary core ideas, cross-cutting concepts, and science practices to explain energy-related phenomena and it demonstrates the role Rasch modeling can play in validating assessments that measure the complex science learning outcomes described in the NGSS. The results can inform best practices in how to measure students' multidimensional science understanding feasibly and efficiently.

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

Achieve. (2019a). Task annotation project in science. Retrieved from https://www.achieve.org

Achieve. (2019b). 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

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Netherlands: Springer.

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.

Herrmann Abell, C.F., Hardcastle, J.M., & DeBoer, G.E. (2022, March). Exploring the Comparability of Multiple-Choice and Constructed-Response Versions of Scenario-Based Assessment Tasks. *Paper presented at the NARST 2021 Annual Conference*.

Linacre, M. (n.d.). *DIF - DPF - bias - interactions concepts*. Winsteps.Com. https://www.winsteps.com/winman/difconcepts.htm

Linacre, J. M. (2022). WINSTEPS Rasch measurement computer program. Version 5.2.4. Beaverton, Oregon: 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.

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.

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.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.