Validating a Claim-Evidence-Science Idea-Reasoning (CESR) Framework for use in NGSS assessment Tasks

Joseph M. Hardcastle[1]

Cari F. Herrmann-Abell[2] and George E. DeBoer[1]

[1]AAAS Project 2061, [2]BSCS Science Learning

Presented at the 2021 NARST Virtual Annual Conference

Abstract

We developed assessment tasks aligned to the Next Generation Science Standards (NGSS) that require students to use argumentation and explanation practices along with disciplinary core ideas and crosscutting concepts to make sense of energy-related phenomena. Scoring rubrics were created to evaluate students' ability to make accurate claims, cite evidence, use relevant science ideas, and combine those elements to formulate well-reasoned arguments and explanations. We present an analysis of data to investigate the validity and reliability of our rubrics. Due to school closures caused by the COVID-19 pandemic, data were collected using Amazon Mechanical Turk (MTurk). The MTurk data were scored by two researchers to evaluate the inter-rater reliability. Data were then analyzed using Rasch modeling. Results show that rubric categories associated with stating claims, citing evidence, applying science ideas, and formulating coherent, well-reasoned arguments and explanations fit well to the Rasch model, and that rubric categories followed a hierarchy of difficulty. In this hierarchy, applying science ideas and formulating well-reasoned statements were more difficult than citing evidence, which were all more difficult than stating a claim. The ability to locate a student along this hierarchy allows for our tasks to be used to better understand a student's ability to write arguments and explanations of energy-related phenomena.

1. Subject

The *Next Generation Science Standards* (NGSS Lead States, 2013) calls for instruction that integrates multiple dimensions of science. NGSS describes these dimensions as: (1) science and engineering practices (SEPs), (2) crosscutting concepts (CCCs), and (3) disciplinary core ideas (DCIs). To evaluate this new instructional approach, new assessments are needed that engage students in all three dimensions of science.

The National Research Council (NRC, 2014) recommends that NGSS-aligned assessments be designed to allow students to demonstrate the use of science and engineering practices (SEP) in the context of disciplinary core ideas (DCI) and crosscutting concepts (CCC), provide information that situates students' knowledge on learning progressions, and include tools to help teachers interpret and use students' responses to adapt instruction. To meet these assessment goals, they have suggested using sets of interrelated items where the individual items may target one, two, or three dimensions. When taken as a whole these assessments should provide a complete picture of students' three-dimensional science understanding.

Following the recommendations of the NRC, we developed NGSS-aligned assessment tasks that focus on measuring late elementary, middle, and high school students' ability to make sense of energy-related phenomena. These tasks present one or more phenomena, usually embedded in a scenario, followed by a series of interrelated constructed-response and multiple-choice items. In this paper, we outline the procedures we used for developing and validating rubrics for each of the items within those tasks that have students engage in constructing explanations and writing arguments. We present an analysis of data that indicates that the categories in the rubrics are progressively more difficulty for students. We also discuss how the difficulty of rubric categories depends on both students' knowledge of the DCIs and CCCs along with their explanation and argumentation ability. The progression in difficulty and multidimensional nature of the rubrics provide teachers with helpful information to support students in constructing more sophisticated explanations and arguments about energy-related phenomena.

Framework. The claim-evidence-reasoning (McNeill & Krajcik, 2011) framework has become a popular and effective tool for evaluating students' ability to write arguments and explanations. This framework is based on the idea that a student's argument or explanation can be deconstructed into a *claim* that answers the proposed question or gives a stance in an argument, *evidence* that is based on relevant observations and/or data to support the claim, and *reasoning* that uses logic and scientific theories to link the cited evidence to the claim. Rubrics based on this framework have successfully shown that students distribute themselves along a hierarchy of difficulty for those three components. Specifically, Gotwals and Songer found a hierarchy in difficulty starting with writing a claim being the least difficult and providing reasoning being the most difficult for students to include in their writing (Gotwals & Songer, 2013).

Researchers have also begun to extend the CER framework to further stratify students based on the type of reasoning they use. Jin et al. found that reasoning could be differentiated into weak and strong reasoning elements in rubrics (Hu Jin, Yan, Mehl, Llort, & Cui, 2020) while Osborne et al. have shown that students' reasoning can be separated by the uses of general and scientific reasoning (Osborne et al., 2016). General reasoning does not require domain-specific knowledge. In contrast, scientific reasoning requires the use of scientific principles, laws, or theories. We followed a similar approach; however, instead of looking at different types of reasoning, we included a category called "Science Ideas" in the CER framework. Our framework then has four

categories: claim, evidence that supports the claim, stating or using relevant science ideas, and reasoning. For our "reasoning" category, we expect students to tie together the claim, evidence, and science ideas into a coherent argument to support an explanation. The "states or uses science ideas" category was included for two reasons. First, it provides a more direct measure of whether students' writing shows evidence of understanding of specific DCIs. This is helpful when one of the goals is to assess a student's DCI knowledge. Second, we found that many students state or use science principles in their response but don't use those principles to formulate a coherent explanation for the observed phenomenon. These students would not receive the reasoning point in our scoring but could be awarded a point for knowing the relevant science ideas. This allows us to account for students who have some understanding of an underlying science principle but may not be able to coherently reason using the principle. We term this framework Claim, Evidence, Science idea, and Reasoning (CESR).

2. Procedure

Task Development. In accordance with the NRC's recommendations on assessment (NRC, 2014), tasks were developed using the construct-centered approach of Construct Modeling (Wilson, 2004). First, we chose three energy topics: (1) transfer of energy by forces and conservation of energy, (2) thermal energy transfer and dissipation, and (3) energy and chemical reactions. For these topics we specified the appropriate level of understanding of the three dimensions we could expect for each grade band using the NRC *Framework* (NRC, 2012) and the appendixes to NGSS. We then used the framework from the Task Annotation Project in Science (Achieve, 2019) as guidelines during task development. This framework suggests that NGSS-aligned assessments should: (1) focus on real-world phenomena, (2) require students to engage in sense making, (3) require students to use both disciplinary core ideas and science practices, (4) be comprehensible to students, and (5) support the intended purpose and use of the assessment.

We then searched for phenomena that required students to engage with the targeted energy ideas, SEPs, and CCCs. Phenomena were selected with the goal that they would be familiar and engaging to a wide range of students. Once a phenomenon was chosen it was used to describe a scenario that students would engage with by using the targeted three dimensions. Students answered a set of related multiple-choice and constructed-response items, all of which moved the students progressively through a sense-making process that resulted in a final resolution to the problem that had been introduced in the task. While individual items varied in their alignment to the targeted dimensions, taken together the items provide a comprehensive picture of students' ability to engage in SEPs using DCI and CCC knowledge. In addition, some items within a task were not necessarily designed to include all four categories of the CESR rubric. Tasks went through multiple rounds of pilot testing, review by a panel of experts, and revision (Hardcastle, Herrmann-Abell, & DeBoer, 2019; Herrmann-Abell, Hardcastle, & DeBoer, 2020).

Rubric Development. Rubrics for tasks were created at the item level, i.e., at the level of each multiple choice or constructed response question. In this proposal, we outline how rubrics were created for the explanation and argumentation items.

We began rubric development by first drafting an ideal response. We then deconstructed that response into finer-grain statements that could be grouped into the different rubric categories, (1) a statement of a claim, (2) relevant evidence, (3) statement or use of relevant science ideas, and

(4) coherent and logical reasoning that links science ideas and/or evidence to the claim. We refer to these evidence statements as "elements."

These elements were revised, clarified, and generalized so we could consistently identify them in a wide range of student responses. One significant revision involved adding elements for the reasoning category. Initially, we gave students credit for "reasoning" only if they made a claim, supported it with evidence, and then related the claim and evidence to a relevant science idea. During pilot testing, we found that very few students received the reasoning point because many of them were not using both science ideas and evidence to justify their claims. In effect they were saying: "You can see why the claim is true from the evidence." Or "You can see why the claim is true because it is consistent with this science idea." Because both of these could be reasonable and convincing arguments, we decided to give students credit if they used evidence-based reasoning that focused on using evidence and logic to justify the claim and, separately, idea-based deductive reasoning that focused on using science ideas to justify the claim. Table 1 shows an example rubric using the CESR framework.

Table 1: *Example Rubric using the CESR framework*

Item Context	Students watch a video of a newton's cradle and are asked to write an explanation for why the balls eventually stop swinging.
Student makes a **claim**	• The balls stop swinging because energy is transferred away from the balls/the cradle.
Student lists **evidence**	• The balls reach a lower height each time they swing. • A sound is heard as the balls hit one another.
Student either states or uses a general **science idea**	• Moving slower means less kinetic energy (i.e. the slower moving ball has less kinetic energy). [*links speed and energy*] • A lower height means less gravitational potential energy (i.e. the ball that reaches the lower height has less gravitational potential energy). [*links height and energy*] • When two objects interact, each one exerts a force on the other that can cause energy to be transferred to or from the object (i.e.: as the balls swing, they interact with the air {air resistance or friction} causing energy to transfer from the balls to the air.)[*links forces and energy transfer*]. • Sound is an indicator of energy transferred to the surroundings (i.e. the sound the balls make when they hit transfers energy to the surroundings). [*links sound and energy transfer*] • When there is a change in kinetic energy, there is some other change in energy at the same time (e.g. if the cradle has less energy, there must be an increase in energy in the surroundings). [*conservation*]
Student uses **reasoning** to link evidence and science ideas to the claim	• The balls stop swinging because energy was transferred away from the cradle by the sound made when the balls collide resulting in less and less energy to swing. • The balls stop swinging because energy was transferred from the balls to the air as the ball interacts with the air (because of air resistance/forces) resulting in less and less energy to swing.

Data Collection. To evaluate the usefulness of the rubrics for scoring student responses, we administered the tasks to approximately three hundred participants using the online crowdsourcing system Amazon Mechanical Turk (MTurk), a distributed workforce that can complete online assignments. Members of the MTurk workforce were able to view a brief summary of the assignment and then decide whether they wanted to participate. We set up three different assignments, one for each energy theme. The assignments consisted of completing a set of either five or six tasks. Participants were provided a link that sent them to our online assessment utility.

Because the goal of testing this group was to evaluate the usefulness of the rubrics, we wanted responses from individuals who had been exposed to high school science and who had some familiarity with the targeted energy ideas. For respondents to be selected, they had to be from the U.S., be between 18 and 25 years old, and have a high school diploma or equivalent. We obtained 105 respondents for each set of tasks. Of the 315 total respondents, 308 completed all of their tasks and were paid, and seven were dropped from the data set for not completing the tasks or not correctly following instructions. The sample of respondents was 50% female, 49% male, and 1% who did not identify either gender. The sample included 58% White, 14% Asian, 10% Black, 9% Hispanic, 1% Pacific Island, 1% American Indian, 6% who identified with two races or ethnicities, and 1% who did not identify with any of the listed races or ethnicities. Approximately 10% of the sample had only a high school diploma, 36% had some college but no degree, 7% had an associate degree, 44% had a bachelor's degree, and 3% indicated they had obtained a graduate degree. Approximately 77% of the respondents had taken at least one physical science course in college or graduate school.

3. Findings and Analysis

Scoring. The first step in the scoring process was to evaluate each response and identify the presence or absence of individual rubric elements. To evaluate the scoring reliability, all of the responses were scored by two researchers and the percentage match and Cohen's kappa were calculated for each rubric element. An acceptable kappa reliability (> 0.70) was achieved for most rubric elements.

If the kappa reliability was found to be below 0.70, the researchers reviewed the scoring and found that in most cases their scoring matched more than 90% of the time. When there was high matching but relatively low kappa, it was because very few respondents received points for these elements. This is consistent with the fact that Cohen's kappa will be lower in cases of where a large proportion of the students are answering incorrectly, and few students answer an correctly (high prevalence) (Byrt, Bishop, & Carlin, 1993). In the end, all scoring mismatches were reviewed by the researchers so that a final decision on scoring could be made. After reconciling the scores at the element level, a dichotomous CESR category-level score was assigned. The response received a point for the category if it included at least one of the elements from that category. For example, to get a point in the evidence category, a response would have to cite at least one piece of the evidence listed as an element under the evidence category.

Rasch analysis. We used Rasch analysis to estimate item and person 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 CESR category was treated as a dichotomous item in the Rasch analysis. A separate Rasch analysis was run for

each of the three sets of tasks because there was no overlap among the sets and, therefore, no linking items.

Table 2 summarizes the item fit statistics for each of the three sets of tasks. The item separation indices, which indicate the number of levels into which items can be reliably separated, were high, about four for each topic. This indicates a wide range of item difficulties for the tasks.

Table 2: *Summary of Rasch Item Fit Statistics*

	Thermal Energy		Kinetic Energy and Forces		Chemical Reactions and Energy	
	Median	SD	Median	SD	Median	SD
Standard error	0.36	0.23	0.38	0.10	0.36	0.11
Infit mean-square	0.90	0.24	0.94	0.18	0.98	0.13
Outfit mean-square	0.83	0.37	0.95	0.50	0.93	0.27
Point-measure correlation	0.51	0.13	0.53	0.13	0.43	0.13
Separation index	4.13		4.28		3.84	
(Reliability)	(.94)		(.95)		(.94)	

Figures 1, 2, and 3 show the Wright maps comparing the item measures for the CESR categories for each set of tasks. For fifteen of the nineteen items that included a claim element, the claim rubric element was least difficult element. Evidence and science idea elements were usually the next most difficult rubric categories with reasoning being the most difficulty rubric category for the majority of items. While some items did not include all rubric categories, the overall results suggest a progression in difficulty with claims being relatively easy, followed by citing evidence, then stating and using science ideas, and finally applying coherent reasoning being the most difficult aspect. This progression is consistent with other researchers who have used similar CER based rubrics (Gotwals & Songer, 2013; Hu Jin et al., 2020; Hui Jin, Mehl, & Lan, 2015; Osborne et al., 2016).

While many items seem to follow this hierarchy in difficulty, several exceptions to this hierarchy were observed. For example, in task 1 item 4 in Figure 2, the evidence category is more difficult than the reasoning category, and the science idea category is easier than the claim category. To further explore these cases, we examined the rubric elements and responses for each item where rubric categories didn't follow the hierarchy in difficulty observed in other items. We found that rubric categories vary in difficulty due to a variety of reasons including how much knowledge of energy ideas is required, the relative complexity of practice required, and/or the amount of scaffolding provided in the task. For example, several items had relatively easy claim elements possibilty due to the fact that the claim only required students to affirm a given claim or agree/disagree with a statement about the claim. In contrast, the most difficult claim elements required students to compose their own original claim. For the evidence category, lower difficulty was found in tasks with relatively simple data or observations while higher difficulty was found in tasks that tended to require some understanding of energy ideas to parce the data or require students to identify evidence from several different data sources. Science ideas were found to vary in difficulty due to the amount of scaffolding in the task. Heavily scaffolded tasks where the application or use of a science idea was highly suggested via text, images, or previous items in the task were relatively easy while items without such scaffolding were more challenging for students.

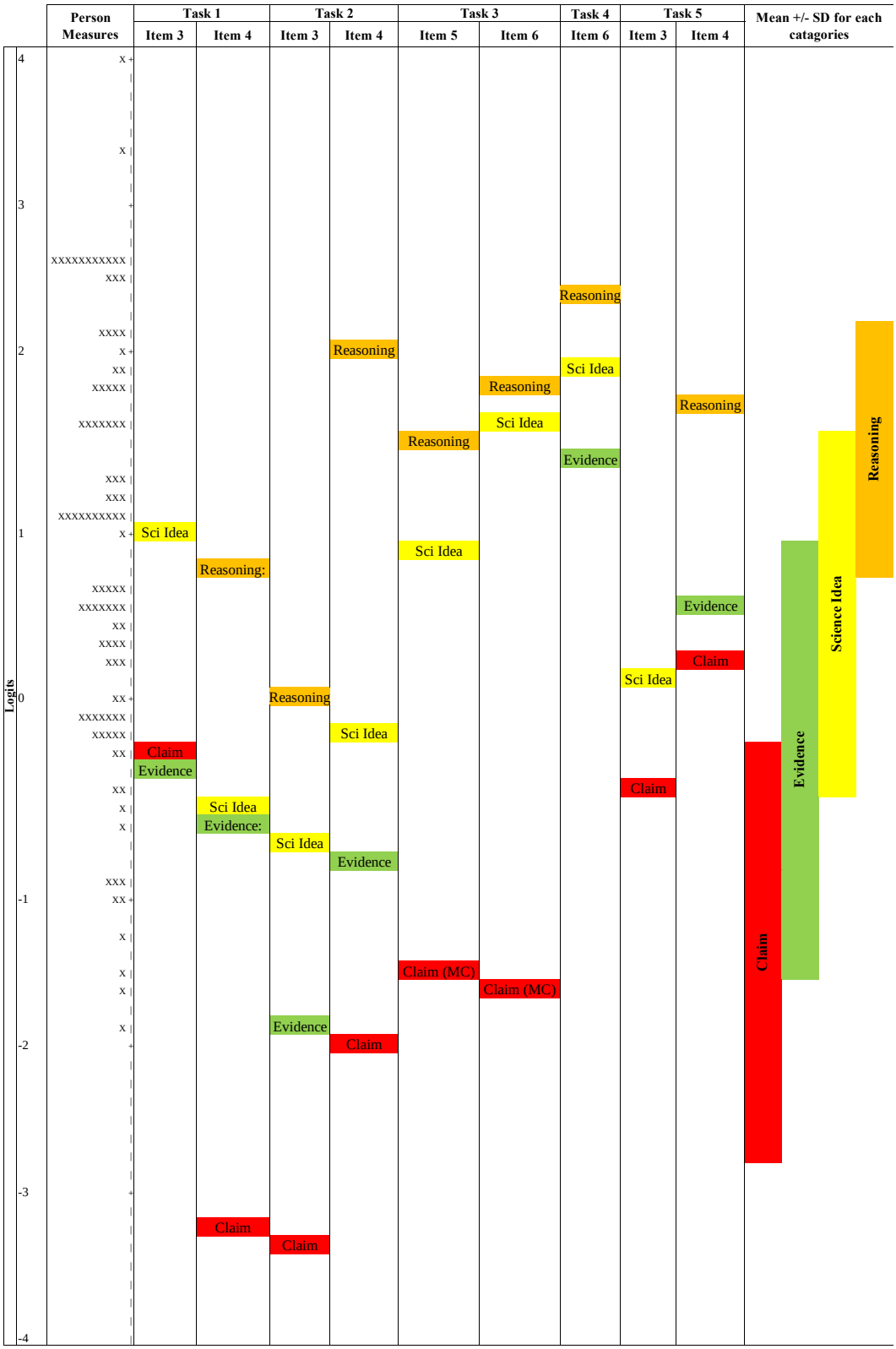

Figure 1: *Wright Map showing the difficulties of Claim, Evidence, Science Idea (labeled as Sci Idea in the figure), and Reasoning categories for nine items from the set of tasks on Chemical Reactions and Energy.*

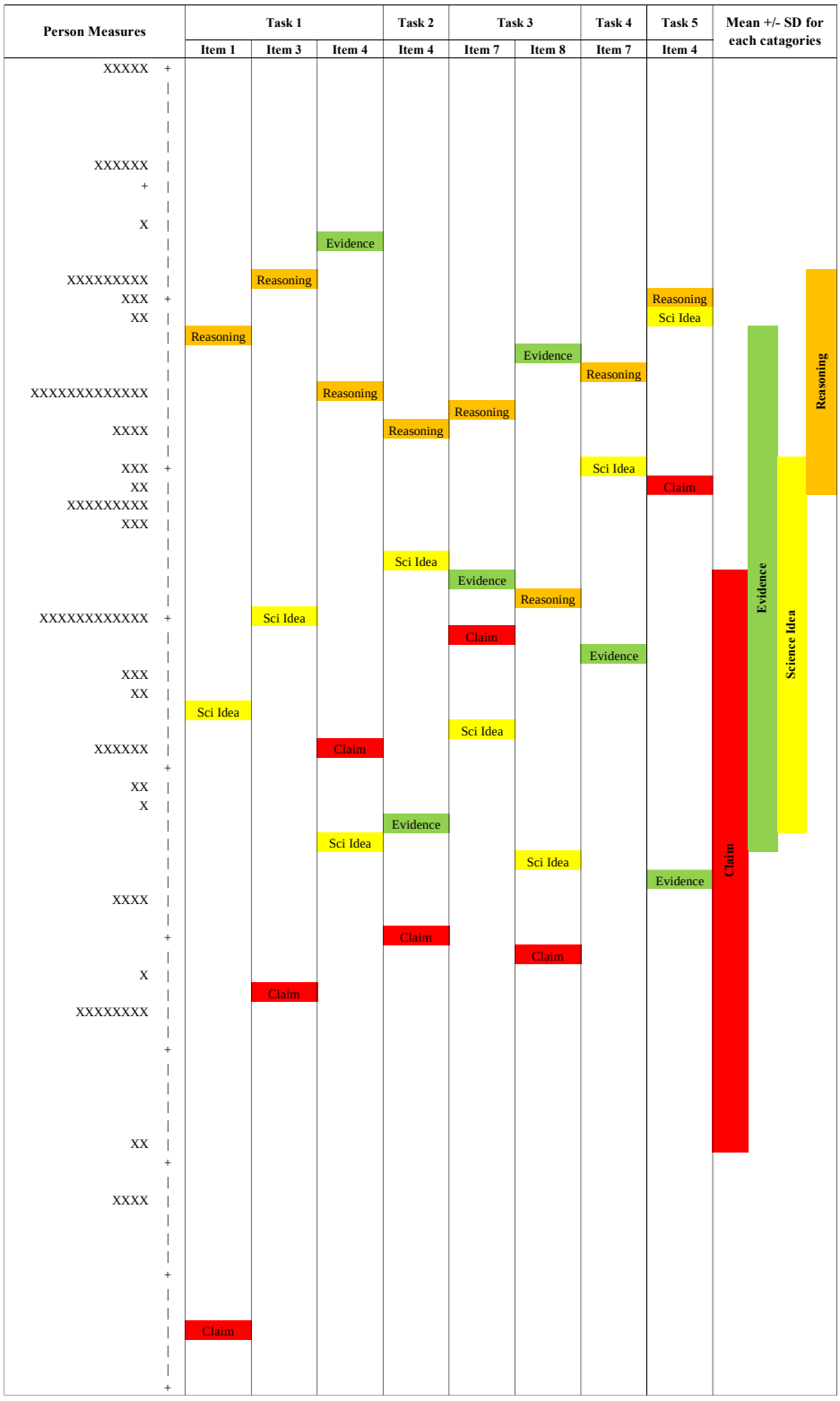

Figure 2: *Wright Map showing the difficulties of Claim, Evidence, Science Idea (labeled as Sci Idea in the figure), and Reasoning categories for eight items from the set of tasks on Kinetic Energy and Forces.*

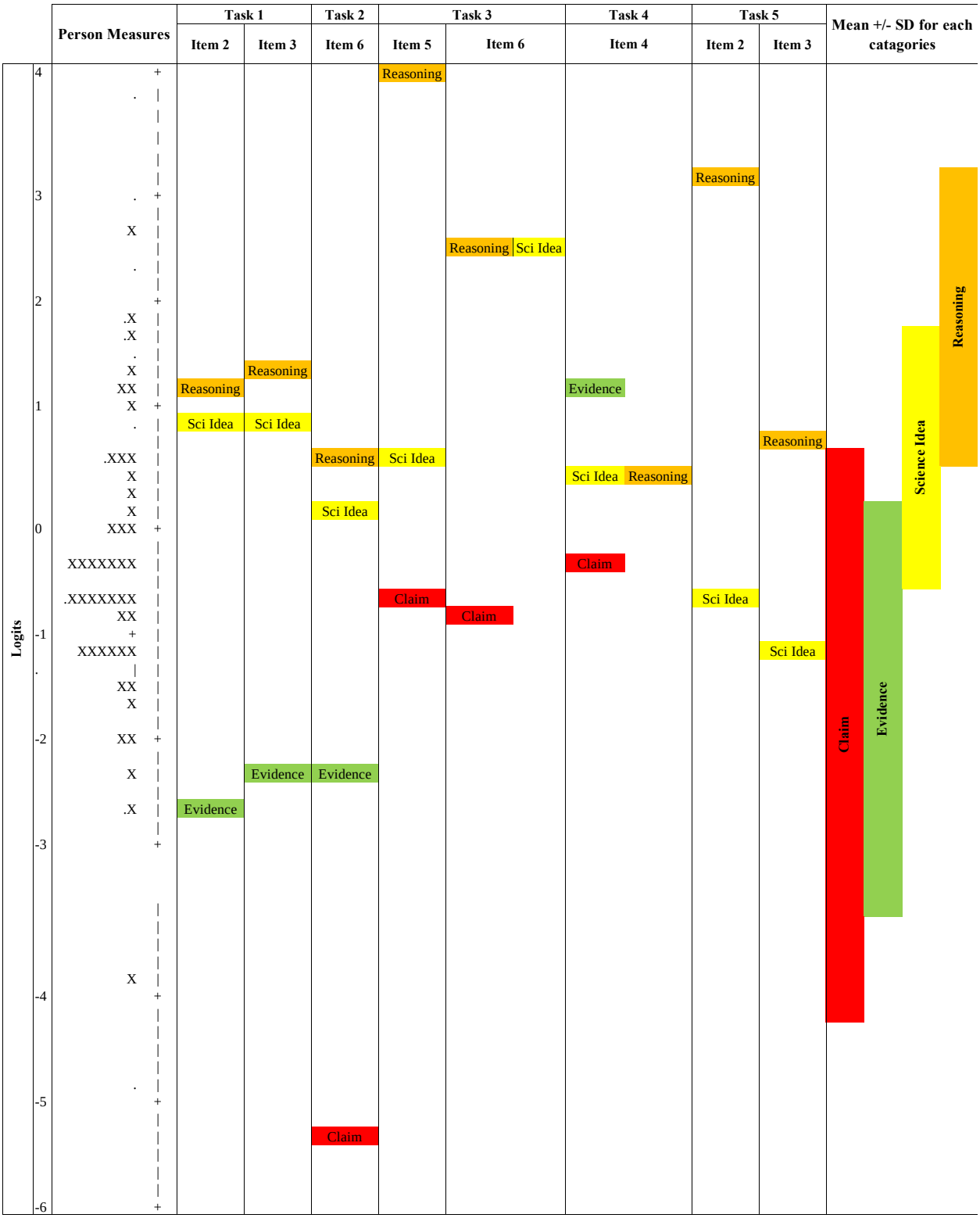

Figure 3: *Wright Map showing the difficulties of Claim, Evidence, Science Idea (labeled as Sci Idea in the figure), and Reasoning categories for nine items from the set of tasks on Thermal Energy.*

Overall, our examination of the difficulties of rubric categories indicated that for many items claim, evidence, science ideas, and reasoning followed a progression in difficulty from easiest to hardest. However, the amount of scaffolding provided, the complexity in the data presented, or the level of practice required can significantly influence the difficulty of the catagories relative to one another.

4. Conclusion

We present an analysis of MTurk data used to investigate the validity of using a claim-evidence-science idea-reasoning framework (CESR) for scoring written arguments and explanations in NGSS-aligned assessment tasks. The data showed that the claim, evidence, science idea, and reasoning categories fit well to a Rasch model and spanned a large difficulty range. For most items the categories followed a progression, with writing and identifying claims being relatively easy and reasoning that was based on the evidence and/or science ideas being the most difficult. Stating and applying science ideas in an argument or explanation was found to be easier than writing complete and coherent reasoning, but more or just as challenging as citing evidence. Some exceptions to this progression were found, and an examination of the rubrics of these exceptions showed that categories could vary in difficulty due to the content knowledge required, the sophistication in the practice required, and the amount of scaffolding provided within the task.

Our results provide support that the tasks are able to measure students along a progress in their ability to write arguments and explanations using energy ideas. Our results also show that the addition of a science idea category to the CER framework provides a distinct category that allows for more direct measures of students' application and statement of relevant science ideas. These science idea categories were also distinct in their difficulty, with most being more difficult than citing evidence but easier than providing coherent reasoning. Their inclusion thus allows the rubric to more accurately locate students who may not be able to write logical and coherent reasoning statements but are able to cite evidence and relevant science ideas.

While our finding provides evidence for the validity of these rubrics for use in NGSS-aligned assessments, there are several limitations to the study that should be highlighted. One is the collection and analysis of adult data instead of data from the intended student population. This limitation was due to difficulties in acquiring student data during the COVID-19 pandemic. We are currently collecting data from students in grades 4 through 12 in order to validate these rubrics with the target student population. In addition, this study was limited to tasks focusing on the topic of energy and does not present any data on the generalizability of the findings to other topics.

5. Significance

Our results pose a new framework for analyzing arguments and explanations that should be of interest to the NARST community. The inclusion of a specific "statement or use of science idea" category in the CER framework allows for a more direct measure of students' content knowledge when scoring written explanations and arguments. Our results also indicate that it is easier for students to state or use science ideas in their explanations than it is for them to write coherent, logical reasoning statements, suggesting that even when they know science ideas that are relevant to the problem, they are still unfamiliar with how to incorporate those ideas into a

coherent argument or explanation. The approach that we have taken may serve as an avenue for other NARST researchers to further study explanation and argumentation learning progressions. Lastly, our approach of integrating DCI, CCC, and SEPs into a rubric framework and analyzing them as a single dimension showcases one method for trying to design, analyze, and evaluate an assessment where the three dimensions have been united into a single construct.

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

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423–429.

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

Hardcastle, J., Herrmann-Abell, C. F., & DeBoer, G. E. (2019). Assessing Students' Ability to create and use Models to Explain Energy-Related Phenomena. In *Paper presented at the NARST 2019 Annual Conference*. Retrieved from https://eric.ed.gov/?id=ED595699

Herrmann-Abell, C. F., Hardcastle, J., & DeBoer, G. E. (2020). Developing NGSS-Aligned Tasks to Assess Elementary School Students' Ability to Explain Energy-Related Phenomena. In *Paper presented at the Annual Meeting of the American Educational Research Association (2020)*. Retrieved from https://eric.ed.gov/?id=ED605234

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.

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