4-2023

Measuring Claim-Evidence-Reasoning Using Scenario-based Assessments Grounded in Real-world Issues

William L. Romine
Wright State University - Main Campus, william.romine@wright.edu

Amy Lannin

Ankita Agarwal
Wright State University - Main Campus

Maha Kareem

Emily Burwell

Measuring Claim-Evidence-Reasoning Using Scenario-based Assessments Grounded in Real-world Issues

William Romine[1], Amy Lannin[2], Ankita Agarwal[3], Maha Kareem[2], Emily Burwell[1]

[1]Department of Biological Sciences, Wright State University, Dayton OH USA

[2]College of Education and Human Development, University of Missouri, Columbia MO USA

[3]Department of Computer Science, Wright State University, Dayton OH USA

Abstract

Improving students' use of argumentation is front and center in the increasing emphasis on scientific practice in K-12 Science and STEM programs. We explore the construct validity of scenario-based assessments of claim-evidence-reasoning (CER) and the structure of the CER construct with respect to a learning progression framework. We also seek to understand how middle school students progress. Establishing the purpose of an argument is a competency that a majority of middle school students meet, whereas quantitative reasoning is the most difficult, and the Rasch model indicates that the competencies form a unidimensional hierarchy of skills. We also find no evidence of differential item functioning between different scenarios, suggesting that multiple scenarios can be utilized in the context of a multi-level assessment framework for measuring the impacts of learning experiences on students' argumentation.

Introduction

The ability to make a claim, support it with evidence, and communicate one's stance using quantitative reasoning is a key component of science literacy aligning with scientific practices (NGSS Lead States, 2013). Claim-Evidence-Reasoning (CER) is an accepted framework for measuring and teaching argumentation (Gotwals & Songer, 2010). However, understanding of how to measure CER, and how students express CER in the context of socioscientific issues (SSIs) needs further work. In particular, given the increasing focus on integrating mathematical reasoning into science classes (NGSS Lead States, 2013), work is needed to integrate a focus on quantitative reasoning into assessments of CER. The purpose of this study is two-fold. First, we describe the creation of scenario-based assessments of CER which integrate quantitative reasoning and explore the construct validity of these assessments for inferring students' levels of CER. Second, we utilize these analyses and qualitative inspection of students' responses to better understand middle school students' mastery of CER.

Background

Berland and McNeil (2010) proposed a progression for argumentation that attempted to account for increasing complexity of both the argumentation product and process. This was built upon by Osborne et al. (2016), which proposed a framework to build and test a learning progression for argumentation using Toulmin's (1958) model. Osborne et al. (2016) use cognitive load theory

(Sweller, 1994) to posit that progressively higher amounts of intrinsic cognitive load are needed to develop and critique arguments with increasingly complex structure; hence the progression moves from lower levels, which focus on simply identifying a claim, to the highest levels where students are evaluating multiple competing arguments and proposing alternative explanations (Osborne et al. 2016). Most recently, a learning progression in argumentation was proposed by Deane et al. (2019): argumentation starts with making a claim and seeing multiple perspectives (Level 1). This then buttresses the process of finding evidence and supporting ideas (Level 2), which in turn buttresses effective communication of the argument in writing (Level 3). Argument evaluation and critique comprise the highest level (Level 4).

Argumentation should not be presented as taking sides in a debate, but more on the premise of a dialogue with evidence and multiple perspectives (Harris, 2017). Although understanding how students negotiate multiple competing arguments is beyond the scope of the present study, we nonetheless seek to use a learning progression measurement perspective which is tied more closely to the original CER framework of Gotwals and Songer (2010) to understand how middle school students construct arguments in the context of real-world social issues. We hypothesize that identifying the purpose of an argument or making a claim is the least complex within the process of argumentation. Students can then begin engaging in the more complex process of identifying the types of evidence that might be needed to support the claim. Once evidence is gathered, students are able to engage in communicating how the evidence supports the claim through the lens of their disciplinary understanding and quantitative reasoning ability.

Methods

Responses from 107 middle school students were analyzed. They hailed from multiple school districts in the Midwestern United States. Each student completed one of two scenarios: *Charleston's Flooding Problem (n = 74)* or *Farm Pollution (n = 33).* These contained 5 parallel items (ordinal 1-4 scale) which measured ability to establish purpose, provide evidence, use quantitative reasoning, understand content, and communicate in writing.

In the interest of facilitating measurement of CER in the context of multiple argumentation tasks embedded within an SSI, we were interested in the following aspects of validity: (1) reliability of the tasks in generating a measure for CER, (2) the efficacy of the tasks in providing a unidimensional CER measure, (3) the efficacy of the tasks in capturing a wide range of CER ability, and (4) measurement consistency of the tasks across multiple real-world scenarios. Reliability and unidimensionality of the tasks were investigated using the Rasch partial credit model (Masters, 1982). The Rasch model is a philosophical approach to evaluation of construct validity in that it models the ideal that the probability of a student achieving a certain level on a competency should be proportional only to the difference between the student's ability and the difficulty of that competency (Wright & Stone, 1979). Concordance of the competency scores to that assumption was evaluated through mean squares fit, where values are expected to range between 0.5-1.5 (Wright et al. 1994). Principal components analysis (PCA) on the model residuals was used to test the assumption of unidimensionality, where a first eigenvalue below 2 is indicative of a unidimensional measure (Raiche, 2005). Ordinal logistic regression was used to evaluate uniform and non-uniform differential item functioning (DIF) (Swaminathan & Rogers,

1990) between the two distinct scenarios. Significant uniform DIF indicates that the competency difficulty changes depending on the scenario, and non-uniform DIF indicates that the efficacy of the competencies to discriminate between high and low levels of CER changes with respect to the scenario. The null hypothesis of no difference was evaluated at the 95% confidence level.

Results

Open-ended responses from 107 students were collected and scored by multiple raters in the context of a scoring event in which raters analyzed and discussed anchor responses together in order to facilitate agreement, and then proceeded to score responses independently using a rubric (Table 1). Each student completed one of two scenarios: *Charleston's Flooding Problem (n = 74)* or *Farm Pollution (n = 33).* The interested reader is encouraged to contact the authors for copies of these SBAs. These contained 5 parallel competencies scored using a rubric developed by the research team (Table 1) on an ordinal 1-4 scale which measured ability to: (1) establish purpose, (2) provide evidence, (3) use quantitative reasoning, (4) understand content, and (5) communicate in writing.

Table 1. Rubric for Scenario-based Assessment of Argumentation.

	4 – Capstone	3 – Milestone	2 – Milestone	1 - Benchmark
Establishing purpose/stating claim	Position/claim is clear and takes into account the complexities of an issue. Other viewpoints are explicitly used in the explanation of the position.	Position/claim acknowledges or somewhat takes into account complexities of an issue. Other viewpoints are acknowledged.	Position/ claim is stated but is simplistic or unclear. Other points of view may be acknowledged.	Position/ claim is absent or off-topic.
Evidence Use of sources to explore issues and analyze evidence	Comprehens-ively analyzes or synthesizes information taken from sources(s) Evidence and viewpoints are referenced and questioned	Analyzes or synthesizes information taken from sources(s) Evidence and viewpoints are taken as mostly fact, with little questioning	Takes information from sources with minimal or no evaluation Viewpoints are taken as fact, without question	Information is not taken from source(s) Writer does not include or question viewpoints

Domain-specific content and vocabulary	Presents appropriate/ accurate use of science/math/ literacy concepts AND discipline-specific vocabulary	Generally accurate but may have some missing ideas or some misconcep-tions of concepts **OR** discipline-specific vocabulary	Response based on misconceptions Minimal or unclear use of domain-specific vocabulary	No clear reference to domain-specific content or vocabulary
Reasoning Interpretation and analysis of sources	Effectively demonstrates reasoning using accurate explanation of information Makes appropriate inferences and conclusions based on and referencing data	Demonstrates reasoning in explaining data and information Makes mostly appropriate inferences and conclusions based on that information	Demonstrates some reasoning in attempts to explain data Makes somewhat appropriate inferences and conclusions based on that information but may make some incorrect conclusions	Reasoning is not yet evident Data is not used or may be merely listed.
Written communication Development of ideas and clarity of expression	Effectively develops ideas and skillfully communicates meaning to readers with clarity and fluency Is almost error-free	Develops ideas that generally convey meaning to readers The writing has few errors	Is developing ideas. Expressions not consistently clear Writing may include some errors	Ideas not yet developed or clear Writing may have many errors

*Adapted from University of Missouri's STEM Literacy Project, Title II, Part A of the Improving Teacher Quality Grant, *Missouri Department of Higher Education* and based on VALUE Rubrics - Association of American Colleges and Universities

The following example of a low CER response shows that this student is able to make a claim. However, the response shows limited understanding of the discipline-specific content in the SBA and limited use of data-driven reasoning.

> *The statement above that says the town Charleston is considerably flooding is inaccurate because it is missing crucial points to prove that there has been considerable water precipitation. One point they fail to put was how much precipitation has increased by from 2005 to this point. They also forgot to mention which type of precipitation was the main cause of the flooding and how they could solve it.*

This response can be contrasted with the following high CER response which shows clear articulation of the claim as well as detailed reference to and understanding of the quantitative data toward both supporting the students' claim and falsifying alternative claims.

> *Charleston's flooding problems are due to the increases in developed land areas with impermeable surfaces. Impermeable surfaces do not allow water to go through to the soil; this will increase water runoff. If the services are distributed throughout the city, the water will flow with change to run towards areas that are permeable (man-made surfaces that allow water to flow through such as baseball fields or golf courses) and naturally occurring areas (parks, forest, fields), causing flooding to occur. In the 40 years of recorded data, the following changes have occurred: It is evident that the vast increase in developed-impermeable surfaces as compared to changes in the developed permeable in natural areas is the over-riding factor in the flooding problem. The local meteorologist has stated in error that the flooding problem is due to increased precipitation.*

As a measure of argumentation within the lens of CER, the scale showed acceptable reliability ($r_{person} = 0.96$) and unidimensionality (1^{st} eigenvalue from PCA on Rasch residuals = 1.57 < 2). All five competencies displayed satisfactory fit with the Rasch partial credit model and had point biserial correlations above 0.7 (Table 2). It was interesting that the competency focusing on integration of content understanding had higher-than-expected fit (infit = 0.53, outfit = 0.36); a mark of high discrimination, which indicates potential bias in favor of students with higher levels of CER (Masters, 1988). Although this is a sign of potential multidimensionality of the item, this was not picked up in the PCA on Rasch residuals, which means there is no evidence that it measures specific factors extraneous to argumentation. Additionally, the ordinal logistic regression procedure revealed no significant uniform or non-uniform DIF between the two scenarios. This was confirmed by p-values above 0.05 for the scenario factor (difficulty) and the scenario-by-measure interaction (discrimination) in the models for the responses on each competency. This indicates that any biases we see in the assessment are likely due to scores on the competencies themselves as opposed to the scenarios in which they are embedded.

The Wright map (Figure 1) shows a difficulty spread of over 2 logits, meaning the competencies challenge students across a wide range of CER ability. Difficulty was defined as the center of the scoring scale; the point at which a student with an ability level at that scale location had an equal chance of scoring at the highest (capstone) and lowest (benchmark) levels on the rubric (see Table 1). Establishing purpose was the easiest competency for students to meet: 53.5% of the students had ability levels that were equal to or above the difficulty level of this competency. The most difficult of the competencies was the integration of quantitative reasoning into arguments, which only 40.2% of the students met or exceeded.

Table 2. Rasch difficulty and fit indices. '% Achieved' indicates the percentage of students who met or exceeded the difficulty level of the competency. Competencies are listed from least to most difficult.

Competency	Difficulty	SE	Infit	Outfit	% Achieved
Purpose	-1.34	0.28	0.97	0.99	53.3
Evidence	-0.49	0.30	1.34	1.09	48.6
Content	0.44	0.29	0.53	0.36	43.9
Communication	0.48	0.26	0.70	0.48	43.9
Quantitative	0.90	0.28	1.05	1.36	40.2

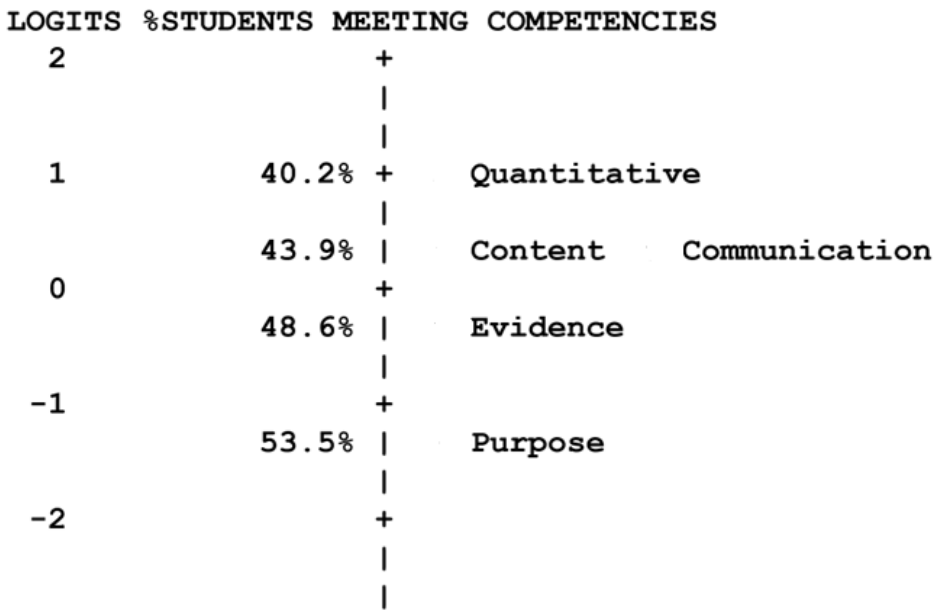

Figure 1: Person-item map (Wright map) showing the difficulty of the competencies in logits (right side) and the percentage of students who met or exceeded the difficulty level of that competency (left side).

Discussion

The CER scale (Table 1), in the context of the two SBAs used, has the desirable properties of high reliability and unidimensionality which make it promising for teachers or researchers needing a parsimonious argumentation measure. Nonetheless, the closer-than-expected fit with the Rasch model for the Content and Communication competencies suggests the tasks may also be measuring content knowledge and writing ability in addition to the core CER construct. This is common in assessments designed to assess higher-level literacy skills which themselves are scaffolded upon related skills (Masters, 1988). Whether or not this is a problem is a matter of perspective. On one hand, a student's content knowledge and ability to communicate can be viewed as necessary for argumentation. But on the other hand, we would like to utilize measures for argumentation that are as independent as possible from other literacy competencies such as writing ability and content knowledge which are positively correlated with CER in order to facilitate accurate inferences regarding the true correlation of these constructs and causal factors from experimental studies. Although these biases are small and may not be harmful to the scale, they nonetheless deserve attention both for future SBA development and training of scorers to recognize and attempt to mitigate these types of biases.

Finally, the data suggest that the measures are repeatable across multiple types of SSIs. Although only two SSIs were compared in this study, and hence more work is needed to substantiate this claim, it is nonetheless promising in that opens up the opportunity to develop scenarios that are unique to the disciplinary context of a specific curricular intervention. In addition, the ability to equate the tasks across multiple scenarios creates opportunities for application of multi-level assessment frameworks which have been shown to be informative in evaluating outcomes associated with SSI-based curricula (Sadler, et al. 2013).

Conclusion

Understanding how students process and communicate scientific information is more important than ever given that the focus on combatting misinformation is increasing (Sharon & Baram-Sabari, 2020). Toward understanding CER as a progression, scholars interested in learning progressions may be interested in seeing significant overlap with Osborne et al. (2016) in that the competencies tend to progress from less complex declarative processes (stating a claim or purpose) to more complex practices (communication, application of content, and use of quantitative reasoning). This work builds on Osborne et al. (2016) and Deane et al. (2019) in our framing of argumentation in terms of negotiation of SSIs. Further, the SSI-based scenarios encourage students to use data to negotiate information and competing perspectives which is central in the NGSS practices and a key aspect of literacy that can be gained from SSI-based curricula in general (Romine, Sadler, & Kinslow, 2017; Romine, Sadler, Dauer, & Kinslow, 2020).

Acknowledgements

This research was funded by National Science Foundation DRK-12 grant #2010312. The views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education*, *94*(5), 765-793.

Deane, P., Song, Y., van Rijn, P., O'Reilly, T., Fowles, M., Bennett, R., ... & Zhang, M. (2019). The case for scenario-based assessment of written argumentation. *Reading and Writing*, *32*, 1575-1606.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education*, *94*(2), 259-281.

Harris, J. (2017). *Rewriting: How to do things with text (2nd ed.*). Boulder, CO: Utah State University Press.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, *25*(1), 15-29.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

NGSS Lead States. (2013). *Next generation science standards: For states, by states.* Achieve, Inc. on behalf of the twenty-six states and partners that collaborated on the NGSS. Washington, DC: National Academy Press.

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*, *53*(6), 821-846.

Raiche, G. (2005). Critical eigenvalue sizes in standardized residual principle components analysis. *Rasch Measurement Transactions, 19*, 1012.

Romine, W. L., Sadler, T. D., & Kinslow, A. T. (2017). Assessment of scientific literacy: Development and validation of the Quantitative Assessment of Socio-Scientific Reasoning (QuASSR). *Journal of Research in Science Teaching*, *54*(2), 274-295.

Romine, W. L., Sadler, T. D., Dauer, J. M., & Kinslow, A. (2020). Measurement of socio-scientific reasoning (SSR) and exploration of SSR as a progression of competencies. *International Journal of Science Education*, *42*(18), 2981-3002.

Sadler, T. D., Romine, W. L., Stuart, P. E., & Merle-Johnson, D. (2013). Game-based curricula in biology classes: Differential effects among varying academic levels. *Journal of Research in Science Teaching*, *50*(4), 479-499.

Sharon, A. J., & Baram-Tsabari, A. (2020). Can science literacy help individuals identify misinformation in everyday life? *Science Education*, *104*(5), 873-894.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295-312.

Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge University Press, Cambridge.

Wright, B.D., & Stone, M.A. (1979). *Best test design.* Chicago, IL: MESA Press.

Wright, B.D., Linacre, J.M., Gustafson, J.E., & Martin-Loff, P. (1994). Reasonable mean square fit values. *Rasch Measurement Transactions, 8*(3), 370.