Practical Assessment, Research, and Evaluation

Volume 24 *2019* Article 8

2019

Using Rater Cognition to Improve Generalizability of an Assessment of Scientific Argumentation

Katrina Borowiec

Courtney Castle

Follow this and additional works at: https://scholarworks.umass.edu/pare

Recommended Citation

Borowiec, Katrina and Castle, Courtney (2019) "Using Rater Cognition to Improve Generalizability of an Assessment of Scientific Argumentation," *Practical Assessment, Research, and Evaluation*: Vol. 24 , Article 8.
Available at: https://scholarworks.umass.edu/pare/vol24/iss1/8

This Article is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Practical Assessment, Research, and Evaluation by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 24 Number 8, November 2019 ISSN 1531-7714

Using Rater Cognition to Improve Generalizability of an Assessment of Scientific Argumentation

Katrina Borowiec, *Boston College*
Courtney Castle, *Woodrow Wilson Graduate School of Teaching and Learning*

Rater cognition or "think-aloud" studies have historically been used to enhance rater accuracy and consistency in writing and language assessments. As assessments are developed for new, complex constructs from the *Next Generation Science Standards (NGSS)*, the present study illustrates the utility of extending "think-aloud" studies to science assessment. The study focuses on the development of rubrics for scientific argumentation, one of the NGSS Science and Engineering practices. The initial rubrics were modified based on cognitive interviews with five raters. Next, a group of four new raters scored responses using the original and revised rubrics. A psychometric analysis was conducted to measure change in interrater reliability, accuracy, and generalizability (using a generalizability study or "g-study") for the original and revised rubrics. Interrater reliability, accuracy, and generalizability increased with the rubric modifications. Furthermore, follow-up interviews with the second group of raters indicated that most raters preferred the revised rubric. These findings illustrate that cognitive interviews with raters can be used to enhance rubric usability and generalizability when assessing scientific argumentation, thereby improving assessment validity.

Scoring rubrics are routinely used in educational assessment. Popham (1997) defines a rubric as a "scoring guide used to evaluate the quality of students' constructed responses" (p. 72). Rubrics provide qualitative descriptions for each scoring category (Moskal, 2000) and contain three components: criteria which explain what aspects of the response are to be evaluated; definitions which provide a description of the qualitative characteristics of responses in each scoring level; and a scoring strategy (Popham, 1997). In general, there are two scoring strategies: a holistic rubric instructs raters to provide one overall assessment of a student's performance on an item, while an analytic rubric asks raters to provide separate scores for multiple dimensions of a student's response (Popham, 1997; Wolfe & Song, 2016). In the context of scientific argumentation, a holistic rubric might instruct raters to provide one score for the overall argument, while an analytic rubric might ask raters to provide separate scores for the claim, evidence, and reasoning. Some rubrics might combine holistic and analytic scoring by asking raters to provide multiple scores along specific dimensions, in addition to an overall holistic judgment of quality (Moskal, 2000).

Rating Quality

Rating quality is an important consideration when evaluating the validity and reliability of students' scores on constructed response items. Wolfe and Song (2016) define "rating quality" as "the degree to which a set of scores is precise and unbiased" (p. 109). Therefore, measurement error is minimal when the rating quality is high, since there is a high degree of alignment between students' actual scores and the criteria specified in the rubric (Wolfe & Song, 2016).

Sources of systematic measurement error attributable to raters are known as "rater effects" (Wolf

& Song, 2016, p. 109). Zhang (2013) identified six major areas of rater effects:

1. *Scale shrinkage* occurs when raters use only a subset of possible scores.
2. *Inconsistent scoring* happens when raters provide erratic scores.
3. *Halo effects* occur when raters make generalizations about a student's performance based on other responses provided by that student.
4. *Stereotyping* occurs when raters' hold biases toward a particular group.
5. *Perception differences* occur when raters' scores are not independent due to comparison of responses among students.
6. *Rater drift* happens when raters apply different scoring criteria over time.

An alternative model for understanding rater effects is Wolfe and Song's (2016) Rater Quality Framework which includes three categories of rater effects. The first category is *rating context*, which includes the location of the rater training and scoring; rater monitoring procedures; and the format of the scoring rubric. The second category is the *characteristics of students' response*, such as the essay topic, student handwriting, and raters' biases toward particular populations of test-takers. The third and final category is *rater characteristics*, which includes raters' prior scoring experience, raters' content knowledge, raters' mood during scoring, and raters' cognition and understanding of the rubric.

Testing programs monitor rater performance throughout the testing process. Two frameworks can be utilized to evaluate rater performance: rater agreement and rater accuracy (Wolfe & Song, 2016). Rater agreement is evaluated using inter-rater reliability indicators (e.g., percent agreement, Cohen's Kappa, and the ICC) to evaluate consistency among raters. In comparison, using the rater accuracy framework, content experts will score a set of validation items, and these items will be included in the set of responses scored by raters. The rater's scores on these validation responses are compared to expert scores as a measure of accuracy. Measures of quality from the rater agreement framework provide a measure of reliability, while those from the rater accuracy framework provide a measure of validity.

Rater Cognition

While automated scoring procedures are commonly used for selected response items, human raters are still frequently used for constructed response items. Even when automated essay scoring is used, this typically occurs after using human raters to calibrate the scores (Wolfe & Song, 2016). Given the important role of raters in assessment, it is critical that they understand how to properly employ the scoring guide or rubric.

Assessment professionals rely on four assumptions about raters' understanding and application of the scoring guide or rubric to justify the validity of scores (Myford, 2012). First, assessment professionals assume that all raters use the scoring guide or rubric in a similar way; thus, students' scores are not dependent on the particular rater assigned to their response. When this assumption is violated, students with the same ability will receive systematically different scores, depending on the rater who scored their response. Second, raters are assumed to understand the rubric and its categories and apply them appropriately. Third, assessment professionals assume that raters are not swayed by construct-irrelevant factors. Finally, raters are assumed to score responses consistently over time and evaluate each student's response independently. Yet, as Myford (2012) explains, the assessment field is in its infancy with respect to collecting the types of information needed to verify these assumptions.

Research designed to evaluate raters' understanding of scoring guides and rubrics is referred to as rater cognition research, which involves "gaining an understanding of raters' thought processes as they score different types of performances and products, striving to understand how raters' mental representations and the cognitive strategies and rating styles they employ influence their judgments" (Myford, 2012, p. 48). Bejar (2012) traced the first major phase of research on rater cognition to Fechner's (1897; as cited by Bejar, 2012) early evaluation of the processes people use to evaluate art. Bejar identified Edgeworth (1890; cited in Bejar, 2012) as the first person to recognize that raters' judgements were prone to error. The second major phase of rater cognition research involved "think-aloud" studies, in which raters were asked to verbalize their thoughts as they scored students' responses.

Two models for understanding rater cognition are Bejar's (2012) Rater Cognition Model and Crisp's (2012)

Behavioral Response Model. Bejar views rater cognition as the process by which raters receive scoring training, encode the scoring rules into a "mental scoring rubric" in their mind, and use this mental scoring rubric to code responses (p. 4). Scoring problems occur when raters misunderstand the subtleties in the scoring guide and encode the scoring guide incorrectly. Alternatively, Crisp's model includes six behavioral processes which occur iteratively. During the first process, *Planning and Orientation*, raters familiarize themselves with the specific topic being assessed and remind themselves about the features they should evaluate in the response. For the second process, *Reading and Understanding*, raters read and decipher meaning from the response. During the *Task Realization* stage, raters evaluate whether the response demonstrates an effort to respond to the task. The fourth process is *Social and Emotional*, in which raters might express an emotional reaction to the student's essay or express personal musings about why a student responded in a particular way. The fifth process involves Concurrent Evaluations, meaning that the rater evaluates aspects (e.g., grammar) of the response as they read the essay. Finally, the sixth process, *Overall Evaluation/Score Consideration*, involves the rater applying the scoring criteria to assign a final score. These processes are not sequential in this model, since the rater is presumed to proceed through her evaluation in an iterative process (Crisp, 2012).

Previous studies of raters' experiences scoring constructed response items have primarily focused on general expository writing (Vaughan, 1991), tests for English language learners such as the TOEFL or other placement tests (e.g., Barkaoui, 2010; Cumming, 1990; Cumming et al., 2000), and foreign language tests (e.g., Deygers & Van Gorp, 2015). For instance, Vaughan (1991) conducted a "think-aloud" study in which nine raters were each asked to score six essays. The results indicated that raters varied with respect to the characteristics of the essays that they focused on (e.g., grammar, organization), with some raters focusing on construct-irrelevant features of the essays, including handwriting, whether the essay was boring or amusing, and whether the essay was offensive. Moreover, raters sometimes made assumptions about the characteristics of the student writing the essay (e.g., whether the student was an English language learner). Cumming, Kantor, and Powers (2002) found differences in raters' focus on features of students' essays, depending on whether the rater was a native English speaker.

Scientific Argumentation

Scientific argumentation is a critical component of scientific literacy (McNeill & Krajkik, 2008), and the NRC's (2012) *Framework for K-12 Science Standards* includes "engaging in argument from evidence" (p. 49) as one of eight essential science practices. Argumentation has been described as "the attempt to establish or prove a conclusion on the basis of reasons" (Norris, Phillips, & Osborne, 2008, p. 90).

Berland and McNeill (2010) adapted Toulmin's (1958) argumentation framework to propose a model for scientific argumentation that includes four components: the claim, evidence, reasoning, and rebuttal. A claim is a student's assertion or conclusion provided for a given prompt (Berland & McNeill, 2010; McNeill & Krajkik, 2008). *Evidence* are scientific data (i.e., information obtained firsthand from an investigation or secondary data) used to support a claim. *Reasoning* is the rationale for why a claim is logical, and involves creating a link between the claim, evidence, and relevant scientific principles (Berland & McNeill, 2010; McNeill & Krajkik, 2008). The *rebuttal*, which provides evidence and reasoning to explain why a counterclaim(s) is implausible, is the fourth component in their argumentation framework.

The progression of student argumentation can be evaluated in terms of the complexity of their scientific discourse (Berland & McNeill, 2010). Berland and McNeill explain that students' scientific arguments should be evaluated with respect to whether they include the necessary structural components of an argument and the appropriateness of the content. The structural components include the rationale and the rebuttal. The rationale of more advanced arguments will include both evidence and reasoning, while less complex arguments tend to provide only evidence. Rebuttals are typically only observed during more advanced stages of students' argumentation progression, and are typically observed in students in grades five through 12. (This study, which focuses on assessing argumentation in upper elementary students, does not include the rebuttal component.)

When evaluating the content components of an argument, Berland and McNeill (2010) note that more advanced arguments will include a causal explanation of varying levels of complexity. Additionally, more advanced arguments will contain evidence, reasoning, and a rebuttal that are appropriate for the given situation,

without any irrelevant information. More complex arguments will include evidence, reasoning, and rebuttals that are not only appropriate, but are also sufficient, meaning that "the quantity or complexity of the evidence, reasoning, and/or rebuttal is able to convince an audience of the claim" (p. 774).

Research Problem

Raters' ability to apply scoring rubrics as intended by assessment developers is critical to the validity of test scores. To maximize scoring reliability, it is important to understand how raters understand the construct as represented by the rubric, and where they experience confusion or misinterpretation. Examining rater cognition is one way to refine a rubric, by uncovering and addressing discrepancies between raters' understanding and the intended operationalization of the construct. Studies of rater cognition have been utilized to refine language and writing assessment for decades (e.g., Cumming, 1990; Cumming, Kantor, & Powers, 2002; Vaughan, 1991; Zhang, 2016).

In the current study, cognitive interviews were used to provide information about raters' interpretation and use of a scoring rubric to rate scientific arguments. Assessment of scientific practices like argumentation (National Research Council [NRC], 2012; NGSS Lead States, 2013) is less common than assessment of science content, so less is known about how to validly and reliably assess these constructs. In particular, interrater reliability on constructed response items assessing scientific argumentation has been shown to be poor (Castle, 2018). Cognitive interviews were used to inform rubric revisions, and the generalizability, interrater reliability, and accuracy of raters' scores using the original (Rubric 1) and revised (Rubric 2) rubrics were then compared to determine whether there was an improvement.

This study examined three related research questions (RQs):

RQ1) Can cognitive interviews provide insight about how raters make judgments while scoring the quality of scientific arguments?

RQ2) Did rubric modifications based on this information improve the generalizability, interrater reliability, and accuracy of argumentation scores?

RQ3) Did rubric modifications based on this information improve raters' reported ease using the rubric?

The first and third research questions were investigated using cognitive interviews with raters. A generalizability theory framework was used to evaluate reliability (Brennan, 2000; Shavelson, Webb, & Rowley, 1989) in tandem with the intraclass correlation coefficient (ICC), a standard measure of reliability, while percent agreement with expert scores was used to evaluate rater accuracy (Wolfe & Song, 2016).

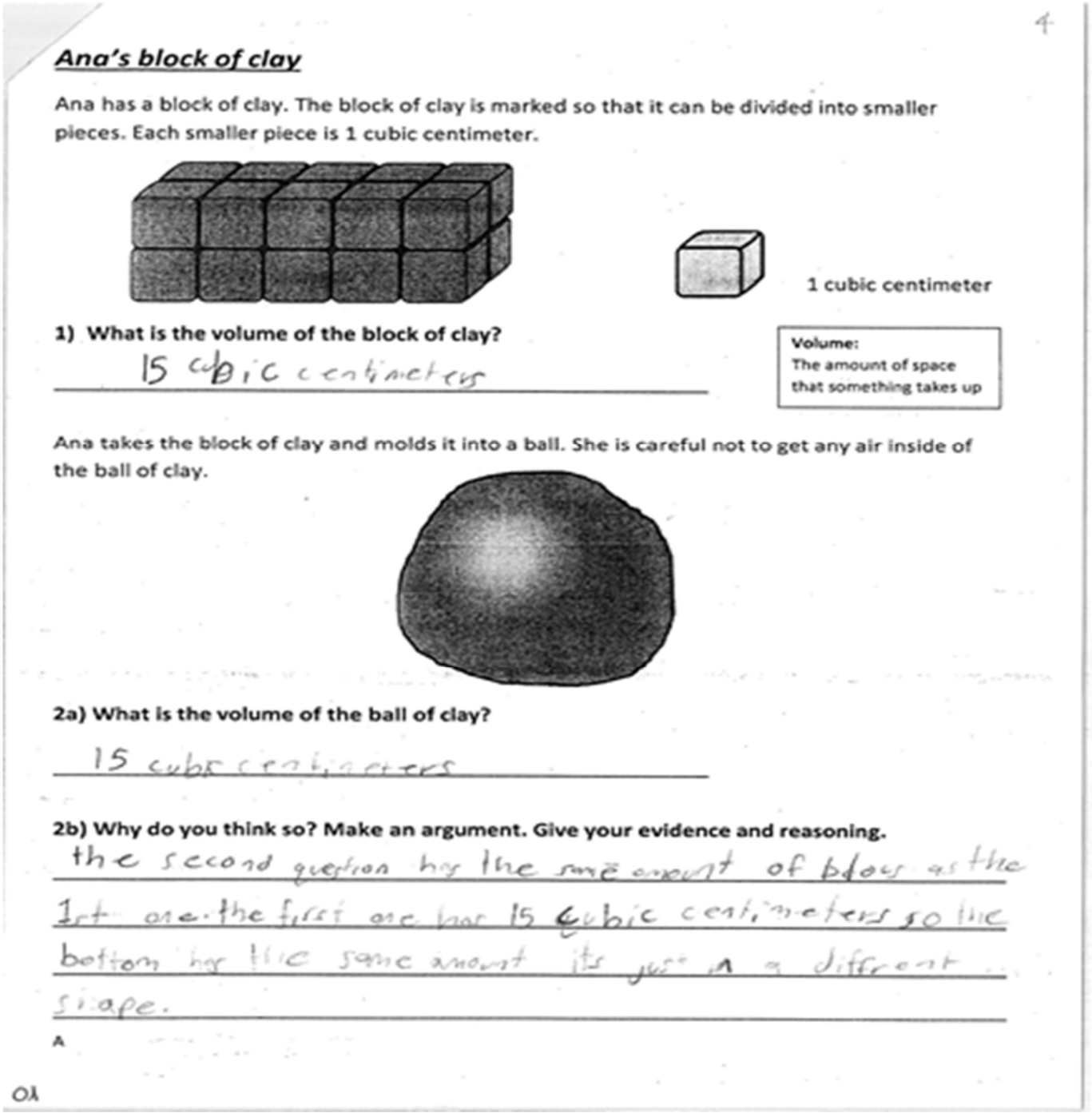
Methods

Instrument

The items examined in this study come from an assessment measuring elementary (grades 4-6) students' understanding of concepts and practices from *the Next Generation Science Standards* (NGSS Lead States, 2013). All items were multidimensional, multicomponent items (National Research Council, 2014) assessing students' understanding of one Disciplinary Core Idea (Matter), one Crosscutting Concept (Scale, Proportion, and Quantity), and one Science and Engineering Practice (SEP; Engaging in Argument from Evidence). The three dimensions were assessed together in the context of a common scenario grounded in a scientific phenomenon, but with a separate prompt for each dimension. An example item can be found in Figure 1. All items were vetted by content experts prior to administration.

Study Design and Analysis

Overview. The overall research study was comprised of three sub-studies: the Initial Cognitive Interview Study, the Generalizability and Rater Accuracy Study, and the Follow-up Cognitive Interview Study. These three sub-studies were based on foundational research from Castle's (2018) dissertation. Table 1 presents an overview of the research design. Each sub-study will be discussed in greater detail below.



4

Ana's block of clay

Ana has a block of clay. The block of clay is marked so that it can be divided into smaller pieces. Each smaller piece is 1 cubic centimeter.

1 cubic centimeter

1) What is the volume of the block of clay?

15 cubic centimeters

Volume:
The amount of space that something takes up

Ana takes the block of clay and molds it into a ball. She is careful not to get any air inside of the ball of clay.

2a) What is the volume of the ball of clay?

15 cubic centimeters

2b) Why do you think so? Make an argument. Give your evidence and reasoning.

the second question has the same amount of blocs as the 1st one. the first one has 15 cubic centimeters so the bottom has the same amount its just in a different shape.

A

01

Figure 1. Item 1. Students are first asked to calculate the volume of a block of clay, then to make a claim about the volume after the block is rolled into a ball and provide an explanation. This student's explanation reads, "the second question has the same amount of blocs [sic] as the 1st one. the [sic] first one has 15 cubic centimeters so the bottom has the same amount its [sic] just in a different shape."

Table 1. Research Design Overview

Sub-Study	Research Question Addressed	Analytic Approach
Pre-Study (Castle, 2018)	RQ0. What is the interrater reliability when scoring scientific arguments?	Intraclass correlation coefficients
Initial Cognitive Interview Study (Sub-Study 1)	RQ1. Can cognitive interviews provide insight about how raters make judgments while scoring the quality of scientific arguments?	Cognitive interviews
Generalizability and Rater Accuracy Study (Sub-Study 2)	RQ2. Did rubric modifications based on this information improve the generalizability, interrater reliability, and accuracy of argumentation scores?	Generalizability analysis, Intraclass correlation coefficients, percent agreement with expert raters
Follow-up Cognitive Interview Study (Sub-Study 3)	RQ3. Did rubric modifications based on this information improve raters' reported ease using the rubric?	Cognitive interviews

*Note: One group of raters participated in the pre-study and the first sub-study, while a separate group of raters participated in sub-studies 2 and 3.

Pre-study. Sub-studies 1 through 3 are based on Castle's (2018) dissertation research, for which 11 items were administered via matrix sampling to a pilot sample of 369 students in grades 4-6 from several northeastern public-school districts. Responses were scored by a group of seven raters, such that two raters scored each item. All raters were graduate students in educational measurement at an institution in the northeast United States and each received a stipend for their efforts. An initial analysis revealed that interrater reliability was low on the Engaging in Argument from Evidence dimension; absolute ICC's ranged from 0.49 to 0.78 for two raters. The observed low interrater reliability was the impetus for the sub-studies that followed, beginning with the Initial Cognitive Interview Study.

Initial cognitive interview study. The first sub-study investigated RQ1. Cognitive interviews involve participants recounting their thoughts from an experience of interest, allowing researchers to gather critical information about their cognitive processes and interpretations. Cognitive interviewing has its roots in cognitive psychology (Ericsson & Simon, 1993), and it can be used as a tool to investigate a great variety of research problems. A common application of the cognitive interviewing technique is in the field of measurement, where it is used to improve the quality of survey instruments (Willis, 1999). In educational measurement the cognitive interview has been used with great success to support the development and validation of knowledge and attitudinal assessments (e.g., Almond et al., 2009). The exercise helps assessment and survey developers determine whether an item contains any sources of confusion that could inhibit an examinee or survey respondent from completing the task in the intended manner.

Popular cognitive interviewing tactics include think-alouds (Ericsson & Simon, 1993) and probing (Willis, 1999). In this study, the cognitive interviews utilized a combination of think-aloud and probing techniques. The think-aloud method involves minimal input from the interviewer, therefore reducing opportunities for the interviewer to introduce bias. Furthermore, the open-ended nature of the method leads to a possibility of unanticipated information from the interviewee, and the interviewee's verbalization occurs as they experience the item which makes their stream of consciousness a more "pure" data source than a retroactive report (Beatty & Willis, 2007; Willis, 1999). The open-endedness can also be a drawback if participants have difficulty providing the focus and/or amount of detail desired by the interviewer (Beatty & Willis, 2007; Willis, 1999). Therefore, a semi-structured interview format (Arksey & Knight, 1999) was used to ensure that certain topics were covered, while providing the raters and interviewers flexibility to address new areas of interest that arose during the interview.

Cognitive interviews were conducted with five raters who scored scientific argumentation items from the aforementioned instrument. The interviews took part after scoring was finished. Participants received a $10 Amazon gift card for their participation. During the

think-aloud portion of the interview, each rater was presented with three student responses which they had previously scored. (None of the raters reported remembering their previous scoring decisions.) They were asked to examine the response, and verbalize their thought process as they considered how they would score it. All raters were asked questions about their experience including:

1) "Can you tell me what you were thinking about when you scored this student's response?"
2) "How did you choose score X?"
3) "Why didn't you choose score Y?"
4) "Can you identify any ways in which the scoring guide could be improved, based on your experience scoring this student's response?"

Additional follow-up questions were asked based on each rater's specific responses. For instance, if evidence was mentioned, the rater might have been asked, "Why do you consider <text from student response> evidence?" Each interview lasted between 30 minutes and one hour.

With each participant's permission, all interviews were audio recorded and reviewed by the researchers to distill themes. Based on the analysis, the rubrics were modified to rectify common areas of confusion. The original and revised rubrics can be found in Appendix A and Appendix B, respectively.

Generalizability and rater accuracy study. The pre-study had some design constraints that prevented an examination of rater effects; notably, there were no items scored by more than two raters. Thus, the dataset was not suitable for comparisons across raters. To answer RQ2, a follow-up generalizability and rater accuracy study was conducted to examine how rubric modifications impacted scoring.

This study was designed and analyzed in accordance with generalizability theory (Brennan, 2000; Shavelson, Webb, & Rowley, 1989), which allows researchers to evaluate the breakdown of score variance due to multiple sources such as raters and items, apart from error variance. Four new raters scored 84 responses on four items and received a small stipend for their efforts. The raters were graduate students in education at an institution in the northeast United States. These raters scored all responses twice; using the original rubric (Rubric 1), and the revised rubric (Rubric 2). The modified rubric, based on results of the cognitive interviews, provided greater clarity in the form of new scoring categories and detailed examples. Thus, the carry-over effect from the modified rubric was expected to be greater than that of the original rubric. Therefore, all raters used Rubric 1 first, then rescored the same dataset using Rubric 2. Interrater reliability for both rubrics was measured by an intraclass correlation coefficient (ICC (2,1); Shrout & Fleiss, 1979) and calculated with SPSS (IBM Corp., 2017).

mGENOVA version 2.1 (Brennan, 2001) was used to conduct a generalizability study (g-study), including the computation of variance and covariance components, absolute and relative error variances, and generalizability coefficients. A multivariate design was used, $p^{\bullet} \times i^{\bullet} \times r^{\bullet}$, with two random facets of generalization (item and rater) both fully crossed with one fixed facet (rubric). This means that all levels of each random facet were the same across both levels of the fixed facet; the items and raters were exactly the same for both rubrics. The item and rater facets were random, indicating that each item and rater is considered to be randomly sampled from a universe of items and raters, each with an equal probability of being selected. The rubric facet is fixed, however, meaning that these two particular rubrics (the original and modified rubrics) are the only levels of interest for this variable.

Additionally, the accuracy of raters' responses was evaluated by comparing raters' scores to expert scores. (The 'experts' in question were the authors.)

Follow-up cognitive interview study. After scoring was completed, the second set of raters participated in another round of cognitive interviews to investigate their perception of the two rubrics. Raters received a $10 Amazon gift card for participating in the interview.

Similar to the first round of interviews, raters were asked to "think-aloud" as they scored three items using the first rubric and then again using the second rubric. After answering the questions from the first interviews, the second group of raters was asked to make general judgments about each rubric, such as:

1. "Which of the scoring rubrics did you generally prefer and why?"

2. "Did it take more time to score responses using one of the rubrics compared to the other?"

When appropriate, the interviewer used some additional probes to clarify or react to the raters' comments. One author conducted all the interviews, each lasting between 30 minutes and one hour.

With each participant's permission, all interviews were audio recorded and reviewed to distill themes.

Results

Initial Cognitive Interview Study

Based on the interviews with the raters from the initial study, five central themes emerged:

1) Difficulty separating correctness from the quality of argumentation.
2) Determining whether mathematical reasoning counts as "reasoning."
3) Reasoning and evidence intertwined or implicit within the argument.
4) Students' (in)appropriate use of causal language.
5) Importance of the examples provided in the rubric.

Theme 1: Separating correctness from the quality of argumentation. The original rubric (Rubric 1) required raters to score arguments with respect to the quality of their evidence and reasoning. The rubric explains that "Evidence and reasoning do not necessarily have to support a correct answer, but they should support the chosen answer." The intention was for raters to separate the correctness from the argumentation; the rationale being that students can provide a quality argument even if the underlying conceptual understanding is flawed. Once they had assigned a score to the quality of the argumentation, raters were also asked to assign a separate score to the correctness of the argument.

Nevertheless, all five raters expressed difficulty separating the correctness of the response from the quality of argumentation. For instance, Helen describes essentially disregarding the rubric guidelines due to the importance she placed on correctness:

> *If the evidence was too flawed, if the reasoning was too flawed, even if they gave the evidence, I just couldn't go there….For me, I guess I have to get a sense that they understand the concept. Just going on the pure rules of evidence and reasoning, even if there's evidence and reasoning provided, I don't think most of the time I would go with a '3' score.*

Helen would generally not award a '3' unless the response was conceptually correct, even when both evidence and reasoning were present. In contrast, Emily recognized that the correctness of students' responses could hinder her judgment of quality reasoning or evidence. She described mentally course-correcting herself when she felt she might be veering off track:

> *I know when I was doing this, even when they got the wrong answer I was always like "No, that's fine." So it wasn't that they necessarily got penalized for getting a [wrong] answer but just because I didn't see their reasoning because I didn't identify with it, it was harder to give them credit.*

Similarly, Matt described "actively fighting against" the tendency to look for the correctness of the student's response: "Whenever I saw reasoning, I was like let's be sure this is reasoning even though it could be wrong."

Theme 2: Mathematical reasoning. When students' responses included mathematical reasoning, the raters disagreed on how to score the response. Two raters indicated that they considered mathematical equations to be reasoning. For instance, Luke remarked, "The evidence from the first piece, the bowl weighing five grams, the observation there. The reasoning would be the mathematical reasoning. 7 minus 5 is 2. How you get there." In contrast, Matt considered numbers evidence: "Whenever I saw numbers, I took that as evidence, because they were referring back to the information they had."

Theme 3: Reasoning and evidence intertwined or implicit. All five raters noted that sometimes reasoning and evidence were intertwined or implicit, making it challenging to judge when they were present. Emily described the difficulty in scoring a student's response, when the distinction between reasoning and evidence was unclear: "Sometimes it's difficult to decide if something is reasoning or evidence, and you don't want to double-barrel it and make it both because that doesn't seem like a fair application of the rubric either." Eva noted that she "didn't think it was necessary for the kids to give evidence as explicit as the ones that are in the example." In other words, the evidence was less

explicit than one might assume based on the rubric. Similarly, Emily described how implicit reasoning led her to draw inferences in students' responses: "The reasoning is sort of implicit in this, because they're saying all of the things that lead you to believe they must have reasoned to arrive at the answer, but they don't explicitly articulate the reasoning that happened." Furthermore, when the reasoning was implicit, Luke expressed a desire for more nuanced scoring options:

> *Indirectly she's saying they weigh the same….Between her mentioning the difference in the size, but then implying that there's no difference in weight and if they were different in size, one would weigh less, the shorter one. I struggled with these. It's definitely between a 2 and 3…. If I could give this a 2.5, this would be a 2.5.*

The desire for more nuanced scoring—an option between a '2' and a '3'—suggests a level of rater frustration with the process of scoring responses with implicit evidence or reasoning.

Moreover, the issue of correctness coincided with separating evidence and reasoning. For example, Emily described how the correctness of students' responses impacted her decisions regarding whether and how to separate evidence and reasoning:

> *This could be a case where whether or not the student ultimately gets to the correct answer biases whether or not I'm willing to let it count as more than one thing [evidence or reasoning]. Because I feel like part of the reason I was okay with this being both [evidence and reasoning] was because they were right. So it's a lot easier to defend. "Oh, yeah, they ended up on the right track. This is good." Whereas if they ended up with something totally wrong, they almost have to give you a written indication of what their reasoning is because you can't put it together because it's wrong.*

Theme 4: Causal language. Four out of the five raters considered causal language an indicator of reasoning. Yet, the raters recognized that linking causal language with reasoning could lead to faulty assumptions about the student's argumentation skills. For instance, Emily explained:

> *This is another one where I'm really hung up on the function of the word "because," and whether that constitutes whether they have reasoned or if they are just parroting an observation and happened to just have used the word "because." I feel like that's impossible to know. I could see a case for saying whether there is reasoning or not depending on how the word "because" hits you at the time that you read it.*

Since the raters do not have access to the cognitive process students utilized when responding to the item, they were faced with the difficulty of trying to discern students' motivations for using particular syntactic structure. This challenge perhaps explains why Matt noted, "I tried to avoid looking at the 'because,' generally, because anything with a 'because' looks like reasoning." "Since" and "so" were other examples of causal language discussed by the raters.

Theme 5: Importance of the examples. All five raters relied heavily on examples as scoring category reference points. As Emily noted, "I know when I was doing these I was looking very carefully at the examples." Helen and Luke described using the examples to distinguish between two scoring categories:

> *With the ones where I was struggling and thinking "Oh. Am I going to give this kid a '1'? Oh. Wait a minute. Maybe I should give them a '2,' because it kind of fits with the guidelines provided by a '2.'" So I definitely relied on the examples. I frequently went back to it and looked at the examples in particular to say, 'Is this kind of like what that example was?'...I think having the examples is very helpful.*
> -Helen

> *I think the examples would help too. Sometimes that would sway me too. If I was a little undecided, the piece that would push me in one direction or the other would be an example and how closely it would relate to the [student's response].*
> -Luke

As previously described, raters struggled with disentangling the correctness of students' responses from the quality of their argument. Eva hypothesized that part of the challenge might be related to the examples provided: "I don't know if it's a reflection of having less examples of incorrect evidence and reasoning that made it harder to put kids there." It might have been harder for raters to recognize incorrect evidence and reasoning since there were fewer examples of this scenario in the rubric. This comment highlights the extent to which raters relied on the examples to guide their scoring decisions, especially when the decision was not straightforward.

Five key modifications were made to the rubric based on information from the think-alouds:

1) The rubric was changed from a holistic format to an analytic format, for which raters were instructed to evaluate evidence and reasoning separately.

2) Raters were asked to score whether the science was correct prior to evaluating the quality of the argument. The intention was to indulge raters' instinct to prioritize correctness, so they could subsequently focus on the argumentation.

3) Clearer directions regarding what constitutes evidence and reasoning were provided for each item. For example, the rubric specified that mathematical reasoning should be interpreted as reasoning.

4) A note was added reminding raters to critically evaluate whether causal language actually constituted reasoning.

5) Annotated examples were added explaining why specific responses should receive certain scores.

Generalizability and Rater Accuracy Study

To analyze the effect of the rubric revisions, an analysis of the generalizability (g-study), reliability, and accuracy was conducted to compare the two rubrics. A subset of responses from 84 students to four items was rated by four new raters for the generalizability and rater accuracy study.

Before comparing the original and revised rubrics, we compared the reliability of the pool of raters from the pre-study and initial cognitive interviews with the pool of raters from the generalizability, rater accuracy study, and follow-up cognitive interviews to determine whether notable differences existed between the groups. Among the second group of raters, interrater reliability for Rubric 1 showed an overall decrease relative to the pre-study raters. Absolute ICC's in the pre-study ranged from 0.49 to 0.78 (Castle, 2018), while absolute ICC's in the g-study ranged from 0.23 to 0.60. However, it should be noted that the ICC in the pre-study was calculated for two raters, while the ICC in the g-study was calculated for four raters. Since ICCs tend to decrease as the number of raters increases, we did not interpret this as a difference between the two groups of raters. When Rubric 1 ICCs were compared for pairs of raters in the g-study and rater accuracy study, they tended to be similar in size to the ICCs reported in the pre-study.

When comparing Rubric 1 to Rubric 2, ICCs were higher for the revised rubric, with a few exceptions (Table 2). Two of the four items (Items 2B and 3) demonstrated ICC decreases associated with one of the argument subscores, suggesting that there may be item-specific effects on rater judgment; some items may have unique factors that make rater judgments on either the evidence or reasoning subscore more difficult, leading to less shared variation in scores

Table 2. Intraclass Correlation Coefficients (ICC) for Rubric 1 and Rubric 2

	Rubric 1		Rubric 2	
	ICC Absolute	ICC Consistency	ICC Absolute	ICC Consistency
Overall Argument Score				
Item 1	0.312	0.340	0.459	0.519
Item 2A	0.405	0.462	0.528	0.573
Item 2B	0.231	0.266	0.283	0.403
Item 3	0.602	0.637	0.555	0.593
Evidence Score				
Item 1	--	--	0.484	0.514
Item 2A	--	--	0.467	0.502
Item 2B	--	--	0.102	0.156
Item 3	--	--	0.635	0.645
Reasoning Score				
Item 1	--	--	0.293	0.320
Item 2A	--	--	0.410	0.435
Item 2B	--	--	0.328	0.362
Item 3	--	--	0.437	0.485

The variance component accounting for the largest amount of variation in scores was the residual variance component (person × rater × item), in both rubrics. The next largest variance component was the person × item interaction, suggesting that patterns of responses to the different items depended on individual student differences relative to the item. These differences could relate to content exposure, familiarity with the item context, or language fluency. This interaction accounted for a greater percentage of variance in the second rubric. The third largest variance component was the person effect, or universe score variance. This variance component represents the amount of variance attributable to differences between individuals, and is generally the main component of interest in any test score. The percentage of variance attributable to individuals was larger for the modified rubric, suggesting that scores from the second rubric contained more information about the relative performance of

examinees. The variance components associated with the rater facet tended to decrease in size on the modified rubric—the percentage of variance attributable to rater and person × rater both decreased, suggesting that the rubric revisions were associated with slightly less variation in raters' categorization of student responses. It appears that raters tended to agree more when the modified rubric was used. There was also a slight decrease in the amount of rater variation associated with the different examinees. The percentage of variance attributable to the item × rater term increased substantially with the revised rubric, indicating that raters were more likely to interpret the new scoring guidelines differently for different items when using the modified rubric.

The covariances (Table 3) indicate how much variance is shared between the two rubrics. Because all effects are linked (i.e., all persons, raters, and items were the same across both rubrics), the covariance indicates how much the person, rater, and item effects tend to vary in the same way across both rubrics. The largest covariance is from the person × item effect (0.111), indicating that differences in responses due to characteristics of the person × item interaction tended to manifest in the same way across both rubrics. There was also a large covariance in the person effect across rubrics (0.089), indicating that variation in the universe score tended to manifest similarly across both rubrics. The rater covariance (0.044) indicates that raters tended to make similar judgments, regardless of which rubric they used. All of these covariances are similar in size to the rubric-specific variance components, indicating that much of the observed variation on one rubric manifests similarly on the other rubric. The item, person × item, person × rater, and residual covariance (person × item × rater) all tended to be smaller in size than the rubric-specific variance components, suggesting that there was less shared variation between the two rubrics on these effects.

Both the generalizability coefficient ($\boldsymbol{E}\rho^2$) (Brennan, 2003) and the dependability coefficient (Φ) were higher for the revised rubric, suggesting that the observed scores were more generalizable among items and raters under the revised rubric (Shavelson, Webb, & Rowley, 1989). One limitation of the revised rubric is the increase in the error variance. As Brennan (2000) notes, increased generalizability is not always associated with smaller error variances, and vice versa.

Table 3. Variance and Covariance Components, Error Variances, and G- Coefficients

	Rubric 1	Rubric 2	Covariance
Variance Components			
Person	0.072	0.103	0.089
Item	0.028	0.000	0.004
Rater	0.044	0.039	0.044
Person x Item	0.106	0.144	0.111
Person x Rater	0.016	0.011	0.002
Item x Rater	0.000	0.044	0.000
Person x Item x Rater (Error)	0.220	0.215	0.053
Total	0.486	0.552	
Proportion of Variance			
Person	14.77%	18.65%	
Item	5.77%	0.00%	
Rater	9.07%	7.13%	
Person x Item	21.84%	26.14%	
Person x Rater	3.25%	1.92%	
Item x Rater	0.00%	7.92%	
Person x Item x Rater (Error)	45.29%	39.07%	
Total	100.00%	100.00%	
Error Variances			
Relative	0.044	0.052	
Absolute	0.062	0.065	
G-coefficients			
Generalizability coefficient ($\boldsymbol{E}\rho^2$)	0.619	0.664	
Dependability coefficient (Φ)	0.535	0.614	

An additional analysis compared the raters' scores to expert scores. All raters tended to have higher agreement with the expert score under the modified rubric (Table 4). This was true of both the Evidence and Reasoning subscores, compared to the overall Argument score under the original rubric. However, when evidence and reasoning were combined into a single score (by totaling the subscores), raters tended to have similar accuracy with both rubrics.

Follow-up Cognitive Interview Study

The primary purpose of the first round of cognitive interviews was to identify aspects of the rubric that could be improved. While the researchers were interested in potential improvements to the rubric in the second study, the primary purpose was to compare the ease of raters' experiences using both rubrics.

Table 4. Percent of Rater Scores Matching Expert Scores, Averaged Across All Items

	Rubric 1		Rubric 2		
	Argument total	Argument rescored*	Evidence	Reasoning	Sum of Evidence and Reasoning
Rater 1	70.54%	72.32%	80.36%	62.50%	68.75%
Rater 2	69.64%	73.21%	82.14%	80.36%	72.32%
Rater 3	56.25%	58.04%	71.43%	72.32%	59.82%
Rater 4	51.79%	57.14%	58.04%	78.57%	56.25%
Average across raters	62.05%	64.73%	72.99%	72.77%	64.29%

*Note: The original argument score was rated on a scale from 0 to 3, whereas the argument score for the revised rubric ranged from 0 to 2. The lowest two categories in the original argument score for Rubric 1 were collapsed for comparisons with Rubric 2. A "0" or "1" in Rubric 1 had the same meaning as a "0" in the revised rubric.

After the first round of cognitive interviews, several revisions were made to the rubric. Most of these changes enhanced the interpretability of the rubric (i.e., the ease with which raters could understand the scoring categories and make decisions). However, one major change was made to the scoring rules. Whereas the original rubric had asked that raters provide a single score for the overall argument on a four-point scale, the revised rubric asked that raters provide separate scores for evidence and reasoning – each on a two-point scale.

Three out of the four raters preferred the second (revised) rubric, while one rater preferred the first rubric. Raters commented on several aspects of the rubrics that swayed their individual preferences.

Theme 1: Ease of use. Overall, raters remarked on the ease of use afforded by each rubric. For instance, Marissa noted that the increased clarity in the second rubric made it easier to make decisions: "I actually felt like the second one [rubric] was a little clearer for me and made it a little easier to make decisions." Lucy echoed Marissa's general sentiment: "I remember the second rubric being easier to work with and easier to score." In comparison, Sophia preferred the first rubric because of the freedom it awarded her in making decisions: "I liked the first rubric, because it gave a lot of room to be able to make my own judgments."

Theme 2: Point system. Two of the raters commented on the number of points available in each scoring guide. The first two excerpts below are from Marissa, who preferred the second rubric overall. In particular, Marissa explained that the first rubric contributed to greater uncertainty in her scoring, due to what she described as the "grey area" between weak and adequate evidence:

> *I feel like I spent more time hemming and hawing over where does this fit into. Is it a three, a two, or a one? And then separating the two out, because there was also that issue with what if the evidence is strong, but the reasoning is a little funky. Does that go into a '2'? Does that go into a '1'? I just felt like it gave more area too. So maybe it was more time consuming. I don't know if that's better, but for me as a scorer it just meant more uncertainty.*

> *The first one I guess was "nicer" because you could get some credit for giving weak evidence. Like I felt that gave me a grey area that was actually more difficult for me to decide whether this is good evidence; this is weak evidence. I don't know. Even though the line was still between good evidence and weak evidence. With the other one it was just easier to say 'Oh this is good evidence, give them a point. This is like whatever … no…don't give them the point.' That one or zero was a little easier for me.*

In contrast, Sophia preferred the first rubric, since there were more points available. In the second rubric, there was no in-between score for "so-so" reasoning or evidence, which Sophia found somewhat frustrating:

> *There were more points for the argument in the first rubric. There were some where the evidence was so-so and the reasoning was so-so, but you couldn't quite give them a definite 0, because that's the same as getting a blank. Because the kid actually tried a little bit.*

Theme 3: Evidence and reasoning as separate dimensions. Following the Initial Study, one major change to the second rubric was the separation of evidence and reasoning into separate scoring dimensions. Lucy and Rachel preferred the separation of the two dimensions:

> *I think when we had the second training before introducing the second rubric the feedback that we had on the first rubric was that it was sometimes hard to distinguish between evidence and reasoning. And I think the second rubric made that clearer.* -Lucy

> *I probably liked when evidence and reasoning was separate. Just because sometimes reasoning was dominating the evidence, or the other way. …It was getting complicated to decide whether it was evidence or reasoning [when using the first rubric].* -Rachel

Sophia, in contrast, believed that the separation of evidence and reasoning made the scoring process more

challenging, due to the relational nature of the two dimensions:

> *Evidence and reasoning, those two concepts, should be relational, and so for the first rubric, you could definitely see that they were meshing together, whereas for the second rubric you definitely just saw separate point systems for each. I couldn't wrap my head around how to separate those two.*

Theme 4: Separating correctness from the quality of argumentation. In the first rubric, raters were instructed to score the argument itself before the correctness, whereas the order was reversed in the second rubric. Given that the raters in the first study expressed difficulty separating the correctness from the argumentation, placing the correctness score earlier in the scoring process was hypothesized to help raters focus on the argumentation without distraction. Marissa specifically commented on this change to the rubric: "I think separating out correctness and scoring that separately is helpful, because once that's out there you can focus on other things." In effect, the revision was functioning as intended for Marissa. At the same time, Marissa still experienced some challenges ignoring the correctness when scoring the argument:

> *I tried very hard not to let [the correctness of the response impact my scoring] …. I remember some very clear moments where I was like "Wow, this is a such a great, well-constructed argument that's totally wrong." And I gave them all the points for the argument and reasoning. So there were instances where I did that. But then I wonder with some of these border cases. So I don't know. But I did make a conscious effort not to be swayed by whether it was correct or not.*

Lucy expressed similar difficulty separating the correctness from the quality of argumentation:

> *That was pretty hard for me as a scorer [to separate the correctness from the argumentation], because I don't think when I was taking exams or been younger, I would have thought that was a possibility that you could just write a wrong answer.*

Theme 5: Importance of examples. In the Initial Cognitive Interview Study, the raters described relying heavily on the examples when making scoring decisions. All the raters in the Follow-up Cognitive Interview Study also relied on the examples. Even Sophia who preferred the first rubric overall remarked that the layout of the examples was helpful in the second rubric. In the second rubric, there was an explanation for why certain scores were assigned next to each example. As Marissa noted:

> *The examples were more helpful [in the second rubric]. The way they were laid out. The way they were explained….I really liked examples, especially if there is common weird things. So sometimes you will see that a few times and you're like 'They're using that causal language, but it's not what they think it means.'…I think having the examples and then having this spread next to it is helpful too.*

Sophia's comment above also alludes to the issue of "causal language" potentially erroneously signaling reasoning. The second rubric, including the examples provided, was designed to address areas of confusion identified from the first rubric. Thus, the second rubric warned the raters that the presence of causal language should not automatically be equated with reasoning

Discussion

Rater cognitive interviews provide valuable information about how raters interpret scoring rubrics (Cumming, 1990; Cumming, Kantor, & Powers, 2002; Vaughan, 1991; Zhang, 2016), and this information can be used to clarify rubrics. In this study, rubric modifications included better descriptions of key aspects of the students' responses (Moskal, 2000), more examples, and separation of a holistic argument score into reasoning and evidence subscores. In other words, the scoring procedure changed from holistic to analytic following rubric modifications (Popham, 1997; Wolfe & Song, 2016).

Using Bejar's (2012) Rater Cognition Framework, the reliability and accuracy of raters' scores relies on the assumption that raters' have properly mentally encoded the rubric. Accordingly, the updated descriptions and examples were intended to help raters interpret the rubric, thus make more accurate and consistent decisions about whether or not student responses fit into a particular scoring category. From another perspective, using Wolfe and Song's (2016) Rater Quality Framework, the rubric modifications were designed to reduce rater effects due to rater context (i.e., the format of the rubric) and rater characteristics components (i.e., raters' understanding of the rubric), but not the characteristics of students' responses. According to the g-study results, the rubric revisions did appear to decrease the amount of variation attributable to raters, and improve the generalizability of scores, although differences in rater behavior were small (indicated by the

large covariance in rater effects between the two rubrics). Interrater reliability and rater accuracy, measured by agreement with expert raters, also improved. All of these measures indicate better reliability and validity, signaling that raters were able to make more accurate and consistent judgments with the revised rubric (Wolfe & Song, 2016).

Using analytic subscores narrowed the scope of rater focus to one aspect of the response at a time. Subscore ratings tended to be more accurate, and had higher interrater reliability and generalizability, indicating that they may be easier for raters to understand and apply. Rater interviews confirmed that most raters tended to prefer the analytic rubric, finding it easier to make decisions between different scoring categories. The current finding is congruent with Klein and colleagues' (1998) study comparing analytic versus holistic scoring methods in science performance assessment, in which they found higher reliability for the analytic rubric. Similarly, Jönsson and Balan (2018) found higher agreement for a writing assessment when raters applied an analytic rubric.

One of the major sources of rater confusion was the difficulty of distinguishing between the "correctness" of a student's argument and the quality of the construction of the argument. This is an example of what Zhang (2013) referred to as a halo effect, in which raters are biased towards giving similar judgments to multiple observations from the same respondent. In this case, raters' perception of the accuracy of the students' conceptual understanding interfered with their judgment of the quality of students' argumentation. Although the rubric modifications included measures to better help raters distinguish between these constructs (e.g., a greater variety of examples including more examples with well-constructed arguments based on incorrect understanding, and a separate "correctness" score), raters still expressed bias towards arguments that contained evidence of "correct" understanding. This example demonstrates the difficulty of evaluating multidimensional science aligned with the NGSS, in which demonstration of science practices and crosscutting concepts co-occur with demonstration of content fluency (NRC, 2014). As multiple interrelated dimensions are assessed in the same context, it can be difficult for raters to make judgments that disentangle these dimensions. This is further complicated because conceptual understanding is a common and well-defined construct; thus, raters unintentionally rely on the familiar judgment of conceptual understanding as a proxy for a more difficult judgment about students' mastery of scientific practices.

Limitations

Several limitations should be considered when evaluating these findings. First, the number of raters included in both the initial (n=5) and follow-up (n=4) interviews was relatively small.

Second, modifications to Rubric 2 included clarifying key components of Rubric 1. For instance, Rubric 2 included a clarified definition for evidence and reasoning in the context of each item. Therefore, the risk of carry over effect was greater when moving from Rubric 2 to Rubric 1, so all raters were asked to score responses using Rubric 1 first and then Rubric 2. This design raises the possibility that rater reliability and accuracy might have improved when using Rubric 2 because raters were more familiar with the argumentation construct and had more experience scoring. Moreover, rater fatigue might have influenced raters' scoring for the second rubric, although we suspect that this was not an issue since multiple days passed before raters proceeded to the second rubric.

Third, this study addresses only one of eight NGSS SEPs, exploring some of the complications that arise in assessing scientific argumentation. Although all eight practices describe skills necessary for doing science (rather than knowing science), they are also distinctly complex in their own ways; therefore, results may not generalize to other practices.

Conclusion

Rater cognition, previously used to enhance rater accuracy and consistency on assessments of language and writing (Cumming, 1990; Cumming, Kantor, & Powers, 2002; Vaughan, 1991; Zhang, 2016), may also prove a valuable tool in assessing scientific practices. This represents one of the first studies to extend the practice of rater "think-aloud" studies to science assessment. As assessment developers design tests and rubrics for new, complex constructs from the *Next Generation Science Standards* (NGSS Lead States, 2013), cognitive interviews will provide an important tool to improve assessment validity (i.e., accuracy) and reliability, since they produce evidence that raters indeed

understand the construct and the associated scoring procedures as intended.

To support assessment validity and reliability, rater cognitive interviews should be used as a regular part of the assessment development and validation processes. Moreover, given the prior success of rater cognition studies in language and writing assessment and the increased reliability and accuracy observed in this study, we recommend that cognitive interviews be used to improve rubric clarity in a more diverse range of fields and purposes.

Performance assessment has been identified as a potential solution to address calls for more authentic assessments of student learning (Guha, Wagner, Darling-Hammond, Taylor, & Curtis, 2018; Lane & Stone, 2006; Linn, 1993; Wiggins, 1990). However, there are concerns about the generalizability of scores obtained from performance assessments, given the reliance on rater judgment to score responses (Davey, Ferrara, Holland, Shavelson, Webb, & Wise, 2015; Lane & Stone, 2006; Linn, 1993). Accordingly, cognitive interviews could provide a strategy for exploring raters' understanding of performance assessment rubrics, leading to rubric improvements. While the present study included only one rubric modification, multiple iterations may lead to even higher reliability and generalizability. Although conducting cognitive interviews require additional time and resources, the long-term impact on scoring validity prove beneficial.

References

Almond, P. J., Cameto, R., Johnstone, C. J., Laitusis, C., Lazarus, S., Nagle, K., Parker, C. E., Roach, A. T., & Sato, E. (2009). White paper: Cognitive interview methods in reading test design and development for alternate assessments based on modified academic achievement standards (AA-MAS). Dover, NH: Measured Progress and Menlo Park, CA: SRI International.

Arksey, H., & Knight, P. T. (1999*). Interviewing for social scientists: An introductory resource with examples*. Thousand Oaks, CA: Sage.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7(1), 54-74.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2), 287-311.

Bejar, I.I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, 31(3), 2-9.

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.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, 24(4), 339-353.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Brennan, R. L. (2003). Coefficients and indices in generalizability theory. Center for Advanced Studies in Measurement and Assessment, CASMA Research Report, 1, 1-44.

Castle, C. (2018). Measuring multidimensional science learning item design, scoring, and psychometric considerations (Doctoral dissertation). Boston College, Chestnut Hill, MA.

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice*, 31(3), 10-20.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7(1), 31-51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, 86(1), 67-96.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). TOEFL 2000 writing framework. Princeton, NJ: Educational Testing Service.

Davey, T., Ferrara, S., Holland, P. W., Shavelson, R., Webb, N. M., & Wise, L. L. (2015). *Psychometric considerations for the next generation of performance assessment.* Washington, DC: Center for K-12 Assessment & Performance Management, Educational Testing Service.

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, 32(4), 521-541.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Guha, R., Wagner, T., Darling-Hammond, L., Taylor, T., & Curtis, D. (2018). *The promise of performance assessments: Innovations in high school learning and college admission*. Palo Alto, CA: Learning Policy Institute.

IBM Corp. (2017). *IBM SPSS Statistics for Windows*, Version 25.0. Armonk, NY: Author.

Jönsson, A., & Balan, A. (2018). Analytic or holistic: A study of agreement between different grading models. *Practical Assessment, Research, and Evaluation*, 23(12), 1-11.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., ... & Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, 11(2), 121-137.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (pp. 387 - 431). Lanham, MD: Rowman & Littlefield Publishers.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, 15(1), 1-16.

McNeill, K. L., & Krajcik, J. (2008). Inquiry and scientific explanations: Helping students use evidence and reasoning. In Luft, J., Bell, R. L., & Gess-Newsome, J. (Eds.), Science as inquiry in the secondary setting (pp. 121-134). Arlington, VA: NSTA press.

Moskal, B. M. (2000). Scoring rubrics: What, when, and how? *Practical Assessment, Research, and Evaluation*, 7(3), 1-5.

Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues and Practice*, 31(3), 48-49.

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.

Norris, S., Philips, L., & Osborne, J. (2008). Scientific inquiry: The place of interpretation and argumentation. Science as inquiry in the secondary setting. In Luft, J., Bell, R. L., & Gess-Newsome, J. (Eds.), *Science as inquiry in the secondary setting* (pp. 87-98). Arlington, VA: NSTA press.

Popham, W. J. (1997). What's wrong-and what's right-with rubrics. *Educational Leadership*, 55, 72-75.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44(6), 922.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind. In L. Hamp-Lyons (Ed.), *Assessing second language writing* (pp. 111-126). Norwood, NJ: Ablex.

Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research & Evaluation*, 2(2), 1-3.

Willis, G. B. (1999). Cognitive interviewing: A "how to" guide. Presented at the Annual Meeting of the American Statistical Association.

Wolfe, E.W. & Song, T. (2016). Methods for monitoring and document rating quality. In H. Jiao & R.W. Lissitz (Eds.), *The next generation of testing* (pp. 107-142). Charlotte, NC: Information Age Publishing.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections*, 21(2). Princeton, NJ: ETS.

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing*, 27, 37-53.

Acknowledgments:

The authors would like to thank Michael Russell, Professor of Measurement, Evaluation, Statistics, and Assessment at Boston College, for his ongoing support, advice, and mentorship; Lisa Keller, Associate Professor of Research, Educational Measurement, and Psychometrics at UMass Amherst, for her statistical guidance; the NCME and PARE proposal reviewers for their feedback; and our raters, for without whom this study would not be possible.

Note:

A previous version of this paper was presented at the 2019 National Council on Measurement in Education Conference (NCME) in Toronto, Canada.

Citation:

Borowiec, K.& Castle, C. (2019). Using Rater Cognition to Improve Generalizability of an Assessment of Scientific Argumentation. *Practical Assessment, Research & Evaluation*, 24(8). Available online: http://pareonline.net/getvn.asp?v=24&n=8.

Corresponding Author

Katrina Borowiec
Boston College

email: katrina.borowiec [at] bc.edu

Appendix A. Rubric 1 (Original Rubric)

<table>
<tr><th colspan="2">Question 2b (Argument)</th></tr>
<tr><th>Score</th><th>Description of Response</th></tr>
<tr><td>3</td><td>Student supports argument with both evidence (observations about the block and/or the ball, or a statement that such an observation cannot be made) and reasoning (a clear link between the evidence and the conclusion being made, either based on the volume of solids being invariant during shape change, or some other belief). Evidence and reasoning do not necessarily have to support a correct answer, but they should support the chosen answer.
Examples:
<table>
<tr><td><u>Correct evidence and reasoning</u>
Because Ana took the rectangle that was 5 cm and turned it into a ball, so it’s just a different shape, but same volume.
Because it is still the same clay so they take up the same amount just a different shape.
The block of clay was 20 cc so the ball must also be 20 cc.</td><td><u>Incorrect evidence and reasoning</u>
Because the ball is taller than the block so it takes up more space.
The circle of clay is probably lighter, because it has no edges and is rolled up, making the material smaller and lighter.</td></tr>
</table></td></tr>
<tr><td>2</td><td>Student supports argument with either evidence (observations about the block and/or the ball, or a statement that such an observation cannot be made) or reasoning (a clear link between the evidence and the conclusion being made, either based on the volume of solids being invariant during shape change, or some other statement of belief). Evidence or reasoning do not necessarily have to support a correct answer, but they should support the chosen answer.
Examples:
<table>
<tr><td><u>Correct evidence</u>
It’s still the same thing as before but a different shape.
Because it was 20 cc before she crumpled it.
She didn’t add or take away any clay.</td><td><u>Incorrect evidence</u>
The ball is squashed.
I don’t know because I can’t measure it.
I think that the ball is 4 wide, 3 height, and 2 for length.</td></tr>
<tr><td><u>Correct reasoning</u>
Changing the shape won’t change the volume.</td><td><u>Incorrect reasoning</u>
Because when you make something into a ball or crumple something up you make it a little bit bigger than it already is.
Taller things take up more space.</td></tr>
</table></td></tr>
</table>

1	Student supports a claim with irrelevant evidence (evidence that does not support their previous responses), weak evidence (appeal to authority, personal experience, tautological reasoning, or vague references to data), and/or reasoning that doesn't support their answer. **Evidence or reasoning do not necessarily have to support a correct answer, but they should support the chosen answer.** Examples: <u>Weak evidence</u> Because I did it in class.
0	Statements that do not offer any evidence or reasoning, e.g., "I don't know" or "It just does".
Missing	1a) and/or 2a) are answered, but 1b) is blank.
Blank	1a), 2a), 1b), and 2b) are all blank.

Appendix B. Rubric 2 (Revised Rubric)

Question 2b – Engaging in Argument from Evidence	
Correct Principle	1 = Correct (Volume is invariant with reshaping) Examples: Changing the shape doesn't change the volume. It's the same material just a different shape. It's still the same material. She didn't add anything or take anything away.
	0 = Unclear or missing
	-1 = Incorrect (Volume may change with reshaping) Examples: The block is longer so it takes up more space. Crumpling things up makes them bigger. [Estimates the height and length of the ball.]
	Blank = 1a), 2a), 1b), and 2b) are all blank.
Evidence	1 = Provides explicit, relevant evidence in support of a claim. Evidence should be explicit scientific data, which supports a claim. "Scientific data are information, such as observations and measurements…provided to the students" (Berland & McNeill, 2010, p. 772). Evidence should be an observation or measurement of a physical quality of object(s). In this item, evidence is most likely to be a statement about the volume of the clay block, but it may also include specific statements about the shape of the clay, or other observable physical attributes of the clay. Personal experience is not valid evidence. Mentioning the "amount" should not be considered evidence, since it does not specify a direct observation; referring to the "amount" can count as reasoning. The student's claim is their answer to the previous question, unless the previous answer is missing. Their claim may be repeated within the argument (or, stated for the first time if the previous answer is missing). If the student utilizes mathematical expressions containing numbers that represent the volume of the block and/or ball, these numbers may be considered evidence. *Evidence and reasoning may be woven together in one statement, such that they are inextricably linked in the student's argument. Use your best judgment to determine whether evidence and reasoning are present, based on the descriptions above and below. Do not make large inferences about what the student meant; when in doubt, place the burden of proof on the student.*

	0 = Does not provide explicit, relevant evidence in support of a claim.
	Missing = 1a) and/or 2a) are answered, but 1b) is blank.
	Blank = 1a), 2a), 1b), and 2b) are all blank.
Reasoning	1 = Provides appropriate reasoning in support of a claim. The reasoning clearly articulates the logic behind the claim. If evidence is present, the reasoning may provide a rationale for why the evidence supports the claim. Students often use words like "because," "so," "since," etc. which we may falsely attribute to causal reasoning. If a student uses causal language, carefully evaluate the content of their argument. Students may use these words to repeat a claim or provide evidence. In this case, this causal language should not be taken as an indicator of student reasoning. Referring to the "amount" can count as reasoning. If the student utilizes mathematical expressions to demonstrate the logic behind their argument, this may be considered reasoning. *Evidence and reasoning may be woven together in one statement, such that they are inextricably linked in the student's argument. Use your best judgment to determine whether evidence and reasoning are present, based on the descriptions above. Do not make large inferences about what the student meant; when in doubt, place the burden of proof on the student.*
	0 = Does not provide relevant reasoning in support of claim.
	Missing = 1a) and/or 2a) are answered, but 1b) is blank.
	Blank = 1a), 2a), 1b), and 2b) are all blank.

Example	**Scores with explanations**
On the first problem Ana add 20 cubes = 20 cc. Then she made a ball. It is still 20cc because nothing got lost or added only the formation changes.	Correct Principle: 1 (In this example, the student says "It is still 20cc because nothing got lost or added only the formation changes." This statement suggests that the student understands that changing the shape of the object will not change its volume.) Evidence: 1 (The student's claim is that the volume of the ball of clay is 20 centimeters. The student's evidence is the first block was 20 cc, because there were 20 cubes.) Reasoning: 1 (The student supports their claim and evidence by explaining that changing the shape will not change the volume of an object: "It is still 20cc because nothing got lost or added only the formation changes.")
Because the ball is taller than the block so it takes up more space.	Correct Principle: -1 (This student's response indicates that he or she thinks that if you change the shape of an object, its volume will also change: "The ball is taller." Thus, the student does not seem to understand that the volume will remain the same regardless of shape.) Evidence: 1 (Although not explicitly stated, the student's claim is that the volume of the ball will be greater than the volume of the block. Thus, in this case, the student's evidence is in the form of an observation about the shape of the ball compared to the block: "The ball is taller than the block.") Reasoning: 1 (Although the student's claim and reasoning are incorrect, the student does provide reasoning. The student argues that taller objects will take up more space, and therefore, have more volume.

Because she rolled the same amount of clay.	Correct Principle: 1 (From this response, it seems clear that the student understands that the volume of an object will not change just because its shape has changed. Thus, the student seems to understand the scientific principle.) Evidence: 0 (In this example, the claim is implied: the volume of the ball of clay is the same as the volume of the block of clay. However, the student does not provide any evidence to support this claim.) Reasoning: 1 (The student supports their implied claim with the statement that the ball is the "same amount.")
I think so because about 11 full cubes could fit in the ball. If not, the answer would be 9 because you could also fit 9 full cubes in that ball.	Correct Principle: -1 (In this example, the student estimates the number of cubes that would fit in the ball, which suggests that he or she does not understand that the volume of the ball will be the same as the volume of the block. Thus, the student does not seem to understand the Correct Principle.) Evidence: 1 (The student's claim is that the volume will be 11. The student's provides evidence in the form of an observation based on visual estimation of the ball of clay.) Reasoning: 0 (The student does not provide reasoning in this example.)
I think Ana's ball of clay is 80g because 1 cubic centimeter of sugar is 2g and there is 40g + 40g = 80g.	Correct Principle: -1 (The provides false evidence and reasoning that is unrelated to the ball of clay question. The student does not seem to understand the scientific concept.) Evidence: 0 (The student's claim is that the "ball of clay is 80g." The student supports this claim with false/irrelevant evidence: "1 cubic centimeter of sugar is 2g.") Reasoning: 0 (In this example, the student provides mathematical reasoning, but the reasoning is irrelevant and based on false evidence.)