Studia paedagogica
vol. 24, n. 4, 2019
www.studiapaedagogica.cz
https://doi.org/10.5817/SP2019-4-3

MEASURING PRODUCTION AND COMPREHENSION OF WRITTEN ARGUMENTS IN UPPER-ELEMENTARY GRADES

ALINA REZNITSKAYA,
IAN A. G. WILKINSON

Abstract

We describe a systematic process of developing measures of argument production and comprehension. These measures, designed for students in upper-elementary language arts classrooms, are called Writing Argument and Reading Argument. We discuss the rationale and theoretical framework for the measures, describe pilot and validation studies, and present initial findings to support the reliability, validity, and usability of these measures. Our results showed that both measures had acceptable inter-rater reliability. The correlations among Writing Argument, Reading Argument and an established reading comprehension test were moderate, which highlights the importance of task-specific competencies. The performance on both measures was not associated with ethnicity of the students. Gender was a significant predictor, with girls performing better than boys. Teachers found both measures to be pedagogically useful. Although some teachers initially struggled with learning how to use the scoring rubrics, they generally found the scoring for both tasks to be informative for their practice.

Key words

writing argument, reading argument, measures, validation, elementary grades

Introduction

Helping students develop the ability to formulate and comprehend arguments is increasingly seen as one of the key purposes of schooling (Kuhn, Hemberger, & Khait, 2016; Lipman, 2003; Newell, Beach, Smith, VanDerHeide, Kuhn, & Andriessen, 2011). This position is now reflected in major national and international educational policy documents (e.g., National Governors Association, 2010; Partnership for 21st Century Skills, 2012). For example, the Common Core Standards initiative in the US (National Governors Association, 2010, p. 25) stresses the importance of writing and reading arguments across the curriculum, and considers argumentation to be "broadly important for the literate, educated person living in the diverse, information-rich environment of the twenty-first century".

Unfortunately, the widespread recognition of the value of argumentation has not generated enough efforts to understand how this construct can best be measured. Published, standardized tests of argumentation skills and related abilities (e.g., reasoning and critical thinking) often lack sufficient validity evidence and are rarely designed for elementary school students (Hughes, 1992; Poteet, 1989; Sutton, 1992). This is regrettable, considering that elementary-age children are developmentally poised to engage in argumentation and can improve their skills through such engagement (Mercer, 2011; Reznitskaya, Anderson, Dong, Li, Kim, & Kim, 2008; Stein & Trabasso, 1982). Lack of quality measures for younger students is also problematic given that major policy documents describe argumentation skills as an intended educational outcome for students from the early grades.

In addition to standardized measures, there are several custom-made instruments designed by researchers who study argumentation development (e.g., Chambliss & Murphy, 2002; Kuhn & Crowell, 2011; Means & Voss, 1996). Although these tools offer interesting insights into how argumentation can be measured, they are developed to answer specific research questions and have limited information about their psychometric properties. Further, the analysis of student performance often involves coding or diagramming student responses, making these measures impractical for classroom teachers. Considering the recognized importance of argumentation, we need better researched and more practical tools that can provide teachers and researchers with meaningful diagnostic information about student progress. In this paper, we describe a systematic process aimed at developing and validating measures of argumentation, called *Writing Argument* and *Reading Argument*.

Theoretical Framing

Our thinking about what is meant by competency in argumentation has been informed by multiple strands of research. First, based on schema-theoretic views of cognition (Anderson & Pearson, 1984; Gick & Holyoak, 1983; Rumelhart & Ortony, 1977; Wolfe, Britt, & Butler, 2009), we assumed that knowledge of argumentation can be represented as an abstract mental structure we call an *argument schema* (Reznitskaya & Gregory, 2013). Because schemas serve a variety of functions, affecting perception, comprehension, learning, inferencing, and remembering, students with developed argument schemas are expected to perform better on a variety of tasks, such as when writing and reading arguments. That is, better performers will make use of relevant 'slots' in their argument schemas that get activated once the task is recognized as requiring this type of knowledge.

To further specify what constitutes a developed argument schema, we drew on Toulmin's (1958) original work on argument structure, in which he described the general layout of an argument and identified unique functions of six argument elements, such as claim (position), data (evidence), warrant, qualifier, grounds (evidence for warrant), and rebuttal. The Toulmin model has been widely adopted by researchers in education (Rapanta, Garcia-Mila, & Gilabert, 2013), as it offers a useful heuristic for analyzing the quality of arguments.

However, the Toulmin model was largely designed to represent an argument from a single viewpoint (Healy, 1987). For example, even the rebuttal in the Toulmin model "is anticipated from the (single) perspective of the advocate, and so lacks the dynamic force of criticisms coming from a wholly other perspective" (Healy, 1987). In contrast, similar to other scholars (Anderson et al., 2001; Graff, 2003; Kuhn, 1991; Nussbaum & Ordene, 2011), we view argumentation as a dialogic process of testing competing ideas against each other with the goal of moving towards the most reasonable answer. That is, even a solitary engagement in argumentation, such as the writing of an argumentative essay, represents an interaction and struggle with those who think differently in an effort to figure out what is most reasonable to conclude. Such a view of argumentation implies that consideration, refutation, and integration of alternative viewpoints become important elements in a developed argument schema.

Competency in argumentation also includes the knowledge of criteria that distinguish good arguments from poor ones. To better understand these criteria, we reviewed the models of argument evaluation developed by scholars of argumentation, logic, reasoning, and critical thinking (Ennis, 1996; Govier, 2010; Hollihan & Baaske, 1973). We identified and grouped key criteria

into four main categories, and presented them in a way that was accessible to elementary school teachers (Reznitskaya & Wilkinson, 2017):

1. Clarity: We are clear in the language and structure of our arguments.
2. Acceptability: We use reasons and evidence that are well examined and accurate.
3. Logical validity: We are logical in the way we connect our positions, reasons, and evidence.
4. Alternative perspectives: We explore and evaluate different perspectives.

These strands of prior research – schema-theoretic views of cognition, modified Toulmin model, and criteria for evaluating the quality of arguments – helped to us to construct the measures of argumentation and informed our expectations about student performance. For example, students with advanced argument schemas are expected to construct arguments in writing that use the language and structures appropriate for this type of discourse, contain acceptable reasons and evidence, have logical connections between positions and reasons (warrants), and take into account alternative viewpoints. Similarly, such students are expected to engage more productively with written texts that contain arguments. According to Govier (1987, p. 233), recognizing the presence of an argument is "something quite elementary and yet elusive to many not encouraged to think about reasoning, argumentation, and the justification of claims. It is the sense that reasoning is going on, that there is an inference made from some propositions to others, and that this inference can be critically scrutinized". Thus, when reading the arguments of others, students with advanced argument schemas will be more likely to look for common argument elements and to ask evaluative questions: What is the author's position? Is the taken position supported by reasons? Are alternative perspectives taken into account?

Developing Measures of Argument Production and Comprehension

To develop the measures, we followed a systematic, multi-step process of test construction proposed by Crocker and Algina (1986). We began with a literature review to define the constructs, as described in the previous section. We also examined existing measures of similar constructs (e.g., Chambliss & Murphy, 2002; Means & Voss, 1996; Nussbaum & Kardash, 2005; Page-Voth & Graham, 1999; Phillips & Patterson, 1987) to identify effective assessment strategies and formulate hypotheses to be tested in validity studies. We chose to use a constructed-response format for both measures because instruments that allow for greater flexibility in participant responses are considered to be more compatible with contemporary

theories of learning and instruction (Shepard, 2000) and more suitable for measuring complex cognitive behaviors, such as argumentation (Ennis, 2003; Halpern, 2003). Fixed-choice tests of argumentation and related constructs have been criticized for obscuring the thinking process that underlies the response. In contrast, open-ended tasks help to generate rich information on student performance in a direct manner, allowing one to determine not only what students think, but also how they think (Halpern, 2003).

Over the next three years, we developed an initial set of measures and conducted several pilot studies, during which we iteratively administered versions of the reading and writing tasks to students in 18 fifth-grade classrooms. Some of the pilot studies were conducted as part of a larger research project, during which we designed and tested a professional development program for Grade 5 teachers to help them promote argumentation (e.g., Reznitskaya & Wilkinson, 2015; Wilkinson et al., 2017). The pilot studies allowed us to examine student performance, make revisions to the measures, and develop scoring rubrics. Additional data to inform revisions came from interviews conducted with 20 students for the *Writing Argument* task and 25 students for the *Reading Argument* task. We also conducted two focus-group interviews with five fifth-grade teachers to get their insights and reactions in order to better understand the use of the measures in a typical classroom.

Revisions included changing the stimulus materials to make them more controversial and relevant to students' interests, modifying vocabulary to make the texts more accessible, and revising the instructions and format of the tasks. For example, making instructions for *Writing Argument* tasks clearer and more detailed helped to reduce floor effects and increase variability in student performance. This is consistent with previous studies, which show that more precise and carefully-worded instructions improve students' production of arguments (Nussbaum & Kardash, 2005; Page-Voth & Graham, 1999). Similarly, we made several changes to the format and instructions of the *Reading Argument* task. The *Reading Argument* task was initially designed as a recall task that required students to read and recall a given text that contained arguments on both sides of the issue. Our examination of students' recall responses showed poor performance and low variability, indicating that the task was too difficult for the students. We investigated whether performance could be improved by having students recall the text orally, rather than in writing. Our analysis indicated that the quality of recalls was not influenced by communication modality. As a result, we left the written modality of response unchanged, while allowing students to keep the original text and requiring them to restate the arguments from the text in their own words, rather them to recall them from memory. This proved to be an effective revision than alleviated floor effects and increased variability of responses.

Criterion Measures

As mentioned earlier, our review of existing measures of argumentation and related constructs revealed a near absence of psychometrically-sound instruments that target elementary school students. However, in our search of suitable criterion measures, we found the *Test of Inference Ability in Reading Comprehension* (Phillips & Patterson, 1987) to be a promising choice. This is a standardized test of reading, during which students read full-length passages and answer multiple-choice questions after each paragraph. The test requires students to engage in close, attentive reading and critical thinking about the text (Norris, 1995; Norris, Leighton, & Phillips, 2004). However, the test was intended for use with students in grades 6, 7, and 8 and had only moderate levels of reliability. We decided to administer the *Test of Inference Ability in Reading Comprehension* to 191 students from 10 fifth-grade classrooms in a pilot study to learn more about its psychometric properties and its suitability for younger students. The results of our analysis using the Rasch model showed that the test was somewhat easy for the target group of grade 5 students. More problematically, the internal consistency reliability was too low, .62, for the test to be used as a criterion measure.

Upon further review of literature, we chose the *Gates-MacGinitie Reading Test (GMRT)* (MacGinitie, MacGinitie, Maria, & Dreyer, 2002) as a psychometrically-sound measure that should relate to student performance on argumentation tasks. The *GMRT* is a highly regarded standardized test of vocabulary and reading comprehension. Independent reviewers conclude that *GMRT* is a well-developed instrument that is supported by convincing evidence of reliability and validity (Johnson, 2005; McCabe, 2005). KR-20s and alternate form reliabilities range from .85 to the .95. Scores on the *GMRT* correlate highly with those on other standardized tests of cognitive and school-related abilities, as well as with course grades (MacGinitie et al., 2002; Oka & Paris, 1987).

We expected to see positive correlations between students' performance on the *GMRT* and our measures of argumentation. Although very few researchers have investigated the relationship between argumentation skills and other variables, especially for students in elementary grades, research indicates potential connections between argumentation skills and other cognitive abilities (Means & Voss, 1996; Weinstock, Neuman, & Glassner, 2006).

Final Measures

This section describes the final measures of argument production and comprehension, and the scoring procedures we used to conduct reliability, validity, and practicality studies.

Writing Argument

This task has been used in our previous research (e.g., Reznitskaya, Anderson, McNurlen, Nguyen-Jahiel, Archodidou, & Kim, 2001; Reznitskaya, Kuo, Clark, Miller, Jadallah, Anderson, & Nguyen-Jahiel, 2009), but we made several revisions to the stimulus materials, task instructions, and scoring rubrics based on the information obtained from the pilot studies described earlier. In the final version, the stimulus was a short story (776 words), in which a boy, named Jack, faces the moral dilemma of whether or not to tell on his classmate Thomas, who cheated in a Pinewood Derby race. Students were read the story, as they followed along, and then asked to write an essay in response. Students had 25 minutes to complete the essay, which was shown to be ample time in our pilot studies. The task instructions are shown in Figure 1.

> Please write a letter to your teacher explaining whether or not YOU think Jack should tell on Thomas.
>
> - Remember to clearly state your opinion[1] and support your opinion with reasons and evidence.
> - Remember also to think about how other people might disagree with your opinion, and how you would respond to them.
> - Don't forget your conclusion.
>
> Do your best and write as much as you can. You can go back and re-read the story if you like.

Figure 1
Instructions for Writing Argument task.

To score student performance on the *Writing Argument* task, we developed an analytic rubric with three categories that captured the criteria for quality arguments described earlier. The maximum score in each category is 3 and the minimum is 0. Figure 2 presents the rubric description for the highest score.

[1] We used the word 'opinion' (rather than claim, position, or point of view) because it was the one best understood by students in our pilot studies.

Category	Score: 3
Clarity	The essay has clear language and a coherent structure (e.g., position-support-alternative-restatement of position). The entire essay is well focused.
Support for Chosen Position	The writer provides a clear position that is strongly supported by reasons (e.g., there are 4 reasons, some of which are elaborated, or there are more than 4 reasons.) All reasons are clearly relevant and accurate.
Alternative Perspective	The writer provides one elaborated reason or more than one reason for an alternative perspective. The writer explains why the chosen position is more reasonable than the alternative using accurate and relevant reason(s).

Figure 2
Scoring rubric for the maximum score on the Writing Argument task

Several researchers have recently criticized typical approaches to analyzing the production of arguments for focusing too heavily on the structural elements that are either present or absent in student responses (Chinn, Duncan, Hung, & Rinehart, 2016; Nussbaum, 2011; Rapanta, Garcia-Mila, & Gilabert, 2013). That is, most assessment frameworks currently follow the Toulmin model (1958) by awarding credit for the key elements of an argument, such as positions and reasons, while largely disregarding the content of these elements. Although there are currently no clear solutions to address these concerns, we aimed to improve the validity of our measures by supplementing the analytic rubrics for scoring written arguments with a list of *relevant* and *acceptable* statements, supporting and opposing a given position. For the *Writing Argument* task, the list was developed in our earlier studies using code-based analysis of student responses (Reznitskaya et al., 2001; Reznitskaya, Kuo, Glina, & Anderson, 2009). This list was designed to address two criteria of argumentation quality: acceptability of reasons/evidence and logical validity.

Students received four individual scores on the *Writing Argument* task. The first three scores corresponded to the analytic categories listed in Figure 2 and included *Writing Argument-Clarity, Writing Argument-Support, and Writing Argument-Alternative*. In addition, we used a summary score, *Writing Argument-Total* by adding student scores on all three categories.

Reading Argument

The stimulus for this task was a short passage (482 words), called "Are Zoos Good Places for Animals?" In the passage, two fictional characters present opposing arguments on this issue. Students read the passage individually and were asked to list key argument elements (i.e., positions and reasons) from the passage in writing. Based on the information from our pilot studies, students were given 30 minutes, and the original text was left with the students as they worked on this task. The instructions for listing key argument elements are shown in Figure 3.

Please read the directions carefully.
There are two opinions discussed in the story.

Opinion #1
In one sentence, in your own words, please state one of the opinions discussed in the story:
In your own words, please list and number all the reasons for this opinion:

Opinion #2
Now, in one sentence, in your own words, please state the other opinion discussed in the story:
In your own words, please list and number all the reasons for this second opinion:

Figure 3
Instructions for Reading Argument task.

To score student performance on *Reading Argument,* we developed an analytic rubric, in which we listed positions and reasons for each side of the issue from the original text. We then compared positions and reasons from the original text to the corresponding statements listed in a student response. If a given statement contained the same key terms and expressed the same meaning, we scored it as present. To increase the validity of scoring, we made a distinction in our rubric between main reasons and their elaborations, such as related evidence and examples. For instance, in the original text, one of the main reasons for supporting zoos was that they "help to protect rare animals." This reason was elaborated with a statement that "mountain gorillas have been saved from extinction by being bred in zoos." Students who listed both ideas received credit for the main reason and for its elaboration.

There were four summary measures for *Reading Argument* task:

- *Reading Argument-Position*, or the total number of positions from the original text listed in a student response.
- *Reading Argument-Reasons*, or the total number of main reasons for both positions from the original text listed in a student response.
- *Reading Argument-Elaborations*, or the total number of elaborations (i.e., evidence and examples) from the original text listed in a student response.
- *Reading Argument-Total*, or the total number of positions, reasons, and elaborations from the original text listed in a student response.

Validating Measures of Argument Production and Comprehension

Reliability and Validity Studies

The data for reliability and validity studies was collected as part of the larger research project, conducted over three years in public schools in Ohio and New Jersey. (e.g., Reznitskaya & Wilkinson, 2015; Wilkinson et al., 2017). The long-term goal of this project was to design and test a comprehensive professional development (PD) program to help teachers support the development of students' argumentation skills. Each year constituted a new iteration of the professional development program. We conducted the reliability and validity studies of the final *Writing Argument* and *Reading Argument* measures during the last (third) year of the PD project.

The last year of the PD project was a quasi-experimental study conducted in 26 fifth-grade classrooms from public schools in Ohio (12 classrooms) and New Jersey (14 classrooms). Teachers from seven classrooms at each site, Ohio and New Jersey, were randomly assigned to participate in a year-long PD program (e.g., Reznitskaya & Wilkinson, 2015; Wilkinson et al., 2017). In total, teachers under experimental conditions received 36 contact hours of professional development. The control teachers (i.e., five in Ohio and seven in New Jersey) continued to use their regular instructional methods.

At the end of the school year, after the completion of the PD program in experimental classrooms, we administrated the *Writing Argument* and *Reading Argument* measures to students in the experimental and control classrooms. Although the experimental design was not necessary for the validation of our measures, it did not interfere with the validation process. Moreover, it might have helped to increase variability in student performance, thus allowing us to observe argumentation skills across a broader spectrum.

Participants

Participants were 504 fifth-grade students from 12 classrooms in Ohio and 14 classrooms in New Jersey. Ohio participants came from a large suburban

school district with 19% minority enrollment and 6% of students designated as low income (eligible for free or reduced lunch). New Jersey participants came from three urban districts, with 49% average minority enrollment and 29% of students designated as low income, on average.

Procedure

We administered the *Writing Argument* and *Reading Argument* tasks to all participating fifth-graders using the standardized instructions described above. In addition, we administered the *GMRT* (MacGinitie et al., 2002) to all students as a criterion measure.

Student responses on both tasks were de-identified before scoring. Two scorers at each site independently evaluated student performance on the *Writing Argument* task, since it required a higher level of inference-making from the scorers, compared to the *Reading Argument* task. The training of the scorers was conducted by the authors at their respective sites. During the training, both authors and two scorers independently scored seven sets of six randomly selected student responses (21 from each site, 42 total). After scoring each set, two scorers and the authors from each site met to discuss their scoring choices until at least 80% agreement was reached for each sub-score.

The *Reading Argument* was scored by a single scorer at each site. To assess the reliability of scoring, we randomly selected 55 responses from each site to be scored by two scorers (110 total, 20.18% of the data). The training was similar to that conducted for the *Writing Argument* task. During the training, both authors and a scorer at each site independently scored three sets of six *Reading Argument* responses (9 from each site, 18 total). After scoring each set, the scorer and the authors discussed the discrepancies. The training concluded once at least 80% agreement was reached. Finally, the *GMRT* was scored independently by two scorers at each site using the answer key. The discrepancies between the two scores were resolved by re-checking student answers and the answer key.

Follow-up Study of Teacher and Student Perceptions

Participants

In the year following the completion of the PD project, we recruited eight teachers and 152 students from fifth-grade classrooms in New Jersey to participate in additional research about their perceptions of the *Writing Argument* and *Reading Argument* measures. During the study, we asked teachers to share their reactions to both measures and we assessed students' interest in the readings used as prompts.

All teachers were former participants in the PD project. However, because this study was conducted in the following academic year, they taught a new cohort of students. Participants came from five urban districts, with 62% average minority enrollment and 36% of students designated as low income, on average.

Procedure

Participating teachers received the *Writing Argument* and *Reading Argument* measures and related scoring instructions. The teachers were asked to use both measures in their regular instruction and to complete an online survey to share their reactions to each measure. The survey had seven parallel statements about each of the two tasks to be evaluated with a 4-point ranking scale ranging from "strongly disagree" (1) to "strongly agree" (4). The surveys also included an open-ended question about any additional comments.

Students in participating classrooms completed the *Writing Argument* and *Reading Argument* tasks. Following the completion of each task, students were asked to indicate whether or not they enjoyed reading the stories used as prompts for the measures. We wanted to assess students' enjoyment of the stories used as prompts because affective dimensions of the experience (e.g., topic interest) play an important role in students' production and comprehension of texts (Asher, Hymel, & Wigfield, 1978; Hidi, 2001; Renninger, Suzanne Hidi, & Krapp, 1992).

Results

Reliability of Writing Argument

Table 1 presents reliability evidence for the *Writing Argument* scores, showing inter-rater reliability across two scorers within sites. Most results between two scorers assessed with Tau-b and ICC coefficients are above .70, which is within acceptable levels (Koo & Li, 2016; Stemler, 2004).

Table 1
Reliability of Scoring for Writing Argument

Measure	**OSU n=256**		**MSU n=243**	
	Tau-b	ICC[1]	Tau-b	ICC[1]
Writing Argument-Clarity	.63	.80	.58	.77
Writing Argument-Support	.73	.88	.55	.73
Writing Argument-Alternative	.83	.94	.79	.90
	Pearson r		Pearson r	
Writing Argument-Total	.87		.76	

[1] Two-way mixed, consistency, average measures intraclass correlation

Because the estimates displayed in Table 1 show consistency, rather than agreement between the scorers, we also compared means for the four summary measures of *Writing Argument*. Table 2 displays the means at both sites, which show high agreement between scorers.

Table 2
Means for Writing Argument

Measure	**OSU N=234**		**MSU n=252**	
	Scorer 1	**Scorer 2**	**Scorer 1**	**Scorer 2**
Writing Argument-Clarity	2.41	2.30	2.35	2.35
Writing Argument-Support	2.55	2.56	2.41	2.49
Writing Argument-Alternative	1.04	0.92	0.87	0.79
Writing Argument-Total	6.00	5.78	5.63	5.63

Reliability of Reading Argument

Table 3 presents Pearson correlations between two scorers and the means for each scorer at two sites. Taken together, the evidence shows high inter-rater reliability and agreement on all summary measures of *Reading Argument*. Note that the correlation for *Reading Argument-Position* was likely to be suppressed by the low variability on this measure as it only had three possible scores. The exact agreement for *Reading Argument-Position* was 85%.

Table 3
Pearson Correlations and Means for Reading Argument

Measure	*r*	**Mean**	
	n=110	**Scorer 1 (OSU) n=55**	**Scorer 2 (MSU) n=55**
Reading Argument-Positions	.77	1.64	1.63
Reading Argument-Reasons	.96	4.11	4.10
Reading Argument-Elaborations	.95	3.85	3.58
Reading Argument-Total	.97	9.59	9.31

Validity

Table 4 presents Pearson correlations among the *Writing Argument-Total, Reading Argument-Total,* and scores on the GMRT for both sites.

Table 4
Pearson Correlations Among Writing Argument-Total, Reading Argument-Total, and GMRT Test for OSU and MSU

Measure	OSU		MSU	
	Writing Argument-Total	*Reading Argument-Total*	*Writing Argument-Total*	*Reading Argument-Total*
Writing Argument-Total	1	–	1	–
Reading Argument-Total	.33** (n=232)	1	.30** (n=239)	1
GMRT	.47** (n=228)	.42** (n=232)	.28** (n=237)	.37** (n=237)

** Significant at the 0.01 level.

With some minor variations, the results are generally consistent across the two sites and show small to moderate correlations among variables. Although all correlations are statistically significant, the percentage of variance shared between pairs is not large. This implies that each task calls on students to use a specialized set of skills and knowledge.

We also examined the functioning of both measures in relation to gender and ethnicity of the students. Table 5 shows means and standard deviations on writing and reading measures by gender and ethnicity.

Table 5
Means (Standard Deviations) for Writing Argument-Total, Reading Argument-Total Based on Sex and Ethnicity

Measure	OSU				MSU			
	Gender		Ethnicity		Gender		Ethnicity	
	Male n=116	Female n=118	White n=183	Non-White n=51	Male n=131	Female n=121	White n=152	Non-White n=100
Writing Argument-Total	5.51* (1.60)	6.26* (1.46)	5.96 (1.64)	5.62 (1.27)	5.48 (1.52)	5.80 (1.36)	5.70 (1.54)	5.52 (1.29)
Reading Argument-Total	9.38* (3.79)	11.37* (3.27)	10.44 (3.56)	10.25 (4.04)	8.47 (3.89)	9.25 (3.92)	8.77 (4.02)	8.96 (3.77)

** Significant at the 0.01 level.

The independent t-tests showed that gender was significantly associated with students' performance on *Writing Argument* (t=3.73, p<01) and *Reading Argument* (t=4.33, p<01) at OSU, with girls performing better than boys on both tasks.

At MSU, the difference was not statistically significant for either writing (t=1.73, p<.08) or reading (t=1.58, p<.11). At both sites, ethnicity (i.e., white vs. non-white) was not a significant predictor for writing (OSU: t=1.40, p<.16; MSU: t=.97, p<.33) or reading (OSU t=.32, p<.75; t=–0.38, p<.70).

Teacher and Student Perceptions

Table 6 displays responses from eight teachers to a survey about their experiences with the *Writing Argument* and *Reading Argument* measures. Teachers had positive reactions to both tasks and the scoring procedures. They found both tasks useful and relevant to their instructional goals.

Table 6

Means and (Standard Deviations) of Teacher Responses to Survey on the Practicality of Argument Measures (1=Strongly Disagree, 4=Strongly Agree, n=8)

#	**Survey Items**	**Writing Argument**	**Reading Argument**
1	The task was useful for my students	3.75 (0.46)	3.88 (0.35)
2	The task was aligned with instructional goals I have for my students	4.00 (0.00)	3.75 (0.46)
3	The instructions, prompts, and texts for the task were clear, specific, and age-appropriate	3.50 (0.93)	3.38 (0.92)
4	I will use the task in the future	3.75 (0.46)	3.75 (0.46)
5	The scoring rubric for the task was easy to use	3.63 (0.52)	3.50 (0.53)
6	The scoring rubric for the task allowed me to learn important information about my students	3.63 (0.52)	3.88 (0.35)
7	I will use the scoring rubric for the task in the future	3.13 (1.36)	3.38 (0.52)

The positive survey results in Table 6 were also supported by teachers' comments in an open-ended part of the survey, in which teachers responded to the question "Please share any additional thoughts about administering and scoring Writing/Reading Argument tasks and about how we can improve them." For example, teachers found both measures to be interesting, authentic, pedagogically effective, and age-appropriate:

- *I liked how both of the stories were authentic and the kids could relate to them.*
- *I loved the* [writing] *story and the exercise that went with it. I think it really asks the students to think critically and decide what they feel is right or wrong. They*

have to back up their reasoning with evidence, which is something I always require in the writing process, so I absolutely loved that they had to do this. The story was great because it was easy for the students to connect with.

- *The students enjoyed the zoo piece [reading] and listing their reasons to support their opinions.*
- *These were very useful and informative assignments, both of which I will use again… They were both engaging, as well as age/grade-appropriate.*

Teachers generally thought that the scoring rubrics for both tasks were helpful and informative, although some teachers commented on their initial struggle to learn how to use the rubrics:

- *I found the rubrics to be very helpful and easy to use... The [writing] rubric is one design we use quite often and I was more familiar with it.*
- *The [writing] rubric was easy to use, as well as informative. I love to use rubrics that I can teach into. This one was clear and would be easy for students to use to reflect upon their own writing.*
- *The answer key [reading] saved me a lot of time and allowed me to focus my own thinking on the key points of the texts. Also, it created greater levels of consistency in grading.*
- *The* [reading] *rubric seemed complicated at first and took a lot of thought, but once I got the hang of it I found it easy to use.*

As for the students (n=152), they had generally positive reactions to the stories used as prompts for *Writing Argument* and *Reading Argument* measures. The majority of the students (72.00% for *Writing Argument* and 74.6% for *Reading Argument*) found both stories enjoyable, regardless of their gender (Writing χ^2=.17, p<.92, Reading χ^2 =1.66, p<.44, df=2) or ethnicity for *Reading Argument* (χ^2 =2.80, p<.25, df=2). There were statistically significant differences for *Writing Argument*, with non-white students enjoying the story more (χ^2 =9.51, p<.01, df=2).

Discussion

In this paper, we discussed the development and validation of measures of argument production and comprehension. We conducted multiple pilot studies and examined psychometric properties of the measures. Our results show that both measures have acceptable inter-rater reliability. The correlations among *Writing Argument, Reading Argument* and *GMRT* are consistent across two sites and statistically significant. The size of the correlations is small to moderate, which is a finding that requires further investigation. It suggests

that task-specific skills may play a more important role in the quality of performance than we assumed based on the domain-general, schema-theoretic views of cognition.

The performance on *Writing Argument* and *Reading Argument* tasks was similar, regardless of the ethnicity of the students. Girls performed better on the *Writing Argument* and *Reading Argument* tasks at the Ohio (OSU) site, but not at the New Jersey (MSU) site. Although the majority of the students liked the stories used as prompts, non-white students related to the writing prompt more than white students.

Teachers found both measures helpful and practical for use in their classroom. This is encouraging, especially considering that practitioners have difficulty teaching argumentative writing and reading, and often lack classroom resources to support instruction (Newell et al., 2011). Our *Writing Argument* and *Reading Argument* tasks can assist teachers with designing lessons to engage students in argumentation, as well with identifying strength and weaknesses in students' performance. These tasks can also support professional development efforts aimed at helping teachers develop their knowledge of argumentation. This is especially important since, despite recent national requirements in the US to engage students in argumentation from the early grades (National Governors Association, 2010), teachers are often unaware of fundamental concepts, such as the structure of arguments and the criteria for evaluating argument quality (Kuhn, 2005; Newell et al., 2011).

Despite the evidence to support the intended functioning of both measures and positive reactions from practitioners, we recognize the need for continuous revision and testing of the measures. For example, we have begun the work of creating parallel forms for these measures to allow practitioners and researchers to accurately measure students' development over time. In addition, future studies should explore the reasons behind the differential performance and the attitudes towards the writing task based on demographic characteristics such as gender and ethnicity. We also need to examine how argumentation skills relate to other variables, such as epistemic cognition (Mason & Scirica, 2006; Weinstock et al., 2006) and personality characteristics, including the need for cognition and extraversion (Nussbaum & Bendixen, 2003).

Furthermore, our measures are designed to prioritize the assessment of domain-general competencies. They assess argumentation quality across disciplines based on criteria such as the acceptability of reasons and evidence and the logical validity of inferences, which are assumed to be applicable across multiple domains (e.g., Govier, 2010; Hollihan & Baaske, 1973; Nielsen, 2013; Toulmin, 1958). Current research supports the existence of general structures and skills of argumentation that can be used across multiple domains (Fischer, Chinn, Engelmann, & Osborne, 2018; Klahr, Zimmerman,

& Jirout, 2011). At the same time, studies suggests that thinking through a complex problem also relies on domain-specific competencies including "negotiated norms and conventions that shape knowledge claims and argumentation within each disciplinary community" (Goldman et al., 2016, p. 223). Future research should focus more on identifying and measuring both domain-general and domain-specific competencies in argumentation.

We also need to continue exploring alternative frameworks for assessing the quality of argumentation. For example, similar to most other studies (for review see Ferretti & Fan, 2016), the implicit purpose of our *Writing Argument* task was to persuade an opponent that a chosen position was better than alternatives. Writing to persuade is a meaningful task, which is crucial for academic success and important for many professional and personal endeavors (Graff, 2003; Kuhn et al., 2016). Teachers in US schools are expected to engage students in persuasive writing, and this is the skill that is emphasized in national policy documents and routinely assessed on standardized tests (National Governors Association, 2010). However, engaging in argumentation can serve a variety of other purposes, such as developing a deeper understanding of a problem (i.e., inquiry) or reaching a compromise (i.e., negotiation) (Walton, 1998). Assessment tasks that address different purposes for argumentation may offer students distinct opportunities to practice, learn, and demonstrate relevant skills.

Also, our *Writing Argument* and *Reading Argument* measures were still largely based on the Toulmin model (1958), which defines the structure and core elements of an argument. By focusing on the structure, rather than the content of arguments, such assessments reveal some aspects of reasoned argumentation but may omit other important features, such as accuracy, relevance, coherence, and persuasiveness (Chinn et al., 2016; Driver, Newton, & Osborne, 2000; Erduran, Simon, & Osborne, 2004; Newell et al., 2011; Nussbaum & Ordene, 2011). In a powerful example, Chinn et al. (2016) used a written argument against vaccination to demonstrate that a seriously flawed reasoning can remain undetected, and even receive high scores based on the presence of key structural elements.

In this study, we tried to address the limitations of the Toulmin model by generating task-specific lists of acceptable and relevant statements against which we compared student performance. Such enhancements are helpful, but they fall short of providing a theoretically-sound and transparent assessment of argumentation quality. We need to continue searching for new analytic approaches that can offer a more comprehensive and precise assessment of various skills involved in comprehension and production of arguments.

References

Anderson, R. C., Nguyen-Jahiel, K., McNurlen, B., Archodidou, A., Kim, S.-y., Reznitskaya, A., Tillmanns, M., & Gilbert, L. (2001). The snowball phenomenon: Spread of ways of talking and ways of thinking across groups of children. *Cognition and Instruction, 19*(1), 1–46.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson, R. Barr, M. L. Kamil & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 255–291). New York: Longman.

Asher, S. R., Hymel, S., & Wigfield, A. (1978). Influence of topic interest on children's reading comprehension. *Journal of Reading Behavior, 10*(1), 35–47.

Chambliss, M. J., & Murphy, P. K. (2002). Fourth and fifth graders representing the argument structure in written texts. *Discourse Processes, 34*(1), 91–115.

Chinn, A. C., Duncan, R. G., Hung, L. C.-C., & Rinehart, R. W. (2016). *Epistemic criteria and reliable processes as indicators of argument quality in science students' argumentation*. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont: Wadsworth.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287–312.

Ennis, R. (1996). *Critical thinking*. Upper Saddle River: Prentice Hall.

Ennis, R. (2003). Critical thinking assessment. In D. Fasko (Ed.), *Critical thinking and reasoning* (pp. 293–313). Cresskill: Hampton Press, Inc.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of toulmin's argument pattern for studying science discourse. *Science Education, 88*(6), 915–933.

Ferretti, R. P., & Fan, Y. (2016). Argumentative writing. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research, Second edition*. New York: Guilford Press.

Fischer, F., Chinn, C. A., Engelmann, K., & Osborne, J. (Eds.). (2018). *Scientific reasoning and argumentation: The roles of domain-specific and domain-general knowledge*. New York: Routledge.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive psychology, 15*(1), 1–38.

Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M., Greenleaf, C., Lee, C.D., Shanahan, C., & Project, R. (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist, 51*(2), 219–246.

Govier, T. (1987). *Problems in argument analysis and evaluation*. Providence: Foris.

Govier, T. (2010). *A practical study of argument*. Belmont: Wadsworth Publishing Company.

Graff, G. (2003). *Clueless in academe*. New Haven: Yale University Press.

Halpern, D. F. (2003). The "how" and "why" of critical thinking assessment. In D. Fasko (Ed.), *Critical thinking and reasoning* (pp. 331–354). Cresskill: Hampton Press, Inc.

Healy, P. (1987). Critical reasoning and dialectical argument: An extension of Toulmin's approach. *Informal Logic, 9*(1), 1–12.

Hidi, S. (2001). Interest, reading, and learning: Theoretical and practical considerations. *Educational Psychology Review, 13*(3), 191–209.

Hollihan, T. A., & Baaske, K. T. (1973). *Arguments and arguing: The products of human decision making*. Prospect Hights: Waveland.

Hughes, J. N. (1992). Review of the Cornell critical thinking tests. In J. J. Kramer, J. C. Conoley & L. L. Murphy (Eds.), *The eleventh mental measurements yearbook*. Lincoln: The Buros Institute of Mental Measurements.

Johnson, K. M., (2005). Gates-MacGinitie reading tests fourth edition forms S and T. In B. S. Plake & R. A. Spies (Eds.), *The sixteenth mental measurements yearbook* (Vol. 16). Lincoln: Buros Institute of Mental Measurements.

Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science, 333*(6045), 971–975.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Kuhn, D. (1991). *The skills of argument*. Cambridge: Cambridge University Press.

Kuhn, D. (2005). *Education for thinking*. Boston: Harvard Education Press.

Kuhn, D., & Crowell, A. (2011). Dialogic argumentation as a vehicle for developing young adolescents' thinking. *Psychological Science, 22*(4), 545–552.

Kuhn, D., Hemberger, L., & Khait, V. (2016). *Argue with me: Argument as a path to developing students' thinking and writing*. New York: Routledge.

Lipman, M. (2003). *Thinking in education*. New York: Cambridge University Press.

MacGinitie, W. H., MacGinitie, R. K., Maria, K. & Dreyer, L. G. (2002). *Gates-MacGinitie reading tests*. Rolling Meadows: Riverside Publishing.

Mason, L., & Scirica, F. (2006). Prediction of students' argumentation skills about controversial topics by epistemological understanding. *Learning and Instruction, 16*(5), 492–509.

McCabe, P. P. (2005). Review of the Gates-MacGinitie reading tests. In R. A. Spies & B. S. Plake (Eds.), *The Sixteenth Mental Measurements Yearbook*. Lincoln: Buros Institute of Mental Measurements.

Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction, 14*(2), 139–178.

Mercer, N. (2011). Reasoning serves argumentation in children. *Cognitive Development, 26*(3), 177–191.

National Governors Association. (2010). *Common core state standards: Appendix A. research supporting key elements of the standards*. Washington DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

Newell, G. E., Beach, R., Smith, J., VanDerHeide, J., Kuhn, D., & Andriessen, J. (2011). Teaching and learning argumentative reading and writing: A review of research. *Reading Research Quarterly, 46*(3), 273–304.

Nielsen, J. A. (2013). Dialectical features of students' argumentation: A critical review of argumentation studies in science education. *Research in Science Education, 43*(1), 371–393.

Norris, S. P. (1995). Format effects on critical thinking test performance. *The Alberta Journal of Educational Research, 41*(4), 378–406.

Norris, S. P., Leighton, J. P., & Phillips, L. M. (2004). What is at stake in knowing the content and capabilities of children's minds? A case for basing high stakes tests on cognitive models. *Theory & Research in Social Education, 2*(3), 283–308.

Nussbaum, E. M. (2011). Argumentation, dialogue theory, and probability modeling: Alternative frameworks for argumentation research in education. *Educational Psychologist, 46*(2), 84–106.

Nussbaum, E. M., & Bendixen, L. D. (2003). Approaching and avoiding arguments: The role of epistemological beliefs, need for cognition, and extraverted personality traits. *Contemporary Educational Psychology, 28*(4), 573–595.

Nussbaum, E. M., & Kardash, C. M. (2005). The effects of goal instruction and text on the generation of counterarguments during writing. *Journal of Educational Psychology, 9*(2), 157–169.

Nussbaum, E. M., & Ordene, V. (2011). Critical questions and argument stratagems: A framework for enhancing and analyzing students' reasoning practices *The Journal of the Learning Sciences, 20*(3), 443–488.

Oka, E. R., & Paris, S. G. (1987). Patterns of motivation and reading skill in underachieving children. In S. J. Ceci (Ed.), *Handbook of cognitive, social, and neuropsychological aspects of learning disabilities* (pp. 115–145). New Jersey: Lawrence Erlbaum Associates.

Page-Voth, V., & Graham, S. (1999). Effects of goal setting and strategy use on the writing performance and self-efficacy of students with writing and learning problems. *Journal of Educational Psychology, 91*(2), 230–240.

Partnership for 21st Century Skills. (2012). *A Framework for 21st Century Learning.* Retrieved from http://www.p21.org/index.php

Phillips, L. M., & Patterson, C. C. (1987). *Test of inference ability in reading comprehension*. Newfoundland: Institute for Educational Research and Development.

Poteet, J. (1989). Review of the Ennis-Weir critical thinking essay test. J. C. Conoley & J. J. Kramer (Eds.), *The tenth mental measurements yearbook* (pp. 289–290). Lincoln: The Buros Institute of Mental Measurements.

Rapanta, C., Garcia-Mila, M., & Gilabert, S. (2013). What is meant by argumentative competence? An integrative review of methods of analysis and assessment in education. *Review of Educational Research, 83*(4), 483–520.

Renninger, K. A., Suzanne Hidi, & Krapp, A. (Eds.). (1992). *The role of interest in learning and development.* Mahwah: Lawrence Erlbaum Associates.

Reznitskaya, A., Anderson, R. C., Dong, T., Li, Y., Kim, I., & Kim, S. (2008). Learning to think well: Application of argument schema theory. In C. C. Block & S. Parris (Eds.), *Comprehension instruction: Research-based best practices* (pp. 196–213). New York: Guilford Press.

Reznitskaya, A., Anderson, R. C., McNurlen, B., Nguyen-Jahiel, K., Archodidou, A., & Kim, S. (2001). Influence of oral discussion on written argument. *Discourse Processes, 32*(2 & 3), 155–175.

Reznitskaya, A., & Gregory, M. (2013). Student thought and classroom language: Examining the mechanisms of change in dialogic teaching. *Educational Psychologist, 48*(2), 114–133.

Reznitskaya, A., Kuo, L., Clark, A., Miller, B., Jadallah, M., Anderson, R. C., & Nguyen-Jahiel, K. (2009). Collaborative reasoning: A dialogic approach to group discussions. *Cambridge Journal of Education, 39*(1), 29–48.

Reznitskaya, A., Kuo, L., Glina, M., & Anderson, R. C. (2009). Measuring argumentative reasoning: What's behind the numbers? *Learning and individual differences, 19*(2), 219–224.

Reznitskaya, A., & Wilkinson, I. A. G. (2015). Professional development in dialogic teaching: Helping teachers promote argument literacy in their classrooms. In D. Scott & E. Hargreaves (Eds.), *Sage handbook of learning* (pp. 219–232). London: Sage Publications.

Reznitskaya, A., & Wilkinson, I. A. G. (2017). *The most reasonable answer: Helping students build better arguments together*. Boston: Harvard Education Press.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99–136). Hillsdale: Erlbaum.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 1–14.

Stein, N. L., & Trabasso, T. (1982). Children's understanding of stories: A basis for moral judgment and dilemma resolution. In C. J. Brainerd & M. Pressley (Eds.), *Verbal processes in children: Progress in cognitive development research* (pp. 161–188). New York: Springer-Verlag.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4), 1–19.

Sutton, R. E. (1992). Review of the New Jersey test of reasoning skills. In J. J. Kramer, J. C. Conoley & L. L. Murphy (Eds.), *The eleventh mental measurements yearbook* (pp. 606–608). Lincoln: The Buros Institute of Mental Measurements.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Walton, D. (1998). *The new dialectic: Conversational contexts of argument*. Toronto: University of Toronto Press.

Weinstock, M. P., Neuman, Y., & Glassner, A. (2006). Identification of informal reasoning fallacies as a function of epistemological level, grade level, and cognitive ability. *Journal of Educational Psychology, 98*(2), 327–341.

Wilkinson, I. A. G., Reznitskaya, A., Bourdage, K., Oyler, J., Nelson, K., Glina, M., Drewry, R., Kim, M.-Y. (2017). Toward a more dialogic pedagogy: Changing teachers' beliefs and practices through professional development in language arts classrooms. *Language & Education, 31*(1), 65–82.

Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009). Argumentation schema and the my side bias in written argumentation. *Written Communication, 26*(2), 183–209.

Corresponding authors
Alina Reznitskaya
The College of Education and Human Services, Department of Educational Foundations, Montclair State University
Email: reznitskayaa@mail.montclair.edu

Ian A. G. Wilkinson
Curriculum and Pedagogy, Faculty of Education and Social Work, The University of Auckland
E-mail: ian.wilkinson@auckland.ac.nz