Learning and Individual Differences 19 (2009) 219–224

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

# Measuring argumentative reasoning: What's behind the numbers?

Alina Reznitskaya [a,*], Li-jen Kuo [b], Monica Glina [a], Richard C. Anderson [c]

[a] Montclair State University, United States
[b] Northern Illinois University, United States
[c] University of Illinois at Urbana-Champaign, United States

ARTICLE INFO

*Article history:*
Received 3 May 2008
Received in revised form 29 October 2008
Accepted 3 November 2008

*Keywords:*
Argumentation
Alternative assessment
Scoring rubrics
Reliability
Validity

ABSTRACT

The aim of this paper is to develop a more thorough, empirically-based understanding of the differences in measurement of written argumentation when alternative scoring frameworks are employed. Reflective compositions of 127 elementary school children were analyzed using analytic and holistic scales. The scales were derived from Argument Schema Theory, an explicit model of argumentation development. We investigated the relationships among the different scales, as well as their relative reliability and efficiency. The scores derived using analytic and holistic methods have adequate reliability. Although less efficient, analytic scoring allows for gathering more sensitive and detailed information about the differences in student performance. The results suggest that the choice of an analytic framework for measuring argumentation should not be arbitrary, as each scoring method taps into distinct facets of the construct.

Published by Elsevier Inc.

Many educators today strongly advocate for the use of assessment tools that allow respondents to demonstrate their competence with a greater degree of flexibility (Baron, 1996; Strickland & Strickland, 1998). They argue that open-ended formats are more compatible with contemporary theories of learning and instruction (Shepard, 2000) and are better suited for measuring higher-order educational outcomes, such as reasoning (Grounlund, 1998; O'Neil, 1992). Yet, given that each student might react to the task requirements uniquely, open-ended formats present challenges in relation to the meaningfulness, consistency, and efficiency of scoring.

This paper addresses the need for developing a more thorough, empirically-based understanding of the potential differences in the measurement of reasoning with open-ended formats when alternative scoring frameworks are employed. Two scoring methods were investigated: analytic and holistic. We explored the conceptual relationships among different scoring methods and subscales, and analyzed the reliability and efficiency of each method.

## 1. Measuring argumentative reasoning

The literature on reasoning is full of alternative definitions for this complex construct, as well as its composite skills (e.g., Ennis, 1986; Kuhn, 1992; Voss & Means, 1991). In this paper, we adopt the idea that reasoning is "a flow of propositions within a discourse of reasoned argumentation" (Anderson et al., 2001, p. 2). Following Vygotsky and others (Mead, 1962; Vygotsky, 1962), argumentation is viewed as thoroughly situated and negotiated within a particular social context.

To further understand the construct of argumentative reasoning, we rely on a theoretical model called Argument Schema Theory (AST) (Reznitskaya & Anderson, 2002). According to AST, argumentative knowledge is represented through a skeletal mental structure we call an *argument schema*. Based on normative models of a rational argument (e.g., Toulmin, 1958; Walton, 1996), we propose that a developed argument schema will include such elements as the position, reasons, grounds, warrants, backing, modifiers, counter-arguments, and rebuttals. It contains an understanding of the rhetorical organization of an argument, its properties, functions, and conditions for use. Importantly, an argument schema is more than a simple collection of individual elements. Rather, the elements and their relationships are supported through a set of beliefs, which constitute an "explanatory framework" (Mishra & Brewer, 2003) for the schema. The explanatory framework for an argument schema is the insight into the function and value of a rational argument as a means for choosing among alternatives. According to Govier, such insight is "something quite elementary and yet illusive 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" (Govier, 1987, p. 233).

An argument schema can be further broken down into recurrent patterns, or *argument stratagems* (Anderson et al., 2001). Argument stratagems are specific rhetorical and reasoning moves utilized in argumentation. For example, in a study examining children's debates based on fictional stories they had read (Anderson et al., 2001), we noted that children often used such expressions as "in the story, it

* Corresponding author. Montclair State University, University Hall 2193, Montclair, NJ 07043, United States. Tel.: +1 973 655 4080.
*E-mail address:* reznitskayaa@mail.montclair.edu (A. Reznitskaya).

1041-6080/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.lindif.2008.11.001

**Table 1**
Research questions.

| Question | Description |
|---|---|
| 1 | What is the relationship between holistic and analytic scoring of argumentation? |
| 2 | What is the overall structure of argumentative ability? |
| 3 | Does the use of different scoring methods result in consistent substantive interpretations when educational treatments are compared? |
| 4 | How different are interrater reliability ratings for analytic versus holistic scoring methods? |
| 5 | Which scoring method of written argumentation is more cost-effective? |

said" or "on page 23, she said," to explicitly mark information as coming from the story in order to enhance its credibility and to add to the persuasive force of their arguments. We labeled this stratagem with the general form "In the story, it said [EVIDENCE]." The capitalized, bracketed part of the stratagem will change in response to contextually different scenarios. However, the underlying purpose, form, possible consequences, and objections to this stratagem will remain the same. The use of appropriate stratagems may help students generate arguments that are both more elaborate and better focused (Reznitskaya, Anderson, & Kuo, 2007; Reznitskaya et al., 2001).

The present study applied AST to developing two methods for scoring performance on an open-ended task designed to measure reasoning abilities of elementary school students. Consistent with our definition of reasoning and the related theoretical framework, the task required students to reflect on a dilemma facing a character in a fictional story. The story described an age-appropriate ethical problem, providing information that could be used to support contrasting resolutions. Thus, reasoning was tested through engaging students in argumentative discourse within a context-rich ill-structured situation. Students responded to the task individually and in writing, following a general prompt.

Scoring methods typically used to evaluate argumentation performance on open-ended written tasks can be grouped into two broad categories: analytic and holistic. Generally, analytic scoring focuses on one characteristic of performance at a time, while holistic scoring relies on a single overall score that takes into account the entire response. Studies using both scoring methods often utilize Toulmin's model of argument to evaluate student performance (e.g., Chambliss, 1995; Martttunen, 1992; McCann, 1989). This model also provides an important starting point for AST and the ensuing scoring frameworks by incorporating claims (positions) and data (reasons) as the key elements of an argument schema. Notably, Toulmin's model does not explicitly include counterarguments, which are central to our concept of argumentation. Following scholars within the social learning tradition (Bakhtin, 1986; Vygotsky, 1981), we view reasoning as inherently dialogical. Even the solitary construction of an argument is modeled after a dialogue with others that is focused on the consideration and evaluation of alternatives. In our theory and related scoring procedures, we expanded Toulmin's model to incorporate and emphasize counterarguments. Two other additions to Toulmin's model include the explanatory framework and argument stratagems described earlier.

While both analytic and holistic scoring methods are often used in studies of argumentation (e.g., Crowhurst, 1987; Freedman & Pringle, 1988), there has been no systematic examination of their differences. Yet, each method is based on certain assumptions regarding the attribute being measured, and these assumptions need to be examined empirically. This study analyzed different scoring methods and compared their reliability and efficiency. Table 1 lists specific research questions addressed in this study.

## 2. Method

The data for this study comes from our previous research in which we conducted a quasi-experiment to investigate the effectiveness of two instructional methods intended to promote the development of argumentation (Reznitskaya et al., 2007). The instructional methods included group discussions of controversial issues raised in children's readings and explicit instruction in argumentation. 127 elementary school students, 56 boys and 71 girls, participated in the study. 77% of study participants were European Americans, 27% were African Americans, and 5% were from other ethnic groups.

Classrooms that were matched on demographic and school ability variables were assigned to one of the three treatment conditions, as summarized in Table 2.

Having completed their respective treatments, students wrote a reflective essay in response to an 820-word story. In the story (McNurlen, 1998), an unpopular boy named Thomas wins the school Pinewood Derby race, but he breaks the rules by not making his car by himself. He confides to his classmate, Jack, that he has received help from his older brother in making his car. The students were asked to write an essay reflecting on whether or not Jack should tell on Thomas. Children were given 40 min to work on the essay.

Student essays were transcribed, and each composition was given an anonymous identification code to keep us blind to the treatment when we evaluated students' responses. The first author performed analytic and holistic scoring of all essays using QSR (1999) NVivo computer software, which allows for a code-based analysis of qualitative data. The second and third authors assisted with the interrater reliability analysis of 30 randomly selected student responses and recorded the time it took to score the essays using both methods.

### 2.1. Analytic scoring

In analytic scoring, we wanted to separately evaluate each proposition made by students in their essays. The coding was done in several steps elaborated in Table 3.

**Table 2**
Description of treatment conditions.

| Treatment condition | Description of activities | Number of students |
|---|---|---|
| Discussion | Students took part in group debates about their readings using a pedagogical model called Collaborative Reasoning (CR)[a]. During CR discussions, students take positions on the dilemma brought up in their readings, provide supporting reasons for their positions, use story information and personal experience as evidence, present counterarguments to their peers, and respond to the counterarguments others offer. | 41 |
| Discussions + Lessons | Students participated in CR discussions and received explicit instruction in argumentation delivered in two scripted lessons. During the lessons, children were presented with the definition, purpose, and uses of an argument. They discussed five parts of an argument, including position, reasons, supporting facts, counterarguments, and rebuttals. In addition, children in the Discussions + Lessons condition studied specific argument stratagems related to each part of the argument. For example, the teacher introduced reasons with rhetorical forms "The first reason is [REASON FOR]," "The second reason is [REASON FOR]," and counterarguments with "Some people might say [REASON AGAINST]." | 47 |
| Control | Students in the Control condition had their regular language arts instruction and did not participate in any group discussions using CR method or lessons on argumentation. | 39 |

[a] CR is an established instructional practice with a developing empirical base (Anderson, Chinn, Waggoner, & Nguyen, 1998; Chinn, Anderson, & Waggoner, 2001; Clark et al., 2003).

**Table 3**
Analytic coding procedure.

| Step | Description |
|---|---|
| 1 | Each composition was parsed out into distinct idea units. As defined by Mayer (1985), an idea unit "expresses one action or event or state, and generally corresponds to a single verb clause" (p. 71). For example, the following essay contains 8 idea units separated with a slash (/) with related key verbs underlined. I don't think Jack should tell./Thomas is poor/and might not have very much./He also probably has never won anything./Next, he doesn't have many friends./Next, he probably doesn't get good grades/and he didn't know how to make that race car./Those are the reasons why I think he shouldn't tell. |
| 2 | Each idea unit was evaluated in terms of its relevance to the main issue of whether or not Jack should tell on Thomas. Idea units that were not logically or explicitly linked to the main issue were coded as Irrelevant. For example, statements such as "the story was interesting" or "Thomas should give his trophy to his brother" were coded as Irrelevant. |
| 3 | A unique code was assigned to each relevant argument that was advanced by the students on both sides of the issue. In this way, we eventually compiled a list of all propositions for and against Jack's telling on Thomas. Once additions to the list became infrequent, the list was used to assign a unique code to each distinct and acceptable argument advanced by students in their essays. |
| 4 | Relevant propositions supporting each side of the issue identified in Step 3 were grouped into five categories: Textual, Affective, Hypothetical, Abstract, and Contextualizing. |

**Table 4**
Coding categories with selected examples and frequencies of use.

| Category | Description | Examples of propositions for and against telling on Thomas | Total number of idea units |
|---|---|---|---|
| Textual | Propositions that referenced information presented in the story. | Thomas did not build the car (For). Thomas made some of the car (Against). | 637 |
| Affective | Propositions that appealed to probable feelings and emotions related to the decision of whether or not to tell on Thomas that were not described in the story[a]. | Jack will feel better (For). Thomas will be angry at Jack (Against). | 76 |
| Hypothetical | Propositions that extended the story world and described characters' probable actions and their consequences. This category did not include the imagined emotions because all groupings in our coding system were designed to be mutually exclusive. | There could be another race (For). Thomas might beat Jack up (Against). | 268 |
| Abstract | Propositions representing generalized, abstract moral principles and rules, as well as prescriptions as to the 'right' actions of story characters. | Cheating is wrong (For). It is important to keep a secret (Against). | 352 |
| Contextualizing | Statements that reframed the situation by considering the circumstances under which the situation will or will not hold (i.e., a qualifier) or by comparing a given situation to a new one that is similar in important respects (i.e., an analogy)[b]. | It is like having someone do your homework for you (For). It is like having someone check your homework (Against). | 43 |

[a] We wanted to separately examine such statements because the importance of empathy is highlighted in several theoretical models of moral decision-making, where morality is viewed as more than detached reasoning about what is fair and right (e.g., Gilligan, 1982; Lipman, 2003; Noddings, 1992).
[b] Although not often present in the arguments of young children (Means & Voss, 1996), qualifiers and analogies may signify a more advanced level of sophistication in argumentation.

The coding resulted in five formative categories representing different types of justifications used by students to support their claims. These categories, along with related descriptions and selected examples, are presented in Table 4. Table 4 also includes the frequency of each category, measured by the total number of idea units in student essays.

The categories listed in Table 4 were derived from the data, although we consulted several frameworks described by other researchers to assist with the identification of specific groupings of student arguments that have theoretical and practical significance (Reznitskaya et al., 2001; Gleason, 1999; Means & Voss, 1996). We then developed five summative subscales reflecting various aspects of student performance, displayed in Table 5. According to AST, people with a developed knowledge of argumentation, or an advanced argument schema, will be able to generate arguments that 1) contain multiple reasons (*Fluency*), 2) consider a problem from divergent standpoints (*Flexibility*), 3) include consistent consideration of opposing views (*Alternative*), 4) logically or explicitly relate to the main issue (*Focus*), and 5) incorporate useful rhetorical moves appropriate for argumentation, or argument stratagems (*Form*).

When coding argument stratagems in relation to the fifth subscale, *Form*, we noticed that some children utilized rhetorical patterns in their essays that were uncharacteristic of argumentation (i.e., "my favorite character is," "I can make a connection," and "if I were to

**Table 5**
Summative analytic subscales.

| Subscale | Description |
|---|---|
| Fluency | Reflects the total number of positions taken on the main issues and the relevant arguments generated on both sides. |
| Flexibility | Reflects the variety in types of proposed arguments. A composition was assigned a one-point credit for using each of the categories presented in Table 4. If the same category (i.e., Textual) was used to support an alternative perspective on the issue, an additional credit was given. |
| Alternative | Reflects the ability to consider opposing perspectives on a given issue. Students received one point for presenting an alternative position and one point for each argument advanced to support the alternative. |
| Focus | Reflects the ability to write argumentative prose with a consistent control over this discourse genre. The score assigned to each person represents a proportion of the content coded as Irrelevant and Repetitive over Relevant Content. A lower value on this variable represents a better focused composition. |
| Form[a] | Reflects formal aspects of argumentative discourse. According to AST, a person with a developed argument schema will have a repertoire of various argument stratagems. For example, explicitly labeling a proposition, by using the stratagem "But some people might say [REASON AGAINST]", not only enhances the cohesion of a composition, but may also prompt one to think of an alternative reason. Expanding on our previous work with oral and written arguments of elementary school children (Anderson et al., 2001; Reznitskaya et al., 2001), we examined the occurrence of the following five stratagems: "The reason is [REASON]," " In the story it said [EVIDENCE]," "For example [EXAMPLE]," "Some people might say [OPPOSING POSITION/REASON]," and "In conclusion [MAIN CLAIM]." |

[a] When coding argument stratagems in relation to *Form*, we considered the underlying concept or function rather than the exact wording. For example, the following statements received credit, although the surface forms of the stratagems differed from those introduced during the explicit instruction in Discussions + Lessons condition: "On page 3, it said [EVIDENCE];" "Probably the only reason why anyone else would want to disagree is [OPPOSING REASON];" "This is why I think [MAIN CLAIM]." Thus, children in all treatment conditions could receive credit for using organizational text markers appropriate for argumentation, whatever the exact wording.

change one thing"). These textual markers are more representative of a general "literature-response" discourse schema, which children are likely to learn through their regular language-arts instruction. Research on text-processing indicates that students often apply familiar discourse structures to new reading and writing tasks, even when the use of these structures is counterproductive (Scardamalia & Bereiter, 1986; van Dijk & Kintsch, 1983). While we awarded credit to students who used appropriate argument stratagems in their writing (see Table 5 for details), we subtracted points on the *Form* variable for textual markers uncharacteristic of argumentation, given that they were followed by content not related to the question of whether or not Jack should tell on Thomas.

### 2.2. Holistic scoring

With *Holistic* scoring, we wanted to simultaneously focus on several macro-level features of students' compositions, thus getting an indication of an overall schematic structure acquired by the students and employed in their essays. We modeled our 7-point rating scale on several rubrics previously used by researchers (Freedman & Pringle, 1988; NAEP, 2000), while making it consistent with our theoretical assumptions regarding argumentation, expressed through AST. For example, in contrast with several rubrics used by others (e.g., NAEP, 2000), we did not evaluate students' writing in terms of word choice, spelling, or punctuation because our goal was to assess the quality of reasoning rather than general writing skills. The holistic scoring criteria are displayed in Table 6.

## 3. Results and discussion

We conducted several analyses of the data using SPSS software. To examine the relationship between holistic and analytic scoring of argumentation (Question 1, Table 1), we performed step-wise regression analysis, using *Holistic* scoring as a dependent variable and five analytic subscales as predictors. Multicollinearity statistics for all predictors were well within acceptable range, with the largest variance inflation factor being 2.2. *Flexibility*, was shown to be a significant predictor ($p<.01$) and accounted for 31% of the variance in holistic ratings, followed by *Form* ($p<.01$), which accounted for 12%. None of the other subscales contributed significantly to explaining the variance in *Holistic* scoring. These results indicate that holistic ratings capture only some aspects of argumentative writing, although both substantive and structural dimensions are being represented in the overall score. The moderate overlap between holistic ratings and analytic sub-scales suggests that there are unique facets of performance that are tapped into through the use of different scoring frameworks. In future studies, we plan to further clarify the differences in meanings captured by alternative scoring methods, by, for example, interviewing the raters about their decision-making process during the scoring.

**Table 6**
Holistic scoring rubric.

| Score | Description |
|---|---|
| 7 | The essay contains five argument components: positions, supporting reasons, opposing reasons, elaborations, and rebuttals;<br>There is a consistent discussion of opposing perspective;<br>The essay is well-structured and focused;<br>No irrelevant information is included, repetition is low;<br>The essay contains organizational signals appropriate for argumentation. |
| 6 | The essay states a clear position on the issue supported by elaborated reasons;<br>There is a consistent discussion of opposing perspective;<br>The essay is well focused. |
| 5 | The essay states a clear position on the issue supported by elaborated reasons;<br>There is some consideration of alternatives/qualification of chosen position, but it is not well-developed;<br>There is little or no attempt at reconciling the alternative positions;<br>The essay may contain irrelevant or repetitive information. |
| 4 | The essay contains a position on the issue supported by 4 or more distinct or elaborated reasons, which are often presented in a list-like fashion;<br>Alternative perspectives are not discussed;<br>The essay is better focused than category 3, although it may contain organizational signals inappropriate for argumentation. |
| 3 | The essay contains a position on the issue supported by 4 or more distinct or elaborated reasons (often in a list-like fashion);<br>Alternative perspectives are not discussed;<br>There is a lot of irrelevant and/or repetitive and/or inconsistent information;<br>The essays may contain organizational signals inappropriate for argumentation. |
| 2 | The essay contains a position on the issue supported fewer than 4 reasons;<br>The reasons are not elaborated;<br>Alternative perspectives are not discussed;<br>The essays may contain organizational signals inappropriate for argumentation. |
| 1 | The essay is underdeveloped, containing 2 or fewer distinct reasons;<br>The essay may contain irrelevant information;<br>Alternative perspectives are not discussed;<br>The essays may contain organizational signals inappropriate for argumentation. |

**Table 7**
Rotated component loadings.

| | Component | |
|---|---|---|
| | Content | Organization |
| Flexibility | **.90** | −.13 |
| Fluency | **.86** | .19 |
| Alternative | **.85** | −.01 |
| Form | .16 | **.82** |
| Focus | .13 | **−.80** |

*Note*: Salient loadings (those over the absolute value of .70) are in boldface type.

Next, we performed an exploratory principle component analysis (PCA) of five analytic subscales to examine the overall structure of argumentative ability (Question 2, Table 1). The analysis resulted in a clear 2-component solution, accounting for 74% of the variance. Table 7 displays rotated component loadings using varimax rotation.

The first component, which we named *Content*, accounted for 47% of the variance. It was dominated by variables related to the extensiveness of student compositions, including the number of reasons, the variety of reasons, and the consideration of alternatives. The second component, which we called *Organization*, accounted for 27% of the variance. It primarily reflected structural aspects, including consistent control over the discourse genre (i.e., the lack of irrelevant or repetitive information) and the presence of organizational signals appropriate for argumentation. Thus, written argumentation appears to be a multifaceted construct with two independent dimensions related to substantive and organizational aspects of discourse. Future studies should further examine the generalizability of the two-dimensional structure of argumentation, as this finding may have important implications for the teaching of argumentation. For example, will the same structure hold for different tasks, scoring methods, or age groups?

In our next analysis (Question 3, Table 1), we examined whether the answers to substantive research questions would be affected by the choice of a scoring framework. We compared the differences among the treatment conditions using 1) holistic scoring and 2) analytic component scores, *Content* and *Organization*, generated through PCA analysis described above. The component scores were computed using the regression procedure available in SPSS. With this method, the score represents a sum of the products of component loadings (Table 7) and the respondents' standardized scores on original variables.

We used Analysis of Variance (ANOVA) to examine the effects of an instructional treatment with three levels (i.e., Discussions, Discussions + Lessons, and Control) on students' argumentative writing. Using *Holistic* scoring, the educational method was not a statistically significant factor at $\alpha = .05$.

Analytic component scores, *Content* and *Organization*, were also analyzed as dependent variables using ANOVA.[1] The effects of the

[1] We used two univariate ANOVAs because PCA generates components that are orthogonal to each other. To adjust for the possibility of inflated Type I error rates, α was set to .05/2 = .025.

**Table 8**
Interrater reliability for holistic and analytic scales[a].

| Scale | Pearson $r$[b] |
|---|---|
| Holistic | .78 |
| Flexibility | .86 |
| Fluency | .89 |
| Alternative | .74 |
| Focus | .78 |

[a] Reliability was not examined for the *Form* subscale because the coding for *Form* was done using text search functions in NVivo and did not involve a subjective judgment by the raters.

[b] Correlation coefficients for the *Holistic*, *Alternative*, *Focus*, and *Flexibility* subscales are likely to be underestimated due the restriction of range resulting from low variability of these variables.

educational method were statistically significant for both variables ($p<.01$). Students in the Discussions condition performed better on the *Content* variable than students in the other two conditions ($p<.01$). The difference between Discussions + Lessons and Control conditions was not statistically significant for the *Content* variable. On the *Organization* variable, students in the Discussions + Lessons condition performed better than students in the other two conditions. The difference between the Discussions and the Control conditions was not statistically significant for the *Organization* variable.

With the use of analytic component scores *Content* and *Organization*, statistical results were fully consistent with our previous substantive interpretations related to the effectiveness of the educational methods designed to promote argumentation development (for more discussion of substantive interpretations, please refer to Reznitskaya et al., 2007; Reznitskaya et al., 2001). Holistic scoring, however, produced different conclusions. Based on the observed low power of the analysis using holistic scoring as a dependent variable ($\beta=.07$) and the low variability of assigned scores (SD = 1.2), we suggest that holistic scoring had limited potential for detecting the differences among the treatments designed to promote argumentation development.

In our next analysis, we considered the interrater reliability of different scoring methods (Question 4, Table 1). Thirty essays were randomly selected for the reliability analysis. In order to eliminate carry-over effects between two scoring methods, the second author rescored the essays using the holistic method, while the third author rescored the same essays using the analytic method. Both scorers were given written instructions and scoring criteria, as well as an oral explanation of how to apply them. 10 essays that were not part of the final reliability analysis were used for training. During the training, the raters scored the essays independently, and then discussed their scoring decisions with the first author in order to acquire a better understanding of the scoring systems. Reliability estimates for both scoring methods were quite high, especially considering the low variances of the original variables. Table 8 displays Pearson correlation coefficients for the holistic and analytic subscales.

Finally, we assessed the efficiency of both scoring systems by recording the time it took to code, rate, and check individual compositions (Question 5, Table 1). Analytic scoring was a much more time-consuming procedure, with the average time to score an essay equaling approximately 6 min. In contrast, scoring an essay using the holistic method averaged about 1.5 min. It is important to note, however, that despite lower efficiency, analytic scoring offers the opportunity to *separately* examine different dimensions of argumentative ability, thus providing richer diagnostic information. For example, using analytic scoring one can generate separate summaries of substantive and organizational qualities of argumentative writing (i.e., Flexibility, Fluency, Alternative, Focus, Form variables).

To conclude, this study compared data-analytic strategies used to summarize important features of written argumentation. We based our scoring frameworks on theoretically-driven aspects of argumentation that constitute the sources of students' difficulties, including developing a variety of relevant reasons, considering alternatives, and using text structures and linguistic markers appropriate for argumentation (Gleason, 1999; Kuhn, 1991; NAEP, 1999). Our analytic and holistic scoring schemes were similar to those used by other researchers of argumentative reasoning in many respects, including categories used in analytic scoring (e.g., Bensley & Haynes, 1995; Crowhurst, 1987; Means & Voss, 1996) and criteria for evaluating different levels of performance used in holistic scoring (e.g., Hidi, Berndorff, & Ainley, 2002; Knudson, 1992; Yeh, 1998). Our results may have important implications for other researchers who examine argumentation development by gathering rich verbal data on open-ended argumentative tasks and then transforming it into numerical form. While this approach can bring about many benefits typically associated with numbers, as opposed to qualitative analysis alone, the potential advantages of quantification may be lost when we express student performance in numbers without a thorough understanding of what these numbers mean. This study helped to clarify the interpretation of scores generated through the application of different scoring frameworks. Our findings suggest that by changing a scoring system, the essence of the attribute being measured may be fundamentally altered.

## References


Anderson, R. C., Chinn, C., Waggoner, M., & Nguyen, K. (1998). Intellectually stimulating story discussions. In J. Osborn & F. Lehr (Eds.), *Literacy for all: Issues in teaching and Learning* (pp. 170–186). New York: Guilford.

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

Bakhtin, M. M. (1986). *Speech genres and other late essays (V. W. McGee, Trans.)*. Austin, TX: University of Texas Press.

Baron, J. B. (1996). *Performance-based student assessment: Challenges and possibilities.* Chicago: University of Chicago Press.

Bensley, A. A., & Haynes, C. (1995). The acquisition of general purpose strategic knowledge of argumentation. *Teaching of Psychology*, *22*, 41–45.

Chambliss, M. J. (1995). Text cues and strategies successful readers use to construct the gist of lengthy written arguments. *Reading Research Quarterly*, *30*(4), 778–807.

Chinn, C. A., Anderson, R. C., & Waggoner, M. A. (2001). Patterns of discourse in two kinds of literature discussion. *Reading Research Quarterly*, *36*(4), 378–411.

Clark, A., Anderson, R. C., Kuo, L., Kim, I., Archodidou, A., & Nguyen-Jahiel, K. (2003). Collaborative reasoning: Expanding ways for children to talk and think in school. *Educational Psychology Review*, *15*(2), 181–198.

Crowhurst, M. (1987). *The effects of reading instruction and writing instruction on reading and writing persuasion*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, ERIC Document Reproduction Service No. ED 281 148.

Ennis, R. H. (1986). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 127–148). New York: Freeman and Company.

Freedman, A., & Pringle, I. (1988). Why students can't write arguments. In N. Mercer (Ed.), *Language and literacy from an educational perspective*, *Vol. 2*. (pp. 233–242). Milton Keynes, UK: Open University Press.

Gilligan, C. (1982). *In a different voice.* Cambridge, MA: Harvard University Press.

Gleason, M. M. (1999). The role of evidence in argumentative writing. *Reading & Writing Quarterly*, *14*, 81–106.

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

Grounlund, N. E. (1998). *Assessment of student achievement.* Boston, MA: Allyn & Bacon.

Hidi, S., Berndorff, D., & Ainley, M. (2002). Children's argument writing, interest and self-efficacy: An intervention study. *Learning and Instruction*, *12*, 429–446.

Knudson, R. (1992). The development of written argumentation: an analysis and comparison of argumentative writing at four grade levels. *Child Study Journal*, *22*(3), 167–183.

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

Kuhn, D. (1992). Thinking as argument. *Harvard Educational Review*, *62*(2), 155–177.

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

Martttunen, M. (1992). Commenting on written arguments as a part of argumentation skills — Comparison between students engaged in traditional vs. on-line study. *Scandinavian Journal of Educational Research*, *36*(4), 289–302.

Mayer, R. E. (1985). Structural analysis of science prose: Can we increase problem solving performance? In B. K. Britton & J. B. Black (Eds.), *Understanding of expository text* (pp. 65–87). Hillsdale, NJ: Erlbaum.

McCann, T. M. (1989). Student argumentative writing knowledge and ability at three grade levels. *Research in the Teaching of English*, *23*(1), 63–77.

McNurlen, B. (1998). *Pine Wood Derby.* Champaign, IL: Center for the Study of Reading.

Mead, G. H. (1962). *Mind, self, and society from the standpoint of a social behaviorist.* Chicago: University of Chicago Press.

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.

Mishra, P., & Brewer, W. F. (2003). Theories as a form of mental representation and their role in the recall of text information. *Contemporary Educational Psychology*, *28*, 277—303.

NAEP (1999). *National Assessment of Educational Progress Writing Report Card for the Nation and the States.* Retrieved December 8, 2004, from http://nces.ed.gov/nationsreportcard//pdf/main1998/1999462.pdf

NAEP (2000). Scoring of twelfth-grade persuasive writing. *The Nations Report Card: NAEP Facts*, *5*(3).

Noddings, N. (1992). *The challenge to care in schools: An alternative approach to education.* New York: Teachers College Press.

O'Neil, J. (1992). Putting performance assessment to the test. *Educational Leadership*, *49*(8), 14—19.

QSR (1999). *QSR NVivo [Computer software].* Victoria, Australia: Qualitative Solutions and Research.

Reznitskaya, A., & Anderson, R. C. (2002). The argument schema and learning to reason. In C. C. Block & M. Pressley (Eds.), *Comprehension instruction* (pp. 319—334). New York: Guilford.

Reznitskaya, A., Anderson, R. C., & Kuo, L. (2007). Teaching and learning argumentation. *Elementary School Journal*, *107*(5), 449—472.

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.

Scardamalia, M., & Bereiter, C. (1986). Research on written composition. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 778—803). New York: Macmillan.

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

Strickland, K., & Strickland, J. (1998). *Reflections on assessment.* Portsmouth, NH: Boynton/Cook Publsihers.

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

van Dijk, T., & Kintsch, W. (1983). *Strategies of discourse comprehension.* Orlando, FL: Academic Press.

Voss, F. J., & Means, M. L. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction*, *1*, 337—350.

Vygotsky, L. S. (1962). *Thought and language.* Cambridge, MA: MIT Press.

Vygotsky, L. S. (1981). The genesis of higher-order mental functions. In J. V. Wertsch (Ed.), *The concept of activity in Soviet psychology* (pp. 144—188). Armonk, NY: Sharpe.

Walton, D. (1996). *Argumentation schemes for presumptive reasoning.* Mahwah, NJ: Erlbaum.

Yeh, S. (1998). Empowering education: Teaching argumentative writing to cultural minority middle-school students. *Research in the Teaching of English*, *33*, 49—81.