# ASSESSING MATHEMATICAL ARGUMENTATION THROUGH AUTOMATED CONVERSATION

Gabrielle A. Cayton-Hodges
Educational Testing Service
gcayton-hodges@ets.org

*Mathematical Argumentation skills have historically been overlooked in assessment, but the inclusion of Mathematical Argumentation in the Common Core State Standards (CCSS) as one of the Standards of Mathematical Practice challenges assessment developers to assess this mathematical practice. Explanation and justification of one's own thinking to a specific audience is considered a fundamental part of this mathematical practice. Based on student demonstration of argumentation skills when engaging with peers, we have developed automated conversations with virtual teachers and peers to investigate how alternative conversational patterns influence types of student responses. This technology allows assessment developers an innovative avenue for exploring new task designs that adapt to individual users and produce additional data not found in traditional measures. Preliminary findings from this investigation are presented.*

Keywords: Assessment and Evaluation, Classroom Discourse, Elementary School Education, Technology

## Defining Mathematical Argumentation

Mathematical Argumentation has been defined most generally as "understanding relationships to make connections to new ideas" (Mueller & Maher, 2009). Hunter (2007) defines argumentation in the classroom as collaborative argumentation in which students work together through mathematics discourse "to critically explore and resolve issues which they all expect to reach agreement on ultimately." (Hunter, 2007, p. 3-18). More specifically, argumentation includes making a conjecture, proving a proposition, justifying an inference, or explaining a point. Students have been found to demonstrate argumentation skills when arguing with and asking questions of peers. In addition, explanation and justification of one's own thinking such that it can be understood by a specific audience is considered a fundamental part of this mathematical practice.

## Conversation-Based Assessment

To address the challenge of assessing mathematical argumentation, we turned to the prospects of using Conversation-Based Assessment (CBA). Such automated conversations with virtual agents have been widely used to support student learning in intelligent tutoring systems (ITS) (e.g., Graesser et al., 2004; Halpern, Millis, Graesser, Butler, Forsyth, & Cai, 2012; Millis, Forsyth, Butler, Wallace, Graesser, & Halpern, 2011). Students' interactions with these agents can be used to gather evidence about their knowledge and skills, and provide them with appropriate help (e.g., feedback, scaffolding). The use of CBA, and more specifically, "trialogues" (three-party conversations among a human student and two virtual agents) for assessment is more recent (see Yang & Zapata-Rivera, 2010), but this area of application is a natural fit for assessment purposes due to the underlying requirement for ITSs to assess relevant skills that will enable intelligent and adaptive responses. Leveraging this requirement allows assessment developers an innovative avenue for exploring new task designs that adapt to individual users and include additional data not found in traditional measures (i.e., conversational responses related to specific scaffolding).

Trialogues are one way to create learning environments that can be used to simulate particular learning strategies or social interactions (Butler, Forsyth, Halpern, Graesser and Millis, 2011). This makes this type of environment an ideal one for the assessment of argumentation skills since we are

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.

able to recreate not only the mathematical content learned in the classroom, but also the interactions that accompany them.

In the development of the task that serves as the basis for our automated conversation, we looked at what design principles could be used to structure tasks so that students' collaboration, and their discourse in particular, will be "thought-revealing" (Kelly & Lesh, 2000). Hoover, Hole, Kelly and Post (2000) proposed a set of principles for developing thought-revealing activities: 1. The model construction principle, 2. The reality principle, 3. The construct documentation principle, 4. The construct shareability and reusability principle, and 5. The effective prototype principle.

These principles suggest that a thought revealing task should require the development of "an explicit construction, description, explanation or justified prediction;" (p. 609) involve a situation that requires students to engage in meaningful mathematics; result in the creation of a product that itself provides information about student understanding; require students to produce explanations of process and not just a final product/answer; and result in the creation of an idea that can be referred back to in another context. We developed not only our underlying task, but also structured our conversations around these principles.

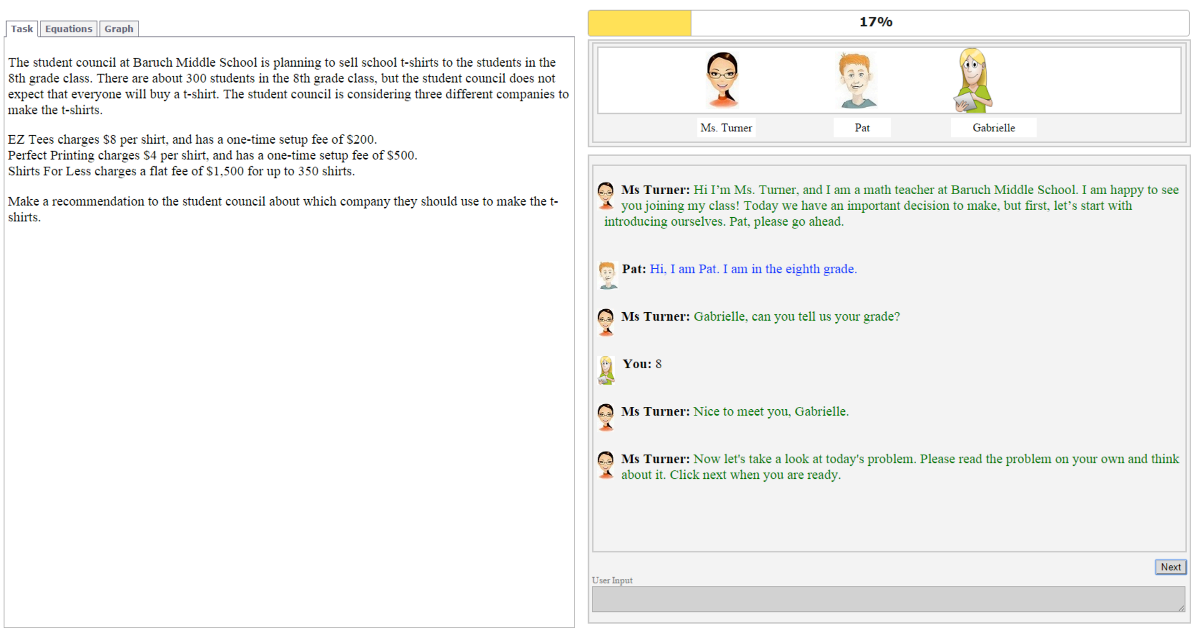


**Figure 5:** Screenshot of CBA prototype.

We developed a CBA that involves students engaging in an automated trialogue with a virtual teacher and virtual peer agents. The trialogue occurs in a simple chat-like interface as the student is led through solving a problem that involves both linear algebra and mathematical argumentation (Figure 1 shows a screenshot of the prototype, including the problem the students were asked to solve). We found that students are able to showcase their skills with mathematical argumentation through explanation, refutation, evidence, and position-taking just as would be the case in a classroom setting. Further, by utilizing automated scoring engines already integrated into the design of the CBA, we were able to come up with scores for mathematical argumentation that are more objective than the subjective scores of classroom observation or teacher rating.

Researchers have explored how best to support students' skills to support deep conversations and question-asking (see Graesser, Ozuru & Sullins, 2010). We aimed to build on this body of literature by continuing onto the next step in the development process, the framing of the questions and prompts to provide the continued support of thought-revealing responses and therefore student

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.

argumentation contained within. For instance, do students provide the most information when countering misconceptions or when responding to direct questioning? Do students respond differently when the question comes from a virtual teacher as opposed to a virtual student? These are the design questions we aimed to better understand through this study.

## The Problem Statement

In drafting the structure of automated conversations, just as in a real-life classroom situation, we must make multiple decisions as to how to query a student to elicit certain information. Sometimes, a very small change in wording of a query may elicit more, less, or different information. However, unlike a real-life situation, we do not have the opportunity to listen to nuances in the student response and ask the question again in a different way. This study looks at small changes to automated questions at three important points in the mathematical argumentation conversation to determine what impact these changes have on student responses.

**Main Research Question:** How do alternative conversational prompts influence the types of math responses gathered from students in the math prototype?

**Sub-questions:**

1. Do students respond differently in a situation where they are first asked to explain in their own words or when they are responding to another student's ideas? [Manipulation 1]
2. Do students respond differently to a question asked by a virtual student versus a virtual teacher? [Manipulation 2]
3. Do students provide a more complete argument when prompted to respond to an unlikely answer or a likely answer? [Manipulation 3]

## Study Design and Procedure

### Sample

The study was conducted with students in 8$^{th}$-grade algebra at four schools in different regions of the U.S.A. We investigated the three research sub-questions simultaneously, using the same sample of students. Students were randomly assigned to one of the eight possible conditions as shown in Figure 2 (numbers shown were planned). We had aimed for a total of 120 students but due to technical difficulties at one school that caused them to end before all students were complete, we ended with 123 records for Manipulation 1, 107 for Manipulation 2, and 74 for Manipulation 3.

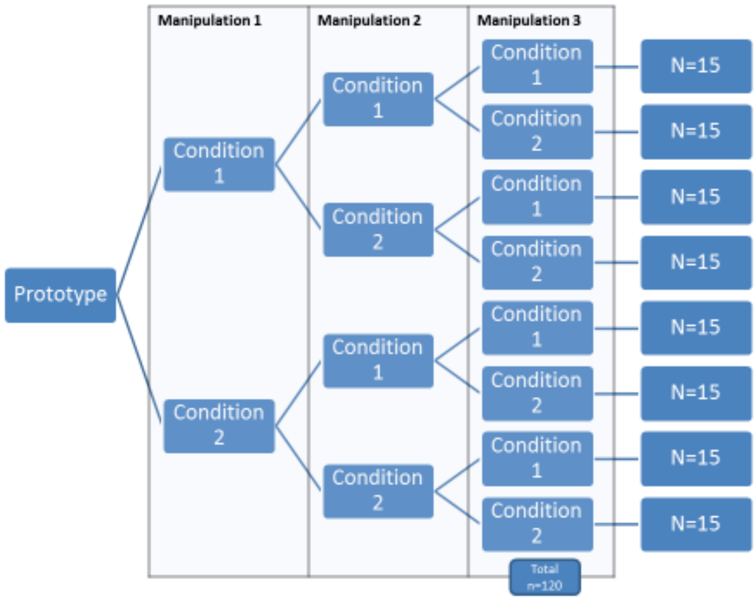


**Figure 6:** Study design.

## Instruments

Students were administered a short pretest focusing on linear algebra skills as a baseline measure. They then engaged in the automated conversation on a computer, with random assignment to one of eight conditions, as described above. They then answered a short post survey about their perceptions of the activity.

## Data Analysis

**Scoring and Analysis.** The pretest items were all automatically scored. Most of the analyses of the conversation data were also automatically scored with the exception of one longer argumentation item (Manipulation 3), which had to be scored by human raters.

**Analysis.** Each of the three manipulations has two discrete conditions that are being compared in their outcomes. Each of those manipulations was analyzed using Chi Squared Tests of Independence.

# Results

## Manipulation 1

Manipulation 1 varied by whether the initial response by the student was in reaction to a misunderstanding by the virtual student, Pat (Condition 1), or whether it was in reaction to a direct question by the virtual teacher, Ms. Turner (Condition 2). In both conditions, we were looking for the student to answer that y is the dependent variable and represents the total cost. In Condition 1, Pat offers an incorrect answer where y is the independent variable and is the cost per shirt. Figure 3 shows the conversation diagram for the first cycle of Condition 1. Condition 2 differs in the opening such that Pat does not offer an [incorrect] answer and instead Ms. Turner directs her question directly to the student.

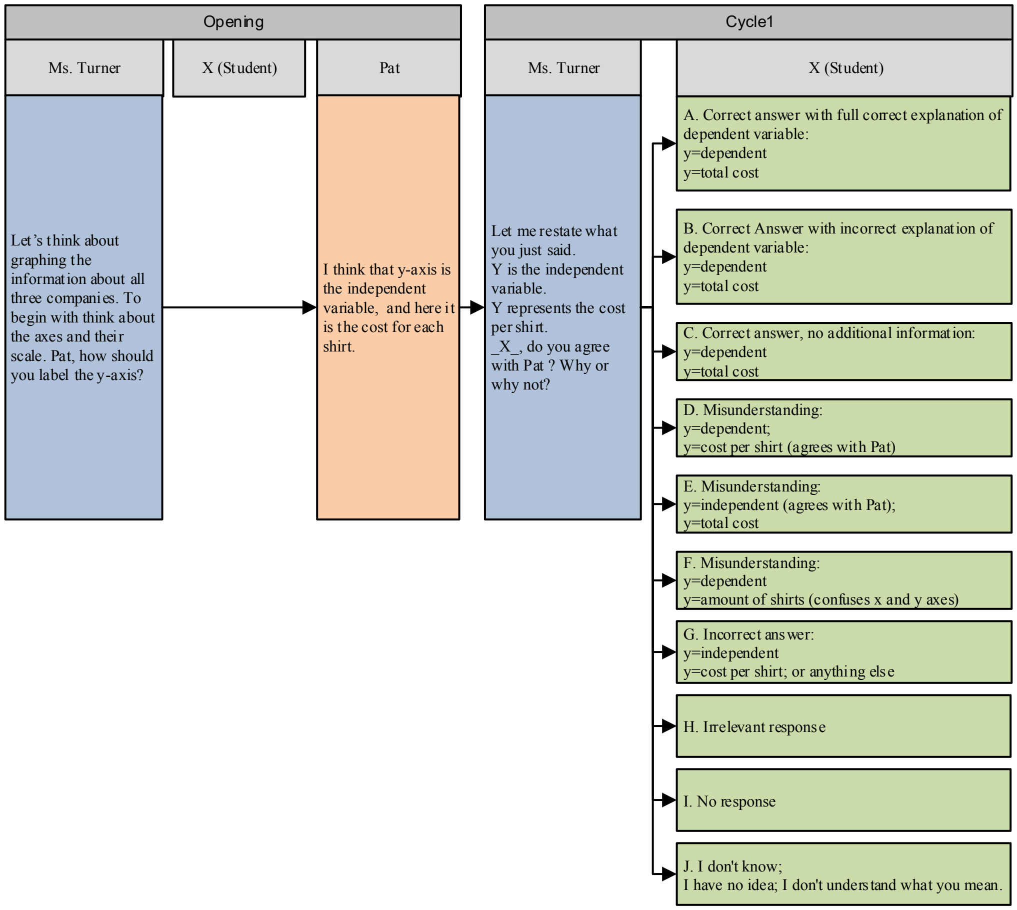


**Figure 7:** Conversation Diagram for Manipulation 1, Condition 1.

First, we looked directly at how the CBA system evaluated the student responses to the manipulated question. The results are shown in Table 1, where A is a completely correct response that leads directly to the end of the conversation, and C-F are variations of which components were correct or partially correct (B is missing as no one went down that path). G is the completely incorrect, but relevant response where the student states the same thing as Pat, that y is independent and the cost per shirt (or other relevant but incorrect labels for *y*). Other represents a path that the system computed as irrelevant, blank (i.e., no response provided by the student), or metacognitive (e.g., "I don't know") (H, I, or J).

**Table 1: Manipulation 1 Cycle 1 Response**

| | **Student Response Evaluation** | | | | | | | | **Pretest Score** |
|---|---|---|---|---|---|---|---|---|---|
| | A | C | D | E | F | G | Other | *Total* | % correct |
| Condition 1 (misunderstanding) | 1 | 2 | 5 | 5 | 2 | 29 | 16 | *60* | 66% |
| Condition 2 (direct question) | 1 | 0 | 8 | 0 | 6 | 5 | 43 | *63* | 69% |
| Total | 2 | 2 | 13 | 5 | 8 | 34 | 59 | *123* | 68% |

Students in Condition 1 were most likely to have both parts of the response incorrect (*G*) while students in Condition 2 were most likely to have a response that was interpreted as irrelevant or blank (*Other*; $\chi^2$=38.9, *p*<.001). This result was interesting as the misunderstanding by Pat was intentionally built into the script to make it clear which pieces of information were relevant to the

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.

question while allowing the student to correct the information with their own response. This approach may have backfired as it indicates that the students given the response by Pat initially were likely to agree with him. On the other hand, more than half of the students that were directly questioned by the teacher with "how should you label the *y*-axis" did not produce any relevant response, indicating that both groups may have been lacking this basic knowledge of how to contextualize linear functions. Students in Condition 1, however, had Pat's answer to copy or restate while the other group did not have any information to use.

To explore this further, we looked at student responses to follow-up questions posed after this initial response in Cycle 1 to see which students eventually arrived at the correct answer through further questioning (Table 2).

**Table 2: Manipulation 1 Final Answer**

| | Correct | Incorrect | Other | *Total* |
|---|---|---|---|---|
| Condition 1 (misunderstanding) | 23 | 21 | 16 | *60* |
| Condition 2 (direct question) | 9 | 11 | 43 | *63* |

None of the students who began down the "Other" paths were able to eventually reach the correct answer. Of the remaining students, approximately half in each conditional eventually reached the correct answer (23/44 students in Condition 1 and 9/20 students in Condition 2), demonstrating that it was only the direct response to the manipulated prompt that caused student differences, there was no further chain reaction to this manipulation.

**Manipulation 2**

For Manipulation 2, the original version of the manipulated question (Condition 1) has Ms. Turner explicitly telling the students to use $y=mx+b$ and asks what $m$ and $b$ represent. In Condition 2, Pat says "I know we're supposed to use $y=mx+b$ for the equations. But I'm not really sure what $m$ and $b$ stand for. [Student], can you help me? What do you think $m$ and $b$ stand for?" This is similar to Manipulation 1 in that the student is responding to either Pat or Ms. Turner, but it also differs from that manipulation in that the question in this case is near-identical. There is no misconception introduced, and in both conditions the student is explicitly asked to define $m$ and $b$ in $y=mx+b$. In this particular instance, the flow chart is nonlinear, that is, the students are expected to say that $m$ is the slope (or cost-per-shirt) and $b$ is the $y$-intercept (or set-up fee) but there is no prescribed order to those two events. As shown in Table 3, there were more respondents in Condition 2 who met both expectations (defined both $m$ and $b$), but the difference between the groups was not statistically significant ($\chi^2=3.13$, $p=0.37$).

**Table 1: Manipulation 2 Results**

| | Only $m$ is defined | Only $b$ is defined | $m$ and $b$ are defined | Neither | *Total* | Pretest |
|---|---|---|---|---|---|---|
| Condition 1 (Ms. Turner) | 11 | 5 | 17 | 17 | *50* | 65% |
| Condition 2 (Pat) | 10 | 4 | 29 | 14 | *57* | 70% |
| *Total* | *21* | *9* | *46* | *31* | *107* | 68% |

After the manipulated question in Manipulation 2, all students were asked to write the equation for one of the companies, which was posed in an identical manner for all students. We investigated whether students' initial responses to the manipulated question led to response differences on this new, non-manipulated question. The results are shown in Table 4. No statistically significant difference in performance was found between the groups ($\chi^2=0.81$, $p=0.27$).

**Table 2: Manipulation 2 and Later Prompts**

| | Full Equation Written Correctly | Incorrect | *Total* |
|---|---|---|---|
| Condition 1 (Ms. Turner) | 30 | 20 | *50* |
| Condition 2 (Pat) | 40 | 17 | *57* |
| *Total* | *70* | *37* | *107* |

**Manipulation 3**

The focus of Manipulation 3 was on the final mathematical argument. This manipulated question asks students to develop an argument for which company should be used for the school fundraiser. They do this by responding to an email from the student council stating they will go with either EZ Tees (Condition 1) or Perfect Printing (Condition 2). We chose these two conditions based on evidence from preliminary data of human trialogue interactions (teacher and two students). In the human trialogues, most triads arrived at the conclusion that Perfect Printing was the best choice of the three companies because it is the cheapest for the greatest range of shirts ordered. Thus, we intended to compare an argument for/against an unlikely choice with an argument for/against a likely choice. However, as can be seen in Table 5, most students in both conditions chose Shirts for Less (SfL), the third company, as the best choice. Therefore, the two conditions were each prompting students to respond to a choice that most thought unideal and, we did not have a condition with the most common choice.

**Table 3: Manipulation 3 Final Argument**

| | EZ Tees | SfL | Perfect Printing | Other | *Total* | *Pretest* |
|---|---|---|---|---|---|---|
| Condition 1 (unlikely choice) | 9 | 16 | 10 | 2 | *37* | *67%* |
| Condition 2 (likely choice) | 4 | 23 | 9 | 1 | *37* | *68%* |

The data seem to indicate that students in Condition 2 (Perfect Printing prompt) were more likely to choose Shirts for Less than those in Condition 1 (EZ Tees prompt), but the difference was not statistically significant ($\chi^2$=3.22, $p$=0.19).

We then scored student arguments along a rubric that was designed to align with an Argumentation Learning Progression (Cayton-Hodges et. al., 2014). The argument was scored 1-5, with 5 being the most complete and convincing argument. Results are shown in Table 6.

**Table 4: Manipulation 3 Argumentation Score**

| **Argument Score** | **1** | **2** | **3** | **4** | **5** | **Total** |
|---|---|---|---|---|---|---|
| Condition 1 (unlikely choice) | 10 | 14 | 8 | 3 | 2 | *37* |
| Condition 2 (likely choice) | 13 | 8 | 4 | 7 | 4 | *36* |

The results indicate that students in Condition 1 may have been writing slightly more proficient arguments than those in Condition 2. However, we achieved only 61% reliability (exact score matches) over multiple scorers in the rubric, so we did not perform inferential statistics on this data. It is clear that, as a whole, the sample did not perform well on the final argument. We see this as indicating a weak performing population, which was also shown in Manipulation 1, which could also be one reason for the discrepancy with the human trialogues, as that population of students was overall quite strong. We plan to investigate this question further using a sample of 9$^{th}$ and 10$^{th}$ grade students to see how the findings compare.

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.

## Conclusion

This study aimed at better understanding the design choices made when developing CBA questions, which could also translate to choices made when encouraging argumentation in the classroom. We found that introducing misconceptions, a common approach to encourage argument in CBA, could actually lead students to repeat the misconceptions later in the assessment as opposed to argue against them. Meanwhile, other changes such as a direct question by a virtual teacher versus a virtual student had little, if any, effect responses from students in our sample.

Finally, we were unable to test the premise of responding to a likely versus unlikely answer since a majority of students in both cases chose a different answer than intended by the problem. This was overwhelmingly true in Condition 2, which was supposed to be the “likely” answer.

Further research on this prototype is ongoing, including increasing sample sizes and assessing students in later grades who should have more command of the material, to see if the results change with a stronger population.

## References

Cayton-Hodges, G. A., Nabors Olah, L., Attali, M., Coppola, E., Marquez, E., Leusner, D. (2014). *Learning Progressions in Mathematics and Competency Model Refinement*. Educational Testing Service. Princeton, NJ.

Butler, H. A., Forsyth, C., Halpern, D. F., Graesser, A. C., & Millis, K. (2011). Secret agents, alien spies, and a quest to save the world: operation ARIES! Engages students in scientific reasoning and critical thinking. *Promoting student engagement*, 1, 286-291.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A. M., et al. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, *36*, 180-193.

Graesser, A., Ozuru, Y., & Sullins, J. (2010). What is a good question? *In Bringing reading research to life* (pp. 112–141). New York, NY, US: Guilford Press.

Halpern, D. F., Millis, K., Graesser, A. C., Butler, H., Forsyth, C., & Cai, Z. (2012). Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning. *Thinking Skills and Creativity*, *7*, 93-100.

Hunter, R. (2007). Can you convince me? Learning to use mathematical argumentation. In Woo, J. H., Lew, H. C., Park, K. S. & Seo, D. Y. (Eds.). *Proceedings of the 31st Conference of the International Group for the Psychology of Mathematics Education*, Vol. 3, pp. 81-88. Seoul: PME.

Kelly, A. E., & Lesh, R. A. (Eds.). (2000). *Handbook of research design in mathematics and science education*. Routledge.

Lesh, R., Hoover, M., Hole, B., Kelly, A., & Post, T. (2000). Principles for developing thought-revealing activities for students and teachers. *Handbook of research design in mathematics and science education*, 591-645.

Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A. C., & Halpern, D. (2011). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, & J. Lakhmi (Eds.), *Serious games and edutainment applications* (pp.169-196). London: Springer-Verlag.

Mueller, M. F., & Maher, C. A. (2009). Convincing and Justifying through Reasoning. *Mathematics Teaching In The Middle School*, *15*(2), 108-116.

Yang, H. C., & Zapata-Rivera, D. (2010). Interlanguage pragmatics with a pedagogical agent: the request game. *Computer Assisted Language Learning*, *23*(5), 395-412.