Use of Automated Scoring and Feedback in Online Interactive Earth Science Tasks

Mengxiao Zhu, Ou Lydia Liu, Liyang Mao, and Amy Pallant

mzhu, lliu @ets.org, limao@ixl.com, apallant@concord.org

Abstract - **In formative assessment, constructed response questions are typically used for scientific argumentation, but students seldom receive timely feedback while answering these questions. The development of natural language processing (NLP) techniques makes it possible for the researchers using an automated scoring engine to provide real-time feedback to students. As is true for any new technology, it is still unclear how automated scoring and feedback may impact learning in scientific argumentation. In this study, we analyze log data to examine the granularity of students' interactions with automated scores and feedback and investigate the association between various students' behaviors and their science performance. We first recovered and visualize the pattern of students navigating through the argument items. Preliminary analyses show that most students did make use of the automated feedback. Checking feedback and making revisions also improved students' final scores in most cases. We also cluster the activity sequences extracted from the time-stamped event log to explore patterns in students' behavior.**

Index Terms - log data analysis; automated scoring and feedback; scientific argumentation; earth science

INTRODUCTION

In formative assessments used in secondary school science courses, constructed response questions (also called items in education studies) are typically used for scientific argumentation, because they allow for a greater degree of construct representation and are more authentic than multiple-choice questions [1]–[3]. However, as scoring constructed-response items is both costly and time-consuming [4], students seldom receive timely feedback while answering these questions. The development of natural language processing (NLP) techniques makes it possible for the researchers to use automated scoring engines to provide real-time feedback to students [5]. As is true for any new technology, it is still unclear how automated scoring and feedback may impact learning in scientific argumentation, especially when the scoring rubrics are complex and go beyond simple categories, such as correct, partially correct, and wrong [6].

In this study, we use online interactive questions and take advantage of log analytical tools to examine the granularity of students' interactions with automated scores and feedback and investigate the association between various students' behaviors and their scientific argumentation performance. We also analyze log data to capture rich information and identify patterns from students' behaviors while interacting with the automated score and feedback. Specifically, this study addresses the following research questions:

How do students use automated scoring and feedback?

Does students' performance improve with automated scoring and feedback?

The benefit of using log-data is two-fold. First, it tracks all the activities conducted by students when they answer questions and interact with the automated scoring and feedback. This complete log enables the researchers to recover all events that happened during students' learning process. Second, log data is automatically collected by the project server without interfering with students, which minimizes the impact of data collection on student activities.

This paper contains results from a pilot study of a multi-year project. For this pilot study, we collected data from 42 students and analyzed related log data. This paper is organized as follows. We first review related works in multiple related research fields; then we introduce the task design and data collection process; next we focus on the analysis of the impact of automated scoring and feedback on learning; then we provide results on student activity patterns by visualizing the activity sequence and clustering analysis; finally we discuss the results and future works.

RELATED WORK

Scientific argumentation

As one of the eight scientific practices important for K-12 students in the NGSS [7], scientific argumentation involves making a claim from evidence about a scientific question based on the generally accepted scientific knowledge and research framework, and explicitly addressing boundaries of the claim's application [8]. Engaging students in scientific argumentation can help students deepen the learning of science concept [9], incorporate science epistemology into their learning [10], [11], engage in collaborative learning [12], [13], and support decision-making [14].

As no scientific evidence can support a claim with 100% accuracy, besides making claims and providing explanations, researchers also suggest looking at how students treat uncertainty in formulating their scientific arguments [15]. Uncertainty in students' argumentation could stem from

students' confidence in their understanding of knowledge and their ability to perform investigations [16]. Uncertainty in students' argumentation can also resemble scientists' treatment of uncertainty focusing on limitations of current scientific knowledge base, experiments, equipment, and models [17]. In this study, we adopt a four-component structure in assessing argumentation, including claims, explanations, student rating of uncertainty, and uncertainty rationale.

Automated Scoring

To reduce the cost and time of scoring constructed response items by human raters, automated scoring engines have been developed and successfully implemented for a variety of item types and content domains, including written essays [18], mathematical equations [19], oral communications [20], as well as science questions [21]. Examples of automated scoring engines include c-rater™ and c-rater-ML offered by Educational Testing Service [22], AutoMark [23], and Intelligent Tutoring Systems such as AutoTutor, ITSPOKE, and TLCTS [24]. If developed appropriately, automated scoring would not only greatly reduce scoring time and cost, but would also make scoring constantly available for test takers and increase scoring consistency [25].

Automated feedback

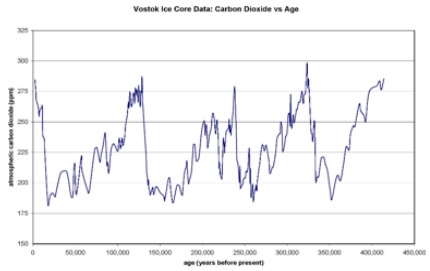
Another advantage of automated scoring is the capability of providing students with instant performance-specific feedback that is not feasible under human scoring. Educational experts consider feedback a critical element to catalyze learning [26]–[28]. Several analysis studies showed that feedback could improve student learning by 0.40 to 0.80 standard deviation units [29], [30].Students are likely to benefit more from immediate rather than delayed feedback [30], [31], because immediate feedback could correct misconceptions before they are encoded into students' learning [32].

Some of the existing intelligent tutoring systems (e.g., AutoTutor) probe students' responses after they type one word to two sentences, while the ideal answer is about three to seven sentences [33]. However, prior research showed that interrupting feedback provided when a student is engaged in problem solving may impede learning [30]. Therefore, the present study takes a different approach by allowing students to complete their answers and providing feedback based on the evaluation of the entire answer. The effectiveness of such feedback will be evaluated based on students' log data.

TASK DESIGN AND DATA COLLECTION

In this study, we developed an online curriculum module on the topic of climate change. The module allows students to use interactive models to explore factors related to the Earth's future climate such as CO_2, albedo, and human produced greenhouse gases.

After exploring an interactive model, students were encouraged to answer a set of argumentation questions (referred to as argument blocks) built within the climate module. The climate module included a total of eight argument blocks and each argument block consisted of a multiple-choice claim, a constructed-response explanation, an uncertainty rating on the five-point Likert scale, and a constructed-response uncertainty rationale. An example of an argument block is shown in Figure I.



Temperature is related to the amount of carbon dioxide in the atmosphere. How does the temperature of 100,000 years ago compare to the temperature of 400,000 years ago?

1. The temperature 100,000 years ago was likely

A. lower than 400,000 years ago
B. the same as 400,000 years ago
C. higher than 400,000 years ago.

2. Explain your answer.

3. How certain are you about your claim based on your explanation?

Not at all certain *Very certain*
1 2 3 4 5

4. Explain what influenced your certainty rating.

FIGURE I
AN EXAMPLE OF AN ARGUMENT BLOCK

Students' answers to the scientific argumentation items in the eight blocks were graded by c-rater-ML, an automated scoring engine develop at Educational Testing Service for scoring short-answer items [34]. C-rater-ML utilizes supervised machine learning and an automated model-building process to produce scoring models for each argumentation item. In particular, explanation items were graded on a 7-point scale (e.g. score of 0 to 6), while uncertainty rationale items were on a 5-point scale (e.g., score of 0 to 4).

In addition to automated scores, automated feedback was also provided to students after they submitted their responses to each argumentation item. The feedback included instructions for revisiting models, and/or hints as well as an acknowledgement of correct responses. The feedback is intended to provide students with the opportunity to review their answers and point them to the relevant instructional steps. Students were allowed and encouraged to revise their responses as many times as they wanted to until they were satisfied with their scores.

In this pilot study, data were collected from 42 students from two Advanced Placement (AP) classes, two honors classes and one college prep class. Most students were in 9th or 11th Grade. Two teachers were assisting with the data collection. They were given two hours of instruction before the classroom runs. All questions in the module were presented to the students in a web browser. The server records three log files while students answer the questions. The first is the final answers for each student, the second is the argument block report on each submission for each student, and the third is the detailed event log with time stamps for each student. Given that we are interested in how students interact with

argument blocks and their activities, we focused on the last two log files. The argument block report includes the scores and feedback for each submission together with the usefulness ratings. The event log data includes time-stamped information on student activities, such as, visiting to a certain page, visiting to a certain question, or submission of answers for automated feedback. In the next two sections, we present details on the analysis on these two files.

IMPACT OF AUTOMATED SCORING AND FEEDBACK ON LEARNING

Students' reaction to automated score and feedback

First, we study how students react to the automated scoring and feedback. After answering all four questions in the argument block, including two multiple choice and two constructed response questions, students can submit the answers to get automated score and feedback. Since scores and feedback were provided only after each submission, the log data do not have enough information to allow us to separate students' activities on different questions. In this analysis we take the argument block as the unit of analysis.

To collect information on students' perceived usefulness of the automated scores and feedback, students were asked to rate the usefulness of the scores and feedback they received after each submission. The rating ranges from "not at all", "somewhat", to "very". Among all ratings, the majority (*78.46%*) are either "somewhat" (*42.13%*) or "very" (*36.33%*). This results show that students consider the automated scores and feedback to be at least somewhat useful most of the time.

Over the eight argument blocks, on average, students made one revision (μ=0.96) for each argumentation block. The maximum number of feedback checking and revising is 24, in which case the student improved his/her score on an argument item from 3 to the highest possible score of 6. A boxplot of the number of times students checked the feedback and attempted to answer the items in the blocks is shown in Figure IV. Despite the differences in the content for the eight argument blocks, the means of the number of attempts for different blocks do not significantly different from each other.

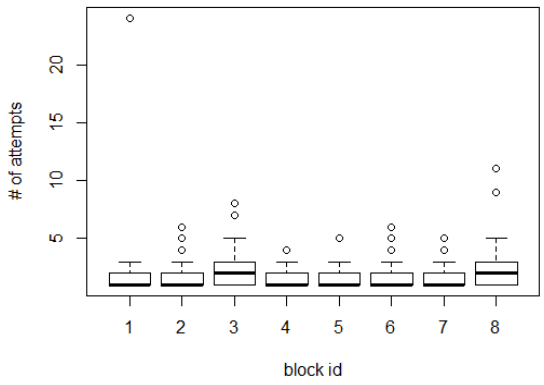


FIGURE II
BOXPLOT OF NUMBER OF ATTEMPTS BY ARGUMENT BLOCK

Given that the eight argument blocks appear in the order of 1 through 8, we also plot the number of attempts for each argument block for all students in Figure IV. Each color represents activities for one student. Overall, the data do not show any general trend. One obvious outlier is the student who tried 24 times for the first argument block. This enthusiasm does not seem to extend over the rest of argument blocks.

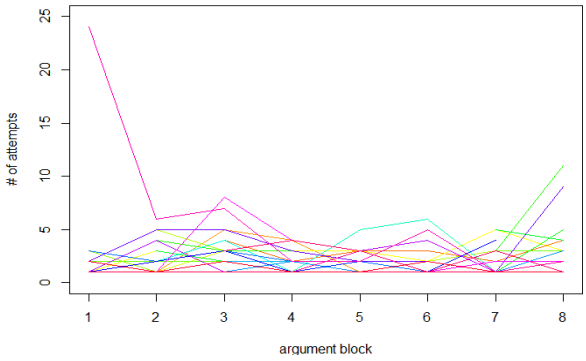


FIGURE IV
NUMBER OF ATTEMPTS BY ARGUMENT BLOCK FOR INDIVIDUAL.

Impact of automated scoring and feedback in assisting learning

Besides how students react to and make use of the automated scoring and feedback, we also investigate their impact in assisting learning. For each argument block, students are allowed and encouraged to revise their answers based on the feedback. We analyze the data to find answers to the research question of whether or not students' performance improved with the provided automated scores and feedback. Since only item 2 and item 4 were scored with feedback, we focus on students' scores on these two items. Several variables were created from the raw data, including initial scores for item 2 and item 4, final scores for item 2 and item 4, score increase for item 2 and item 4, total increase over both items, number of attempts by students, and the usefulness ratings from the students. Figure IV provides the Pearson correlation among these ten variables. Here the unit of analysis is also an argument block.

Here, we discuss some of the interesting findings. Some of the high and positive correlations are due to the way we construct these variables. For instance, the overall total increase (*total_increase*) was calculated as the sum of score increases for item 2 and item 4, and thus is highly correlated with the increase for item 2 (*increase_item2*) *r=0.82* and with the increase for item 4 (*increase_item4*) *r=0.83*. However, the correlation between the increase for item 2 and item 4 is not very high *r=0.36*, indicating that students did not tend to improve on both items.

Since students got multiple chances to submit their answers and to rate the usefulness of the automated scores and feedback, we created two variables to capture the maximum value on the usefulness rating (*usefulness_max*) and the average value over multiple ratings (*usefulness_avg*) for

individual students by argument blocks. Non-surprisingly, these two variables are also highly correlated, *r=0.9*. We then check how the usefulness rating is related to various measures on the items. It is found that the highest correlations for both maximum and average usefulness are with the final scores for item 4 (*final_score_item4*), with *r=0.45* and *0.42* respectively.

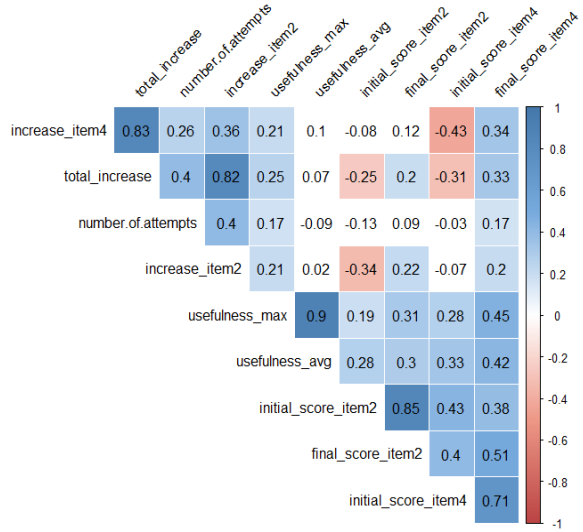


FIGURE IV
CORRELATION AMONG ARGUMENT BLOCK VARIABLES

There are several negative relations in the matrix, for instance, the initial scores are negatively correlated with score increase, *r=−0.34* for item 2 and *r=−0.43* for item 4. The total score increase is also negatively correlated with initial scores for item 2 with *r=−0.25* and for item 4 with *r=−0.31*. This may be because of the ceiling effect, i.e., the scores for item 2 range from 0 to 6 and the scores for item 4 range from 0 to 4. Students who got high scores in the first place have less room to increase. We also find that the initial scores are positively correlated with final scores. For instance, the initial score for item 2 and the final score of item 2 are highly correlated with *r=0.85*, and *r=0.71* for item 4. The reason might also be the ceiling effect.

MINING STUDENT ACTIVITY PATTERNS

Besides how students use automated scores and feedback and their impact on learning, we also explore students' activity patterns while working on the argument blocks. This section provides results on the event log analysis. As described in the section on Task Design and Data Collection, the event log data are time-stamped data on student actions. Due to unknown reasons, some data were missing during data collection. The final dataset include records of 16 students working on 73 argument blocks. The system defined ten activities, including focus in, focus out, answer saved, arg-block submit, exit page, open activity index, submit answer, open activity, clicked in the simulation window, open activity page. In order to recover the activities that we are interested in, we focused on the first four argument block related events. We define the time working on an argument block item as the time between the event focus in and focus out. The system event *arg-block submit* is the indicator of the start of checking the scores and feedbacks, which happens when the students click the submission button for an argument block. We use the id's for argument blocks and student id's to identify record for all 73 cases for student working on argument blocks.

Visualization and state distribution

We use the R package TraMineR [35] to visualize and generate state distribution graphs for our data. Figure VI is the visualization of all 73 sequences sorted by the actions from the beginning of the sequences. The activities are color-coded as shown in the legend. Students spent different amount of time working on the argument blocks, indicated by the length of the sequence. The unit for the y-axis is seconds. The longest sequence stretched over more than 13 minutes, and the short ones are less than 1 minute long.

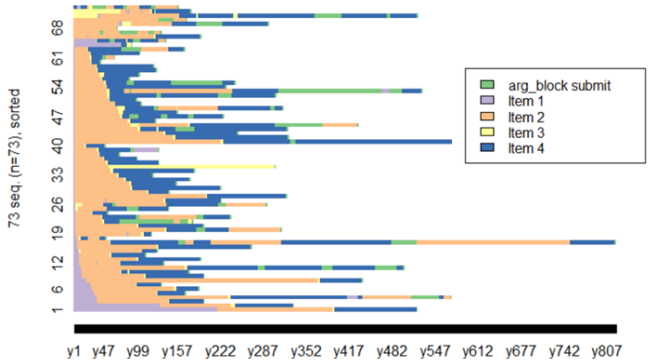


FIGURE VI
VISUALIZATION OF ALL SEQUENCES

Overall, students spent most of time working on item 2 and item 4, shown in orange and blue in Figure VI. This observation is confirmed by the plot of state distribution as in Figure VIII. On average, students did spend most of the time working or revising item 2 and item 4. For the two multiple choice items, item 1 and item 3, they spent less time. The mean time for checking feedback (*arg_block submit* event) is not very high but decent compared to the time spent answering the other two multiple choice questions.

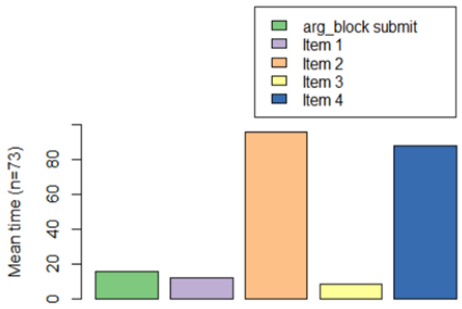


FIGURE VIII

PLOT OF STATE DISTRIBUTION

Clustering

To further explore patterns in students' activities, we conduct cluster analysis on the action sequences. The first step for clustering analysis is to define the similarity/dissimilarity among sequences. As suggested by researchers [36], editing distance can be an effective and flexible way to compare action sequences. Editing distance is defined as the minimal cost of transforming one sequence to another. We adopt one of the editing distances, optimal matching (OM) [35], which allows for both substitution of an element, and the insertion or deletion of an element. Using this method, the program calculates the minimal cost to transform one sequence to another using the above three actions. In TranMineR, this process is completed through dynamic programming.

Two parameters, i.e., a substitution cost matrix and the insertion/deletion costs need to be determined and feed into the program. For clustering, we are less interested in the relative length of the sequence, but more interested in the transitions among different activities, such as working on different items, checking feedback, and making revisions. We define the substitution cost as 2 for all pairs and insertion/deletion cost to be half of the substitution cost. We further normalize the measure by dividing the longer of the two sequences to diminish the impact of sequence length. The reason for not choosing other distance measures, such as Hamming distance or dynamic Hamming, is that OM is the only method that allows us to define the cost for substitution and insertion/deletion and at the same time it works for sequences of different lengths.

We choose hierarchical agglomerative clustering (using *agnes* in the R package *cluster*) [37], because it does not require pre-defined number of clusters and can be cut at different levels to form different number of clusters. For our exploratory study, these features make it the preferred approach. The dendrogram of the clustering using ward method is shown in Figure X. We tentatively cut it into different clusters. While balancing the size of each cluster and the potential to find meaningful patterns, we cut the dataset into three clusters. Visualizations (as shown in Figure XII) of these three chapters show some patterns in different clusters. Cluster 1 contains students who only revised item 4 if they made any revisions. Cluster 3 contains mostly students who revised item 2 and in a lot of cases both item 2 and 4. Cluster 2 includes a mix of students who switch back and forth between items, even before submitting their answers.

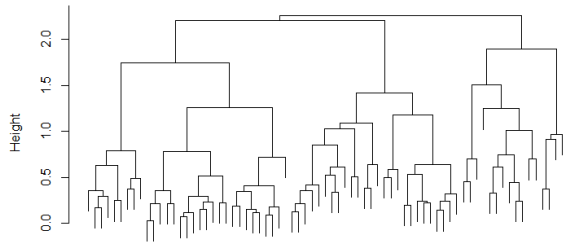


FIGURE X
DENDROGRAM FOR CLUSTERING

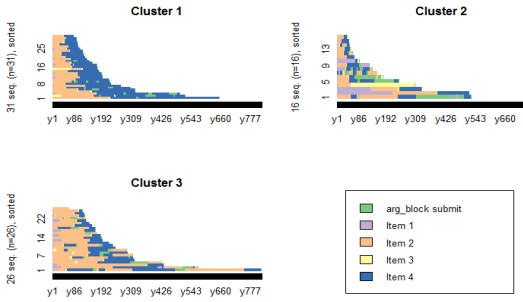


FIGURE XII
VISUALIZATION OF SEQUENCES FOR 3 CLUSTERS

DISCUSSIONS AND FUTURE WORK

In this paper, we summarize our findings in a pilot study on the application of automated scoring and feedback in earth science learning with a focus on the building of scientific argumentation based on evidence. We found that in most cases, students reported positively about the available instant score and feedback. Generally speaking, students were willing to try this new feature and make effort to improve their answers. Consequently, there are observed final score increases on the two constructed response questions under study. The increase on the scores did vary across different students. And the final score on the uncertainty rationale question is most related to students' rating of the usefulness of the automated scores and feedback. By analyzing the time stamped event log, we recovered and visualized students' activities. As expected, students spent most of their time on the two constructed response question. They also spent a decent amount of time checking the score and feedback, and making revisions. Clustering analysis generated three clusters that represent different student activity patterns.

This study also has some limitations. First of all, as part of the pilot study, the sample size is small. With available bigger samples in the next several years, the analysis and findings can be more interesting and provide more information on the impact of automated scoring and feedback on student learning process. Second, log data includes only information recorded by the server. Observations on log data are rather speculative. Future studies can include analysis on designed surveys for students and teachers to collect their feedback and experience with the automated scoring and feedback system. Last, the current design of the online module and the argument block system does not allow us to collect time-stamped data on the details of reading feedback. We had to recover that information from the time-stamps of other events, which may lead to inaccurate measures, mostly underestimation. In the future design, the system can be improved to allow recording at the item level to provide more detailed information.

Despite the above mentioned limitations, this paper showcases the application of log data analysis and machine learning techniques in understanding the impact of automated scoring and feedback on learning. The availability of complex data format and sources calls for analysis methods and

techniques beyond traditional psychometrics [38]. In recent years, researchers made an effort to explore new ways to analyze such data, e.g., [39]–[41]. We hope this study shed some light on this newly emerged and fast growing direction.

ACKNOWLEDGMENT

The work in this paper is funded through the NSF project (#1418019).

REFERENCES

[1] S. Lane, "Validity of High-Stakes Assessment: Are Students Engaged in Complex Thinking?," *Educ. Meas. Issues Pract.*, vol. 23, no. 3, pp. 6–14, Oct. 2005.

[2] H.-S. Lee, O. L. Liu, and M. C. Linn, "Validating Measurement of Knowledge Integration in Science Using Multiple-Choice and Explanation Items," *Appl. Meas. Educ.*, vol. 24, no. 2, pp. 115–136, Mar. 2011.

[3] L. A. Shepard, "Commentary: Evaluating the Validity of Formative and Interim Assessment," *Educ. Meas. Issues Pract.*, vol. 28, no. 3, pp. 32–37, Sep. 2009.

[4] H. Wainer and D. Thissen, "Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction," *Appl. Meas. Educ.*, vol. 6, no. 2, pp. 103–118, Apr. 1993.

[5] C. Leacock and M. Chodorow, "Automated grammatical error detection," in *Automated essay scoring: A cross-disciplinary perspective*, M. D. Shermis and J. Burstein, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 2003, pp. 186–199.

[6] O. L. Liu, C. Brew, J. Blackmore, L. Gerard, and J. Madhok, "Automated scoring for inquiry science assessment: An application of c-rater," in *2013 Annual Conference of the National Council on Measurement in Education*, 2013.

[7] NGSS Lead States, *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press, 2013.

[8] S. Toulmin, *The uses of argument*. New York: Cambridge University Press, 1958.

[9] C. V. Aufschnaiter, S. Erduran, J. Osborne, and S. Simon, "Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge," *J. Res. Sci. Teach.*, vol. 45, pp. 101–131, 2008.

[10] R. A. Jimenez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, "Doing the lesson' or 'doing science": Argument in high school genetics," *Sci. Educ.*, vol. 84, pp. 757–792, 1999.

[11] W. A. Sandoval, "Conceptual and epistemic aspects of students' scientific explanations," *J. Learn. Sci.*, vol. 12, no. 1, pp. 5–51, 2003.

[12] R. A. Duschl and J. Osborne, "Supporting and promoting argumentation discourse in science education," *Stud. Sci. Educ.*, vol. 38, pp. 39–72, 2002.

[13] C. Chin and J. Osborne, "Students' questions and discursive interaction: Their impact on argumentation during collaborative group discussions in science," *J. Res. Sci. Teach.*, vol. 47, no. 7, pp. 883–908, 2010.

[14] D. Kuhn and W. Udell, "The development of argument skills," *Child Dev.*, vol. 74, no. 5, pp. 1245–1260, 2003.

[15] A. Bricker and P. Bell, "Conceptualizations of argumentation from science studies and the learning sciences and their Implications for the practices fo science education," *Sci. Educ.*, vol. 92, pp. 473–493, 2008.

[16] K. E. Metz, "Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design," *Cogn. Instr.*, vol. 22, pp. 219–290, 2004.

[17] H.-S. Lee, O. L. Liu, A. Pallant, K. C. Roohr, S. Pryputniewicz, and Z. E. Buck, "Assessment of uncertainty-infused scientific argumentation," *J. Res. Sci. Teach.*, vol. 51, no. 5, pp. 581–605, May 2014.

[18] M. D. Shermis and J. C. Burstein, *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates, 2003.

[19] T. Risse, "Testing and assessing mathematical skills by a script based system," in *the 10th International Conference on Interactive Computer Aided Learning*, 2007.

[20] D. Higgins, K. Zechner, X. Xi, and D. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Comput. Speech Lang.*, vol. 25, pp. 282–306, 2011.

[21] O. L. Liu, C. Brew, J. Blackmore, L. Gerard, J. Madhok, and M. C. Linn, "Automated Scoring of Constructed-Response Science Items: Prospects and Obstacles," *Educ. Meas. Issues Pract.*, vol. 33, no. 2, pp. 19–28, Jun. 2014.

[22] C. Leacock and M. Chodorow, "C-rater: Automated Scoring of Short-Answer Questions," *Comput. Hum.*, vol. 37, no. 4, pp. 389–405.

[23] T. Mitchell, T. Russell, P. Broomhead, and N. Aldridge, "Towards robust computerised marking of free-text responses," in *the Proceedings of the 6th CAA International Computer Assisted Assessment Conference*, 2002.

[24] A. C. Graesser, "Learning, thinking, and emoting with discourse technologies," *Am. Psychol.*, vol. 66, no. 8, pp. 746–757, 2011.

[25] D. M. Williamson, X. Xi, and F. J. Breyer, "A Framework for Evaluation and Use of Automated Scoring," *Educ. Meas. Issues Pract.*, vol. 31, no. 1, pp. 2–13, Mar. 2012.

[26] P. Black, C. Harrison, C. Lee, B. Marshall, and D. Wiliam, *Assessment for learning: Putting it into practice*. Berkshire, England: McGraw-Hill Education, 2003.

[27] S. Clarke, *Enriching feedback in the primary classroom*. London, England: Hodder and Stoughton, 2003.

[28] J. Hattie, *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge, 2009.

[29] R. Azevedo and R. M. Brenard, "A meta-analysis of the effects of feedback in computer-based instruction," *J. Educ. Comput. Res.*, vol. 13, pp. 11–127, 1995.

[30] V. J. Shute, "Focus on formative feedback," *Rev. Educ. Res.*, vol. 78, pp. 153–189, 2008.

[31] D. Anderson, R. A. Magill, and H. Sekiya, "Motor learning as a function of KR schedule and characteristics of task intrinsic feedback," *J. Mot. Behav.*, vol. 33, pp. 59–66, 2001.

[32] R. E. Dihoff, G. M. Brosvic, and M. L. Epstein, "The role of feedback during academic testing: The delay retention effect revisited," *Psychol. Rec.*, vol. 53, pp. 533–548, 2003.

[33] A. C. Graesser, S. Lu, G. T. Jackson, H. H. Mitchell, M. Ventura, A. Onley, and M. M. Louwerse, "AutoTutor: A tutor with dialogue in natural language," *Behav. Res. Methods, Instruments, Comput.*, vol. 36, no. 2, pp. 180–92, 2004.

[34] J. Z. Sukkarieh and J. Blackmore, "c-Rater: Automatic content scoring for short constructed responses," in *Proceedings of the Twenty-Second International Florida Artificial Intelligence Research Society Conference*, H. C. Lane and H. W. Guesgen, Eds. Menlo Park, CA: AAAI Press, 2009, pp. 290– 295.

[35] M. Gabadinho, A., Ritschard, G., Müller, N. S., Studer, "Analyzing and Visualizing State Sequences in R with TraMineR," *J. Stat. Softw.*, vol. 40, no. 4, pp. 1–37, 2011.

[36] J. Hao, Z. Shu, and A. von Davier, "Analyzing Process Data from Game/Scenario-Based Tasks: An Edit Distance Approach," *JEDM - Journal of Educational Data Mining*, vol. 7, no. 1. pp. 33–50, 29-Jan-2015.

[37] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, “cluster: Cluster Analysis Basics and Extensions. R package version 2.0.3.” 2015.

[38] A. A. von Davier, “Virtual & collaborative assessments: Examples, implications, and challenges for educational measurement,” in *Invited Talk at the Workshop on Machine Learning for Education, International Conference of Machine Learning 2015*, 2015.

[39] A. A. von Davier and P. F. Halpin, “Collaborative Problem Solving and the Assessment of Cognitive Skills : Psychometric Considerations,” no. December. Educational Testing Service (EST) Research Report Series, Princeton, NJ, 2013.

[40] D. Kerr and G. K. Chung, “Identifying Key Features of Student Performance in Educational Video Games and Simulations through Cluster Analysis,” *J. Educ. Data Min.*, vol. 4, no. 1, pp. 144–182, 2012.

[41] Y. Bergner, Z. Shu, and A. Davier, “Visualization and Confirmatory Clustering of Sequence Data from a Simulation-Based Assessment Task,” in *Educational Data Mining 2014*, 2014, pp. 177–184.

AUTHOR INFORMATION

Mengxiao Zhu, Associate Research Scientist, Computational Psychometrics Research Center, Research and Development Division, Educational Testing Service.

Ou Lydia Liu, Director Research, Research and Development Division, Educational Testing Service.

Liyang Mao, Marketing Research Associate, IXL Learning.

Amy Pallant, Senior Research Scientist, The Concord Consortium.