International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: http://www.tandfonline.com/loi/tsed20

Investigating the impact of automated feedback on students' scientific argumentation

Mengxiao Zhu , Hee-Sun Lee, Ting Wang, Ou Lydia Liu, Vinetha Belur & Amy Pallant

To cite this article: Mengxiao Zhu , Hee-Sun Lee, Ting Wang, Ou Lydia Liu, Vinetha Belur & Amy Pallant (2017): Investigating the impact of automated feedback on students' scientific argumentation, International Journal of Science Education, DOI: 10.1080/09500693.2017.1347303

To link to this article: http://dx.doi.org/10.1080/09500693.2017.1347303

Published online: 15 Jul 2017.

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION, 2017
https://doi.org/10.1080/09500693.2017.1347303

Investigating the impact of automated feedback on students' scientific argumentation

Mengxiao Zhu [a], Hee-Sun Lee[b], Ting Wang[a], Ou Lydia Liu[a], Vinetha Belur[a] and Amy Pallant[b]

[a]Educational Testing Service, Princeton, NJ, USA; [b]Concord Consortium, Concord, MA, USA

ABSTRACT

This study investigates the role of automated scoring and feedback in supporting students' construction of written scientific arguments while learning about factors that affect climate change in the classroom. The automated scoring and feedback technology was integrated into an online module. Students' written scientific argumentation occurred when they responded to structured argumentation prompts. After submitting the open-ended responses, students received scores generated by a scoring engine and written feedback associated with the scores in real-time. Using the log data that recorded argumentation scores as well as argument submission and revisions activities, we answer three research questions. First, how students behaved after receiving the feedback; second, whether and how students' revisions improved their argumentation scores; and third, did item difficulties shift with the availability of the automated feedback. Results showed that the majority of students (77%) made revisions after receiving the feedback, and students with higher initial scores were more likely to revise their responses. Students who revised had significantly higher final scores than those who did not, and each revision was associated with an average increase of 0.55 on the final scores. Analysis on item difficulty shifts showed that written scientific argumentation became easier after students used the automated feedback.

ARTICLE HISTORY
Received 29 November 2016
Accepted 22 June 2017

KEYWORDS
Log data analysis; automated scoring and feedback; scientific argumentation; climate change

Introduction

According to the Next Generation Science Standards (NGSS Lead States, 2013), scientific argumentation is one of the eight scientific practices in K–12 education in which students are expected to engage while learning science. Since scientific argumentation is carried out through their written or spoken language, students' ability to construct scientific arguments is usually assessed through constructed-response items. In addition, constructed-response items are associated with a greater degree of construct representation and can more effectively capture students' thoughts than multiple-choice questions with structured responses (Lane, 2005; Lee, Liu, & Linn, 2011; Shepard, 2009). However, human scoring of constructed-response items takes a lot

CONTACT Mengxiao Zhu mzhu@ets.org Educational Testing Service, 660 Rosedale Rd, MS 02-T, Princeton, NJ 08541, USA

of time (Wainer & Thissen, 1993), which makes it practically impossible for individual students in the classroom to get instant feedback. Fortunately, recent decades have witnessed the application of natural language processing (NLP) techniques to autoscoring of written texts, which can allow students to receive instant scores and real-time feedback (e.g. Dzikovska et al., 2013; Ha, Nehm, Urban-Lurain, & Merrill, 2011; Liu, Rios, Heilman, Gerard, & Linn, 2016). While autoscoring applications and related studies are beginning to crop up, the impact of automated feedback on written scientific argumentation is yet to emerge.

To investigate the impact of automated feedback on student learning of scientific argumentation, this study used an online climate change curriculum module that had eight scientific argumentation tasks with which automated scoring and feedback mechanisms were integrated. Since all student interactions with the module as well as their answers and scores were recorded by the server in the background, we analysed log data to answer the first two questions shown below and conducted psychometric analyses to answer the third:

(1) What patterns emerged in students' responses to the automated feedback?
(2) How did students' response patterns relate to their final argumentation scores?
(3) How did item difficulties shift from the initial to last argumentation responses?

By answering the first two questions, we sought to characterise the type of impact the automated feedback can create on student learning of scientific argumentation. We also examined the item difficulty shifts before and after the automated feedback was provided to students because if students were improving scientific argumentation through automated feedback, scientific argumentation tasks should become easier after revisions based on the feedback.

Research background

Scientific argumentation

It is well publicised that secondary school students in the U.S.A. perform around the mid-range on internationally administered science assessments such as Trends in International Math and Science Study (TIMSS) (Gonzales et al., 2008) and Program for International Student Assessment (PISA) (OECD, 2010) (as cited in Cope, Kalantzis, Abd-El-Khalick, & Bagley, 2013). Nationwide concerns over the quality of science education have persisted in the past decades and have resulted in several rounds of standards initiatives. Most recent is the Next Generation Science Education Standards (NGSS) which are conceptually based on a book entitled 'A Framework for K-12 Science Education: Practices, Cross Cutting Concepts, & Core Ideas' (NRC, 2012). NGSS incorporates scientific practices as a means to learn disciplinary core ideas. Among the eight science practices identified in the NGSS scientific argumentation, the process of making a claim using evidence interpreted in light of the established knowledge (Sandoval & Reiser, 2004) within the confines of investigations where theoretical and methodological limitations abound (Staley, 2014) reflects how scientific knowledge is generated, debated, and refined (Bricker & Bell, 2008). Scientific

argumentation is, therefore, viewed as helping students make sense of scientific ideas, instead of just rote memorisation (Aufschnaiter, Erduran, Osborne, & Simon, 2008). It can also aid students to engage in collaborative learning and decision-making (Chin & Osborne, 2010; Duschl & Osborne, 2002; Kuhn & Udell, 2003).

Scientific argumentation in the science classroom has been implemented in different ways, including written argumentation when students conclude their data-based investigations (Duschl & Osborne, 2002) and argumentation discourse when students communicate, debate, and critique one another's arguments in social settings (Kuhn, 2010). The former is known as rhetorical argumentation, while the latter is known as dialogic argumentation. In this study, we focus on the former since automated scoring and feedback are designed to help individual students or groups of students working together write arguments that are more complete, more conceptually elaborated, and more integrated between data and reasoning.

Most scientific arguments include commonly recognisable elements (Sampson & Clark, 2008). The most referenced Toulmin (1958)'s *structural* analysis of argumentation includes a *claim* to the question students are investigating, *data* that are collected and analysed to support their claim, *warrants* that establish explicit links between the data and scientific knowledge, *backing* that theoretically supports the scientific validity of the warrants, *qualifiers* that indicate the strength of the claim, and *conditions of rebuttal* that specify the conceptual, methodological, and contextual boundaries where the claim can be true. To improve students' written argumentation, many instructional approaches focus on claim–evidence–reasoning expressed in claim, data, warrants, and backing (Sampson & Clark, 2008). On the other hand, qualifiers and conditions of rebuttal have been promoted mainly through rebuttals or counterarguments in spoken discourse settings where two or more students or groups of students are involved (Erduran, Simon, & Osborne, 2004; Kuhn & Udell, 2003). Recently, Lee et al. (2014) proposed a scientific argumentation framework that assesses not only claim–evidence–reasoning, but also qualifiers and conditions of rebuttal through scientific uncertainty. Their Rasch analysis results indicate that elaborating uncertainty attribution is, indeed, part of the their proposed scientific argumentation construct, and the inclusion of uncertainty fundamentally distinguishes argumentation where the strength and limitation of the said claim are discussed from the explanation where data–theory coordination is established (Lee et al., 2014). Uncertainty attribution plays a role in formulating scientific arguments because much of the evidence on which arguments are based – such as models and scientific data – are investigator-selected representations of the real-world phenomenon, and scientific knowledge and theory can only explain some parts of the phenomenon (Lee et al., 2014). This means a certain amount of uncertainty will always inherently exist in scientific evidence and, consequently, in formulating scientific arguments (Lee et al., 2014). In this study, students' ratings of uncertainty regarding their arguments, and their attribution of uncertainty sources, are used alongside claims and explanations to form a four-component assessment structure.

Automated scoring and feedback

It is very challenging for teachers to evaluate every student's written text and provide timely feedback to the student for further improvement. However, automated scoring

technologies derived from NLP have made it possible to score students' content-based short-answer written text immediately. These technologies have opened the door for using students' written texts in formative assessment opportunities in the classroom. Some examples of automated scoring engines are c-rater™ and c-rater-ML developed by Educational Testing Service (Leacock & Chodorow, 2003; Liu et al., 2016), AutoMark (Mitchell, Russell, Broomhead, & Aldridge, 2002), SIDE (Mayfield & Penstein Rosé, 2010), EvoGrader (Moharreri, Ha, & Nehm, 2014), and Intelligent Tutoring Systems such as AutoTutor (Graesser, 2011), ITSPOKE (Litman & Silliman, 2004), and TLCTS (Johnson & Valente, 2009). Application of these new technologies stimulates interests in the assessment of student learning that occurs in highly complex cognitive skills which multiple-choice item types cannot effectively capture. Scientific argumentation is one such skill associated with students' generation of their own texts, rather than students' selection of a best answer among the provided.

Providing feedback has been shown to enhance student learning significantly (Azevedo & Brenard, 1995; Shute, 2008); thus, many experts consider feedback as a critical component of learning (Black, Harrison, Lee, Marshall, & Wiliam, 2003; Clarke, 2003; Hattie, 2009). One advantage of automated scoring engines, besides the efficiency they provide for scoring, is that feedback can be coupled with automated scores in order to provide instant real-time feedback, which is not feasible under human scoring. The immediacy of such feedback is noteworthy, as immediate, rather than delayed, feedback is likely to correct misconceptions before they are encoded into students' learning (Dihoff, Brosvic, & Epstein, 2003), and therefore offers greater benefit to students (Anderson, Magill, & Sekiya, 2001; Shute, 2008). In the present study, students are provided feedback on the content of their answers, and the effectiveness of such feedback will be evaluated based on students' score changes and reactions to the feedback.

The research context

The climate change module and the argumentation task blocks

This study used the modified version of the High-Adventure Science 'What is the future of Earth's Climate?' module (http://authoring.concord.org/sequences/47) (hereafter referred to as the Climate change module), an interactive online module developed by the Concord Consortium. The module consists of five activities, each focusing on a specific topic on climate change. The topics addressed in the module are (1) Earth's changing climates as illustrated in data trends over different time scales; (2) how solar radiation interacts with components of Earth's surface and atmosphere; (3) the relationships between ocean surface temperature and levels of atmospheric carbon dioxide and water vapour; (4) feedback of ice and clouds on Earth's temperature; and (5) how solar radiation, Earth's surface, and greenhouse gases interact to cause global warming. Each activity includes six to eight steps and is estimated to last about 45 minutes. On each activity step, students gain knowledge about climate change through reading descriptions of scientific phenomena and investigations, viewing images and videos, analysing data tables and graphs, and running interactive computer models. Students are prompted to respond to various types of questions stated in multiple-choice and open-ended formats throughout the entire module.

The climate change module includes eight scientific argumentation tasks where automated scoring and feedback are embedded. In order to elicit students' argumentation responses, a four-part argumentation format is used. This argumentation format consists of a claim, explanation of the claim, student rating of uncertainty, and uncertainty attribution. Among them, the claim is a multiple-choice item, while the uncertainty rating is on a 5-point Likert scale from very uncertain to very certain. The explanation and uncertainty attribution are constructed-response items. The automated scoring and feedback are designed to work for the explanations and uncertainty attributions. We call this four-component argumentation format an *argumentation block*. The main reason we explicitly separate these four parts was due to the fact that students have difficulties with distinguishing claim, data/evidence, and warrant/reasoning when unguided (Berland & Reiser, 2009). In addition, students can request to view related guidelines (as shown in Table 1) per argumentation block.

These guidelines are first used when students are introduced to argumentation as part of an introductory class activity to the climate change module and appear when students click on the '?' icon near each item.

Figure 1 shows an example of an argumentation block. This argumentation task asks students to predict the future of the average global temperature. The argumentation task posits a scientific question and shows a plot of historical temperature data with three options of future temperature trends. Students are asked to construct arguments in the argumentation block. In this case, Question # 7 (Claim) asks the students to choose the future trend based on the historical observation. Question # 8 (Explanation) asks the students to justify their choices. Question # 9 (Uncertainty rating) and Question # 10 (Uncertainty attribution) ask the students to rate their certainty about their choices and justify that certainty, respectively. Note that the numbering of questions follows the order in which the questions are presented in the climate change module activities.

There are eight argumentation blocks across the five climate change module activities. The activity titles and the page titles where the argumentation blocks are embedded are

Table 1. Guidelines for the four parts in an argumentation block.

Claim guidelines	• A good claim is based on the evidence • Evidence may come from graphs and charts • Evidence may also come from models that you run
Explanation guidelines	• A good explanation will cite specific evidence that backs up the claim • When there is a graph or table, you can cite evidence directly from the source • When there is a model, you can describe what happened in the model • A good explanation combines evidence with scientific knowledge
Uncertainty rating guidelines[a]	• Picking a certainty rating is a way to signal how certain you are with your claim • Your certainty rating can be based on how well the scientific knowledge fits the evidence from models, charts, or graphs • Your certainty rating can also reflect on the quality of the evidence or investigation that produced the evidence
Uncertainty attribution guidelines	• A good certainty explanation will explain why you are certain or uncertain about your response • Some topics are more certain than others. Consider the completeness of the evidence, biases in the evidence, and changes that could affect the trends over time

[a]Note that we used certainty rating, instead of uncertainty rating, to minimise students' confusion.

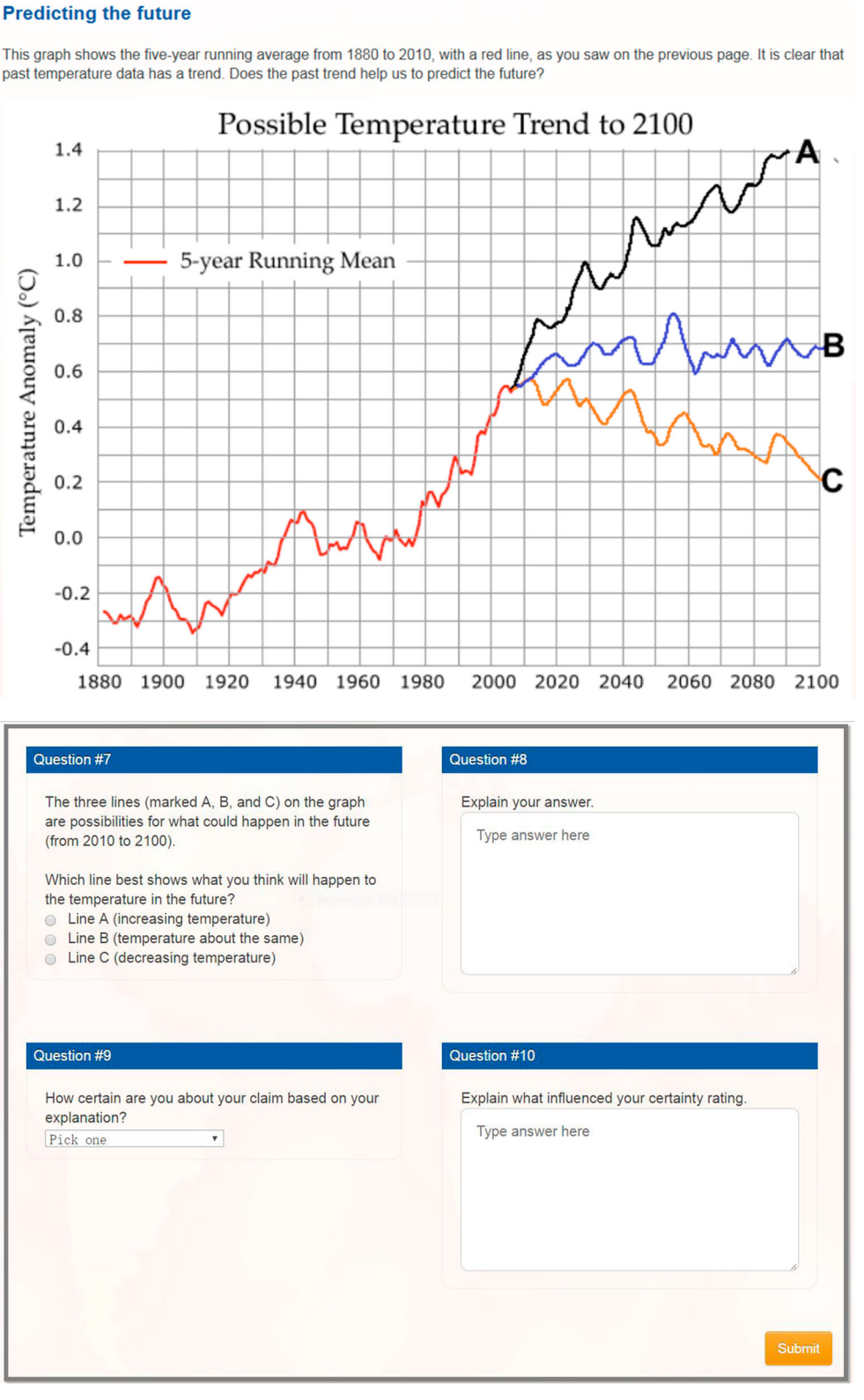

Figure 1. An example of an argumentation block.

Table 2. Eight argumentation blocks in the climate change module.

Argumentation block index	Activity title	Page title
1	Earth's changing climates	Predicting the future
2	Interactions within the atmosphere	Carbon dioxide in the atmosphere
3	Interactions within the atmosphere	Historical carbon dioxide levels
4	Sources, sinks, and feedbacks	Changing ocean temperature
5	Sources, sinks, and feedbacks	Water vapour: a powerful greenhouse gas
6	Sources, sinks, and feedbacks	Combining the effects of carbon dioxide and water vapour
7	Feedbacks of ice and clouds	Arctic sea ice
8	Using models to make predictions	How much reduction?

listed in Table 2. For each argumentation block, students work on the questions, submit their answers to receive the automated scoring and feedback, and then revise and resubmit. Revisions are voluntary, not mandatory. There is no limit on how many times students can resubmit. The climate change module is implemented as part of classroom instruction.

Automated scoring and feedback

Automated scoring and feedback are provided when students construct responses to explanation and uncertainty attribution prompts for each argumentation block. Students' responses are scored immediately following submission to c-rater-ML (Heilman & Madnani, 2013). c-rater-ML is an automated scoring engine designed for scoring short constructed-response answers and developed at the Educational Testing Service. It uses supervised machine learning and automated model-building processes to produce scoring models for the items and has been previously used to score complex constructed-response science items (Liu et al., 2014, 2016). For each argumentation block, explanations are scored on a 7-point scale ranging from 0 to 6, representing:

- Score 0: Blank or off-task responses
- Score 1: Incorrect claim, data, or reasoning is mentioned
- Score 2: Restatement of the claim they chose in the prior multiple-choice claim prompt
- Score 3: Scientifically relevant but not fully elaborated statements related to data or reasoning
- Score 4: Scientifically valid, relevant, and fully elaborated data citation without the mention of reasoning
- Score 5: Scientifically valid, relevant, and fully elaborated reasoning statement without the mention of data
- Score 6: Scientifically valid, relevant, and fully elaborated data and reasoning

Uncertainty attributions are scored on a 5-point scale ranging from 0 to 4 as follows:

- Score 0: Blank or off-task responses
- Score 1: Personal knowledge, experience, or beliefs about the scientific phenomenon, data, or investigation
- Score 2: Nominal mention of 'data' without elaborating the specific patterns or features related to the data mentioned
- Score 3: Scientifically valid elaboration of uncertainty sources associated with the investigation related to the argumentation task

- Score 4: Mention of methodological, theoretical, or contextual limitation of the investigation related to the argumentation task

For further discussion of the uncertainty attribution scoring method, see Lee, Pallant, Lord, and Liu (2017). Scoring rubrics for all argumentation blocks are available (Pallant, Pryputniewicz, Lord, & Lee, 2016). Table 3 summarises the features of the items in an argumentation block.

The first step in building automated scoring models is to obtain as many human scores as possible. In this study, previously collected responses from 1180 students to each of the 16 constructed-response items related to explanations and uncertainty attributions were scored. Scoring rubrics were developed using the uncertainty-infused scientific argumentation framework developed at the Concord Consortium and published and validated by Lee et al. (2014). The data set was scored by multiple experts, including the developer of the scoring rubrics who had over 20 years of science assessment research experience and science content knowledge, and the developer of the climate change module. The inter-rater reliability across the 16 items ranged from 0.82 to 0.96.

c-rater-ML was used to train the models for each of the 16 items using the human-scored responses. For each item, the human-scored responses were randomly split into two sets; a training set included two-thirds of the responses and was used to build the models, while a test set included the remaining third of the responses and was used to validate the model. From the examples in the training set, c-rater-ML extracted several linguistic features, including word sequences, character sequences, semantic and syntactic dependencies, and response length. The c-rater-ML system used support vector regression, similar to multiple regression, to model relationships between linguistic features of the student responses in the training set and the scores. Specifically, a supervised machine learning algorithm analysed the examples in the training set and produced an inferred function, which could then be used to map new responses to scores.

For evaluating model performance on the test set, quadratic weighted kappa (QWK), Pearson's correlation (r), and standardised mean difference (SMD) were computed as measures of agreement between the human and c-rater-ML scores. For the purposes of this study, we used the thresholds suggested by Williamson, Xi, and Breyer (2012) that the satisfactory human–machine agreement with QWK and r must be at least 0.70, and SMD cannot exceed 0.15; and degradation from human–human agreement to human–machine agreement is recommended to be less than 0.10. For all items in this study, the QWK values were above 0.70, correlation coefficients were above 0.70, and the SMD values were less than 0.15. Degradation from human–human agreement to human–machine agreement ranges from 0.04 to 0.22 for QWK and −0.07 to 0.15 for correlation. Although the degradation for the subset of the data is slightly over the suggested threshold of 0.10, we consider overall good human–machine agreement was achieved for two reasons:

Table 3. Argumentation block structure.

Argumentation element	Item type	Scoring	Feedback
Claim	Multiple choice	No scoring	No
Explanation	Constructed response	Automated score of 0–6	Provided
Rating of uncertainty	Likert scale	No scoring	No
Uncertainty attribution	Constructed response	Automated score of 0–4	Provided

Table 4. Explanation score and matching feedback.

Score	Feedback
0	You have not explained your claim yet. Can you include scientific evidence and reasoning that explain your claim?
1	Your claim, evidence, or reasoning was either inconsistent with scientific views or was unclear. Can you modify or elaborate your explanation?
2	You made a claim without an explanation. Can you include scientific evidence and reasoning that support your claim?
3	You identified some climate-related factors associated with temperature. Can you include evidence and reasoning that explain the associations?
4	You included evidence for your claim. Can you elaborate how or why the evidence supports your claim?
5	You included scientific reasoning that explained your claim. Can you add evidence to support your reasoning?
6	You included evidence and reasoning to support your claim. Great job!

Table 5. Uncertainty attribution score and matching feedback.

Score	Feedback
0	You have not explained your certainty rating. Have you compared the strengths and weaknesses of the evidence that you used to support your claim?
1	Your personal beliefs, experiences, and attitudes can influence your certainty rating. How do the strengths and weaknesses of the scientific evidence affect your certainty rating?
2	You mentioned that either the data or the model affected your certainty rating. Can you be more specific about how the data or model influenced your rating?
3	You mentioned specific evidence and knowledge that influenced your certainty rating. Have you also considered the strengths and limitations of the data and models related to this question?
4	You recognised strengths and limitations of knowledge and evidence related to the current investigation. Excellent!

(a) only 25% of the responses were double-scored by two raters and (b) those two raters were experienced scorers for those items in the Climate change module, leading to a fairly high human–human agreement. In conclusion, the automated scoring models showed good human–machine agreements for all explanation and uncertainty rationale items (Mao et al., 2016), and c-rater-ML was able to closely approximate human scores given to complex science items for the purpose of student learning.

Automated feedback was then designed to address what students were missing or not elaborating in their constructed responses in order to help them formulate a competitive argument. Since each score represents ways students incorporate or do not incorporate a particular set of data, reasoning, and uncertainty sources of attribution, feedback statements were developed for each score across the eight argumentation blocks. Details of feedback statements for corresponding score categories are presented in Tables 4 and 5. Generally, each feedback statement includes two parts. The first part is the evaluation of the current argumentation response and the second part provides suggestions on how to improve the response to receive higher scores. Since there are no additional argumentation elements to include, the feedback statements for the highest score categories include only the first evaluation part. Students decide whether to make revisions and resubmit and can revise as many times as they desire. At each round of revision, students are also asked to rate the usefulness of the feedback on the 3-point scale of 'Not at all', 'Somewhat', and 'Very'.

Log data

While students work on the questions in the browser, all students' responses to activity prompts including the argumentation blocks are automatically recorded by the server.

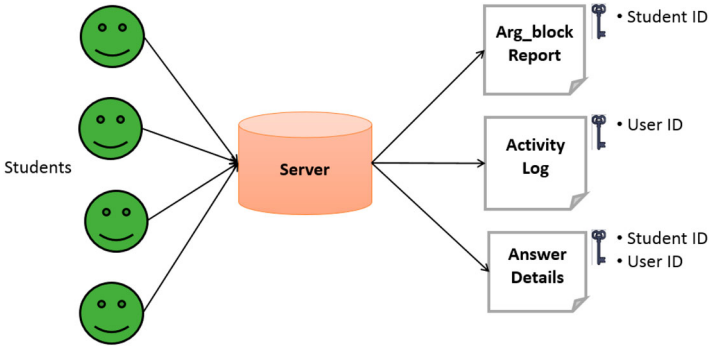

Figure 2. Structure of log data.

Each student is assigned to a Student ID and a User ID as unique keys, and the Student ID and User ID are one-on-one matched. The current system design records all information in three files, which can be cross-checked with each other using these unique keys for students. The file 'Arg-block Report' includes details on each argumentation block submission and has Student ID as the key. The file 'Activity Log' has time-stamped data on all interactions students had with the system, such as opening a certain page, start to answer a question, or submit the answers. It uses the User ID as the key. Finally, the file 'Answer Details' contains information on the final responses to all questions in the module and the report on the progress. Both Student ID and User ID are included in this file. Structures of log files are illustrated in Figure 2. Analyses in this study focus on the data provided in the first file, because it is the main file that captures the interactions of the students with the automated feedback for the eight argumentation blocks.

When a student clicked the submit button inside the argumentation block, the student's responses to all four argumentation prompts were saved as a single row of record and added into the Arg-block Report file. Students could not submit one argumentation prompt at a time. This record contains the time-stamp of that submission, the student ID, as well as the answers, scores, and feedback. We aggregated the data for each argumentation block submission because each submission was made as a whole for the argumentation block. Thus, we used the sum of explanation and uncertainty attribution scores as the score representing the quality of the argumentation submitted. We then aggregated the argumentation data for each student and generated variables including the total number of argumentation blocks completed, the mean initial scores for all argumentation blocks finished by students, the mean final scores for all argumentation blocks finished by students, the mean number of revisions, the mean score increase, the mean time spent on revisions, and the average rating for the usefulness of the feedback.

Method

Participants

The data used in this study were collected in Fall 2015 from 11 classes taught by 3 teachers in 3 high schools in the U.S.A. These teachers implemented the climate change module as part of their classroom teaching. The climate change module was delivered online through

a web-portal connected to the curriculum server residing at the Concord Consortium. The students in small groups worked under the teachers' guidance on the climate change module activities. Even though students worked collaboratively during the module activities, they produced individual responses to the argumentation prompts. A total of 183 students participated in this study. Among them, 45% were females, 49% were males, and 6% did not report gender. There were 87% Whites, 2% African Americans, 1% Hispanics, 1% Asians, 1% Pacific Islanders, and the rest either chose not to answer or reported to belong to another category. In terms of grades, there were 40% 9th graders, 8% 10th graders, 26% 11th graders, 19% 12th grader, and 7% of the participants did not report their grades.

Data analyses

To answer the first research question on the patterns of students' revisions, we conducted frequency analysis to examine students' submission and resubmission actions, independent sample *t*-tests to compare initial scores of students who revised with that of those who did not revise, and Pearson correlation and regression analysis to see if the number of revisions was related to scores.

To address the second research question on the relationship between students' response change and their final scores, we used independent sample *t*-tests to compare mean initial and final scores for students who revised. We also used linear regression analysis to see if the number of revisions predicted score increases.

To answer the third research question on item difficulty shifting, we performed analyses based on classical test theory (CTT) and item response theory (IRT). Item statistics including item difficulty, corrected item-total correlations, and Cronbach's alpha reliability were examined. Items were designed to measure a single-dimension construct (climate change); thus, we conducted a parallel analysis to evaluate if the unidimensional assumption was met. Parallel analysis (PA) is employed to detect the number of components or factors to retain from principal factor analysis (Horn, 1965), and various studies indicate that PA shows the least variability and sensitivity to different factors (Glorfeld, 1995; Zwick & Velicer, 1986); thus, it was recommended to use it to determine the unidimensionality in this study. The Rasch Partial Credit Model (PCM) was used for analysis after confirming the unidimensionality of the data in ConQuest. Lastly, we compared the item difficulties between the initial and last responses.

Results

Argumentation block completion

We first report descriptive statistics on how many students completed and revised the argumentation blocks. Each of the 183 students in our data set had the chance to work on the eight argumentation blocks, but not all of them completed all the argumentation blocks. On average, each student completed 6.57 argumentation blocks. Figure 3 shows a bar chart of the number argumentation blocks completed by students. Even though only 68 students (37.16%) completed all eight argumentation blocks, 162 (88.52%) students completed more than half of the argumentation blocks.

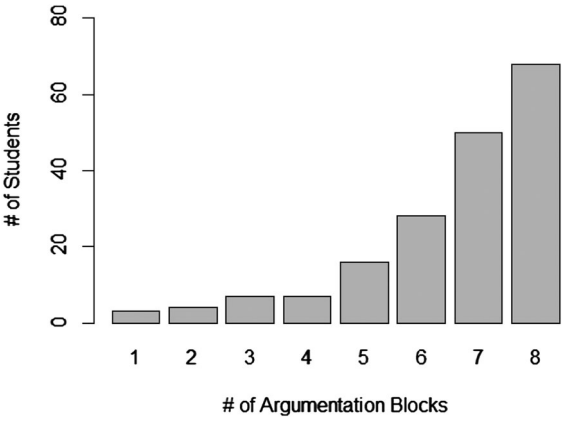

Figure 3. The distribution of the number of argumentation blocks completed.

Student revision patterns

To identify how students' responses changed after receiving immediate feedback statements, we focused on variables related to argumentation revisions. Students' revisions showed that students made use of the feedback. Overall, 77% of 183 students made revisions to one or more argumentation blocks, whereas only 23% did not make any revisions. The distribution of the number of revised argumentation blocks is shown in Figure 4. Since not all students completed six, seven, or eight argumentation blocks that appeared towards the end of the module, the number of students who revised six, seven, or eight argumentation blocks decreased dramatically. Among the 1202 argumentation blocks completed by the 183 students, 32% were revised after receiving the feedback. At the individual student level, among students who made at least one revision, each student revised an average number of 2.08 argumentation blocks, with 1.59 revisions on each block. Each round of revision took an average of 11.79 seconds before resubmission.

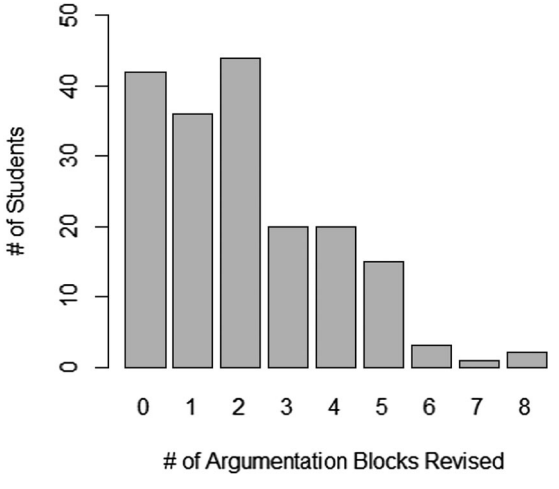

Figure 4. The distribution of the number of argumentation blocks revised.

The students were asked to indicate the usefulness of the feedback after they received the automated scores and feedback. They chose from three options of 'Not at all', 'Somewhat', or 'Very' and these responses were coded as 0, 1 and 2, respectively. Over all resubmissions, the average of these usefulness questions was 1.25, which was between 'Somewhat' and 'Very'.

We studied whether there was a significant difference between students who revised and those who did not in terms of initial argumentation scores calculated by combining explanation and uncertainty attribution scores. An independent sample t-test was conducted over all completed initial argumentation blocks. There was a significant difference in the mean initial scores of students who revised ($M = 5.20$, SD $= 1.81$) and students who did not ($M = 4.44$, SD $= 1.88$); $t(64) = -2.31$, $p = 0.02$. This suggested that students who received higher combined argumentation scores on their initial submissions were more likely to follow suggestions in the automated feedback and revise (Figure 5).

We further explored how initial argumentation scores were related to the number of students' revisions and the amount of time spent on revisions for students who revised. Among the 141 students who made at least one revision, their initial argumentation scores were not significantly related to the number of revisions made ($r = 0.002$, $p = 0.98$), nor was the mean time spent on each revision ($r = -0.34$, $p = 0.10$). Results suggested that students' initial argumentation scores predicted neither the number of revisions they would make, nor the time they spent on revisions.

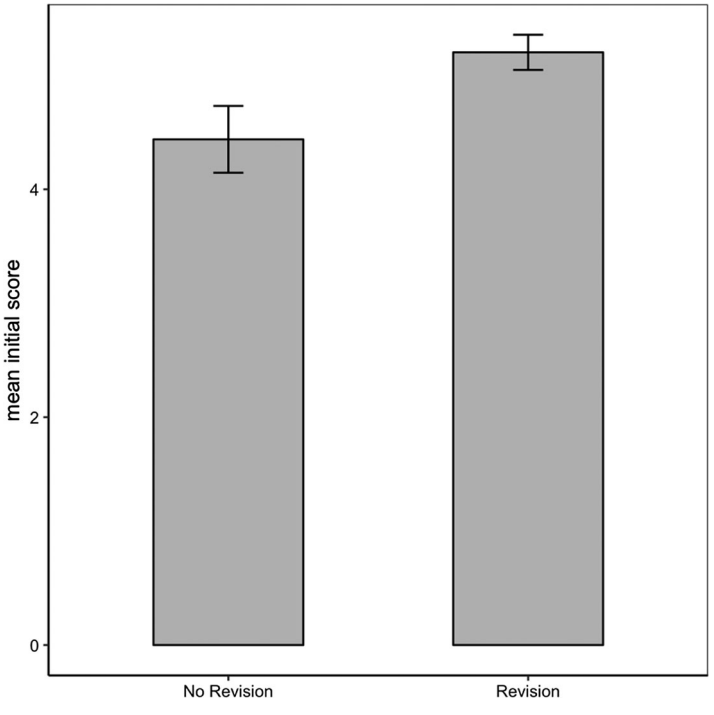

Figure 5. Comparison of the mean initial scores for groups with and without revision.

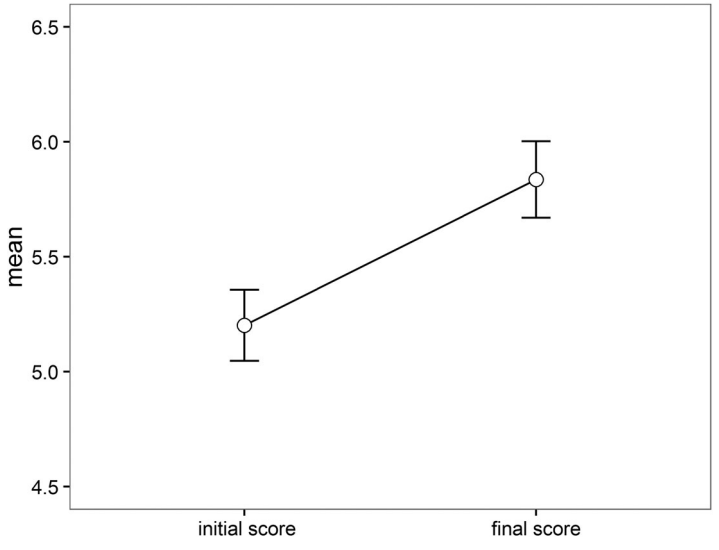

Figure 6. Comparison of the average final scores for groups with and without revision.

Impact of revisions on scores

To examine how revisions impacted students' final argumentation scores, the *t*-test for paired samples was conducted to compare initial and final argumentation scores across all of the completed argumentation blocks for students who made revisions. As shown in Figure 6, the mean initial argumentation score ($M = 5.20$, $SD = 1.81$) was significantly lower than the mean final argumentation score ($M = 5.84$, $SD = 1.95$); $t(136) = -11.41$, $p < .01$, and the effect size (Cohen's d) was 0.96. Results suggested that the students produced more scientifically competent arguments after revisions as compared to their initial arguments.

Linear regression analysis was conducted to test if the number of revisions predicted argumentation score increase. Again, we aggregated the data at the student level. We used the average argumentation score increase over all completed argumentation blocks for each student as a dependent variable and the average number of revisions as an independent variable. The results of the regression indicated that the average number of revisions significantly predicted the average argumentation score increase, $\beta = 0.55$, $t(136) = 6.69$, $p < .001$. That is, each revision resulted in an average of 0.55 increase on the final scores. The average number of revisions also explained a significant proportion of variance in the average score increase, $R^2 = 0.25$, $F(1136) = 44.81$, $p < .001$.

Item difficulty shifting

Detailed psychometric information of the climate change module is available in Mao et al. (2016) with a larger student sample size. To sum up, all of the explanation and uncertainty rationale items across the eight argumentation tasks were at the medium to high item difficulty levels (0.28–0.47), were moderately to highly correlated with the total score (0.35–0.61), and had moderate to high discrimination (0.35–0.61). The Cronbach alpha of

explanation and uncertainty attribution items across the eight argumentation tasks was 0.87, indicating satisfactory reliability. The parallel analysis showed that the first eigenvalue of the real data was about five times larger than the second eigenvalue. The first factor accounted for 25% of the variance in student scores. In addition, the magnitude of the second eigenvalue was not distinguishable from the rest of the eigenvalues. This result suggested that the assessment was approximately unidimensional. In terms of the item fit, the unweighted mean square (MNSQ), sensitive to off-target unexpected responses, was 0.85–1.40. The weighted MNSQ, sensitive to on-target responses, was 0.88–1.33. Both scores with an expected value range from 0.70 to 1.40 would suggest that items meet the PCM fit (Smith, 2000).

To investigate item difficulty shifting before and after revisions, we applied CTT to calculate the item difficulty for explanation and uncertainty attribution items across the eight argumentation tasks. In other words, we averaged student scores for every explanation and uncertainty attribution item, and compared the item difficulties between the initial and final responses. The Cronbach alpha reliability across the eight explanations and eight uncertainty attributions in this study was 0.92, indicating a satisfactory reliability that is aligned with the previous findings from the Mao et al. (2016) study.

Figures 7 and 8 show that students' final scores were consistently higher than their initial scores. This suggested that students' arguments improved after receiving the automated feedback. The increase in average scores between the initial and final responses showed that the item difficulty decreased for the final response for all items compared to the initial response. For example, on the item that asks students to predict the future of Earth's climate based on a line graph (argumentation block 1), a student's initial response was 'it's continually rising, and shows no sign of slowing down, decreasing, or evening out', and was scored 3 out of 6. After receiving the automated feedback, this student revised his response multiple times. His last response was revised to be 'we're releasing a lot of carbon into the atmosphere, population is increasing with more people releasing greenhouse gases and using more fossil fuels for energy', and this response received a score of 5.

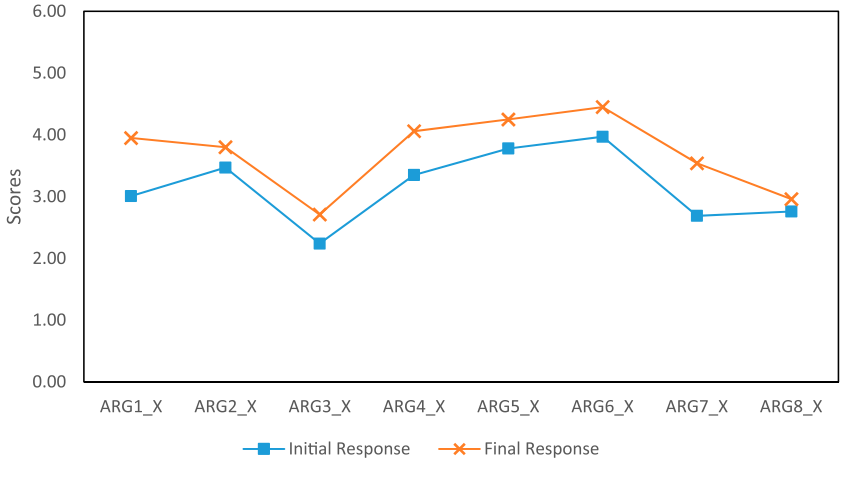

Figure 7. Change in average initial and final scores for eight explanation items.

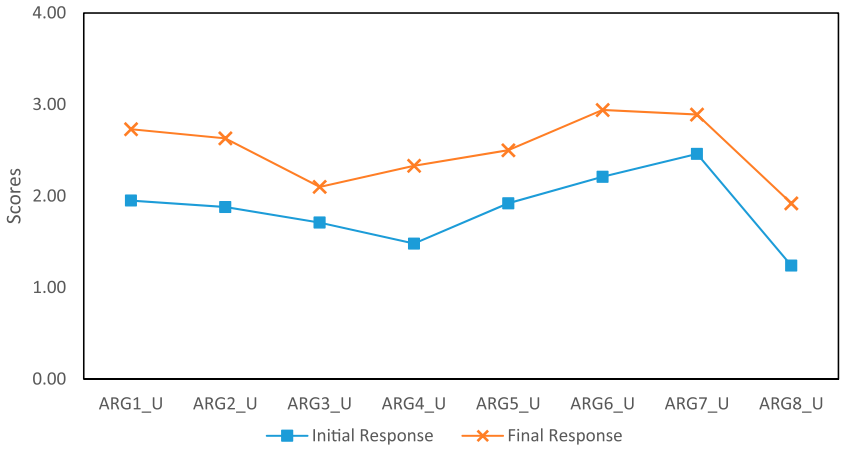

Figure 8. Change in average initial and final scores for eight uncertainty items.

Conclusions

This research used an online climate change module with automated scoring and feedback features in order to study the impact of automated feedback on students' scientific argumentation responses. We found that the majority of the students revised after receiving the automated feedback, and students with higher initial scores were more likely to revise. For those students who revised, their scores significantly improved after revision. Furthermore, psychometric analysis showed that the assessment was unidimensional, and had reasonable reliability and item-test correlations. Item difficulty shifting analysis found that items became easier after students received the feedback and made revisions. This study provides evidence on the positive impact of automated feedback on student scores in scientific argumentation. As future steps, we plan to explore ways to provide customised feedback based on both the scores and the detailed content of the answers, and to study students' activities after receiving the feedback beyond making revisions. Both directions are towards better assisting students' learning experience.

Discussions

Despite these encouraging findings, we also noted two areas that need further studies and improvements in our feedback design. First, in comparing students who revised and those who did not, we found that the students who revised had higher initial scores. It is somewhat counter-intuitive as we expected that students who received lower scores would be more likely to revise to get higher scores. There are several potential explanations to the observed results. For instance, the students who received lower scores may be those who were not motivated to improve. This lack of motivation we speculate could be due to the lack of interest in the topics, or because it was a low-stake assignment. Another possible explanation is that those who got lower scores did not find the feedback helpful, or lacked skill or knowledge on how to improve the response. On the other hand, students who received higher scores might be more interested in the topics, motivated to perform well in school, and have more knowledge or skill to utilise automated

feedback provided to them. Furthermore, correlation analyses showed that students' initial score was not a significant predictor of the number of revisions they made. On the other hand, students with lower initial scores were less likely to make revisions. It is possible to revise the feedback to offer stronger motivation to these students to revise, and to provide additional support to offset the lack of knowledge or skills necessary for the argumentation tasks. For instance, more detailed instructional steps or more detailed prompts contextualised to the argumentation tasks can be provided to students in the feedback.

Second, the impact of the automated scoring and feedback on students' argumentation revisions was investigated in this study mainly through data, which recorded students' responses and interactions with the online curriculum module. Though these analyses led to a discovery of some interesting revision patterns, our explanations of why these observed patterns occurred were speculative. Recognising this limitation, we also collected screencast videos of a subset of students working in groups of two to three. Screencast videos captured the audio of students as well as their computer screen while they were working on the module. The analysis of the screencast videos will give more details on how students processed automated feedback, made decisions on whether to revise, and what additional information or conversation contributed to revisions of their arguments. We hope that further analysis of the screencast data can provide evidence that can explicate the patterns identified in this paper.

Implications for science education

While the instructional focus on scientific argumentation is growing, it is difficult to enact it in real classroom settings where students are likely to need a lot of guidance from teachers. As the path for student learning towards scientific argumentation can be divergent, how to support students' diverse developmental trajectories is at the core of designing instructional support systems. Scientific argumentation allows students to justify claims using data in light of their understanding of the established knowledge and to reflect on limitations of investigations through which the data are produced (Allchin, 2012; Lee et al., 2014). While proper scaffolding is necessary to facilitate student learning (Quitana et al., 2004), it is practically impossible to expect teachers to help each and every student in real time on every argument he/she writes (McNeill & Pimentel, 2010). The automated feedback we tested in this study provides an exemplar for a customised scaffolding system that provides immediate support for individual students adjusted to their progress. Preliminary findings show that this automated feedback system can enhance students to develop scientific arguments when integrated into an online curriculum module.

Future directions of research

The development of automated scoring and feedback to improve students' written scientific argumentation in the context of science instruction is a complex endeavour. This requires numerous components to work seamlessly. Due to this complexity, we have been carrying out our design-based research cumulatively by testing one additional feature at a time to learn how students responded to the added feature. This study was conducted with automated feedback that facilitated students to include argumentation

elements such as data and reasoning. The next design cycle involves replacing the current automated feedback with contextualised feedback, where feedback statements are rephrased to reflect specific data and reasoning involved in each argumentation task. After testing contextualised feedback, we will conduct a quasi-experimental study involving the automated feedback we tested in this study and the contextualised feedback we will design. These two feedback conditions will be randomly assigned at the class level. The study will provide information about how two different types of feedback work for each argumentation task. Based on this information, we will finalise the automated feedback statements.

In addition, this study has not explored the activities students did after they received the automated feedback. For instance, it would be interesting to analyse whether students went back and checked the information provided on the same page, or how they navigated through the system to find more information. As these additional activity details were recorded in the log data, we plan to further analyse the log data in the future studies. Last, we will further explore what potential reasons could explain the item difficulty shifting.

Limitations

One limitation is that the current feedback was not customised to each argumentation task. As mentioned in the previous section, we plan to improve the feedback statements in future studies. Another limitation is that all four argumentation elements were submitted as a unit. As a result, student responses to different argumentation elements in the same argumentation block were not independent from each other. This limitation is due to the design and technical requirements of the automated feedback. For example, if we collect claim, explanation, uncertainty rating, and uncertainty attribution separately, what students do can no longer be considered as an argumentation activity. Separating different items might enable detailed studies on the automated feedback on segregated argumentation elements, but this decision will lose construct validity where all four-part argumentation tasks all at once represent students' current ability to argue with data based on their reasoning and consideration of limitations. In addition, the artificial separation of four-part argumentation prompts creates unnecessary technical and analytical challenges in the way argumentation blocks work. A third limitation is that only 25% of student responses were double-scored by two raters, which might result in exceeding the 0.10 degradation threshold between human–human and human–machine agreements. In the future, we would like to use the stratified sampling approach to select a full spectrum of student responses for human raters and meet the degradation threshold.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This material is based upon work supported by the National Science Foundation under grant number 1418019. Any opinions, findings, and conclusions or recommendations expressed in

this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

ORCID

Mengxiao Zhu http://orcid.org/0000-0003-3596-5585

References

Allchin, D. (2012). Teaching the nature of science through scientific errors. *Science Education*, *96* (5), 904–926.

Anderson, D., Magill, R. A., & Sekiya, H. (2001). Motor learning as a function of KR schedule and characteristics of task intrinsic feedback. *Journal of Motor Behavior*, *33*, 59–66.

Aufschnaiter, C. V., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching*, *45*, 101–131.

Azevedo, R., & Brenard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research*, *13*, 111–127.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26–55.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Berkshire: McGraw-Hill Education.

Bricker, A., & Bell, P. (2008). Conceptualizations of argumentation from science studies and the learning sciences and their implications for the practices of science education. *Science Education*, *92*, 473–498.

Chin, C., & Osborne, J. (2010). Students' questions and discursive interaction: Their impact on argumentation during collaborative group discussions in science. *Journal of Research in Science Teaching*, *47*(7), 883–908.

Clarke, S. (2003). *Enriching feedback in the primary classroom*. London: Hodder and Stoughton.

Cope, B., Kalantzis, M., Abd-El-Khalick, F., & Bagley, E. (2013). Science in writing: Learning scientific argument in principle and practice. *E-Learning and Digital Media*, *10*(4), 420–441. doi:10.2304/elea.2013.10.4.420

Dihoff, R. E., Brosvic, G. M., & Epstein, M. L. (2003). The role of feedback during academic testing: The delay retention effect revisited. *The Psychological Record*, *53*, 533–548.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education*, *38*, 39–72.

Dzikovska, M., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., … Dang, H. T. (2013). SemEval-2013 Task 7: The joint student response analysis and 8th recognizing textual entailment challenge. *Second joint conference on lexical and computational semantics (*SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)* (pp. 263–274). Retrieved from http://www.aclweb.org/anthology/S13-2045

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

Glorfeld, L. W. (1995). An improvement on horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, *55*(3), 377–393. doi:10.1177/0013164495055003002

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2008). *Highlights from TIMSS 2007: Mathematics and science achievement of U.S. Fourth- and eighth-grade students in an international context (NCES 2009-001 revised)*. Washington, DC:National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://nces.ed.gov/pubs2009/2009001.pdf

Graesser, A. C. (2011). Learning, thinking, and emoting with discourse technologies. *American Psychologist*, *66*(8), 746–757.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. *CBE Life Sciences Education*, *10*(4), 379–393. doi:10.1187/cbe.11-08-0081

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

Heilman, M., & Madnani, N. (2013). ETS: Domain adaptation and stacking for short answer scoring. *Second joint conference on lexical and computational semantics (*SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013), 2 (SemEval)* (pp. 275–279). Retrieved from http://www.aclweb.org/anthology/S13-2046

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Johnson, W., & Valente, A. (2009). Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. *AI Magazine*, *30*(2), 72–83. Retrieved from https://www.researchgate.net/profile/Andre_Valente/publication/221606431_Tactical_Language_and_Culture_Training_Systems_Using_Artificial_Intelligence_to_Teach_Foreign_Languages_and_Cultures/links/0046352964ed029fb0000000.pdf

Kuhn, D. (2010). Teaching and learning science as argument. *Science Education*, *94*(5), 810–824.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development*, *74*(5), 1245–1260. doi:10.1111/1467-8624.00605

Lane, S. (2005). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice*, *23*(3), 6–14. doi:10.1111/j.1745-3992.2004.tb00160.x

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389–405. doi:10.1023/A:1025779619903

Lee, H.-S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education*, *24*(2), 115–136. doi:10.1080/08957347.2011.554604

Lee, H.-S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in Science Teaching*, *51*(5), 581–605. doi:10.1002/tea.21147

Lee, H.-S., Pallant, A., Lord, T., & Liu, O. L. (2017). Articulating uncertainty attribution as part of critical epistemic practice of scientific argumentation. *Proceedings of the computer-supported collaborative learning conference (CSCL '2017)*. Philadelphia, PA: International Society of the Learning Sciences.

Litman, D. J., & Silliman, S. (2004). ITSPOKE: An intelligent tutoring spoken dialogue system. *Proceedings of the human language technology conference: 4th meeting of the North American chapter of the association for computational linguistics (HLT/NAACL)* (pp. 233–236). Retrieved from http://dl.acm.org/citation.cfm?id=1614027

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19–28. doi:10.1111/emip.12028

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233. doi:10.1002/tea.21299

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulhollan, M., Lee, H.-S., & Pallant, A. (2016). *Validation of automated scoring for a formative assessment of students' scientific argumentation*. Manuscript submitted for publication.

Mayfield, E., & Penstein Rosé, C. (2010). An interactive tool for supporting error analysis for text mining. *Proceedings of the NAACL HLT 2010: Demonstration Session* (pp. 25–28). Retrieved from http://www.aclweb.org/anthology/N/N10/N10-2007

McNeill, K. L., & Pimentel, D. S. (2010). Scientific discourse in three urban classrooms: The role of the teacher in engaging high school students in argumentation. *Science Education*, *94*(2), 203–229.

Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002). *Towards robust computerised marking of free-text responses*. Proceedings of the 6th CAA international computer assisted assessment conference, Loughborough.

Moharreri, K., Ha, M., & Nehm, R. H. (2014). Evograder: An online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, *7*(15). doi:10.1186/s12052-014-0015-2

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

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

Pallant, A., Pryputniewicz, S., Lord, T., & Lee, H.–S. (2016). *Uncertainty-infused scientific argumentation rubrics*. Concord, MA: The Concord Consortium.

Quitana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., … Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *The Journal of the Learning Sciences*, *13*(3), 337–386.

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education*, *92*(3), 447–472. doi:10.1002/sce.20276

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, *88*, 345–372.

Shepard, L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice*, *28*(3), 32–37. doi:10.1111/j.1745-3992.2009.00152.x

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*, 153–189.

Smith, E. V. (2000). Metric development and score reporting in Rasch measurement. *Journal of Applied Measurement*, *1*(3), 303–326. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12029173

Staley, K. W. (2014). Experimental knowledge in the face of theoretical error. In M. Boumans, G. Hon, & A. C. Petersen (Eds.), *Error and uncertainty in scientific practice: History and philosophy of technoscience* (pp. 39–56). London: Routledge.

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

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103–118. doi:10.1207/s15324818ame0602_1

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432–442. doi:10.1037/0033-2909.99.3.432