An Application of Automated Scoring and Feedback to Support Student Writing of Scientific Arguments

Mengxiao Zhu
Educational Testing Service
mzhu@ets.org

Lydia Ou Liu
Educational Testing Service
lliu@ets.org

Hee-Sun Lee
The Concord Consortium
hlee@concord.org

Abstract

We applied machine learning-based automated text scoring techniques to provide immediate feedback to students who are writing scientific arguments as part of their classroom activities. Based on previously collected and hand-scored student responses, the scoring models were trained and validated for scientific argumentation tasks requiring constructed responses. In so doing, the c-rater-ML engine extracted a set of feature variables from the hand-scored student responses. Empirical studies showed that most students actively interacted with the feedback by making revisions. When students made revisions, their scientific argument scores tended to increase significantly.

Research Background

Providing immediate feedback on individual students' short written responses used to be considered time-consuming and impractical for classroom applications (Gibbs & Simpson, 2004). However, the rapidly-developing automated scoring techniques in recent years make it possible to process student responses in real time. Powerful automated scoring engines equipped with properly-validated scoring models have the ability to provide machine scores with the quality comparable to human raters and with reduced errors and biases (Williamson, Xi, & Breyer, 2012; Zhang, 2013).

In learning and testing, the recent decades witnessed the applications of automated scoring in various domains, such as English, Mathematics, and Science, for the purpose of evaluating writing quality, written content, and speech (e.g., Burstein & Marcu, 2002; Sukkarieh & Blackmore, 2009; Higgins, Zechner, Xi, & Williamson, 2011). However, limited studies focused on leveraging the power of automated scoring on providing immediate feedback for constructed response items in real-world classroom settings. Many existing studies are often constrained by the simplified scoring rubrics (only generating feedback on correct/incorrect instead of multi-level scores), or by the length of the responses that can be processed.

This work was based on an automated scoring engine, c-rater-ML (Heilman & Madnani, 2013) adopted to process students' written scientific arguments as part of writing tasks in a climate change curriculum for secondary school students. Multi-level scoring rubrics were enacted, and the immediate feedback system utilized the automated scores so that feedback is appropriate in addressing shortcomings associated with each score. A series of studies were conducted to explore how students reacted to the automated scoring and feedback as well as to investigate the impact of the automated feedback on the improvement in student arguments after revisions.

Automated Feedback and High-Adventure Science

This study was carried out when students learned one of the High-Adventure Science (HAS) interactive online modules developed by the Concord Consortium, the automated scoring and feedback version of the "What is the future of Earth's Climate?" module (climate change module hereafter, https://authoring.concord.org/sequences/476/). This learning module was designed based on the concepts of using authentic science practices in classrooms (Chinn & Malhotra, 2002; Lee & Songer, 2003) for the learning of scientific argumentation, and integrated scientific narratives, figures and graphs, as well as interactive simulations (the screen shot of one of the argumentation tasks is in Figure 1). The climate change module addresses how factors such as greenhouse gases, ocean, sea ice, and human activities can impact the Earth's future climate.

This module consisted of six one hour long activities with a total number of eight scientific argumentation writing task blocks. Each argumentation block follows a unique scientific investigation context and contains a set of four items to elicit students' written arguments: claim, explanation, uncertainty rating, and uncertainty attribution (Lee et al., 2014). Among these four, claim and uncertainty rating are multiple-choice items, and the other two are constructed response items. The four-item design of the scientific argumentation block was based on the studies that emphasized the importance of not only putting together scientific claims justified by data and reasoning but also considering uncertainty of the evidence-based claims due to the limitations of the investigation context (Buck, Lee, & Flores, 2014; Manz, 2015). The uncertainty-infused scientific argumentation assessment framework covers evaluations on both the abilities to make claims and the ability to evaluate the uncertainty of the claims.

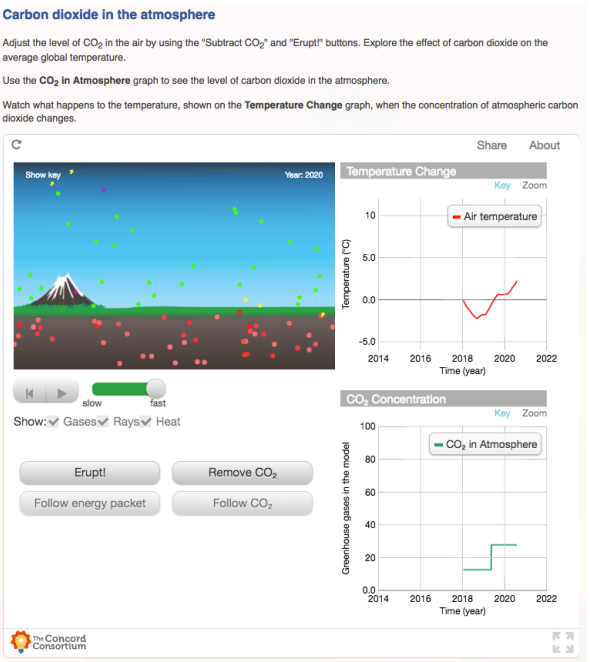

Figure 1. Screen shot of one scientific argumentation task.

Since the uncertainty-infused scientific argument blocks only yield two constructed responses from students on the explanation of the claim (scored on a 7-point scale ranging from 0 to 6) and the uncertainty attribution (scored on a 5-point scale ranging from 0 to 4). Therefore, these two items were automatically scored. Based on these scores, feedback statements were provided (see Figure 2). For the eight argumentation blocks, there are 16 constructed response items, and automated scoring models were built for each item. Response data were collected from 1,180 students, which were then scored by multiple domain experts based on the rubrics (Mao et al., 2018). The automated scoring models using the c-rater-ML engine were then trained and validated using feature variable extracted from the human scored responses, which adopted the support vector regression model (Smola & Schölkopf, 2004). Satisfactory c-rater-ML models were developed with high agreement with the human scores (more details on the model performance can be found in Zhu et al., 2017). The feedback was then developed based on the detailed rubric and the different score levels.

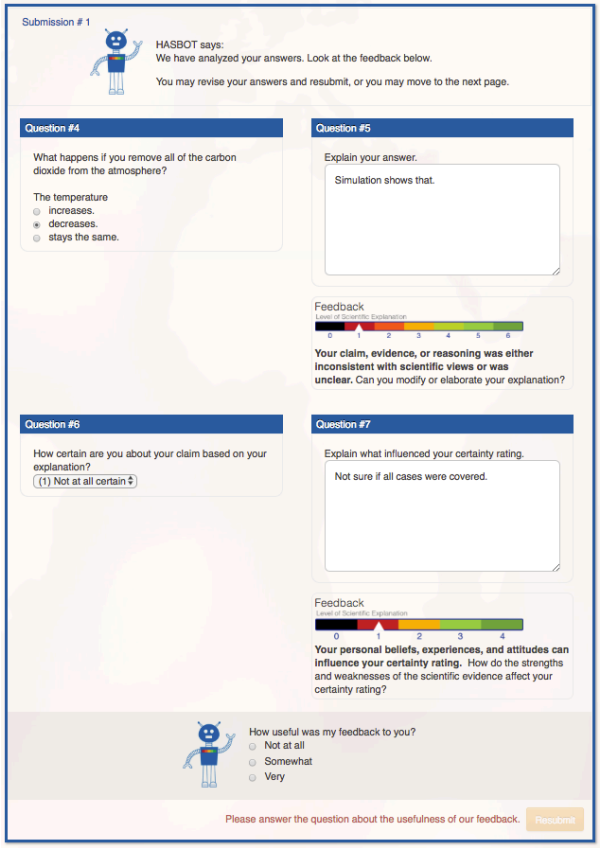

Figure 2. Screen shot of scores and feedback to the uncertainty-infused scientific argument block.

Data Collection and Findings

For this study, two rounds of data collections were conducted. The first round included 183 students from 11 classes at three high schools, and the second round included 374 students from 22 classes at eight schools in the United States. The climate module was integrated with classroom teaching. Since the climate module was delivered online, all student responses and activities were recorded with timestamps by the curriculum server. From the log data, we collected information on the scores of student responses as well as how students interacted with the automated feedback.

On the tendency of making revisions, for data from the first round dataset, 77% of 183 students made revisions to at least one argumentation block in the climate module, and 89% of the 374 students revised in the second round dataset. Further analyses on the subset of students who made revisions showed that there were significant increases on the final scores compared with the initial scores for both rounds of datasets (for the first dataset $M_{ini} = 5.20, SD_{ini} = 1.81, M_{final} = 5.84, SD_{final} = 1.95, t(136) = -11.41, p < .01$; for second dataset $M_{ini} = 5.90, SD_{ini} = 1.22, M_{final} = 6.75, SD_{final} = 1.11, t(333) = 22.47, p < .001$).

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 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.

References

Buck, Z. E., Lee, H.-S., & Flores, J. (2014). I am sure there may be a planet there: Student articulation of uncertainty in argumentation tasks. *International Journal of Science Education*, *36*(14): 2391–2420.

Burstein, J. C., & Marcu, D. (2002). Automated evaluation of discourse structure in student essays. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 200– 219). Mahwah,NJ: Lawrence Erlbaum.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, *86*(2): 175–218.

Gibbs, G., & Simpson, C. (2004). Conditions Under Which Assessment Supports Students' Learning. *Learning in Teaching in Higher Education*, *1*(1): 3–31.

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), 275–279.

Higgins, D., Zechner, K., Xi, X., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, *25*: 282–306.

Lee, H.-S., & Songer, N. (2003). Making authentic science accessible to students. *International Journal of Science Education*, *25*(8): 923–948.

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

Manz, E. (2015). Representing student argumentation as functionally emergent from scientific activity. *Review of Educational Research*, *85*(4): 553–590.

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H. –S., & Pallant, A. (2018). Validation of automated scoring for formative assessment of students' scientific argumentation in climate change. *Educational Assessment, 23*(2): 121-138.

Smola, a J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*: 199–222.

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

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.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *ETS R & D Connections*, (21): 1–11.

Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education,* 39(12): 1648–1668.