Unpacking Why Student Writing Does Not Match Their Science Inquiry Experimentation in Inq-ITS

Haiying Li, Janice Gobert, and Rachel Dickler
haiying.li@gse.rutgers.edu, janice.gobert@gse.rutgers.edu, rachel.dickler@gse.rutgers.edu
Rutgers University

Abstract: Science assessments should evaluate the full complement of inquiry practices (NGSS, 2013). Our previous work has shown that a large proportion of students' open responses did not match their scientific investigations (Li et al., 2017a). The present study both unpacks and compares the sub-components underlying students' performance for experimenting to their written open responses. These findings have implications for the assessment of inquiry practices, design of real-time scaffolding, and teachers' instruction of science.

Keywords: science inquiry assessment, educational data mining, natural language processing

Introduction

The Next Generation Science Standards (NGSS Lead States, 2013) are driving the need for assessments that can accurately measure students' inquiry practices, including: asking questions, planning and carrying out investigations, analyzing and interpreting data, warranting claims, constructing explanations, and communicating findings. Many researchers seek to develop assessments of students' inquiry practices. In particular, several assessments have been developed to capture students' scientific explanations, argumentation, and communication competencies (Liu et al., 2016; McNeill et al., 2006).

Inq-ITS (Inquiry Intelligent Tutoring System; www.inqits.com) is an online inquiry environment for middle school science in which students engage in both experimental and communicative NGSS practices. In Inq-ITS, students' experimental actions are captured in log files which are then automatically analyzed in real-time using patented algorithms (Gobert et al., 2016a; Gobert et al., 2016b). Students' written explanations are constructed in the format of claim, evidence, and reasoning, and are recorded as part of the explaining findings stage of each Inq-ITS virtual lab. Access to student performance in terms of both their actions and writing allows for not only capturing the complement of students' inquiry practices, but also for identifying any potential discrepancies in student performance between their "doing" and "communicative" inquiry practices. Specifically, a study by Li et al. (2017a) found inconsistencies between students' doing and writing for almost half of the participants who engaged in the Inq-ITS Density virtual lab. In the study, "doing" referred to the actions students took as they engaged in virtual science inquiry investigations, such as asking questions, planning and carrying out investigations, and analyzing and interpreting data (NRC, 2012). "Writing" referred to the construction of written scientific explanations containing argumentative components based on the results of students' virtual investigations. The findings from Li et al. (2017a) imply that assessments capturing only students' "doing" or "writing" may result in false positives when students are adept at parroting what they have read or heard but do not understand the science content or inquiry practices. Assessments may also result in false negatives when students who are skilled at science cannot articulate what they know in words.

Present study

The present study further explored the extent to which students' written scientific explanations reflected their doing during an experiment at a more fine-grained level by using specific science inquiry practices (i.e. experimental interpreting and experimental warranting) assessed within the Inq-ITS system as the units of analyses. This study investigated three research questions: (1) To what extent do students' competencies in communicative practices reflect their competences in experimental practices? (2) What distribution is displayed in terms of high versus low competency in experimental practices and high versus low competency in communicative practices? (3) To what extent does high versus low competency in experimental and writing practices mutually affect students' performance on experimental practices or writing practices alone?

293 middle school students (the same students from Li et al. (2017a, 2017b)) completed one Inq-ITS density virtual lab. We performed K-means cluster analyses ($K = 2$) on the sub-components of experimental and writing practices, respectively, and classified students into low versus high for each practice (resulting in four quadrants: Low (experimentation) –Low (writing), Low–High, High–Low, and High–High). We performed the Chi-square analysis and multivariate general linear model on experimental interpreting scores and written interpretation scores, as well as on experimental warranting and written warranting to examine the distribution of

students among the four quadrants. Multivariate general linear models (GLM) were performed to examine the extent to which the performance on experimental versus written interpretation and experimental versus written warranting practices differed among the four quadrants.

Findings and implications

Results of the linear regression for experimental and written interpretations showed that only one sub-component of experimental interpreting (i.e. interpreting IV) significantly predicted the written interpretation scores, $B = 1.30$, $t(293) = 2.78$, $p = .006$, $R^2 = .194$. Results of the linear regression for experimental and written warrants showed that two sub-components of experimental-warranting (i.e. the number of single trials ($B = 1.18$, $t(293) = 3.58$, $p < .001$) and all controlled trials ($B = -.34$, $t(293) = -2.13$, $p = .034$)) significantly predicted the written warranting scores, $R^2 = .056$.

Results of the Chi-square analysis for experimental and written interpretations showed that experimenting and writing were not independent, χ^2 (1, $N = 293$) = 21.77, $p < .001$. Results of the Chi-square analysis for experimental and written warranting showed that experimentation and writing were not independent, χ^2 (1, $N = 293$) = 4.56, $p = .033$. More than 30% of the total students exhibited discrepancies between experimental and written interpretation performance, and approximately 60% exhibited discrepancies between experimental and written warranting performance.

Results of the multivariate general linear model for interpreting revealed a statistically significant difference between experimental and written interpretations across the four groups, $F(6, 578) = 285.01$, $p < .001$; $\eta^2 = .747$. Tests of between-subjects effects indicated that group had a statistically significant effect on both experimental scores (F (3, 289) = 1085.37; $p < .001$; $\eta^2 = .918$) and written interpretation scores (F (3, 289) = 182.95; $p < .001$; $\eta^2 = .655$). Results of the multivariate general linear model for warranting also revealed a statistically significant difference in experimental and written warranting scores among the four groups, $F(6, 578) = 510.26$, $p < .001$; $\eta^2 = .841$. Tests of between-subjects effects indicated that group had a statistically significant effect on both experimental scores (F (3, 289) = 509.06; $p < .001$; $\eta^2 = .841$) and written warranting scores (F (3, 289) = 511.89; $p < .001$; $\eta^2 = .842$).

Results of the study revealed discrepancies between students' performance on inquiry practices through unpacking relations between students' experimental and communicative practices. The results of this study will significantly enhance research on teaching and the science of learning for the following two reasons. First, this study unpacks the complexity of scientific writing based on students' actions while conducting investigations. This study will inform teachers and researchers of the relationship between what students do during science inquiry and write accordingly. If students successfully engage in an experimental practice, can they report/reflect on what they have done as per NGSS (2013) expectations? Second, this study will promote the improvement of teaching methods for science inquiry in order to address students who demonstrate discrepancies between their experimental doing and explanatory writing performance.

References

Gobert, J. D., Baker, R. S., & Sao Pedro, M. A. (2016a). Inquiry skills tutoring system. *U.S. Patent No. 9,373,082.* Washington, DC: U.S. Patent and Trademark Office.

Gobert, J., Sao Pedro, M., Betts, C., & Baker, R.S. (2016b). Inquiry skills tutoring system (alerting system). *US Patent No. 9,564,057.* Washington, DC: U.S. Patent and Trademark Office.

Li, H., Gobert, J., & Dickler, R. (2017a). Dusting off the messy middle: Assessing students' inquiry skills through doing and writing. In E. André, R. Baker, X. Hu, M. Rodrigo, & B. du Boulay (Eds.), *Artificial Intelligence in Education* (Vol. 10331, pp. 175-187). Cham: Springer.

Li, H., Gobert, J., & Dickler, R. (2017b). Automated assessment for scientific explanations in on-line science inquiry. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining* (pp. 214-219). Wuhan, China: EDM Society.

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.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *The Journal of the Learning Sciences*, *15*(2), 153-191.

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

Acknowledgements

The research reported here was supported by Institute of Education Sciences (R305A120778) to Janice Gobert.