Int J Artif Intell Educ (2017) 27:665–670
https://doi.org/10.1007/s40593-017-0158-8

# Preface: Special Issue on Multidisciplinary Approaches to AI and Education for Reading and Writing

**Rebecca J. Passonneau[1] · Danielle McNamara[2] · Smaranda Muresan[3] · Dolores Perin[4]**

Published online: 3 November 2017


For the past decade, a majority of students in the United States have failed to meet grade level standards for reading and writing. Reports from the National Center for Education Statistics in 2004 through 2013 document that only a small percentage of middle school and high school students meet national standards of proficiency in reading and writing (Persky et al. 2003; Salahu-Din et al. 2008; National Center for Education Statistics (NCES) 2012; Glymph 2010; 2013; Glymph and Burg 2013). To achieve proficiency, students must master multiple kinds of verbal and reasoning skills, and this mastery takes years to develop. Much of this development happens in the years from middle school through college, where reading and writing instruction are secondary to disciplinary instruction. This includes reading and writing in STEM subjects, which receives less attention that more formal STEM skills, but is an important aspect of science education (Norris and Phillips 2003; Yore et al. 2004). Intelligent tutoring systems (ITS) for STEM subjects are being supplemented with

✉ Rebecca J. Passonneau
rjp@cse.psu.edu

Danielle McNamara
danielle.mcnamara@asu.edu

Smaranda Muresan
smara@ccls.columbia.edu

Dolores Perin
perin@tc.edu

[1] Pennsylvania State University, State College, PA 16801, USA

[2] Arizona State University, Tempe, AZ, USA

[3] Columbia University, New York, NY 10027, USA

[4] Teachers College of Columbia University, New York, NY, USA

similar systems for reading or writing instruction in STEM subjects, whether through ITS or other technologies.

Corrective measures for students' poor reading and writing skills should at least include increased time for students to practice, attempts to diagnose and provide guidance regarding the skills where they are weak, and targeted feedback while they practice (Kintsch 1990; Kellogg 2008; Johnstone et al. 2002; Kellogg and Raulerson III 2007). However, teachers and schools typically lack the resources to provide such supports. Furthermore, recent surveys indicate that teachers in the disciplines feel ill-prepared to provide instruction in reading and writing skills (Kiuhara et al. 2009; Graham et al. 2014; Gillespie et al. 2014). An increasingly practical option is to develop automated methods for evaluation, guidance, feedback, and instruction in reading and writing, to be deployed in a range of educational technologies. Two lines of research that exemplify this vision are automated methods to apply educational rubrics for reading and writing skills, and digital environments that support source-based writing, meaning writing based on comprehension of source texts.

The editors of two thematically-linked issues of IJAIED bring together five papers on AI applied to stem writing skills, and five papers on automated rubrics, source-based writing, or their combination. The five papers in the first of the two thematically related issues address learning to write in STEM subjects, and focus on STEM-specific tasks pertaining to writers' use of evidence, ability to construct arguments or explanations, or students' level of engagement with science subject matter (Barstow et al.; Klebanov et al.; Rahimi et al.; Tansomboon et al.; Wiley et al.). Three papers address the analysis of explanation and argumentation in science writing (Rahimi et al., Tansomboon et al., Wiley et al.). The paper by Klebanov et al. on engagement with the subject matter has some methodological similiarities to the papers in the following issue that address application of NLP techniques to automating educational rubrics. The five papers in the second of the two thematically related issues, however, more directly address how best to exploit NLP techniques to develop automated rubrics that address multiple aspects of student essays (Knight et al.; Passonneau et al.; Perin & Lauterbach; Rahimi et al.; Vajjala). A common thread linking many of these papers is the desirability of methods that could ultimately provide automated diagnostic feedback for students or teachers. Three of these papers (Passonneau et al.; Perin & Lauterbach; Rahimi et al.) as well as the one by Weston-Sementelli et al. focus on *source-based* or *text-based* writing tasks where the students read one or more texts in preparation for the writing task, and where their performance reflects their competence in both reading comprehension and writing. Very specific issues are also addressed such as formative guidance in legal writing (Knight et al.), writing skills of low-skilled adults in community colleges (Perin & Lauterbach, and indirectly Passonneau et al.), and second language learning (Vajjala).

## Reading and Writing for STEM Subjects

The paper by Barstow et al. presents a study of college-level writing instruction designed to fill gaps in research regarding the use of argument diagraming tools, and

is foundational for subsequent design of intelligent tutoring systems. They contrast the writing performance of a control group with a group who used a domain-general diagramming tool and a group who used argument diagramming specific to the domain of psychology. Both groups with argument diagramming had a significantly greater number of relevant citations, and examples of opposing evidence. The latter had a greater degree of validity of supporting and opposing citations.

To address the low retention rate of students in STEM subjects, the paper by Klebanov et al. investigates the use of NLP in a writing intervention that relies on utility value. College students have been found to have higher motivation for subject matter that has utility value, meaning direct relevance to them. In the utility value intervention, students articulate in writing how their courses relate to their lives. Administration of the intervention has depended on costly training of research assistants to assign utility value scores to student essays. This paper asks whether NLP can be used to automatically identify students with low utility value scores, who would then get additional instruction in articulating utility value.

Evaluation of young students' integration of reading comprehension and writing is investigated by Rahimi et al., who apply NLP to a rubric for text-based writing. The experiments address two dimensions of the rubric: students' use of evidence, and organization of ideas in support of a claim. As they note, AES methods often rely on easily observed features that act as proxies for higher-level writing skills, such as word counts and word length, and assign holistic scores that do not lend themselves well to diagnostic analysis. In contrast, this study develops features that directly represent components of the evidence and organization rubrics. In comparison to baseline models that rely on proxy features, or that adopt existing methods for analysis of text coherence, e.g., Barzilay and Lapata (2005) and Morris and Hirst (1991), the methods developed here perform better and generalize well across datasets.

Tansomboon et al. investigate adaptive guidance for students' short answers to science questions within a web-based science inquiry curriculum. The first of two studies found that students demonstrated better learning if automated guidance was personalized (e.g., with use of students' names) and transparent (i.e., students were provided age-appropriate explanations of how the computer selected feedback statements). The effect was significant, but only for a school that comprised more students with low prior knowledge. The second study compared two kinds of specific guidance, asking students to revisit the ideas and prompting them on ways to plan a revision, and found similar learning gains.

The paper by Wiley et al. considers the requirements for reliable scoring of text-based explanation in science. Information on global warming was distributed across ten texts that middle and high school students read before writing their essays to explain the causes of global warming. The paper applies and compares two types of manual assessment: scores based on a concept map of the ideas, and scores based on causal chains of ideas. The scores had high interrater reliability, and were predictive of performance on a comprehension test. The paper also discusses automating the assessment using Coh-Metrix and LSA, combined with machine learned classifiers for concept detection and causal connections.

## Automated Rubrics and Support for Source-Based Writing

Writing skills are argued to be particularly important for the legal profession in the paper by Knight et al., which presents work on participatory design of a web application, AWA (Academic Writing Analytics) that aims to provide automatic guidance to law school students, to compensate for the lack of sufficient instructor time to provide detailed feedback on drafts from the large numbers of students in law programs. On the question of whether an existing rhetorical parser from NLP can be tuned to automtically select important sentences, a small sample of sentences (N=90) judged by a human expert had very positive results. On the question of whether students would find a tool that highlights key sentences in their own writing helpful, the results were mixed, indicating a need to provide explanations to the students along with the highlighting.

Summarization of source texts is often used as an instructional strategy for reading comprehension and writing (Graham and Perin 2007). A defining characteristic of a summary is selection of important content from source texts. To create a reliable rubric to evaluate students' summaries is costly. Automated methods to evaluate machine-generated summaries are designed to rank summarization systems on multiple summarization tasks, and are not accurate enough to assess an individual summary. The paper by Passonneau et al. presents a manual method to analyze the content quality and coverage of summaries written by students or machines that depends on construction of a content model derived from summaries written by proficient individuals (a wise crowd). The wise-crowed method correlates well with a rubric designed to rate summaries written by community college students. Two automated NLP methods to apply wise-crowd content assessment also correlate well with the educational rubric.

Perin and Lauterbach address automated essay scoring of text-based summaries and text-based persuasive esssays for a specific population, low-skilled adults. They test three Coh-Metrix indices that had been found sufficient to predict human-scored writing quality of average-performing college students (McNamara et al. 2010). While these indices are not good predictors for the low-performing adults, ten other Coh-Metrix indices were identified that were predictive for this population. The resulting measures had very high variance, which the authors interpret to reflect that this population had diverse kinds of poor writing.

Automated Essay Scoring (AES) systems constitute a relatively well-developed commercial technology for evaluation of students' written essays (Burstein et al. 1998; Burstein 2003; Edelblut and Change 2004; Foltz et al. 2013; Landauer et al. 2003; Plakans and Gebril 2013; Rock 2007; Rudner et al. 2006). They typically assign scores to essays on a 3-point or 4-point scale, and agree with human scorers as well as human scorers agree with each other. The paper by Vajjala asks two questions about the design of such systems, including application to second language learners: what linguistic features are most general across different data sets? Does the first language predict writing proficiency in the second language?

Weston-Sementelli et al. compare the effectiveness of ITS strategy instruction for source-based writing, which includes reading comprehension, when students are given only reading comprehension strategies, only writing strategy instruction,

or a combination. Both content quality and writing quality are evaluated, and the combination is found to be significantly more effective than either strategy alone.

As a whole, the ten papers in these two issues point to the near-term feasibility of automated systems that can provide guidance and feedback to students and teachers on students' reading and writing skills, and promote engagement of students with the subject matter they write about. As these technologies mature, they could help create educational systems that promote the development of reading and writing skills throughout the course of our students' educations, and raise the proficiency of all students.

We give special thanks to the Editors-in-Chief for their careful review of the whole process, with guidance and feedback to ensure that the normal IJAIED review processes were followed. We were grateful for the detailed instructions, monitoring and support in this important aspect. It has meant that all papers received three high quality reviews from experts in the field, with a comprehensive metareview from an assigned editor who had no conflict of interest. We thank the IJAIED Associate Editors who handled the papers where the editors had conflicts, namely the papers by Passonneau et al., Perin & Lauterbach and Weston-Sementelli et al.

## References

Barzilay, R., & Lapata, M. (2005). Modeling local coherence: an entity-based approach. In *43rd Annual Meeting on Association for Computational Linguistics (ACL '05)* (pp. 141–148).

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M. (1998). Enriching automated essay scoring using discourse marking. In stede, M., Wanner, L., Hovy, E., Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M. (Eds.) *Workshop on Discourse Relations and Discourse Marking* (pp. 15–21): Association for Computational Linguistics.

Burstein, J. (2003). The e-rater® scoring engine: automated essay scoring with natural language processing. In Shermis, M.D., & Burstein, J. (Eds.) *Automated Essay Scoring: a Cross-Disciplinary Perspective* (pp. 107–115). Hillsdale: Lawrence Erlbaum Associate, Inc.

Edelblut, P., & Change, K. (2004). The impact of MY access!™ use on student writing performance: a technology overview and four studies from across the nation. In *Annual Meeting of the National Council on Measurement in Education*.

Foltz, P.W., Streeter, L.A., Lochbaum, K.E., Landauer, T.K. (2013). Implementation and applications of the intelligent essay assessor. In Shermis, M., & Burstein, J. (Eds.) *Handbook of Automated Essay Evaluation* (pp. 68–88). New York: Routledge.

Gillespie, A., Graham, S., Kiuhara, S., Hebert, M. (2014). High school teachers use of writing to support students' learning: a national survey. *Reading and Writing*, *27*(6), 1043–1072.

Glymph, A. (2010). The nation's report card: Reading 2009. Technical Report NCES 2010-458 National Center for Education Statistics (NCES), Washington, DC.

Glymph, A. (2013). The nation's report card: Reading 2012. Technical Report NCES 2012-457 National Center for Education Statistics (NCES), Washington, DC.

Glymph, A., & Burg, S. (2013). The nation's report card: a first look: 2013 mathematics and reading. Technical Report NCES 2014-451 National Center for Education Statistics (NCES), Washington, DC.

Graham, S., Capizzi, A., Harris, K., Hebert, M., Morphy, P. (2014). Teaching writing to middle school students: a national survey. *Reading and Writing*, *27*(6), 1015–1042.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*, 445–476.

Johnstone, K.M., Ashbaugh, H., Warfield, T.D. (2002). Effects of repeated practice and contextual-writing experiences on college students' writing skills. *Journal of Educational Psychology*, *94*(3), 305–315.

Kellogg, R.T. (2008). Training writing skills: a cognitive development perspective. *Journal of Writing Research*, *1*(1), 1–26.

Kellogg, R.T., & Raulerson III, B.A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review*, *14*(2), 237–242.

Kintsch, E. (1990). Macroprocesses and microprocesses in the development of summarization skill. *Cognition and Instruction*, *7*(3), 161–195.

Kiuhara, S.A., Graham, S., Hawken, L.S. (2009). Teaching writing to high school students: a national survey. *Journal of Educational Psychology*, *101*(1), 136–160.

Landauer, T., Laham, D., Foltz, P. (2003). Automatic essay assessment. Assessment in Education: principles. *Policy and Practice*, *10*, 295–308.

McNamara, D.S., Crossley, S.A., McCarthy, P.M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86.

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, *17*(1), 21–48.

National Center for Education Statistics (NCES) (2012). The nation's report card: Writing 2011.

Norris, S.P., & Phillips, L.M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*(2), 224–240.

Persky, H.R., Daane, M.C., Jin, Y. (2003). The nation's report card: Writing 2002. Technical Report NCES 2003-529 National Center for Education Statistics (NCES), Washington, DC.

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: source text use as a predictor of score. *Journal of Second Language Writing*, *22*(3), 217–230.

Rock, J.L. (2007). The impact of short-term use of Criterion on writing skills in 9th grade. Report Research Report RR-07-07 Educational Testing Service, Princeton, JN.

Rudner, L.M., Garcia, V., Welch, C. (2006). An evaluation of intellimetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, *4*(4), 1–21.

Salahu-Din, D., Persky, H.R., Miller, J. (2008). The nation's report card: Writing 2007. Technical Report NCES 2008-468 National Center for Education Statistics (NCES), Washington, DC.

Yore, L.D., Hand, B.M., Florence, M.K. (2004). Scientists' views of science, models of writing, and science writing practices. *Journal of Research in Science Teaching*, *41*(4), 338–369.