RESEARCH BRIEF

Contributions to Research on Automated Writing Scoring and Feedback Systems

Since 2017, experts at the RAND Corporation and the University of Pittsburgh have been conducting research to advance automated writing scoring and feedback systems for elementary school student writers. This brief summarizes some of the contributions of the work, which have been documented in greater detail in peer-reviewed articles.

Background and Problem Space

Writing is an important skill. Specifically, analytic text-based writing—which focuses on analysis and interpretation of fiction or informational texts, using evidence from the source text to support claims and interpretations—is critical to academic success and college readiness (Wang et al., 2018). As such, it is emphasized in national and state English language arts and writing standards. Students' opportunities to practice and learn analytic text-based writing are limited, however. Elementary grades have not historically focused on analytic text-based writing, teachers report feeling underprepared to teach writing, and the time needed to assess student writing is burdensome. When students do write, they rarely receive substantive feedback and rarely engage in cycles of revision that require them to apply feedback to strengthen their work.

The Promise and Perils of Automated Writing Scoring and Feedback Systems

Automated essay scoring (AES) and automated writing evaluation (AWE) systems have potential benefits. AES refers to the use of computer programs to assign a score to a piece of writing; AWE systems provide feedback on one or more aspects of writing and may or may not also provide scores. AES and AWE systems can reduce teachers' time burden related to grading and providing feedback, respectively, enabling teachers to assign more such writing tasks and thus provide more opportunities for students to learn and practice the skills. Moreover, students benefit from the more immediate—and consistent—feedback that AWE systems can offer, which supports cycles of revision.

AES and AWE systems, however, are not without challenges. Four common critiques are as follows:

1. **AES and AWE systems rarely provide guidance for students to improve substantive aspects of their writing.** Thorough assess-

ment of analytic text-based writing requires evaluating the content of students' writing, particularly how they marshal evidence from a source text to support their analysis. Historically, AES systems have tended to provide a holistic score rather than scores on specific dimensions of writing (e.g., content, organization, style, mechanics). Holistic scores often rely on surface aspects of writing (e.g., spelling, grammar, syntax, word count) because these aspects can be more reliably evaluated by widely used natural language processing (NLP) techniques. In recent years, systems have started to attend more to substantive dimensions of writing. Even so, few systems attend to the substantive dimension of evidence use, which is concerned with how the writing is supported by source material. AES systems that attend to evidence use and AWE systems that provide feedback supporting high-quality use of evidence would have the potential to guide instructional next steps or students' revision process.

2. **How well AES and AWE systems assess the targeted aspect of writing is rarely considered.** For the most part, assessments of how well AES and AWE systems perform have focused solely on how closely their scores correspond to human ratings of the same essays. Human raters have long been considered the "gold standard" for assessing student writing. In fact, human raters can be inconsistent in their assessments of writing and disagree with one another. Moreover, high human-computer agreement does not guarantee that the system is meaningfully assessing important aspects of writing, such as evidence use. To have confidence that an AES or AWE system is attending to the writing skills it purports to measure (i.e., it has *construct validity*), researchers need to gather evidence that the computer scores (that an AES system generates or that underlie an AWE system's feedback output) correlate to outcomes with which the skill should be associated. These outcomes include student achievement on standardized tests of those skills and the quality of classroom writing instruction.
3. **AWE systems are seldom designed for integration into teachers' daily work or instructional routines.** Because their primary output is a score, AES systems are typically regarded as providing summative information (i.e., at the end of a task or unit) about students' writing achievement. In contrast, AWE systems are intended to provide actionable feedback to improve students' writing. Despite the potential for widespread use, however, uptake of AWE systems in classrooms has been limited. Such systems are typically designed to be student facing, removing teachers from the interaction instead of helping them participate actively in students' use of the system. Yet, the development of complex skills—such as use of text evidence to support analytic writing—occurs in the context of rich social (student-teacher and/or peer) interactions. It is essential to design AWE systems with classroom implementation and the role of teachers in mind.
4. **AES and AWE systems may be subject to bias.** The underlying algorithms that assess student work could be unfairly advantaging or disadvantaging the work of students in particular demographic groups. Bias could stem from disproportionate representation of demographic groups (e.g., gender, race, socioeconomic status) in the student writing used to develop or train the system.

With these issues in mind, the research team, composed of researchers at RAND and the University of Pittsburgh, developed eRevise, an AES and AWE system, and undertook studies to address these four challenges. Following a brief overview of eRevise and study context and methods, this brief presents four key findings from the research. Each finding addresses one of the four perils of AES and AWE systems that the research team identified.

eRevise

The research team developed eRevise as a system for improving fifth- to seventh-grade students' skill in using text evidence. eRevise uses machine learning and NLP techniques to predict evidence-use scores in students' writing on the Response to Text Assessment (RTA). The RTA is a formative writing task aligned with the Common Core State Standards (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). Students read a nonfiction text and then write an essay in which they evaluate

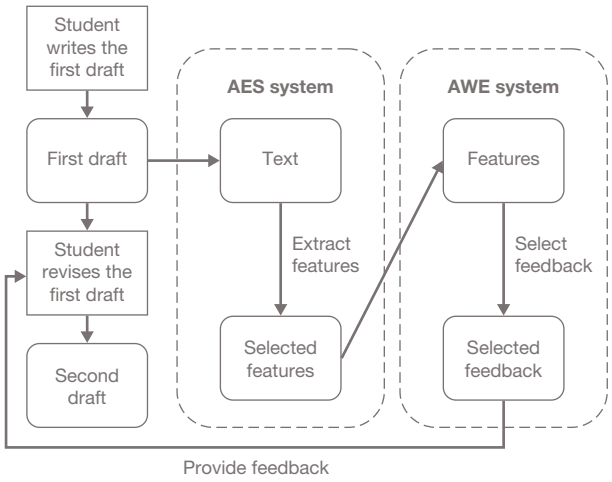
claims made by an author (see Correnti et al., 2012; and Correnti et al., 2013).

The AES score provided by eRevise is based on features of evidence use on typical grading rubrics that humans would use to assess such writing. The features are (1) number of pieces of evidence used; (2) specificity of the evidence provided; (3) concentration, an indication of whether students simply listed evidence or provided elaboration or explanation; and (4) word count. During development, the AES system was trained on more than 1,500 previously collected and manually scored essays. The researchers first used NLP to represent each essay in terms of the four rubric-based features. They then used machine learning techniques to learn a model for predicting an essay's score given only the feature representation as the input. They conducted tenfold cross-validation. This entails randomly dividing the corpus of essays into ten parts, using nine of those parts as a training set and the remaining part for testing, and repeating this ten times (Rahimi et al., 2014).

As for the AWE component, eRevise uses the first two of the features identified above (number of pieces of evidence and specificity of evidence) to select the feedback messages to display to students. For example, students who did not provide enough pieces of evidence receive the following prompt: "Use more evidence from the article." See Figure 1 for a depiction of the architecture of eRevise and Figure 2 for a view of the student interface, with example feedback messages. For more details on eRevise, see Correnti et al., 2020, and Zhang et al., 2019.

FIGURE 1

Architecture of eRevise



SOURCE: Zhang et al., 2019.

Study Contexts and Methods

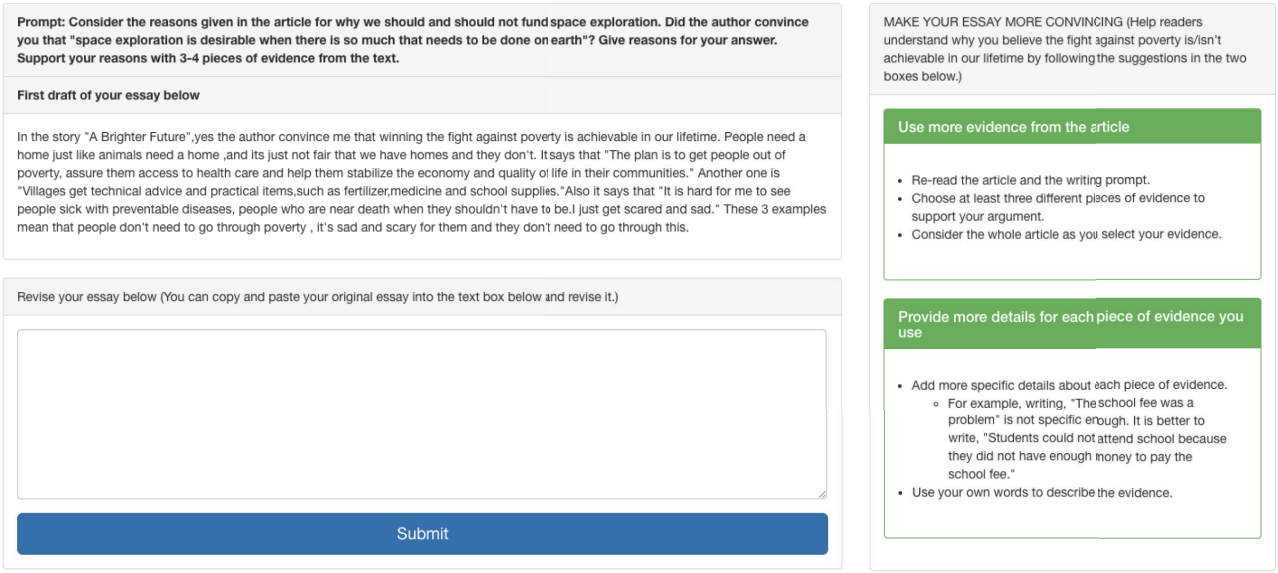
This brief summarizes findings from three studies evaluating eRevise. Study 1 established validity of the eRevise scores (Correnti et al., 2020). It involved 65 sixth-grade language arts classrooms located in 29 different schools in the same large urban district in Maryland. The study examined a total of 1,529 essays. Study 2 evaluated eRevise's performance as a formative feedback system for improving student writing. This study involved fifth- and sixth-grade language arts teachers throughout Louisiana. In the first year, seven teachers participated, and the researchers analyzed essays from 143 students (Zhang et al., 2019). In the second year, 16 teachers participated, and the researchers analyzed 266 sets of first- and second-draft essays (Correnti et al., 2022). Study 3 examined the fairness of the algorithm underlying eRevise using the subset of 735 essays from Study 1 for that was scored by two human raters (i.e., those for which interrater reliability has been established; Litman et al., 2021).

In these studies, the research team examined various aspects of eRevise. They performed human scoring and annotation of a diverse sample of student essays to study the performance and fairness of the algorithms underlying eRevise. They used teacher surveys and writing tasks teachers assigned to students to construct measures of writing instruction that enabled them to examine relationships between students' writing and revision skills in the eRevise system and students' opportunities to practice such skills. Finally, the team surveyed students and interviewed teachers to understand their experiences with the system.

Accordingly, the researchers conducted various analytical procedures. These included calculating correlations to study relationships between students' evidence use score in eRevise and measures of student achievement (Correnti et al., 2020), as well as univariate and multivariate analyses using hierarchical linear models to examine the relationship between evidence-use scores and measures of classroom instructional quality (Correnti et al., 2020). To understand teachers' experiences with eRevise, the team qualitatively analyzed interview responses (Correnti et al., 2022). The evaluation of the extent to which eRevise improved student essays from first to second draft was observational; there was no comparison condition in which classes did not use eRevise or used another tool to revise their

FIGURE 2

Screenshot of eRevise Student Interface with Example Feedback Messages



SOURCE: Zhang et al., 2019.

writing. This limits the claims that can be made about the effectiveness of eRevise.

Findings: Toward Addressing the Perils of Automated Writing Scoring and Feedback Systems

Assess Students' Writing on Specific Features That Indicate High-Quality Text-Based Analytic Writing

To address the critique that AES and AWE systems tend to focus on surface-level aspects of writing, the researchers developed eRevise to focus on evidence use, an important aspect of analytic text-based writing. Furthermore, rather than relying on a general or holistic score for evidence use, thus leaving the construct of evidence use opaque, eRevise uses a rubric-based system to ensure that the features of "good text-evidence use" (e.g., number of pieces of evidence provided, specificity of evidence) are well represented by the scoring algorithm. eRevise uses these feature scores to select the feedback messages that students see. This helps to address the issue that most AWE systems do not provide feedback that reflects students' needs and the notion of strong evidence use and therefore can support students' revision process.

Zhang et al., 2019, provides evidence suggesting that use of eRevise improved students' use of evidence at the feature level and that focusing on the level of features (i.e., using feature scores rather than overall score to select messages and assess students' revision) is desirable because it yields more information than when considering evidence use generally. First, Zhang et al., 2019, showed that, on a 1–4 rating scale, the mean score for overall evidence use improved from 2.62 to 2.72, not statistically significant. Notably, about 20 percent of students already had the maximum score of 4 on their first draft, and for the majority of students, the score did not change. In contrast, scores for the two features used to detect weaknesses in students' evidence use and to provide feedback—number of pieces of evidence used and specificity of the evidence provided—improved significantly from the first to second draft. And improvement (on specificity of evidence) was observable even for the students who already had the maximum overall score of 4 on the first draft (Zhang et al., 2019). The findings suggest the feasibility of assessing a substantive dimension of writing in a fine-grained way that provides information to support students' writing development.

Collect Evidence That the System Assesses the Intended Writing Constructs and Helps Students Improve on Them

In contrast with most AES and AWE systems, which rarely assess performance in terms of whether the system adequately measures the targeted aspect of writing, the researchers gathered evidence about whether eRevise was assessing the evidence-use construct, as claimed. They examined associations between system-generated scores on students' use of text evidence and expected relevant measures of student achievement. Correnti et al., 2020, found a moderate correlation, both at the individual student level and the classroom level, between the eRevise evidence-use score and reading and (to a lesser extent) math achievement scores. This means that students (and classes) with higher scores on the state standardized test of reading (and math) tended to also score higher on the essay, as evaluated by eRevise. (The same associations held for human scores of the essays submitted through eRevise.) This aligns with expectations and helps support the idea that eRevise is measuring the skill that it is designed to measure. Moreover, a reasonable expectation is that students who have had more opportunities to engage in analytic text-based writing (e.g., those are in classes where more such assignments are given and where instructional time is devoted to working on such assignments) would have a higher evidence-use score. Results confirmed this hypothesis, both when students' evidence use was rated by the AES system and when it was rated by a human. Meanwhile, evidence-use scores in written essays were not well correlated with a measure of general reading instruction, which also conforms to expectations. This finding adds support to the idea that eRevise is measuring a distinct writing construct.

A second investigation related to construct validity focused on whether the system is helping students improve as intended. This investigation goes beyond looking only at differences in scores from the first draft to the second draft. Scores could go up simply because students had more time to work on their essays. Or scores could change because the students made revisions that were not aligned with the feedback messages that the system provided. Neither of these scenarios provides evidence that the AWE system works as designed. The researchers examined whether improvements in student essays aligned with the evidence-use features targeted in the feedback messages that eRevise provided (Correnti et al., 2022). Essentially, the second-draft feature scores improved in expected ways. For example, students who received the message to add more pieces of evidence in fact did so—their score for the "number of pieces of evidence" feature increased. Another analysis involved examining what students reported as the one thing they learned from eRevise about using evidence in their writing. In general, students' responses to this open-ended question more closely resembled the feedback messages they received than the messages they did not receive. In other words, students who were asked to provide more explanation were more inclined to say that they learned "to explain how my evidence ties in with my argument" than to say they learned that they need "a lot of evidence."

Design Systems for Integration into Teachers' Daily Work

Teachers participating in studies of eRevise reported that the system was feasible to implement and beneficial insofar as it saved them time, facilitating timely feedback to students and providing students the opportunity to engage in the writing process. Teachers also indicated that the system messages were aligned with their instruction on use of text evidence. However, two-thirds of teachers suggested that the system should be seen as reinforcing the teacher's role, not replacing it in the classroom. In fact, several conveyed that, to get the most out of the system, the teacher should interact with it and the students' writing (Correnti et al., 2022).

Researchers analyzed teacher reports of how they interacted with students during the use of eRevise (Correnti et al., 2022). More than a third of the teachers did not interact with students; they treated the writing task and use of the system as if they were practice for the standardized test. Other teachers responded to student questions (e.g., "Is this enough evidence?") by referring students back to the task and the system. A final group of teachers interacted substantively with student questions. They appeared to use eRevise as a teaching and learning opportunity. For example, teachers elaborated upon or reinforced the system's feedback message. Analysis showed that the extent to which teachers interacted with students during use of eRevise was associated with the extent to which students' evidence-use scores improved from the first draft to the second draft. In short, in classrooms where teachers provided substantive help to any student, the writing improvement score for the class was higher than in classrooms

where teachers did not provide any help or only referred students back to the system. Moreover, the students who asked for substantive assistance and received it saw higher improvement scores. Thus, students in classrooms where teachers provided substantive support benefitted overall, but students who asked specific questions and received substantive support benefitted the most.

Attend to Algorithm Fairness

There are currently three prominent ways to build the underlying algorithms for automatically scoring student essays—feature-based models, neural network models, and a hybrid of the two. Developers typically choose a model based on the extent to which they prioritize prediction accuracy versus the ability to explain what feature(s) or construct(s) the model is assessing. There has been less focus on the algorithmic fairness of these different models. The researchers compared the fairness of these models, examining potential bias based on gender (male versus female), race (Black versus not[1]), and socioeconomic status (as indicated by eligibility for free or reduced-price lunch) (Litman et al., 2021). The research team operationalized fairness in three ways, including overall score accuracy, which measures whether AES scores are equally accurate for all student subgroups. About 48 percent of the essays were written by male students, 68 percent by Black students, and 55 percent by students from low-income families.

Results indicate that different models for scoring student writing expose different types of biases. Overall, based on the analysis, the hybrid model seems to be the fairest. It appears to exhibit bias only with respect to overall score accuracy for gender. Regardless of model and measure, the eRevise algorithm seems most biased with respect to gender. Males tended to produce essays with significantly lower word count. The team explored removing word count as a feature and found that this improved gender fairness but reduced score accuracy.

[1] The research team focused on the contrast between Black students and other students because the Black subgroup was the only race/ethnicity subgroup with sufficient data. Moreover, prior evaluation of another AES system suggested that African American students (particularly males) tend to receive slightly lower scores than from human raters, while scores of students of other races (i.e., White, Hispanic, Asian, American Indian, Other) did not differ (Bridgeman, 2013). Prior examinations of bias among human raters have also shown race-based bias against African American writers (e.g., Johnson and VanBrackle, 2012; Malouff and Thorsteinsson, 2016).

The Future of AES/AWE Systems

The work undertaken by the research team provides an argument for advancing the assessment of substantive dimensions of writing by designing AES and AWE systems that attend to features of writing that lead to actionable feedback to guide students' writing development, consider teachers' role in interacting with the student and the system, and are fair for all student subgroups.

Moving forward, more research is needed to build systems that can assess a wider range of important writing constructs—for example, quality of claims. AES and AWE systems must assess such dimensions to be useful to teachers and students (and researchers of writing) for monitoring students' writing skills toward college and career readiness (Correnti et al., 2020).

With respect to assessing the technical quality of AES and AWE systems, the research undertaken has pushed for more rigorous evaluations of an AWE system's performance. Specifically, developers and researchers need to move beyond human-computer reliability as the primary metric and toward ensuring that the system is assessing important writing features and helping students improve on them (Correnti et al., 2020).

It will take innovative thinking to design systems that deliberately invite and facilitate teacher interactions with students through feedback messages, error corrections, scores, or annotations. Beyond system-provided outputs for teachers (e.g., assessments of student progress) or teacher-accessible dashboards, which are primarily passive supports, developers can consider including discussion protocols that help teachers engage students in meaningful co-examination of system outputs. Teachers could, for example, elicit student understanding of the feedback messages and their revision plan. Developers should also consider ways in which teacher expertise and machine efficiency can intersect (Matsumura et al., forthcoming). For example, teachers may wish to provide additional feedback messages or customize messages given what they know about the student and their learning trajectory.

Finally, developers of AES/AWE systems should attend to algorithmic fairness. Although any bias is undesirable, researchers and developers need to consider the trade-offs between fairness and other important dimensions, such as reliability and explainability. The purpose of the AES or AWE system may

matter; if AES is used for formative purposes, such as to provide feedback to improve teaching and learning, then transparency in how the score or feedback is derived is important. And any approaches to mitigating bias should also consider construct validity—whether removing or changing a feature risks underrepresenting the writing skill that the system is designed to assess and improve (Litman et al., 2021).

References and Related Work

Afrin, Tazin, Elaine Wang, Diane Litman, Lindsay C. Matsumura, and Richard Correnti, "Annotation and Classification of Evidence and Reasoning Revision in Argument Essay Writing," *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, July 10, 2020, pp. 75–84.

Bridgeman, B., "13 Human Ratings and Automated Essay Evaluation," in Mark D. Shermis and Jill Burstein, eds., *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, 1st ed., Routledge, 2013, p. 221.

Correnti, Richard, Lindsay C. Matsumura, Laura S. Hamilton, and Elaine Wang, "Combining Multiple Measures of Students' Opportunities to Develop Analytic, Text-Based Writing Skills," *Educational Assessment*, Vol. 17, No. 2/3, 2012, pp. 132–161.

Correnti, Richard, Lindsay Clare Matsumura, Laura Hamilton, and Elaine Wang, "Assessing Students' Skills at Writing Analytically in Response to Texts," *Elementary School Journal*, Vol. 114, No. 2, December 2013, pp. 142–177.

Johnson, David M., and Lewis Napoleon VanBrackle, "Linguistic Discrimination in Writing Assessment: How Raters React to African American 'Errors,' ESL Errors, and Standard English Errors on a State-Mandated Writing Exam," *Assessing Writing*, Vol. 17, No. 1, 2012, pp. 35–54.

Malouff, John M., and Einar B. Thorsteinsson, "Bias in Grading: A Meta-Analysis of Experimental Research Findings," *Australian Journal of Education*, Vol. 60, No. 3, 2016, pp. 245–256.

Matsumura, Lindsay Clare, Elaine L. Wang, Richard Correnti, and Diane Litman, "Designing Automated Writing Evaluation Systems for Ambitious Instruction and Classroom Integration," in A. Alvi, ed., *Artificial Intelligence in STEM Education: The Paradigmatic Shifts in Research, Education, and Technology*, Milton Park, Abingdon-on-Thames, Oxfordshire, United Kingdom: Taylor and Francis, forthcoming.

National Governors Association Center for Best Practices, Council of Chief State School Officers, *Common Core State Standards: English Language Arts*, Washington, D.C., 2010.

Rahimi, Zahra, Diane J. Litman, Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, and Zahid Kisa, "Automatic Scoring of an Analytical Response-to-Text Assessment," in Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, and Kitty Panourgia, eds., *Intelligent Tutoring Systems, ITS 2014, Lecture Notes in Computer Science*, Vol. 8474, Berlin: Springer Nature, 2014, pp. 601–610.

Wang, Elaine Lin, Lindsay Clare Matsumura, Richard Correnti, Diane Litman, Haoran Zhang, Emily Howe, Ahmed Magooda, and Rafael Quintana, "*eRevis(ing)*: Students' Revision of Text Evidence Use in an Automated Writing Evaluation System," *Assessing Writing*, Vol. 44, April 2020.

Wang, Elaine, Lindsay Clare Matsumura and Richard Correnti, "Student Writing Accepted as High-Quality Responses to Analytic Text-Based Writing Tasks," *Elementary School Journal*, Vol. 118, No. 3, 2018, pp. 357–383.

This Brief Is Based on the Following Articles:

Correnti, Richard, Lindsay Clare Matsumura, Elaine Lin Wang, Diane Litman, Zahra Rahimi, and Zahid Kisa, "Automated Scoring of Students' Use of Text Evidence in Writing," *Reading Research Quarterly*, Vol. 55, No. 3, July/August/September 2020, pp. 493–520 (EP-68092, www.rand.org/t/EP68092).

Correnti, Richard, Lindsay Clare Matsumura, Elaine Lin Wang, Diane Litman, and Haoran Zhang, "Building a Validity Argument for an Automated Writing Evaluation System (eRevise) as a Formative Assessment," *Computers & Education Open*, Vol. 3, December 2022 (EP-68913, www.rand.org/t/EP68913).

Litman, Diane, Haoran Zhang, Richard Correnti, Lindsay Clare Matsumura, and Elaine Lin Wang, "A Fairness Evaluation of Automated Methods for Scoring Text Evidence Usage in Writing," in Ido Roll, Danielle McNamara, Sergey Sosnovsky, Rose Luckin, and Vanya Dimitrova, eds., *Artificial Intelligence in Education, Proceedings of the 22nd International Conference on Artificial Intelligence in Education, 2021, Lecture Notes in Computer Science*, Vol. 12748, Berlin: Springer Nature, 2021, pp. 255–267 (EP-68809, www.rand.org/t/EP68809).

Zhang, Haoran, Ahmed Magooda, Diane Litman, Richard Correnti, Elaine Lin Wang, Lindsay Clare Matsumura, Emily Howe, and Rafael Quintana, "eRevise: Using Natural Language Processing to Provide Formative Feedback on Text Evidence Usage in Student Writing," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, 2019, pp. 9619–9625 (EP-68093, www.rand.org/t/EP68093).

To view this brief online, visit www.rand.org/t/RBA1062-1. The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors. **RAND**® is a registered trademark.

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through grant R305A160245 to the University of Pittsburgh, with subcontract to RAND. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.