Computers and Education Open 3 (2022) 100084

Contents lists available at ScienceDirect

Computers and Education Open

journal homepage: www.sciencedirect.com/journal/computers-and-education-open

Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment

Richard Correnti [a,*], Lindsay Clare Matsumura [a], Elaine Lin Wang [b], Diane Litman [a], Haoran Zhang [a]

[a] University of Pittsburgh/Learning Research and Development Center, 3420 Forbes Ave, Pittsburgh, PA 15260, United States
[b] RAND Corporation, 4570 Fifth Avenue #600, Pittsburgh, PA 15213, United States

1. Introduction

Beginning in the upper elementary grades, text-based argument writing has been increasingly emphasized in U.S. learning standards as critical to college readiness [62,63]. Results of national assessments in the United States consistently show that the very large majority of students do not have proficient writing skills [61], and this is especially the case for text-based writing [45,47,81]. Young writers especially lack familiarity with the discursive features associated with argumentation, such as identifying evidence and explaining how it connects to the claim [66,80,85,86]. Indeed, marshaling effective text evidence in argument writing has proven difficult even for secondary [64], and post-secondary students [22].

Several explanations account for why so many students struggle with text-based argument writing. First, more generally, teachers often do not implement research-based practices for writing instruction that include providing substantive formative feedback on drafts of student essays [8,41,57,69]. Providing timely feedback on drafts of essays is difficult for busy teachers, many of whom are required to keep pace with curriculum guides which require them to address particular content on specified weeks [5]. The reluctance to assign tasks that require students to write across drafts also is reinforced by state accountability policies which, under pressure to ensure that their students meet testing requirements, can lead teachers to assign writing tasks that resemble the content and format of state tests [56,104].

Second, text-based argument writing instruction is rare. Even though recent studies show the payoff for undergraduate students' increased academic achievement in the sciences [71], there is little accumulated knowledge for teaching argumentation even at the college level [33,42]. Text-based argument writing instruction in the elementary grade curricula is also a relatively new addition because curricula have traditionally mostly centered on narrative writing. As a result, many teachers are under-prepared by their undergraduate programs to teach pertinent concepts related to effective argumentation, such as the importance of providing reasons (i.e., warrants) linking evidence to claims as suggested by the Toulmin model (see, e.g., [91]). Research shows that across the elementary and secondary grades, teachers rarely assign tasks that require analysis and use of text evidence [41,58,59]. Surveys reveal that a majority of middle school teachers assign argument writing tasks no more than one or two times per year [27]. In short, classroom supports for text-based argument writing instruction are clearly needed that make critical features of the construct explicit to teachers and students and increase students' opportunities to write and revise their essays in response to substantive feedback.

1.1. Automated writing evaluation systems

Automated writing evaluation (AWE) systems that employ automated essay scoring (AES) technologies to generate personalized feedback to students have been proposed as a way to improve students' classroom writing opportunities (see studies reviewed in [34,88]). AWE systems are intended to serve as formative assessments, broadly, that are intended to provide information that students can use to improve their writing and that teachers can use to increase the quality of their instruction. In other words, they are intended to be *learning tools* for students and teachers alike [82]. AWE systems also are intended to support teachers in their instruction by reducing the burden of grading and providing timely, substantive feedback on students' written responses. In doing so, AWE systems are expected to increase the frequency of students' opportunities to revise their essays in response to substantive feedback.

A persistent criticism of AWE systems, however, is that they have not been designed to meet ambitious writing standards, and this is notably the case with respect to text-based argument writing [9]. The AES technologies that undergird AWE systems have historically leveraged linguistic properties of student writing – for example, syntactic

* Corresponding author.
E-mail address: rcorrent@pitt.edu (R. Correnti).

https://doi.org/10.1016/j.caeo.2022.100084
Received 17 November 2020; Received in revised form 21 February 2022; Accepted 29 March 2022
Available online 1 April 2022

complexity, cohesion, vocabulary, and length – rather than the content of student writing [9,18,65,89]. Unsurprisingly, the positive effects of AWE systems on student writing have mostly been observed in accuracy or linguistic sophistication of responses (studies reviewed in Deeva et al., 2021; [49,50,52]; Ranalli et al., 2016).

Development of AES technologies keyed to source texts has mostly focused on evaluating the quality of students' summaries (see, e.g., [90]) or understanding of subject matter-content (e.g., a concept taught in a curricular unit). In recent years, scoring algorithms have become better at extracting substantive features of writing quality, such as organization, clarity, the presence or absence of argument elements, and subject matter content [25,54,55,67,90,95,97,99,102]. Developing algorithms that capture the content of students' responses (e.g., use of evidence or warrants) continues to be an ongoing area of investigation. Although most of the AES technologies for argument writing have been developed for prompts that are independent of source texts (e.g., [16,105]), we have found it possible to understand evidence-use as a construct when the writing prompt explicitly asks students to use evidence from a source text [106].

1.2. Present study

In the current study, we take up the challenge of automating the assessment of standards-aligned text-based argument writing by investigating the quality of an AWE system – termed *eRevise* – for improving young adolescent students' use of source text evidence in their argument essays. Specifically, we draw on sociocultural theory [48,93] and activity theory (e.g., [70]) to investigate the potential of our system to serve a formative assessment purpose. That is, the degree to which formative feedback delivered through *eRevise*; 1) increases students' understanding of, and use of evidence in their argument essays and 2) fosters teacher-student interactions focused on students' making meaning of the feedback in relation to their own work [1–3].

While the term formative assessment has been used broadly to distinguish between one-shot assessments for evaluative purposes and assessments used in the classroom during instruction, researchers studying formative assessments have implicitly or explicitly aligned themselves with different learning theories [4]. For example, teachers could provide multiple-choice questions during instruction to identify and fill gaps in students' knowledge (see, e.g., [23,24]), a practice that could be seen as aligned with behaviorist theories that characterize learning as the accumulation of small, discrete units of information or skills, often acquired through transmission models of teaching [28]. Immediate feedback may be one mechanism for how formative assessment can influence student performance, but researchers investigating sociocultural theories of formative assessment expand learning outcomes beyond performance to also include students' self-regulation [32,68,100] and identity development [17,70]. These outcomes are theorized to result from dialogic interpersonal interactions. This latter view, built on the ideas of sociocultural learning theorists and focused on dialogic interactions around feedback (e.g., [92]), is where we perceive our activity system for *eRevise* belonging, as student-teacher interactions involve complex judgments about whether and how to implement automated feedback centered on a central tenet of argumentation – evidence use (see, e.g., [15]). In Section 2.2 we describe the theoretical framework that underpins the claims, warrants and sources of evidence for our validity investigation.

2. A validity argument for the use of *eRevise*

The guiding doctrine of a validity argument is how well evidence supports a claim (e.g., [38]). Evidence aligned with the claim provides warrant for a valid inference. While validity arguments have typically been applied to summative assessments, recent work has begun to extend the logic chain to formative assessments [31,35,37,73]. Because this typically involves additional steps for inferences about the proposed interpretation and use of scores from the assessment, the interpretation/use argument is especially important for laying bare theoretical assertions and researcher assumptions so the validity of the evidence can be evaluated in relation to the claim [10]. In essence, the interpretation/use argument states the claim, while the validity argument provides evidence to evaluate the plausibility of the claim [40].

Despite the recognized need to evaluate assessments relative to their intended uses [39], the evaluation of AWE systems has mostly centered around the accuracy of scores (Dikli, 2006 as cited in Chappelle, Cotos & Lee, 2015). Accuracy of AES, for example, the relationship of human-machine ratings, is an important part of construct validity, especially in the case of summative evaluations where scores have consequences for users. For AWE systems, however, a validity investigation needs to consider not just accuracy, but how the system is interpreted and used by participants toward a learning purpose. It is to this end that we focus our work.

2.1. Using activity theory to frame our investigation

To frame our validity argument and attendant investigation, we draw on Pryor and Crossouard's [70] visualization of an activity system for sociocultural theorization of formative assessment. The foundations for activity theory (also referred to as cultural-historical activity theory) are based in the work of Vygotsky and his followers who emphasized the situated and social nature of learning [94]. From this perspective, mental functions occur first as social interactions among and between people (i.e., within communities). These interactions, in turn, are shaped by cultural norms, traditions and institutions and mediated by tools and artifacts (i.e., objects) in the environment (see, e.g., [48,79]). Thus, the social context for learning can be decomposed to include an understanding of structural features of an activity such as the subjects present (e.g., a teacher with a group of students), objects (e.g., a text, curricula or standardized assessment), as well as an understanding of the goals for an activity as perceived by subjects which can influence their interactions (e.g., with learners)[1].

Fig. 1 depicts elements of our formative assessment activity system as applied to *eRevise*. The bottom of the triangle is the *disciplinary norms for text-based argumentation* – the use of evidence and warrants to support claims [91]. On the left side of the triangle is *subjects* – including the goals and values held by teachers for implementing the system - who shape the behavioral scripts (*mediating process*, top of the triangle) teachers employ in their interactions with students around the individualized automated feedback messages (*objects,* right side of the triangle). All of these elements of the activity system around *eRevise* are expected to shape student *outcomes* – their understanding of feedback messages and their application of feedback messages as they revise.

Rooted in this theoretical framework, Table 1 makes explicit our interpretation/use argument (warrants, assumptions, research questions and evidence sources) for the use of *eRevise* as a formative assessment in a classroom activity system.

Disciplinary norms for argumentation (bottom of Fig. 1). In order to develop complex knowledge acquisition and skills for text-based

[1] A good example of how a subject's perceptions of an activity can shape their interactions (with learners) in an activity system is captured in Wertsch's [96] study comparing Brazilian mothers' interactions with children around a puzzle activity with teachers' interactions with children around the same activity. Teachers, perceiving the purpose of the puzzle to be a learning opportunity, encouraged students to complete the task as independently as possible, providing hints only as necessary for completion. Mothers, in contrast, saw the puzzle as a task to be completed and so worked together with their children to finish it. The teachers' and mothers' distinct goals for the activity thus shaped their interactions with the children [26]. In our study, we investigate interactions around the object (automated feedback) as we attempt to understand how subjects interact with the feedback *and with each other* to engage in meaning making about evidence use.

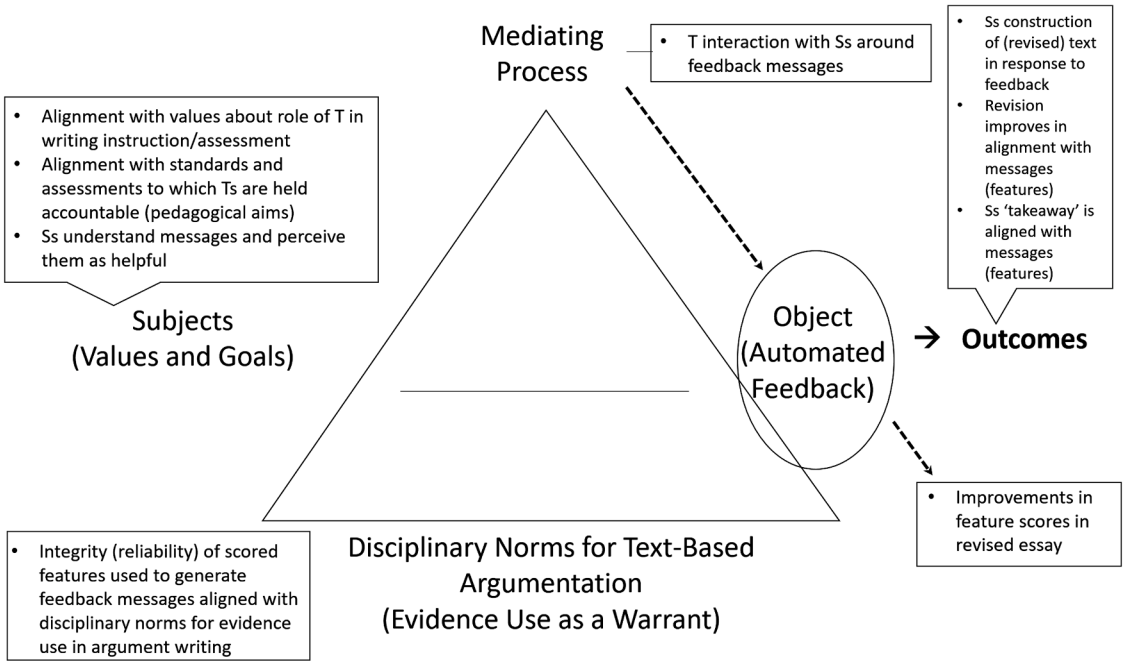

Fig. 1. Components of an Interpretation/Use Argument of a Formative Assessment Activity System with a Focus on Developing Students' Evidence Use Text-Based Argumentation.
Note: T = teacher; Ss = students; 'takeaway' refers to the student response to the question, "What is one thing you learned about using evidence in your writing that you could use again?".

argument writing, we argue that AWE systems should provide information about the content of student writing and reveal strengths and weaknesses of students' abilities linked to disciplinary norms for argumentation. Formative assessments are expected to assist learning by making salient the 'gap' between performance on a task and 'next step' for improvement, and provide scaffolding (e.g., hints or suggestions) for improvement [83,84,93]. Thus, measures of students' performances on evidence use are paramount for understanding where students are and what the imagined 'next step' might be. We begin our validity investigation, therefore, examining the reliability of our automated scoring for the key features for evidence use (see Table 1, RQ1). Next, we gauge the extent to which the automated feature scores facilitated our ability to make meaning out of students' improvements in their essays (see Table 1, RQ2), especially relative to other ways of measuring improvement (i.e., change in rubric scores).

Subjects (top left of Fig. 1). To serve a formative assessment purpose, we argue that teachers must perceive an AWE system to be an authentic learning opportunity for students that is aligned with their pedagogical goals. If teachers see systems as undermining their instructional routines and goals (e.g., the learning standards to which they are held accountable), they are unlikely to implement the system in a way that supports a learning purpose (see Table 1, RQ3 and RQ4) or use the system at all. This is a concern because integration of AWE systems in instruction is critical to their success [30,87].

Students also must be able to make sense of the feedback messages they receive and perceive the information as legitimate. Absent an understanding of the criteria for successful revision (i.e. what the messages are asking them to do to improve their essays) students are unlikely to use the information they receive to successfully revise their argument essays or 'take away' information from the assessment to apply in future writing situations (see Table 1, RQ5). To address the former research questions we draw on teacher interviews, and to address the latter question we rely on student surveys. The interviews and surveys, respectively, describe the extent to which each role group understood the feedback and perceived the feedback as beneficial to students' writing.

Mediating process (top of Fig. 1). To serve a formative assessment purpose, we argue as well that teacher-student interactions should focus on students' making meaning of the feedback in relation to their own work (see e.g., [92]). Indeed, recent work on AWE systems has established that individualized support from teachers is necessary for writing improvement as students often need support to understand the automated feedback messages they receive [43,72,98]. The extent to which teachers provide individualized guidance (for example, use automated feedback messages as a starting point for discussions around writing or clarify and interpret feedback messages with students) significantly impacts students' uptake of feedback messages [11,30,51] and motivation to incorporate AWE feedback in their revision [76,97]. To address these questions, we drew on teacher 'implementation logs' in which teachers documented the questions students asked them and their responses to student queries as a measure of interactivity around student questions. Below we describe how we use this measure to explore the relationship between the mediating process and improvements during revision.

Object (automated feedback) and outcomes (right hand side of Fig. 1). As alluded to earlier, to serve as tools for learning, formative assessments, in this case in the form of automated feedback, must clearly communicate the criteria for successful task performance (e.g., [1]), and be tailored to students' learning needs. In the context of an AWE system such as *eRevise*, the messages must be appropriate to student essays. Otherwise, we might not see improvements in students' essays, nor would we expect improvements aligned with feedback. For our outcomes, as is common practice (see, e.g., [102]), we examined changes in student performance for evidence of student learning attributed to the

Table 1
Validity Argument Framework for *eRevise* as a Formative Assessment.

Warrant/ Inference	Assumption(s)	Research Question	Evidence Source
Aligned with disciplinary norms for argumentation	AWE system captures meaningful features of effective evidence use. Those features can be measured with accuracy. The features are sensitive to meaningful improvements in students' essays.	1) How reliable are the automated scores generated in *eRevise* at identifying features of effective evidence use aligned with disciplinary norms for argumentation (i.e., the number of different evidence focal topics students cited in their essay and the total number of unique and specific references to text-based evidence in students' essays)? 2) Are feature scores sensitive to meaningful improvements in evidence-use?	Feature scores generated in *eRevise* compared to human scores Comparison of feature and rubric scores in identifying improvements
Subjects (Values and Goals)	Teachers perceive *eRevise* as aligned with their pedagogical aims, and work. Students perceive the feedback as interpretable and beneficial.	3) Do teachers perceive *eRevise* as beneficial to their work (i.e., feasible to implement and helpful for their work)? 4) Do teachers perceive *eRevise* as aligned with the standards and assessments to which they are held accountable (aligned with their pedagogical aims)? 5) Do students understand the feedback messages and perceive them as beneficial to their writing?	Teacher interviews Student surveys
Mediating process	Teachers vary in providing active support to students to interpret feedback messages provided in *eRevise*	Independent variable measuring variation in the classroom implementation of *eRevise* as it naturally occurred during implementation [used in RQ9 below].	Teacher implementation logs
Object	Automated feedback based on feature scores of original essay is accurate and meaningful.	***Validity evidence supports our interpretation/use argument for eRevise as a formative assessment.***	Inferences from RQ1-9
Outcome	Essays improve in features of evidence use. Student essays improve in alignment with feedback messages they received. Students' articulated 'takeaway' or learning is aligned with feedback messages.	6) Do students' essays improve in evidence use? 7) Is improvement in student essays aligned with the features targeted in the feedback message they received? 8) What do students believe they learned from using *eRevise*?	Change in rubric/feature scores in student essays from first to final draft Student surveys
Mediating process –> Outcome	Substantive (potentially dialogic) teacher-student interactions support student revision (and retention of concepts).	9) Is there a relationship between substantive teacher interactions and student improvement on feature scores?	HLM model examining relationship between teacher-student interactions and revision improvements

Note: Bolded cell is stated as a claim because it represents the generalized inference (from our interpretation/use argument) we'd like to make from the cumulative evidence gathered in response to research questions 1 through 9.

automated feedback, in general (Table 1, RQ6). We also used the students' improvements in feature scores (revised minus original) to investigate student responsiveness to feedback, specifically, whether the improvements we observed *were aligned with* researcher hypotheses of what we would expect given the feedback messages students were provided - i.e., given their original feature scores and the assumed strengths and weaknesses of evidence-use on their original essay (see Table 1, RQ7).

Given our expressed desire to explore socio-culture theories of student learning and the focus of our activity system on dialogic teacher-student interactions we were interested in outcomes such as shifts in students' understanding of performance expectations because such understanding is likely to contribute to students' self-regulation going forward [32,68,100]. Therefore, we examined how students' experience with our automated formative assessment system influenced their understanding by asking them to articulate what they learned from the formative assessment experience that they will use again in their future (argument-based) writing (Table 1, RQ8). We then inferred from the student responses whether students obtained any generalized understanding(s) from their experience with *eRevise*.

Mediating process influence on outcome (dotted arrows in Fig. 1). Finally, we conducted an empirical test for our hypothesized mediating process. The dotted arrows in Fig. 1 signify our final research question (see Table 1, RQ9) about the relationship between the mediating process (teacher-student interactions) around the object (the automated feedback) and its influence on the outcome (improvements in feature scores). Although our scores constitute a coarse proxy for dialogic interactions, we see this as a nascent empirical test to provide evidence for socio-cultural theorizations (e.g., [1,12,17,70]) of formative feedback on student revision quality.

3. Methods

3.1. Context and participants

Our validity investigation took place in 8 public parishes (i.e., districts) in Louisiana that are representative of the state demographics. As of the 2018–2019 school year, across these parishes, 47% of the students identified as White, 42% African American, 4% Latinx, 2% Asian, and 5% other. About 70% of the students were eligible for free-or reduced-price lunch.

Teachers. 16 English language arts teachers participated in the study. They were selected for their comfort with basic technology and access to a class set of computers to complete the online assessment in *eRevise*. All 16 teachers were white females with at least a Bachelors degree. They averaged 10 years (range = 4–18) of teaching experience. Seven teachers taught fifth grade; eight taught sixth grade; and one indicated she taught both fifth and sixth grade.

Students. The 16 teachers implemented *eRevise* to all students in one of their English language arts classes. The classes averaged 16.6 students (range = 10 to 34). In the end, 266 fifth and sixth grade students completed all data collection (i.e., submitted both a first draft and a revised draft of the essay and completed the post-*eRevise* survey items).

3.2. eRevise and its automated feedback messages

Our AWE system, *eRevise*, was designed to score responses and provide feedback to students on the Response-to-Text Assessment (RTA). Elsewhere, we have described RTA development, administration, and scoring [13–15]. In brief, the assessment used in this pilot is based on a feature article from *Time for Kids* ("A Brighter Future" by Hannah Sachs)

about the Millennium Villages Project, a United Nations-supported effort to eradicate poverty in a rural village in Kenya.[2] The prompt asks students, "Based on the article, did the author provide a convincing argument that 'winning the fight against poverty is achievable in our lifetime'? Explain why or why not with 3–4 examples from the text to support your answer." The RTA rubric for human raters focuses on five dimensions– evidence use, analysis, organization, academic style, and mechanics. Each is scored on a scale from "1=low" to "4=high".

eRevise focuses specifically on the dimension of evidence use. Elsewhere, we provided a validity argument for the automated scoring of this writing construct at the classroom level for research purposes [15]. Aligned with the rubric criteria for this dimension, the automated scoring model that underlies *eRevise* is based on the following four features:

(1) Number of pieces of evidence (NPE): To calculate the breadth of focal topics from the source text that the students used in their essay (NPE), project researchers first defined a list of main topics in the source text (i.e., the *Time for Kids* article) that were then incorporated into the AES system. These four topics correspond to the ways the Millennium Villages project affected the quality of life in a village (i.e., hospital conditions, access to schools, malaria, agriculture). The AES system uses a simple window-based algorithm with fixed window-size to calculate NPE. A window within the essay contains evidence related to a topic if it uses at least two keywords from the list of words for that topic.

(2) Specificity (SPC): For each main topic from the source text, researchers identified a comprehensive list of associated keywords (i.e., specific text evidence/examples). For example, the topic "hospital conditions" included as keywords "water," "electricity," "hospital beds", "medicine," and "doctors" (initially, these aspects were lacking or insufficient in the villages but then improved over time). For each student essay, the AES system used this keyword list to identify matches – i.e., how many (and which) specific pieces of evidence the essay addressed. The system included accounts for the similarity between a word in the student's essay and a word in the topic or keywords list, so students will be credited for evidence that uses slightly different words (e.g., "power" instead of "electricity") or words with different stems. Each phrase containing keyword matches is only counted once to avoid redundancy. To select feedback, we used a measure for unique and specific mentions of evidence associated with the four focal topics of hospital conditions, malaria, agricultural conditions and school (hereafter referred to as SPC_{focal}).

(3) Concentration (CON): High concentration signals listing of evidence without explanation or elaboration and, typically, receives a lower score. Concentration is a binary feature meant to capture a common instance with developing writers – answering a prompt by simply providing unelaborated evidence directly from the source text. To calculate this feature, the AES system counts the number of phrases that contain keyword matches and compares them to the total number of sentences. If there are several keyword matches but fewer than three sentences, the concentration is deemed high.

(4) Word count (WOC): This feature is a proxy for elaboration of thinking and for students using their own language to reason how the evidence supports their main idea versus just letting the evidence speak for itself.

eRevise uses the first two of these natural language processing features generated during automatic scoring of students' first- draft essays to select formative feedback messages on evidence use to guide essay revision. Three levels of feedback messages were available (for full messages see Table A1, Appendix A): Level 1 feedback messages focused on *completeness* (i.e., guided students to provide more evidence) and *specificity (*i.e.*,* guided students to provide more details about the evidence they referenced). Level 2 feedback messages also prompted students to be more *specific*, and, in addition, directed students to *explain* their evidence. Finally, level 3 feedback messages focused students on not only *explaining* the evidence they provided, but also *connecting* it to the overall argument. Elsewhere, we discuss the assumptions and methods used to channel students' essays to each of the three different levels of feedback based on the number of topics (NPE) referred to in their original essay and the number of unique and specific references to source-text evidence for the four focal topics (SPC_{focal}) (see [106] for technical details).

3.3. Procedures/Measures

Participating teachers implemented *eRevise* in late fall 2018. The *eRevise* system is designed for use over two class periods. Students wrote (i.e., typed) their essays on the first day. On a second day (no more than five school days later), students logged into *eRevise* to view the automatically generated formative feedback messages and revise their first drafts. Fig. 2 shows an example screenshot with the formative feedback that students would see on day 2. While *eRevise* generates an automated score in the background, commensurate with our conception of formative assessment, students do not receive the score; they receive only the feedback associated with the scored features.

Teachers were instructed to provide at least 30 min of independent work time on day 1 for students to draft their essay, and on day 2 for them to revise. Actual revision times varied within and across classes. According to *eRevise*'s built-in time log, the average revision time across classes was approximately 25 min (range = 13–57 min).[3] To further understand how *eRevise* was implemented, we asked teachers to keep a detailed record ('implementation log') of the questions students asked during the administration of the formative assessment task (both drafts) and their responses to student questions.

Students completed brief surveys after submitting their first draft and again after the revised draft. Questions on the survey (assessed on a 4-point scale) focused on students' experience with *eRevise*. For example, students were asked about the helpfulness of the feedback message they received, whether they understood what the feedback message was asking them to do, and the extent to which they believed their revised essay had improved from their first draft. Students also responded to an open-ended question, "What is one thing you learned about using evidence in your writing that you could use again?" to gauge the potential of the system to build students' understanding of effective use of text evidence.

[2] Elsewhere we have described key features of our text and administration in order to support measurement of students' analytic thinking and reasoning in response to text (see [14] for an extended discussion). Thus, we chose texts we felt were authentic, complex, and readable, but challenging for the grade level. We did several things to mitigate readability as a potential confound in our measurement strategy. First, we used a lexile analyzer to interrogate the grade-level appropriateness. Second, we define several vocabulary words in call-out boxes for ease of comprehension. Third, the assessments are brief enough that the teacher can read the assessment aloud with students. Fourth, the teacher asks clarifying questions – with standardized language and potential follow-up prompts - during the reading of the text in order to facilitate a literal understanding of the text, from which we expected the students to be able to provide an analytic response in writing.

[3] The elapsed time is a rough estimate of time spent revising. We cannot be certain that students began working as soon as they logged into *eRevise,* nor that they worked without interruption until the time they logged off.

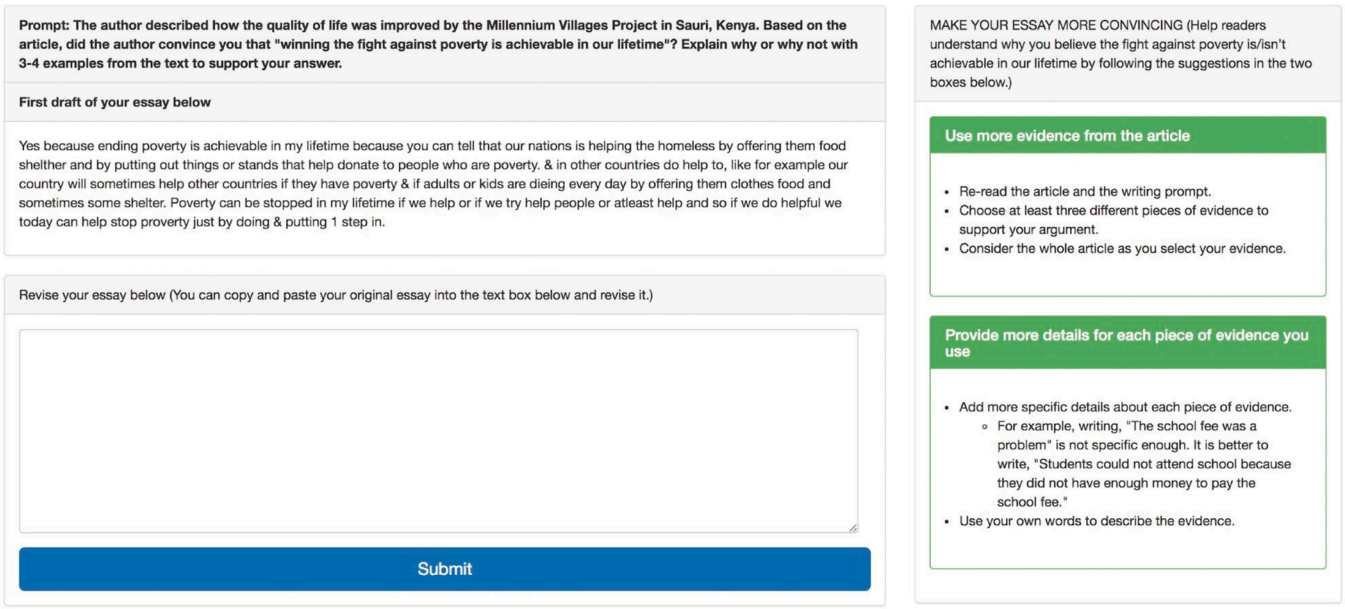

Fig. 2. Screenshot of eRevise essay with associated feedback for a student.
Note: Students had access to the source text if they scrolled down the page.

3.4. Data analyses

RQ1: *How reliable are the automated scores generated in eRevise at identifying features of effective evidence use (*i.e.*, the number of different evidence focal topics students cited in their essay and the total number of unique and specific references to text-based evidence in students' essays)?* To explore the reliability of our automated feature scores, we had a human rater score the two features the *eRevise* AWE system uses to select feedback (NPE and SPC_{focal}) for more than 20% of the students in the sample ($n = 63$). Specifically, the human rater scored both the first-draft essays and the revised-draft essays for these students to assess inter-rater (i.e., computer-human) agreement of the feature scores. The first feature (NPE) was the number of topics from the source text, out of a possible four, that the student marshaled evidence from and referenced in their essay. The second feature (SPC_{focal}) was the number of specific and unique text-based evidence the student referenced from those four focal topics. We examined the intra-class correlation of these continuous measures using SPSS V.26.0.

RQ2: *Are feature scores sensitive to meaningful improvements in evidence-use?* To understand the sensitivity of the feature scores for detecting improvement in student essays, we calculated the number of students who, by our feature metrics (NPE, SPC_{focal}, and WOC as described above), displayed any evidence of improvement. We compared this to the baseline of 41 percent of students who gained at least one point on the AES score based on the evidence-use rubric (see Appendix Table B1 for the human evidence-use rubric). We also examined t-tests for improvement scores for two subgroups of students – first, we examined all students, then we examined only those students whose rubric score did not improve.

RQ3,4: *Do teachers perceive eRevise as beneficial to their work (*i.e.*, feasible to implement, helpful for their work, and aligned with their pedagogical aims)?* To investigate the compatibility of *eRevise* with teachers' instructional context (i.e., consonance with the instructional system), we interviewed all 16 teachers by phone in spring 2019, after their classes had experienced *eRevise*. The 45-minute, semi-structured interview protocol addressed whether students had difficulty understanding and applying the feedback in their revisions, whether the feedback provided was sufficient and aligned with the teachers' pedagogical aims; the pros and the cons of the system; how use of *eRevise* might impact teachers' writing instruction; and how frequently teachers would employ the *eRevise* system in the future if it were available to them.

Interviews were audio-recorded with teachers' permission; subsequently, we generated detailed notes or transcripts for coding. One researcher engaged in multiple readings and performed iterative qualitative coding and analysis [60,101] on the transcripts using Dedoose [19]. Specifically, structural codes reflected the interview topics. Thematic coding emerged from data [60]. We identified themes following established techniques, including clustering, making contrasts, and seeking repeating patterns [6,7,78]. The researcher made transparent the coding scheme, definitions, and example coded excerpts for team discussions and to check for underlying analyst assumptions or biases [21]. Data analysis involved generating counts and percentages of teachers that expressed a given opinion or theme.

RQ5: *Do students understand the feedback messages and perceive them as beneficial to their writing (*i.e.*, what do students believe they learned from using eRevise)?* We examined student surveys to understand how students perceived the feedback provided in *eRevise* and what they self-reported learning from using eRevise that they would apply in future writing situations.

RQ6: *Do students' essays improve in evidence use?* To examine outcome metrics, we first analyzed the data using paired samples t-tests to understand improvement across all students from the initial draft to the revised draft. We examined the breadth (number) of different topics covered among the four focal topics in the source text (NPE). We also examined students' use of text-based evidence within each topic separately (the number of specific and unique uses of evidence identified for each focal topic) and then aggregated across the four focal topics. Higher means on certain topics reveal the evidence students were most likely to select from the text to support their argument – both initially on their first drafts and as they revised.

RQ7: *Is improvement in student essays aligned with the features targeted in the feedback message they received?* We generated analytic hypotheses based on the feedback messages as to which features we would likely 'see' improvement in[4] and conducted a series of between-group comparisons. For example, if students were responsive to the feedback asking them to provide more complete evidence (level 1 feedback), then we would expect practically and statistically significant increases in: specific mentions of evidence from the source text (SPC_{focal}) and in the number of topics mentioned from the source text (NPE). Additionally, we would expect an increase in the density of evidence on focal topics (i.e., the ratio of SPC_{focal} divided by word count) because students received explicit feedback on adding evidence (see Table 4, column labeled "feature-specific hypothesis" for our tests for alignment in relation to the feedback provided).

RQ8: *What do students believe they learned from using eRevise?* Next, we analyzed the similarity between what the student reported learning that they would use next time in their writing and the feedback message provided. We first used natural language processing to represent the meaning of every student text response, as well as the meaning of *eRevise*'s feedback messages (Table 4), in terms of a Term Frequency - Inverse Document Frequency (TF-IDF) vector representation. In this representation, words are modeled as vectors with each dimension in the vector corresponding to a word in the vocabulary and each cell value (co-occurrence count) weighted by multiplying term and inverse document frequencies [36]. We then computed the cosine[5] between the student and the feedback vectors to measure their text similarity. For each student, we compared the similarity between the vectors representing their response and the feedback messages they received, versus the vectors representing their response and the feedback messages they did *not* receive. We regard better alignment (i.e., higher cosine values / text similarity) of students' open-ended response with the feedback message they received as evidence that they had processed (and perhaps acted on and remembered) *eRevise's* feedback.

While this hypothesis test can help us understand statistical significance, it is abstract. Therefore, we further analyzed the data qualitatively for alignment. Specifically, blinded to the feedback messages students received, we coded for which, if any, of the feedback messages are reflected in the student response to the open-ended survey question. We attended to key words and ideas in each message. For example, student responses that used words such as "more evidence" or "different evidence" signaled alignment with feedback message 1. "Details" and "be specific" aligned with feedback message 2. "Explain evidence" and "why" were key words associated with message 3. Finally, "argument", "prove", "elaborate" aligned with message 4. We allowed for double coding (i.e., student responses could align with more than one message). Analysis involved counting the proportion of student responses coded to each message. For the responses that were codable,[6] we report the proportion of responses that aligned with the messages students were provided (i.e., that should have guided their revision) and the proportion of responses that aligned with messages they had not been provided.

RQ9: *Is there a relationship between substantive teacher interactions and student improvement on feature scores?* To investigate the extent to which teachers implemented *eRevise* as a formative assessment, we examined teachers' documentation of student questions and their responses to these questions during classroom implementation. Teachers varied in how they approached their role during assessment – from treating it like practice for a standardized test to facilitating students' understanding of the automated feedback and helping students construct plans for revision based on that understanding. We used this as our main independent variable at the teacher level. Our dependent variable was an 'improvement score' generated from a factor analysis of three composite items: the change in topic breadth (NPE), the change in amount of unique and specific evidence for the focal topics (SPC_{focal}), and the change in word count (WOC). Our main hypothesis was examined in a cross-level interaction between teachers' class-level reports of providing substantive help to students and students' reports of having asked the teacher for help, and the influence of the interaction on the change score. Thus, we examined a series of two-level hierarchical linear models [74] - 1) a fully unconditional model (FUM); 2) a model with group-mean centered student-level covariates where we also examined a random slope for the indicator variable where students said they asked their teacher for help (Model 1); and 3) our final random-intercept random-slope model where we added the cross-level interaction (Model 2 - see Appendix C for the full model description and rationale).

4. Results

4.1. How reliable are the automated scores generated in eRevise at identifying features of effective evidence use? (Aligned with disciplinary norms, RQ1 Table 1)

To examine the reliability between our human rater and the automated features, we examined the intra-class correlation (ICC). The ICCs show excellent agreement for NPE (ICC = 0.844) and moderate agreement for SPC_{focal} (ICC = 0.687) across sixty-three students' first and revised drafts (see Koo and Yi's (2016) guidelines for interpreting ICCs). Given that the focus of our analyses is to explore the utility of these features to represent 'improvement scores', we also examined the intra-class correlation for the human change score and the automated change score for each feature. The intra-class correlations show excellent agreement for both NPE_{chg} (ICC = 0.790) and $SPC_{focal\text{-}chg}$ (ICC = 0.812).

4.2. Are feature scores sensitive to meaningful improvements in evidence-use? (Aligned with disciplinary norms, RQ2 Table 1)

The paired-samples t-tests shown in Table 2 describe improvements in the revised essays across all students by features of evidence-use (n = 266). Table 2 provides evidence of significant improvement on all of the features examined, including NPE, SPC_{focal}, and word count (*ES* range

[4] These predictions were based on the fact that we used the *same features* for understanding improvement that were used to determine the feedback level. Using the same features allowed us to generate testable hypotheses to probe the alignment of feedback messages with improvement scores on those features. In future iterations of *eRevise*, we plan to design a second round of feedback about students' revisions based both on AES-calculated values for these same features, as well as additional features that measure and describe students' revision(s).

[5] The cosine similarity metric is based on the dot-product (a linear algebra operator) of the two vectors, but is modified to normalize for the vector lengths. The normalized dot-product is in fact the same as the cosine of the angle between the two vectors, hence the metric's name.

[6] Some student responses could not be coded for alignment to *eRevise's* feedback messages. We generated the following codes to characterize these responses: Students offered no response, leaving the question blank; Students wrote, "Nothing" or "I don't know"; Student responses pertained to source text rather than use of evidence (e.g., "A lot of people have poverty"); Student responses pertained generally to writing (e.g., "Always reread what you are writing so that it can make sense"); Student responses pertained to grammar or mechanics; Student responses were unclear, for example, responses used "it" without a clear antecedent (e.g., "It helps you write better").

Table 2
Paired-samples t-tests examining change in evidence use from first- to revised- draft.

Feature	Outcome	First Draft M_{pre} (*SD*)	Revised Draft M_{post} (*SD*)	Revised-First Draft M_{diff}	*t*	ES
Count of Focal Topics	**Breadth of Text Evidence (NPE)**	**2.474 (1.183)**	**2.959 (0.997)**	**.49**	**8.10**	**.44**
Count of Unique and Specific Evidence-Use	Malaria-Related Text Evidence (SPC_{mal})	2.316 (1.779)	2.767 (1.827)	.45	5.87	.25
	Hospital-Related Text Evidence (SPC_{hosp})	1.985 (1.622)	2.519 (1.544)	.53	7.80	.34
	School-Related Text Evidence (SPC_{schl})	1.876 (1.729)	2.553 (1.712)	.68	8.47	.39
	Agriculture-Related Text Evidence (SPC_{Agr})	1.083 (1.36)	1.357 (1.378)	.27	4.88	.20
	Cumulative Text Evidence for Focal Topics ($SPC_{focal} = SPC_{mal} + SPC_{hosp} + SPC_{schl} + SPC_{Agr}$)	**7.26 (4.59)**	**9.20 (4.72)**	**1.94**	**10.39**	**.42**
Word Count	Word Count (WOC)	189.823 (106.551)	260.914 (141.996)	71.09	17.13	.57

Note: M_{pre} = sample mean for first draft; M_{post} = sample mean for revised draft; M_{diff} = sample mean change from first to revised draft.
Bolded items = features of evidence use that were used in data-driven approach to channel student essays to context-sensitive feedback messages.

Table 3
Student survey responses within the *eRevise* system.

Question	*M* (*SD*)	Not at all	A little bit	Mostly	Completely
Did you understand the feedback you received?	2.99 (0.90)	12 (5%)	59 (25%)	80 (35%)	80 (35%)
Did you understand how you were supposed to revise your essay based on the feedback you received?	3.07 (0.89)	14 (6%)	41 (18%)	91 (39%)	85 (37%)
		None	A little bit	A Lot	All
How much of the feedback did you use when you revised your essay?	2.86 (0.81)	8 (4%)	70 (30%)	100 (43%)	53 (23%)
How much did you refer back to the original text as you were revising your essay?	2.68 (0.97)	24 (10%)	83 (36%)	64 (28%)	60 (26%)
		None	A little bit	A Lot	
How much better do you think your revised essay is compared to your first draft?	2.43 (0.60)	14 (6%)	104 (45%)	113 (49%)	
The feedback I received was different from what I normally receive	2.13 (0.68)	39 (17%)	119 (52%)	73 (31%)	

Note: Row totals are for 231 students because 35 students were missing on the survey responses within *eRevise*.
despite turning in both drafts of their writing.

from 0.20 - 0.57). For example, the mean number of topics addressed (NPE) per essay increased by nearly one-half, meaning about one out of every two students added evidence on a *new* topic.

Additionally, the feature scores used to measure improvement were more sensitive to revisions in student writing than the 4-point evidence-use rubric. Only 110 students (41%) would have been identified as having improved by 1 or more points on the 4-point evidence-use rubric. Table D1 in Appendix D provides paired-samples t-tests for the remaining 156 students who, using the rubric score, would *not* have been identified as having improved their essay. Table D1 shows that even this group made significant additions to their revised essays in all but two rows of the table. Thus, students improved their essays in ways detectable by change in the feature scores, even when the automated rubric score failed to identify improvement.

4.3. Do teachers see eRevise as beneficial to their work (i.e., feasible to implement and helpful for their work)? (Subject [teacher] values, RQ3 Table 1)

Teachers overall responded very positively to *eRevise*, with nearly all reporting that they would use *eRevise* again. Eighty-seven percent of teachers said they would use it 4–6 times per year or more, while the remainder said they would not use it that often or that their use would depend on the availability of technology. The two most frequently cited benefits teachers mentioned were: 1) the time saved and the ability for students to receive timely feedback (100%); and 2) the opportunity for students to engage in the writing process and receive feedback they may not have otherwise benefitted from (75%).

4.4. Do teachers see eRevise as beneficial to their work (i.e., aligned with their pedagogical aims)? (Subject [teacher] pedagogical aims, RQ4 Table 1)

Teachers reported that the feedback messages were aligned with their instructional goals (72%) and that the system reinforced the feedback they provided to students in their instruction (31%). However, most teachers (66%) also suggested the teacher and system were mutually reinforcing. In other words, most recognized the system would not replace the role of the teacher because to get the most out of the system, the teacher would still need to interact with it and the students' writing (i.e., they saw their interactions with students as a potential *mediating factor* within the activity system). Also, 63% percent of teachers mentioned that at least a few students in their class had difficulty with our feedback. For example, the system asked students to explain their evidence more when the student thought they already had. Teachers suggested the system identify places in the original essay that require revision and/or provide a few specific instances where an explanation would improve the essay.

4.5. Do students understand the feedback messages and perceive them as beneficial to their writing? (Subject [student] perceptions of automated feedback, RQ5 Table 1)

Table 3 shows that students were mostly positive about the feedback they received. Roughly seven out of ten students reported "mostly" or "completely" understanding the feedback, and a similar number reported understanding how they were supposed to revise their essay. Most students felt their revised essay was an improvement from their

Table 4
Changes to Feature Scores Aligned with Feedback Provided to the Student.

Level	# St.	Feedback Heading±	Feature-Specific Hypothesis	Hypothesis Confirmed	Change in feature score 2nd draft – 1st Draft†	Effect Size‡/Change in Ratio (SPC_{focal}:WOC)∓
1	80	*Use more evidence* from the article	1. NPE expected Large Increase	Yes	1.24***	*ES* = 1.27
		Provide more details for each piece of evidence you use	2. SPC_{focal} expected Large Increase	Yes	2.62***	*ES* = 0.82
			3. SPC_{focal} /WOC expected Increase	Yes	0.011**	1:43 to 1:33
2	76	*Provide more details* for each piece of evidence you use	4. NPE expected Moderate Increase	Yes	0.54**	*ES* = 0.78
			5. SPC_{focal} expected Large Increase	Yes	1.95***	*ES* = 1.20
		Explain the evidence	6. SPC_{focal} /WOC expected No Change	Yes	−0.002	*n.s.*; 1:33
3	110	*Explain* the evidence	7. NPE expected No Change	Yes	0.04	*n.s.*
		Explain how the evidence connects to the main idea & elaborate	8. SPC_{focal} expected No Change	No	1.83***	*ES* = 0.30
			9. SPC_{focal} /WOC expected Decrease	Yes	−0.015***	1:25 to 1:29

± See Appendix Table A1 for full feedback messages.
† The mean change in feature score for each row is presented along with results from a paired-samples *t*-test for significance.
‡ Effect size is calculated using the mean change in feature score from 1st to 2nd draft and dividing by the standard deviation.
∓ The ratio at time 1 is provided 1st, while the ratio at time 2 is provided 2nd for hypothesis tests 3, 6 and 9; because the 1st term of the ratio (antecedent) is 1 in both cases, an increase in evidence provided per word is associated with a decrease in the 2nd term (consequent) and vice-versa.
NPE = Number of focal topics student provided evidence for (0–4); SPC_{focal} is the cumulative number of pieces of unique and specific evidence provided on the focal topics; SPC_{focal} /WOC is the ratio of number of pieces of evidence per word in the student's essay.
n.s. = non-significant.
~ $p < .10$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

original version (94% total; 45% felt it was a little bit better; 49% felt it was a lot better). In general, students reported understanding and using the feedback, and they felt their essays improved.

4.6. Do students' essays improve in evidence use? (Student improvement on evidence use [outcome], RQ6 Table 1)

The paired-samples t-tests in Table 2 demonstrated statistically significant improvements in students' revised essays by features of evidence-use. Table 2 also provides descriptive information about which evidence from the text students most frequently used. For example, of the four focal topics (i.e., malaria, hospitals, school, and agriculture), students were most likely to use evidence to compare the number of Africans suffering from malaria before versus after the Millennium Villages Project. The number of specific text-based references related to each focal topic increased from students' first draft to their revised draft; however, evidence related to the topic of schools[7] went up the most (see SPC_{schl} row; $M_{post} = 2.55$; $M_{pre} = 1.87$; $t = 8.47$; $p < .001$; $ES = 0.39$). Meanwhile, students were least likely to mention evidence related to the topic of agriculture (see SPC_{Agr} row; $M_{post} = 1.36$; $M_{pre} = 1.08$; $t = 4.88$; $p < .001$; $ES = 0.20$). In general, students did add specific text-based evidence in their revisions related to the focal topics – with nearly two additional pieces of evidence added (see SPC_{focal} row; $M_{post} = 9.20$; $M_{pre} = 7.26$; $t = 10.39$; $p < .001$; $ES = 0.42$). Students also added general evidence for how conditions improved.[8] On average then, students added over three additional pieces of evidence per essay in their revision ($M_{post} = 17.77$; $M_{pre} = 14.02$; $t = 12.32$; $p < .001$; $ES = 0.47$).

4.7. Is improvement in student essays aligned with the evidence-use features targeted in the feedback message they received? (Student improvement aligned with automated feedback based on researcher hypotheses [outcome], RQ7 Table 1)

We looked at patterns of improvement for students receiving different levels of feedback to investigate alignment of students' revision with the feedback messages they received. Table 4 displays the gist of the feedback for each level along with our analytic hypotheses for each feature. Results suggest that 8 of our 9 hypotheses held. Students from all three feedback levels added evidence to their essays, though we did not anticipate that students receiving feedback level 3 (i.e., explain evidence and connect it to the overall argument) would add more evidence.

In general, the pattern for the number of topics students added to their essay corresponds to the strengths/weaknesses of students' first-draft essays and the feedback they received. A priori hypotheses 1, 4, and 7 in Table 4 suggest a linear decrease from level 1 to level 3 because students receiving level 3 feedback began with essays that addressed a larger number of focal topics. This is precisely what we observed; there was a linear decrease from 1.24 (level 1) to 0.54 (level 2) to 0.04 (level 3) topics added for feedback levels 1 through 3, respectively.

Finally, the ratio of focal topics to the total word count (hypotheses

[7] In our level 1 and level 2 feedback, we provided one example of text-evidence use from the topic 'schools' to demonstrate to students how to be more specific when conveying evidence from the text.

[8] General evidence (not included in Table 2) included evidence that was from the source text but was not part of the four focal topics (i.e., malaria, hospitals, agriculture, and schooling).

Table 5
Qualitative analysis of alignment of student open-ended responses to feedback message received.

Feedback Level	Total *N*	Related to Evidence Use[‡] *N* (% of Total)	Of comments Related to Evidence Use		Examples Related to Automated Feedback Messages
			Student takeaway relates to messages provided *N* (%)	Student takeaway relates to messages NOT provided *N* (%)	
Level 1	64	42 (66)	35 (83)	7 (17)	"I learned that if I fully address the prompt I will be able to get full credit " "I learned that you really do need a lot of evidence in a essay." "that its very important to look back in the story"
Level 2	69	37 (54)	16 (43)	21 (57)	"do it in your own words" "That you have to put enough detail so the reader can understand what you are doing and saying."
Level 3	98	82 (84)	44 (54)	38 (46)	"I will prove my point better when i elaborate." "To explain how my evidence ties in with my argument."
Total	231	161 (70)	95 (59)	66 (41)	

Note: Column totals are for 231 students because 35 students were missing on the survey responses within *eRevise* despite turning in both drafts of their writing.
[‡] There were many reasons why we could not be sure the students' written comments related to feedback messages. We generated the following codes to characterize these responses: Students left the question blank; Students wrote, "Nothing" or "I don't know"; Student responses pertained to the text of the article rather than use of evidence (e.g., "A lot of people have poverty"); Student responses pertained generally to writing (e.g., "Always reread what you are writing so that it can make sense"); Student responses pertained to grammar or mechanics; Student responses were unclear; many responses used "it" without a clear antecedent (e.g., "It helps you write better").

3, 6, and 9 in Table 4) is a proxy for the concentration of the number of unique and specific references to focal topic. We expected that this ratio would increase for students receiving level 1 feedback; that there would be no change for students receiving level 2 feedback; and that the ratio would decrease for level 3 feedback, which guided students to add explanation, not evidence. Our hypotheses about change in the concentration of evidence were confirmed for all three feedback levels. While there was variation in this ratio in the first draft essays, after the revisions, the ratio became roughly similar. On average, students provided one piece of unique evidence per 29–33 words.

4.8. What do students believe they learned from using eRevise? (Student improvement aligned with automated feedback based on student open-ended response[outcome], RQ8 Table 1)

We analyzed students' responses about "…one thing [they] learned about using evidence in [their] writing that [they] could use again". We compared cosine similarity scores between the natural language processing vector representations of the text of students' responses and the feedback messages they saw versus the messages that they did not see. In general, students' responses to this open-ended question were more similar to the feedback messages they received (cosine(A,B) = 0.086) than the messages they did not receive (cosine(A,C) = 0.066).[9] Thus, students' self-reported 'learning' about use of evidence aligned with the feedback they received.

Table 5 provides results from our qualitative coding for alignment between student response and feedback received.

We make three key observations. First, 70% of students responding articulated a 'takeaway' related to evidence use (see column 3, "related to evidence use"). Second, in general, the pattern described by the cosine similarity analysis is confirmed. That is, in the aggregate students' responses were more aligned with the feedback message they were provided (59% of the time). However, the pattern is strongest for students receiving level 1 feedback (83% of the time), followed by level 3 feedback (54% of the time) and finally, level 2 feedback (only 43% of the time). Third, many of the student articulations – e.g., "That you have to put enough detail so the reader can understand what you are doing and saying" – represent generalizations about evidence use aligned with the feedback provided.

4.9. Is there a relationship between substantive teacher interactions and student improvement on feature scores? (Mediating process —> outcome, RQ9 Table 1)

Because our goal in designing *eRevise* was for it to be used for formative assessment purposes, we wondered how teachers interacted with students during its implementation. Our results suggested this varied across teachers. More than a third of the teachers ($n = 7$) did not interact with students at all. In these cases, it appeared as if *eRevise* was being used as practice for the state standardized test – an independent one-draft writing assessment. For these teachers, their implementation log represents only procedural questions from students and corresponding teacher responses. For example, students might ask, "How do I copy and paste?" or "How do I submit?" and the teacher would resolve the problem. Other teachers were slightly less procedural during implementation ($n = 4$). For example, students might ask, "Is this enough evidence?" and instead of answering the question, the teacher would refer students back to the task description. Finally, the last group of teachers ($n = 5$) interacted substantively with student questions. That is, they appeared to use *eRevise* as a teaching and learning opportunity. For example, when a student asked, "Why did it say explain more?", the

[9] Where A is the student response, B is the seen feedback message and C is the unseen feedback message. Given cosine(A,B)>cosine(A,C), the student responses, in the aggregate, were more similar to the seen feedback messages.

Table 6
Hierarchical linear model examining effects of teachers' responses to students' queries during *eRevise* implementation.

	Unconditional Model (FUM)		Student Level (Model 1)		Student Level + Teacher Level (Model 2)	
	Coef.	*s.e.*	*Coef.*	*s.e.*	*Coef.*	*s.e.*
Mean Improvement, γ_{00}	−0.00	.08	−0.01	.09	−0.00	.08
Substantive Response to Questions, γ_{01}					0.20*	.09
Rubric Score on First Draft, γ_{10}			−0.24***	.05	−0.25***	.05
Used Feedback, γ_{20}			0.02	.08	0.03	.08
Re-Read Article, γ_{30}			−0.01	.06	−0.01	.06
How Much Revision is Better, γ_{40}			0.38***	.10	0.39***	.10
Feedback was Different, γ_{50}			−0.11	.08	−0.10	.08
Teacher Helped Me, γ_{60}			−0.10	.16	−0.13	.14
Substantive Response to Questions, γ_{61}					0.41*	.17
I like to write, γ_{70}			0.00	.04	0.00	.04
Approach to Writing, γ_{80}			−0.01	.08	−0.01	.08
Variance Components						
Classroom Variance in means ($\tau_{\beta 0}$)	.062		.071		.046	
(% var. explained from prior model)			(0%)		(21%)	
Classroom Variance in Teacher Helped Me Slope ($\tau_{\beta 6}$)			.085		.004	
(% var. explained from prior model)					(95%)	
Between Student Variance w/in Classrooms (σ^2)	.941		.764		.761	
(% var. explained from prior model)			(19%)		(0.4%)	
Deviance	895.51		747.07		740.94	
(Chi-square; df; p-value from previous model)			(χ^2 = 151.2; df=2; p=.001)		(χ^2 = 6.1; df=2; p=.045)	

Note: FUM = Fully unconditional model.
~ $p < .10$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

teacher reported that they "went over [the student's] writing and discussed that more was needed and that some evidence was the same." The teacher advised the student to "add more information or details to explain your thinking."

To understand the potential consequences of variation in classroom implementation (i.e., the extent to which the assessment appeared to be treated as a formative assessment as opposed to a summative assessment) we constructed a series of hierarchical linear models where students' improvement scores were nested in classrooms. We briefly review findings about the variance components (see bottom panel of Table 6), before we discuss the substantive meaning of associations between covariates and the improvement score.

Findings from the unconditional model reveal significant between-classroom differences in the improvement scores (about 6% of the variance lies between classrooms). As student predictors were added in Model 1, including the first-draft rubric-based score, student-level variance decreased (about 19% of the variance between students was explained) while variance between classrooms increased slightly. Notably, there is a significant reduction in the deviance statistic from the prior fully-nested model, suggesting the explanatory power of these student-level covariates. Model 1 also examined the variance of the slopes for the relationship between being helped by a teacher and the improvement score within classrooms ($\tau_{\beta 6}$= 0.085; $p = .322$). Finally, in Model 2 after adjusting for whether the teacher provided substantive comments on the intercept we explained about 21% of the between-classroom variance in means (reduction in $\tau_{\beta 0}$ from Model 1). More importantly, the addition of teachers' substantive comments as a cross-level interaction on the random slope explained about 95% of the variance between classrooms in the *relationship* between being helped by the teacher leading to a higher improvement score. Once again the deviance statistics revealed significant reduction from the prior fully-nested Model 1 with only the addition of this teacher-level covariate on the random intercept and slope.

Associations between improvement scores and a couple of student-level covariates was observed. For example, students with *lower* rubric scores on the first draft were predicted to have higher improvement scores ($\gamma_{10} = -0.25$; $p < .001$). The model also shows that students made fairly accurate predictions about how much their revision improved. For each scale point on the students' survey response (from 1 "none" to 3 "a lot"), students' improvement scores increased by about 0.4 standard deviations ($\gamma_{40} = 0.39$; $p < .001$). Moreover, the degree to which teachers interacted with their students during implementation of *eRevise* influenced students' improvement scores. This relationship was statistically significant on the classroom mean for improvement scores ($\gamma_{01} = 0.22$; $p = .043$). Thus, in classrooms where teachers provided substantive help to any student the classroom mean was higher by about 0.2 standard deviations. Finally, the cross-level interaction revealed that when students received help in classrooms where teacher interactions with students were more substantive, students' improvement scores were higher ($\gamma_{61} = 0.41$; $p = .027$).

5. Discussion

As natural language processing technologies grow in their ability to assess substantive dimensions of writing quality, we expect that AWE systems such as *eRevise* will proliferate. The potential of AWE systems to meet their intended purpose of supporting teaching and learning, however, is dependent on the degree to which they serve an authentic formative assessment purpose. As reviews of AWE systems have shown, there is a need to generate substantive feedback in response to what students have produced. Thus, one challenge is to design systems with these parameters in mind. A second challenge is to design research studies structured to support validity arguments, which, to date, have been rarely applied to the evaluation of AWE systems [34].

Evidence from our current validity argument for *eRevise* demonstrates promise in both regards. We note that several limitations should be considered in interpreting our results. Most obviously, our study was conducted in a limited number of classrooms, thus attenuating the strength of our inferences regarding students' and teachers' response to the system.[10] Furthermore, we did not have access to individual students' demographic and achievement scores and so were not able to control for potentially important confounding variables (e.g., reading skills) that could impact students' revision independent of the feedback messages. Such variables could also permit investigation into whether *eRevise* has differential impacts on students from different racial and socioeconomic backgrounds. This, along with ensuring that the corpora of essays used as training sets of natural language processing algorithms are drawn from representative samples, is important for ensuring that AWE systems are not biased against or disadvantage particular groups of students (see e.g., [46]). Finally, we examined improvement in our feature scores using Cohen's d_z representing a standardized effect size for within-subjects designs (see, e.g., [44]). Given we might expect improvement from any attempt at revision (i.e., draft 2 feature scores – draft 1), it will be important in the future to explore designs with a comparison group in order to calculate an effect size (Cohen's d_s) for a between-subjects design and/or a difference-in-difference (DID) estimate. This result would demonstrate improvement for the *eRevise* condition beyond what we'd expect under normal (comparison) conditions. These limitations aside, our study contributes to the development of systems that assess substantive dimensions of standards-based writing (i.e., the use of source text evidence) in the early grades, as well as efforts to advance validity arguments for the use of AWE systems as formative assessments more broadly (e.g., [10,73,98]). We discuss our findings relative to the activity system we theorized in Fig. 1, the research questions elaborated in Table 1 and the evidence generated for our claims constituting our interpretation/use argument (IUA).[11]

5.1. Meaning making as a focal inference in IUA for formative assessment

Meaning-making as inferred through model-based hypothesis tests: The object of our activity system was the automated feedback provided by *eRevise,* and our validity argument examined nine research questions. The cumulative evidence from our investigation supports our bolded claim in Table 1 for the potential of *eRevise* as a formative assessment. Central to this claim are the evidence we observed of student meaning-making aligned with disciplinary norms for evidence use in text-based argument writing. For example, we observed an interaction effect in our hierarchical linear model, which examined the role of teacher-student interactions as a mediating process on the quality of student revisions. As other researchers of AWE systems have noted, students often need help understanding automated feedback messages [11,77]. Findings from our hierarchical linear models showed that, in general when teachers provided more substantive support – i.e., they took an active role helping students interpret and use the feedback – students' essays showed larger improvements on the feature scores. For the roughly one-quarter of students who asked their teacher for assistance, there was an *additional relationship* between teachers' substantive help and higher improvement scores. Thus, students in classrooms where teachers provided substantive support benefitted overall, but students who asked specific questions and then received substantive support benefitted the most. All else being constant, students had higher estimated improvement scores when teachers treated *eRevise* as a formative assessment rather than a test of students' independent writing skill.

Our inference, aligned with socio-cultural learning theories, is that students' interaction with a knowledgeable other about the automated feedback better supported them to make meaning of the automated feedback and then make relevant revisions. This is the very essence of formative assessment – teachers facilitating students' interpretation of the feedback in relation to what they wrote to help them 'see' the next steps they need to take to improve their essay.

There is much we did not observe that we would also want to see in order to confirm further theory-based interpretations of our evidence. For example, our theory suggests that these teacher-student interactions promote *student learning* in addition to improved student revisions. To make such an inference, replication studies could seek evidence of transfer by examining students' first draft of a first text-based argument writing task with a cycle of automated feedback and revision to a first draft of a second text-based argument writing task. For students having substantive teacher-student interactions, evidence of greater improvements on students' first drafts could further support inferences about student learning (i.e., confirm that students' generalizations about evidence use from teacher-student interactions during their first writing task were implemented in their next, similar, writing task).

Our findings are consonant with recent calls for more human interaction to be built into automated feedback systems [20,89], and research we described earlier underscoring the importance of interactions around feedback messages (e.g., [92]). We see our empirical results as valuable evidence for the idea that, in certain contexts (i.e., where teachers view the automated feedback as an opportunity to discuss important aspects of evidence use with their students) automated feedback systems could foster the kinds of teacher-student interactions that support successful writing and revision [75]. A question going forward then, is how to design mechanisms for supporting teachers and students to view the automated feedback as an opportunity to scaffold rich conversations at the nexus of targeted constructs in writing such as evidence use as a warrant.

Meaning-making as evidenced in students' articulated learnings: In the absence of a second text-based argument writing task in this study, we asked students what they believe they learned from using *eRevise* that they would carry with them to their next argument writing task. This analysis was limited because the data were captured in written form. Thus, researchers were not able to probe students to expand on vague statements they made. Instead, the analysis included only what students were able to articulate in writing in response to our open-ended survey question. Nevertheless, we find it encouraging that 70 percent of the students presented a well-articulated 'takeaway' about evidence use. These articulations were evidence that students may have learned something about how evidence is used to support their viewpoint that is generalizable. Moreover, we found a statistically significant effect that students' written 'takeaway' was more aligned with the feedback messages they received than the feedback messages they did not receive. Our qualitative analysis provides greater description of this effect which seemed most prominent for students receiving level 1 feedback.

Finally, our argument about student meaning making in response to the automated feedback is also based on researcher hypotheses about the types of improvements we would expect to see in student essays in response to each feedback level. This was a different way to examine patterns in our data to infer whether improvements in feature scores were aligned to the feedback students received. Because 8 of the 9 hypotheses were confirmed, we infer that students were, in general, responsive to the feedback they received – thus, they likely engaged in making sense of the feedback they did receive and revised their essays according to the feedback.

[10] The sample size for students in this study is considered 'large' relative to most other studies of automated feedback systems (see Deeva et al., 2021), but we recommend that these samples be expanded, especially at the classroom level, to further our understanding and to increase power for statistical hypothesis tests.

[11] An interpretation/use argument is the aggregated claim(s) made during a validity investigation that the investigator is attempting to infer. As Kane [39] defines it for assessment scores; "the IUA includes all of the claims based on the test scores (i.e., the network of inferences and assumptions inherent in the proposed interpretation and use)" (p.2).

5.2. Measuring improvement at a grain size that supports the use of eRevise as a formative assessment

Much of the past research on the quality of AWE systems has focused on comparing human and machine-generated rubric scores. We argue, however, that a smaller 'grain size' for assessing change in students' writing (i.e., a feature score) is likely more useful for formative assessment purposes. Our results provide evidence that understanding change in the features (atomistic elements) of evidence use is possible and constructive. Specifically, we found that human raters had moderate to good inter-rater reliability with the AWE system when rating the NPE and SPC features, as well as change in those features. Furthermore, the rubric score was only capable of indicating improvement for 41% of the students, but we established that feature scores *also improved* for students, as a whole, whose scores did not indicate a rubric score change. Feature scores, then, were more sensitive to incremental differences in student writing than rubric scores and provide more specific information about evidence use. They have greater utility for formative assessment purposes because they provide teachers with more information about students' revision efforts and progress. This may be especially important in the context of writing produced by younger students, many of whom struggle to revise effectively [53,103]. We note as well that feature scores can support more precise investigation of the alignment between feedback messages and specific changes in writing quality, which is important for making inferences about the impact of automated feedback messages on students' writing. In all, scores on specific features are a potential avenue for future researchers to explore in evaluating AWE systems intended as formative assessments.

Finally, the inferences above suggest that using feature scores from students' first drafts to channel them to different feedback messages could be beneficial. We think the combination of our design features – i.e., the attempt at automating scoring of substantive features and the hybrid mix of expert-driven feedback combined with a data-driven approach to channel the feedback based on feature scores – contributed to a positive student experience with evidence for student learning about the target domain of evidence use. Designers of AWE systems that seek to be responsive to student-generated text could consider a similar approach.

5.3. Contributions to the design of AWE systems

We designed our AWE system with features that represent salient elements of evidence use for developing adolescent writers. We see the benefits accruing beyond simply greater efficiency and speed of grading – the usual benefits cited for justifying automation. In this study, we sought to demonstrate that an AWE system can encourage content revisions in writing and avoid the common criticism that the system only encourages syntactic and grammatical revisions (see, e.g., [29]). Finally, an essential question for formative purposes is how teachers can integrate automated feedback into their writing instruction to achieve broader goals of, say, developing argumentative writing skills within a process approach to writing. Our vision looking forward is that the feature scores can help teachers construct sociocultural learning opportunities with their students (see, e.g., [11,30]) about these essential components of evidence use and build student understanding of a larger writing construct such as argumentation.

We presented evidence for a validity argument for the use of one writing evaluation system (*eRevise*) and its potential as a formative assessment. While it is impossible to generalize from just one system, we note the congruence of certain design features with critical themes outlined in recent reviews of automated systems. For example, *eRevise* demonstrates the use of the recommended approach of using data to drive context-sensitive feedback to students [20]. In addition, our system provides feedback based on scoring of a substantive dimension (i.e., evidence use), and the feature scores are interpretable. We argue this aids in the measurement of features contributing to clearer hypothesis tests to support the validity argument, while the interpretability of the features aids in fitting into the formative assessment ecology – i.e., aids in channeling essays to appropriate context-sensitive feedback that students and teachers perceived as relevant; while also providing feedback that served to improve student learning in intended ways. Finally, our statistical models provide evidence that a system designed to support or encourage human interactivity around the feedback is likely to see learning gains. We believe *eRevise* could serve as an existence proof for the potential import of these system design characteristics.

6. Conclusion

Our findings confirm the perceived needs for automated feedback systems to design better sociocultural learning processes that include teachers in their design [75]. AWE systems are not likely to be effective if they are seen as burdensome, undermining instruction, and/or treated as summative assessments. Overall, the large majority of teachers indicated that they would use *eRevise* at multiple points in the year if it were available. They appreciated the potential 'time savings' afforded by the system, and perceived the feedback messages and writing task in *eRevise* as aligned with their instruction and their state and district writing standards and so, reinforcing of their instructional messages to students. We interpret these results as evidence of the coherence of *eRevise* with classroom practice and potential to alleviate, not add to, teachers' burden. Looking forward, automated systems that design better ways to support a close partnership between teachers and machines may be the most productive way for advancing the potential of AWE formative assessment systems; especially systems designed to build students' conceptual understanding of argument elements (such as evidence use) through feedback and dialogic teacher-student interactions.

Acknowledgments

The research on eRevise reported here was also supported by the Institute of Education Sciences, U.S. Department of Education through Grant R305A160245 to the University of Pittsburgh. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.caeo.2022.100084.

References

[1] Adie L, van der Kleij F, Cumming J. The development and application of coding frameworks to explore dialogic feedback interactions and self-regulated learning. Br Educ Res J 2018;44(4):704–23.
[2] Aguinis H, Gottfredson RK, Culpepper SA. Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. J Manage 2013;39(6):1490–528.
[3] Aguinis H, Culpepper SA. An expanded decision-making procedure for examining cross-level interaction effects with multilevel modeling. Organ Res Methods 2015;18(2):155–76.
[4] Baird JA, Andrich D, Hopfenbeck TN, Stobart G. Assessment and learning: fields apart? *Assessment in Education: principles*. Policy Practice 2017;24(3):317–50.
[5] Bauml M. Beginning primary teachers' experiences with curriculum guides and pacing calendars for math and science instruction. J Res Childhood Educ 2015;29(3):390–409.
[6] Bernard HR, Wutich A, Ryan GW. Analyzing qualitative data: systematic approaches. Thousand Oaks, CA: Sage Publications; 2016.
[7] Bliese PD, Maltarich MA, Hendricks JL. Back to basics with mixed-effects models: nine take-away points. J Bus Psychol 2018;33(1):1–23.
[8] Brindle M, Graham S, Harris KR, Hebert M. Third and fourth grade teacher's classroom practices in writing: a national survey. Read Writ 2015;29(5):929–54.
[9] Burstein J, Riordan B, McCaffrey D. Expanding Automated Writing Evaluation. Handbook of automated scoring. Chapman and Hall/CRC; 2020. p. 329–46.
[10] Chapelle CA, Cotos E, Lee J. Validity arguments for diagnostic assessment using automated writing evaluation. Lang Test 2015;32(3):385–405.

[11] Chen CFE, Cheng WYEC. Beyond the design of automated writing evaluation: pedagogical practices and perceived learning effectiveness in EFL writing classes. Lang Learn Technol 2008;12(2):94–112.
[12] Chong SW. Reconsidering student feedback literacy from an ecological perspective. Assess Eval Higher Educ 2021;46(1):92–104.
[13] Correnti R, Matsumura LC, Hamilton LS, Wang E. Combining multiple measures of students' opportunities to develop analytic, text-based writing skills. Educational Assessment 2012;17(2–3):132–61.
[14] Correnti R, Matsumura LC, Hamilton L, Wang E. Assessing students' skills at writing analytically in response to texts. The Elementary School Journal 2013;114(2):142–77.
[15] Correnti R, Matsumura LC, Wang E, Litman D, Rahimi Z, Kisa Z. Automated scoring of students' use of text evidence in writing. Reading Research Quarterly 2020;55(3):493–520.
[16] Crossley SA, Varner LK, Roscoe RD, McNamara DS. Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In: International Conference on Artificial Intelligence in Education. Berlin, Germany: Springer; 2013. p. 269–78.
[17] Dann R. Feedback as a relational concept in the classroom. Curriculum J 2019;30(4):352–74.
[18] Deane P. On the relation between automated essay scoring and modern views of the writing construct. Assess Writing 2013;18:7–24.
[19] Dedoose, software, version 8.3.17 (2020). As of June 2, 2020: http://www.dedoose.com.
[20] Deeva G, Bogdanova D, Serral E, Snoeck M, De Weerdt J. A review of automated feedback systems for learners: classification framework, challenges, and opportunities. Comput Educ 2020:104094.
[21] Denzin NK, Lincoln YS. Collecting and interpreting qualitative materials. thousand oaks. CA: Sage Publications; 2003.
[22] Du H, List A. Evidence Use in Argument Writing Based on Multiple Texts. Read Res Q 2020.
[23] Elmahdi I, Al-Hattami A, Fawzi H. Using Technology for Formative Assessment to Improve Students' Learning. Turkish Online J Educ Technol TOJET 2018;17(2):182–8.
[24] Enders CK, Tofighi D. Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. Psychol Methods 2007;12(2):121.
[25] Foltz PW, Rosenstein M. Data mining large-scale formative writing. Handbook of Learning Analytics; 2017. p. 199.
[26] Gallimore R, Goldenberg CN, Weisner TS. The social construction and subjective reality of activity settings: implications for community psychology. Am J Community Psychol 1993;21(4):537–60.
[27] Graham S, Capizzi A, Harris KR, Hebert M, Morphy P. Teaching writing to middle school students: a national survey. Read Writ 2014;27(6):1015–42.
[28] Greeno JC, Collins A, Resnick LB. Cognition and learning. In: Berliner DC, Calfee RC, editors. Handbook of educational psychology. New York: Macmillan; 1996. p. 15–46.
[29] Grimes, D., & Warschauer, M. (2006, April). Automated essay scoring in the classroom. In *Annual Meeting of the American Educational Research Association*, San Francisco, CA.
[30] Grimes D, Warschauer M. Utility in a fallible tool: a multi-site case study of automated writing evaluation. J Technol Learn Assess 2010;8(6).
[31] Gu PY. An argument-based framework for validating formative assessment in the classroom. Front Educ 2021;6:605999. 10.3389/feduc.
[32] Hawe E, Dixon H. Assessment for learning: a catalyst for student self-regulation. Assess Eval Higher Educ 2017;42(8):1181–92.
[33] Harrell M, Wetzel D. Using argument diagramming to teach critical thinking in a first-year writing course. The palgrave handbook of critical thinking in higher education. New York: Palgrave Macmillan; 2015. p. 213–32.
[34] Hockly N. Automated writing evaluation. ELT J. 2019;73(1):82–8. https://doi.org/10.1093/elt/ccy044.
[35] Otter Hopster-den, D Wools, S Eggen, J T, Veldkamp BP. A general framework for the validation of embedded formative assessment. J Educ Meas 2019;56(4):715–32.
[36] Jurafsky, Daniel, & James H. Martin. "Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition", 3rd Edition, Forthcoming. 2022.
[37] Khamboonruang, A. (2020). *Development and Validation of a Diagnostic Rating Scale for Formative Assessment in a Thai EFL University Writing Classroom: a Mixed Methods Study* (Doctoral dissertation).
[38] Kane, M.T. (2006). Validation. *Educational Measurement*, 4(2), 17–64.
[39] Kane MT. Validating the interpretations and uses of test scores. J Educ Meas 2013;50(1):1–73.
[40] Kane MT. Validation as a pragmatic, scientific activity. J Educ Meas 2013;50(1):115–22.
[41] Kiuhara SA, Graham S, Hawken LS. Teaching writing to high school students: a national survey. J Educ Psychol 2009;101(1):136.
[42] Kneupper, C.W. (1978). Teaching argument: an introduction to the toulmin model. College Composition and Communication, 29(3), 237–41.
[43] Koh, W.Y. (2017). Effective applications of automated writing feedback in process-based writing instruction. English Teaching, 72(3).
[44] Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013;4:863.
[45] Lawrence JF, Galloway EP, Yim S, Lin A. Learning to write in middle school? J Adolesc Adult Literacy 2013;57(2):151–61.
[46] Litman D, Zhang H, Correnti R, Matsumura LC, Wang E. June). A Fairness Evaluation of Automated Methods for Scoring Text Evidence Usage in Writing. In: *International Conference on Artificial Intelligence in Education*. Cham: Springer; 2021. p. 255–67.
[47] Lee VE. Using hierarchical linear modeling to study social contexts: the case of school effects. Educ Psychol 2000;35(2):125–41.
[48] Leont'ev A. Psychology and the language learning process. London: Pergamon; 1981.
[49] Li J, Link S, Hegelheimer V. Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. J Second Lang Writing 2015;27:1–18.
[50] Liao HC. Enhancing the grammatical accuracy of EFL writing by using an AWE-assisted process approach. System 2016;62:77–92.
[51] Link S, Mehrzad M, Rahimi M. Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. Comput Assist Lang Learn 2020:1–30.
[52] Lu X. An empirical study on the artificial intelligence writing evaluation system in China CET. Big Data 2019;7(2):121–9.
[53] MacArthur CA. Evaluation and revision. *Best practices in writing instruction* 2018:287.
[54] Madnani N, Burstein J, Elliot N, Klebanov BB, Napolitano D, Andreyev S, Schwartz M. Writing mentor: self-regulated writing feedback for struggling writers. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations; 2018. p. 113–7.
[55] Mao L, Liu OL, Roohr K, Belur V, Mulholland M, Lee HS, Pallant A. Validation of automated scoring for a formative assessment that employs scientific argumentation. Educ Assess 2018;23(2):121–38.
[56] Mathison S, Freeman M. Constraining elementary teachers' work: dilemmas and paradoxes created by state mandated testing. Educ Policy Anal Arch 2003;11:34.
[57] Matsumura LC, Patthey-Chavez GG, Valdés R, Garnier H. Teacher feedback, writing assignment quality, and third-grade students' revision in lower-and higher-achieving urban schools. The Elementary School Journal 2002;103(1):3–25.
[58] Matsumura LC, Garnier HE, Slater SC, Boston MD. Toward measuring instructional interactions "at-scale. Educational Assessment 2008;13(4):267–300.
[59] Matsumura LC, Correnti R, Wang EL. Classroom writing tasks and students' analytic text-based writing. Reading Research Quarterly 2015;50(4):417–38.
[60] Miles MB, Huberman AM, Saldaña J. Qualitative data analysis: a methods sourcebook. 3rd. editor. Thousand Oaks, CA: Sage Publications; 2014.
[61] National Center for Education Statistics. (NCES, 2012). *The nation's report card: writing* 2011 (NCES 2012–470).
[62] National Council of Teachers of English/International Reading Association. Standards for the english language arts (1996/2012). 2022.https://ncte.org/standards/ncte-ira
[63] National Governors Association Center for Best Practices. Council of chief state school officers (NGAC/CCSSO, 2010). common core state standards english language arts standards. washington, DC: national governors association center for best practices. Council of Chief State School Officers; 2022.
[64] Newell GE, Beach R, Smith J, VanDerHeide J. Teaching and learning argumentative reading and writing: a review of research. Read Res Q 2011;46(3):273–304.
[65] Odendahl N, Deane P. Assessing the Writing Process: a Review of Current Practice. ETS Res Memorandum Series 2018.
[66] O'Hallaron CL. Supporting fifth-grade ELLs' argumentative writing development, 31. Written Communication; 2014. p. 304–31.
[67] Palermo C, Thomson MM. Teacher implementation of self-regulated strategy development with an automated writing evaluation system: effects on the argumentative writing performance of middle school students. Contemp Educ Psychol 2018;54:255–70.
[68] Panadero E, Andrade H, Brookhart S. Fusing self-regulated learning and formative assessment: a roadmap of where we are, how we got here, and where we are going. Australian Educ Res 2018;45(1):13–31.
[69] Patthey-Chavez GG, Matsumura LC, Valdes R. Investigating the process approach to writing instruction in urban middle schools. Journal of Adolescent & Adult Literacy 2004;47(6):462–76.
[70] Pryor J, Crossouard B. A socio-cultural theorisation of formative assessment. Oxford Rev Educ 2008;34(1):1–20.
[71] Quintana R, Schunn C. Who Benefits From a Foundational Logic Course? Effects on Undergraduate Course Performance. J Res Educ Eff 2019;12(2):191–214.
[72] Ranalli J. Automated written corrective feedback: how well can students make use of it? Comput Assisted Lang Learn 2018;31(7):653–74.
[73] Ranalli J, Link S, Chukharev-Hudilainen E. Automated writing evaluation for formative assessment of second language writing: investigating the accuracy and usefulness of feedback as part of argument-based validation. Educ Psychol (Lond) 2017;37(1):8–25.
[74] Raudenbush SW, Bryk AS. *Hierarchical linear models: applications and data analysis* methods (Vol. 1). Thousand Oaks, CA: Sage Publications; 2002.
[75] Roschelle J, Lester J, Fusco J. *AI and the future of learning: expert panel report* [Report]. Digital Promise 2020. https://circls.org/reports/ai-report.
[76] Roscoe RD, Allen LK, Johnson AC, McNamara DS. Automated writing instruction and feedback: instructional mode, attitudes, and revising. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 62. Los Angeles, CA: Sage Publications; 2018. p. 2089–93.
[77] Roscoe RD, McNamara DS. Writing Pal: feasibility of an intelligent writing strategy tutor in the high school classroom. J Educ Psychol 2013;105(4):1010.
[78] Ryan GW, Bernard HR. Techniques to identify themes. Field methods 2003;15(1):85–109.
[79] Sannino, A., & Engeström, Y. (2018). Cultural-historical activity theory: founding insights and new challenges. *Cultural-historical psychology*.

[80] Schleppegrell MJ, Achugar M, Oteíza T. The grammar of history: enhancing content-based instruction through a functional focus on language. TESOL Quarterly 2004;38(1):67–93.
[81] Shanahan T, Shanahan C. Teaching disciplinary literacy to adolescents: rethinking content-area literacy. Harv Educ Rev 2008;78(1):40–59.
[82] Shepard LA. The role of assessment in a learning culture. Educ Res 2000;29(7):4–14.
[83] Shepard LA. Linking formative assessment to scaffolding. Educ Leadership 2005;63(3):66–70.
[84] Shute VJ. Focus on formative feedback. Rev Educ Res 2008;78(1):153–89.
[85] Snow CE, Uccelli P. The challenge of academic language. In: Olson DR, Torrance N, editors. The cambridge handbook of literacy. Cambridge, New York: Cambridge University Press; 2009. p. 112–33.
[86] Snijders Tom AB, Bosker Roel J. Multilevel analysis: an introduction to basic and advanced multilevel modeling. sage 1999.
[87] Stevenson M. A critical interpretative synthesis: the integration of automated writing evaluation into classroom writing instruction. Comput Compos 2016;42:1–16.
[88] Stevenson M, Phakiti A. Automated feedback and second language writing. Feedback in second language writing: Contexts and issues; 2019. p. 125–42.
[89] Strobl C, Ailhaud E, Benetos K, Devitt A, Kruse O, Proske A, Rapp C. Digital support for academic writing: a review of technologies and pedagogies. Comput Educ 2019;131:33–48.
[90] Sung YT, Liao CN, Chang TH, Chen CL, Chang KE. The effect of online summary assessment and feedback system on the summary writing on 6th graders: the LSA-based technique. Comput Educ 2016;95:1–18.
[91] Toulmin SE. The uses of argument. Cambridge university press; 2003.
[92] Van der Schaaf M, Baartman L, Prins F, Oosterbaan A, Schaap H. Feedback dialogues that stimulate students' reflective thinking. Scandinavian J Educ Res 2013;57(3):227–45.
[93] Vygotsky LS. Mind in society: the development of higher psychological processes. Harvard University Press; 1980.
[94] Vygotsky L. *Thought and language* (A. kozulin. Trans.). Cambridge, MA; London, England: The MITPress; 1986.
[95] Wan Q, Crossley S, Allen L, McNamara D. Claim Detection and Relationship with Writing Quality. In: *Proceedings of the 13th International Conference on Educational Data* Mining (EDM 2020); 2020. p. 691–5.
[96] Wertsch JV. The zone of proximal development: some conceptual issues. New Dir Child Adolesc Dev 1984;23:7–18. 1984.
[97] Wilson J, Czik A. Automated essay evaluation software in English Language Arts classrooms: effects on teacher feedback, student motivation, and writing quality. Comput Educ 2016;100:94–109.
[98] Wilson J, Roscoe RD. Automated writing evaluation and feedback: multiple metrics of efficacy. J Educ Comput Res 2020;58(1):87–125.
[99] Woods B, Adamson D, Miel S, Mayfield E. Formative essay feedback using predictive scoring models. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining; 2017. p. 2071–80.
[100] Xiao Y, Yang M. Formative assessment and self-regulated learning: how formative assessment supports students' self-regulation in English language learning. System 2019;81:39–49.
[101] Yin RK. Qualitative research from start to finish. New York, NY: Guilford Publications; 2015.
[102] Zhu M, Liu OL, Lee HS. The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. Comput Educ 2020;143:103668.
[103] Wang Lin Elaine, Matsumura Clare Lindsay, Correnti Richard, Litman Diane, Zhang Haoran, howe emily, et al. eRevis (ing): Students' revision of text evidence use in an automated writing evaluation system. Assessing Writing 2020;44.
[104] Wang EL, Matsumura LC. Text-based writing in elementary classrooms: teachers' conceptions and practice. *Reading and*. Writing 2019;32(2):405–38.
[105] Attali Yigal, Burstein Jill. Automated essay scoring with e-rater v.2. Journal of Technology, Learning, and Assessment 2006;4(3).
[106] Zhang H, Magooda A, Litman D, Correnti R, Wang E, Matsmura LC, Quintana R. July). eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing 2019;(Vol. 33, No. 01,:9619–25.