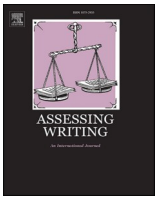
Assessing Writing 52 (2022) 100608

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw



Automated writing evaluation: Does spelling and grammar feedback support high-quality writing and revision?

Kathryn S. McCarthy [a,*,1], Rod D. Roscoe [b,2], Laura K. Allen [c,3], Aaron D. Likens [d,4], Danielle S. McNamara [b,5]

[a] Georgia State University, Atlanta, GA, USA
[b] Arizona State University, Tempe, AZ, USA
[c] University of New Hampshire, Durham, NH, USA
[d] University of Nebraska at Omaha, Omaha, NE, USA

ARTICLE INFO

Keywords:
Writing
Revision
AWE feedback
Writing strategies
Feedback uptake
Mechanics

ABSTRACT

The benefits of writing strategy feedback are well established. This study examined the extent to which adding spelling and grammar checkers support writing and revision in comparison to providing writing strategy feedback alone. High school students ($n = 119$) wrote and revised six persuasive essays in Writing Pal, an automated writing evaluation and tutoring system. All participants received automated strategy feedback after writing the first draft of their essays. Half of the participants were also given access to spelling and grammar checkers while writing. Spelling and grammar feedback on its own had no effect on the quality of students' first draft. Linear mixed effects models revealed improvements from initial draft to revision on most subscales. The addition of spelling and grammar feedback contributed small but significant gains after revision on five subscales (i.e., mechanics, word choice, voice, conclusion, and organization) but no other aspects of the students' essays. Qualitative exploration of exemplar students' revision moves revealed how students incorporated both strategy and spelling and grammar feedback into their revisions. Findings from this study demonstrate that strategy feedback with an opportunity to revise contributed to improved essay quality, but that spelling and grammar feedback provided modest, complementary benefits.

1. Introduction

The widespread use of word processors has made instantaneous feedback on spelling and grammar a common part of the writing experience (e.g., Morphy & Graham, 2012). Modern automated writing evaluation (AWE) systems are often expected to include

* Correspondence to: College of Education & Human Development, Department of Learning Sciences, P.O. Box 3978, Atlanta, GA 30302, USA.
E-mail addresses: kmccarthy12@gsu.edu (K.S. McCarthy), rod.roscoe@asu.edu (R.D. Roscoe), laura.allen@unh.edu (L.K. Allen), alikens@unomaha.edu (A.D. Likens), dsmcnama1@asu.edu (D.S. McNamara).
[1] https://orcid.org/0000-0002-6277-7005.
[2] https://orcid.org/0000-0001-8327-4012.
[3] https://orcid.org/0000-0001-5582-1402.
[4] https://orcid.org/0000-0002-6535-5772.
[5] https://orcid.org/0000-0001-5869-1420.

https://doi.org/10.1016/j.asw.2022.100608
Received 19 February 2021; Received in revised form 21 December 2021; Accepted 26 January 2022
Available online 4 February 2022

spelling and grammar checking tools, and services such as Grammarly promise that the use of their checkers will make "everyone a great writer" (Grammarly, 2019). However, despite the growing list of spelling and grammar checkers available in a variety of languages, as well as the assumption that spelling and grammar feedback is necessary for good writing, there is little empirical work that assesses whether these tools improve overall writing quality.[6] High-quality writing emerges not only from mastery of spelling and mechanics, but from a clear structure and coherence of the content (see Graham & Harris, 2016). Across a number of age groups, the most effective means of improving writing is through instruction on writing *strategies* for planning, drafting, and revising essays for content coupled with opportunities for writing practice with formative feedback (Gillespie & Graham, 2014; Graham & Perin, 2007; Graham, 2006; Graham, Capizzi et al., 2014; Graham, Harris, & Chambers, 2015; Graham, MacArthur, & Fitzgerald, 2013; Kiuhara, Graham, & Hawken, 2009). However, this does not exclude the need for supporting spelling and grammar improvements alongside formative strategy feedback.

Indeed, recent work aimed at supporting students' writing *growth* has emphasized the need for comprehensive, integrated writing support (Graham et al., 2012; Graham, Bruch et al., 2016). Effective writing clearly conveys the author's intended meaning and is appropriate for the given audience and context. Experts on writing instruction emphasize that effective essays are marked by *specific features* that include organization and structure, a clear voice, correct use of genre conventions, as well as "grammar, punctuation, and spelling" (Graham, Bruch et al., 2016, pg. 36). These panel recommendations argue that the most effective way of supporting writing growth is through iterative model-practice-reflect cycles that encourage instruction, practice, and formative feedback along *all* key features of writing (Graham, 2021). Thus, the goal of writing instruction and feedback research should not be to determine which type of feedback is best, but rather to develop a more nuanced understanding of how to best implement integrative writing support.

Although spelling and grammar feedback tools are essentially omnipresent during most writing experiences, prior research on writing instruction and assessment does not provide clear answers regarding the potential effects of providing spelling and grammar feedback *in concert with formative strategy feedback*. Spelling and grammar feedback on their own are unlikely to yield substantial increases in overall writing quality; yet, it is unknown how such targeted feedback may help or hinder writing when provided alongside strategy feedback known to support better writing. To address this gap in the literature, this study examines the effects of providing spelling and grammar checkers above and beyond writing strategy feedback on the quality of initial essays and revisions. Specifically, we contrast the effects of receiving (a) formative strategy feedback alone, versus (b) formative strategy feedback along with feedback on spelling and grammar. The effects of feedback are examined for six essays composed by 119 high school students in terms of holistic scores as well as nine subscales: (1) *grammar, style, and mechanics,* (2) *word choice*, (3) *voice*, (4) *sentence structure,* (5) *introduction paragraph quality*, (6) *body paragraph quality*, (7) *conclusion paragraph quality*, (8) *organization*, and (9) *unity*. As such, this study assesses which aspects of essay revisions are affected by online feedback regarding mechanics, and the extent of strategy feedback uptake on essay revisions.

1.1. Improving writing via strategies and feedback

Less skilled and developing writers tend to *knowledge-tell* (e.g., Scardamalia & Bereiter, 1986) in stream-of-consciousness, transcribing each idea linearly as it comes to them with little regard to the writing goal or the intended audience. By contrast, skilled writers tend to be more strategic, intentional, and recursive. Successful writers engage in (at least) three overarching strategic activities: (a) planning, (b) drafting, and (c) revising (e.g., Flower & Hayes, 1981; Hayes, 2012). Coordinating these processes, enacting underlying skills, and drawing upon relevant knowledge are often challenging to developing writers (McNamara & Allen, 2017) and often lead students to struggle on writing assessments (NAEP, 2012). The central aim of writing instruction is thus to prepare students to understand and navigate these challenges (Harris et al., 2011).

The most successful writing strategy interventions involve providing developing writers with actionable and intentional procedures and "tools" that they can apply to a variety of writing tasks (e.g., Gillespie & Graham, 2014; Graham & Perin, 2007; Graham et al., 2013; Graham, Aitken, et al., 2020; Graham, Capizzi et al., 2014; Graham, Bañales, et al., 2020). Developing writers must also have the opportunity to engage in deliberate practice with the strategies and receive feedback. That is, they must be offered ample time to compose their essays and receive *formative* feedback that helps them to understand how to better plan, draft, and revise their essays. Strategy instruction that targets key writing processes (e.g., planning and revising) leads to strong and positive effects on writing performance (e.g., $ES = 0.82$; Graham & Perin, 2007). Further, engaging in iterative drafting and revision cycles with formative feedback leads to improved essay quality and writing skill (e.g., McNamara & Allen, 2017; Butler & Britt, 2011; Graham & Perin, 2007; Hillocks, 1984; Kellogg & Raulerson, 2007; Midgette, Haria, & MacArthur, 2008, Parr & Timperley, 2010; Proske et al., 2012; Santangelo, Harris, & Graham, 2016).

1.2. Spelling and grammar feedback

Despite consistent research findings on the importance of writing strategy instruction and feedback, many students and instructors tend to focus on mechanical errors such as spelling and grammar (e.g., Otnes & Solheim, 2019; Underwood & Tregidgo, 2006). However, the salience and impact of mechanical errors in writing depends on the audience and the measure. For example, Graham,

[6] It is of note that writing quality can be defined in a number of ways. For our purposes, we operationalize writing quality by having expert raters provide scores on a standardized rubric which includes multiple dimensions related to the content, organization, style, and mechanics of the essay. While such a rubric does not capture all aspects of writing, it is assumed that a standardized rubric captures a generally agreed upon set of criteria.

Aitken, et al.' (2020a), Graham, Bañales, et al.' (2020b) found that children with reading difficulties scored lower on essays, but that norm referenced measures (e.g., standardized tests) tended to be more sensitive to mechanical errors as compared to writing assessments developed by researchers. Along these lines, Crossley et al. (2014) found that there was little influence of mechanical errors on expert raters' evaluations of essays. They asked expert raters with at least 4 years of experience in teaching composition to score 100 student essays using the standardized SAT rubric (1–6). They found that mechanics errors in students' essays were not significantly predictive of the raters' scores.

By contrast, mechanical errors in writing can be highly salient to typical or "everyday" readers (e.g., peers, colleagues, and potential employers; Boland & Queen, 2016; Figueredo & Varnhagen, 2005; Johnson, Wilson, & Roscoe, 2017; Marshall, 1967). For example, Johnson et al. (2017) asked college students to read essays exhibiting spelling and grammar errors versus content and structure errors, or a mix of both. Student raters assigned significantly lower writing quality scores based on spelling and grammar errors. Moreover, spelling and grammar errors also led participants to perceive the essay writers more negatively (e.g., unintelligent, disloyal, and unkind).

Notably, students who place greater value on conventions and mechanics tend to be less skilled writers compared to those who recognize the importance of content and structure (MacArthur, Philippakos, & Graham, 2016). Additionally, there are mixed findings regarding the benefits or detriments of providing feedback and instruction regarding writing mechanics. On the one hand, spelling and grammar feedback and instruction can be beneficial for elementary school students who are developing lower-level literacy skills (Graham & Santangelo, 2014) as well as non-native language learners (e.g., Chodorow, Gamon, & Tetreault, 2010; Heift & Rimrott, 2008). Moreover, several studies have demonstrated that grammar checkers can increase writers' motivation and confidence (Cavaleri & Dianati, 2016; Potter & Fuller, 2008). However, there is reason to suspect that the availability and use of spelling and grammar feedback for proficient speakers (e.g., native language adolescents and adults) may have little effect and even negative effects on writing quality. Drawing attention to these mechanical errors can misdirect students' attention to less important features of writing, biasing their attention to these errors and preventing them from giving their full attention to the content or structure of their essays (Graham & Santangelo, 2014; Morphy & Graham, 2012). Consequently, providing immediate spelling and grammar feedback (e.g., via checking tools) may indirectly hinder writing quality by reinforcing such biased attention. In terms of instruction beyond feedback messages, there is evidence that grammar instruction is not effective in improving writing quality. In several writing interventions, grammar instruction served as a control condition for other instruction types. Indeed, the meta-analysis conducted by Graham and Perin (2007) indicated that explicit grammar instruction had a reliably negative impact ($ES = -0.32$) on writing performance, further suggesting that increasing attention to mechanical issues is ineffective for enhancing writing quality. Importantly, this same meta-analysis highlights that little work has been conducted to examine the effects of spelling feedback or instruction on adolescent and adult writers.

In sum, research suggests that spelling and grammar feedback in isolation may be relatively ineffective for helping students develop their writing skills. However, most AWE systems have the capability of providing formative strategy feedback in addition to (or instead of) spelling and grammar feedback alone. Few or no studies have directly tested the benefits of spelling and grammar feedback within an AWE context (cf., Grimes & Warschauer, 2010; Lin, Liu, & Paas, 2017), and no studies have directly examined the effects of combining spelling and grammar feedback with strategy feedback.

1.3. Automated writing evaluation and feedback

Writing technologies have reshaped the way that people write and the way that writing is taught (Graham, 2021). Advances in technology have not only made spelling and grammar checkers readily available, but also made the assessment and feedback of writing more rapid and personalized. Automated essay scoring (AES) systems originally emerged as a means of assessing student writing in large-scale standardized testing contexts; however, many AES systems have been modified for classroom use such that they also provide formative feedback. In automated writing evaluation (AWE)[7] systems, students engage in cycles of writing, receiving feedback, and revising. The most familiar tools include Educational Testing Service's e-Rater (Attali & Burstein, 2006; see also Hazelton, Nastal, Elliot, Burstein, & McCaffrey, 2021), Vantage Learning's IntelliMetric (Elliot, 2003), and PEG Writing (now MI Write; Page, 2003). However, there are a variety of different AWEs available commercially and in the research sector (Allen & Perret, 2016). In a recent systematic review, Strobl et al. (2019) identified nearly 90 automated writing evaluators. AWEs use natural language processing-driven scoring algorithms (or scoring engines) to provide both summative numeric scores as well as formative feedback for essays and other open-ended responses (see Shermis & Burstein, 2013; see also Allen et al., 2016; Strobl et al., 2019). The types and quality of feedback provided vary from system to system. Notably these systems are not designed to replace instruction. Rather, AWEs are viewed as instructional supplements. They can provide rapid scoring and deliver feedback at scale. An entire class of students can receive one-on-one support and engage in multiple cycles of writing and revisions in a single sitting (Stevenson & Phakiti, 2014). AWEs can also enable multiple iterations of feedback, such that students can address mechanical errors and basic organizational or structural issues before submitting to their instructor. Instructors can then dedicate their energy toward feedback on content (Link, Mehrzad, & Rahimi, 2020; Wilson & Czik, 2016).

Early work in AWEs was conducted primarily with native-speaking high school and college students (Grimes & Warschauer, 2010; Stevenson & Phakiti, 2014). More recently, a relatively large body of research has focused on second language (L2) students learning to

[7] The terms "automated writing system" or "automated writing evaluator" are used here as an umbrella term. Such tools are also often subclassified into specific writing genres (e.g., "automated essay evaluation" and "automated summary evaluation").

read and write in English (e.g., El Ebyary & Windeatt, 2010; Li, Feng, & Saricaoglu, 2017; Li, Link, & Hegelheimer, 2015; Zhang & Hyland, 2018). Research with L2 students has reported on AWEs' accuracy and quality as compared to human feedback (e.g., Attali & Burstein, 2006; Dikli & Bleyle, 2014; Powers, Burstein, Chodorow, Fowles, & Kukich, 2002); how teachers and students perceive the system feedback (e.g., Bai & Hu, 2017; Dikli & Bleyle, 2014; O'Neill & Russell, 2019; Ranalli, 2018); and how students integrate AWE feedback their revisions (Huang & Renandya, 2020; Koltovskaia, 2020). Students tend to find the feedback helpful and tend to be aware of the limitations and fallibility of computer-based feedback (Bai & Hu, 2017;). AWE feedback can lead to improved essay quality and, to some extent, improved writing quality over repeated practice (e.g., Wilson & Roscoe, 2020; Li et al., 2017).

Broadly speaking, AWEs have positive effects on student experience and writing quality (e.g., Shermis et al., 2016). Nonetheless, there are a number of studies that find no effects of AWEs. Differences in findings likely emerge due to the wide variety of manipulations and measures that have been used in these studies. In some cases, AWEs are compared to a no feedback control. In others, AWEs are directly compared to instructor feedback. These studies also vary in terms of the outcome of interest. In some studies, efficacy is measured using error rates that are specific to mechanic issues. In other studies, researchers examine the effect of AWE feedback on distal outcomes including over AWE-generated score, human ratings of overall quality and improved course grades (see Stevenson & Phakiti, 2014, for a review). The differences in findings may also be a function of the AWE in question. Some systems (e.g., Grammarly) primarily provide targeted feedback related to spelling, mechanics, and grammar.[8] By contrast, the Writing Pal offers feedback specific to writing processes and strategies. Other systems (e.g., Criterion) provide feedback at both levels (see Stevenson & Phakiti, 2019). Different types and combinations of feedback are likely to lead to varying results in terms of students' perceptions and more objective outcomes.

Collectively, research to-date suggests that AWEs can serve as a powerful vehicle to support student writing. However, we lack a coherent understanding of which pedagogical approaches embedded within AWEs are most appropriate and yield meaningful learning outcomes (Palermo & Wilson, 2020; Warschauer & Ware, 2006). As such, the present study aims to better understand best practices for writing instruction within the context of automated writing evaluation.

2. The current study

The current study assesses the effects of providing (a) spelling and grammar feedback and/or (b) formative writing strategy feedback within an AWE system.[9] Specifically, we examined how adding spelling and grammar checkers to *The Writing Pal* (W-Pal)[10] AWE and tutoring system affected multiple dimensions of essay quality. High school participants wrote and revised a series of persuasive essays over several sessions. All participants received writing strategy feedback to guide their revisions, but half of the participants were *also* given access to spelling and grammar checking tools while writing and revising.

Given that strategy feedback was provided between initial draft and revision, our analyzes investigated the effects of spelling and grammar feedback generally, but more specifically how they influenced the *improvement* of essay quality from first draft to final draft, above and beyond the increases afforded by the strategy feedback.

Our broadest research question pertained to overall essay quality. Prior research informed several competing hypotheses. First, one hypothesis is that spelling and grammar feedback is *detrimental* (**H1**). During writing, spelling and grammar feedback may be intrusive to the writing process (Morphy & Graham, 2012). In the context of revision, participants tend to rely on less-productive revisions of mechanical errors (Crawford, Lloyd, & Knoth, 2008; Fitzgerald, 1987). By further directing students' attention to these issues, they may neglect more substantive revisions (e.g., ignore the strategy feedback in favor of minor corrections). Further, additional feedback may be "too much of a good thing". For example, McCarthy and colleagues (2018) explored the effects of adding more metacognitive feedback to a literacy tutor and found that adding feedback after every task, in addition to feedback at the end of a full instructional activity showed no positive gains in pre- to post-intervention and that this additional feedback was *harmful* for less-skilled students. Thus, receiving spelling and grammar feedback may be detrimental in the sense that it could impede the use of strategy feedback, resulting in less improvement in the quality of students' essays from initial draft to revision.

By contrast, spelling and grammar feedback could *benefit* writing quality above and beyond strategy feedback alone (**H2**). Giving feedback on spelling and grammar may free up students' time and resources for developing content and structure (Graham & Santangelo, 2014: Morphy & Graham, 2012). In this case, we would expect that students who have spelling and grammar checkers available to them would write higher quality first drafts and have ample time to allocate to substantive revisions suggested by the strategy feedback. If this were the case, we might observe higher scores for initial drafts and/or larger gains from initial drafts to revised drafts.

Finally, a third hypothesis is that there is *no effect* of spelling and grammar feedback (**H3**). Prior work (e.g., Crossley et al., 2014) demonstrates that expert ratings of essay quality are driven by deep features of composition (e.g., the cohesion, or *unity* of ideas across the essay) and spelling and grammar mistakes have minimal impact. By this logic, the addition of spelling and grammar checking tools

[8] At the time of submission, Grammarly's free version offers feedback only on grammar, spelling, and punctuation. The Premium and Business versions do, however, include additional feedback related to tone and clarity. The accuracy and impact of this feedback is unknown and has not been scientifically validated to our knowledge.

[9] Portions of the results from the current study were presented in Allen et al., 2019, McCarthy et al., 2019). The current paper provides a detailed description of the experimental results.

[10] The Writing Pal is currently available without cost to researchers who wish to conduct studies to examine the impact of various types of writing instruction and feedback.

would have no direct effect on essay quality or improvements.

These hypotheses are relatively coarse-grained in the sense that they predict how spelling and grammar relate to overall essay quality. We thus extend our analyzes by assessing more fine-grained components of essay quality – including specific scores for quality of the essays' introductions, bodies, and conclusions as well as their mechanics, sentence structure, and organization. We anticipate that those who are provided spelling and grammar feedback are likely to outperform their peers on the mechanics subscore given that they have access to explicit corrections. However, it is less clear how spelling and grammar feedback might influence the other subscores. Thus, we explore whether spelling and grammar helps, harms, or has no effect on each of these subscores in addition to overall holistic score.

We also recognize that essay scores will be affected by other factors. For example, essay quality tends to vary systemically as a function of writing prompt (Huot, 1990). In addition, essay quality is influenced by individual differences in general writing proficiency as well as by other related literacy skills. Previous work has demonstrated that reading skill is strongly correlated with writing proficiency (Authors, xxxx). To account for this variance, we examine the impact of both prompt and reading skill on the uptake of feedback in the AWE system. In addition to these quantitative analyses of human ratings of essay quality, we examine four exemplar essays to illustrate the various ways that students leverage both strategy and spelling and grammar feedback to revise their essays and how these revisions relate to changes in essay subscores and overall quality scores.

3. Method

3.1. Participants

High school students (n = 121) were recruited from a large metropolitan area in the southwestern United States through flyers as well as radio and social media ads. Two participants did not complete all 6 essays. The 119 participants who completed all essay drafts and are included in the subsequent analyses. One student did not complete the demographic questionnaire. The remaining 118 participants reported an average age of 17 (M = 17.19, SD = 1.28, Range: 13–19). Demographically, 61.89% of participants self-identified as female and 38.13% as male. Participants self-identified as Caucasian (54.23%), Hispanic/Latin American (21.19%), Asian (10.17%), African American (7.62%), or as another race/ethnicity or multiracial (6.78%). The majority of participants (87.28%) identified English as their native language.

Though randomly assigned, the non-native English speakers were split evenly across the two feedback conditions. The majority of L2 students (8) identified Spanish as their native language. Other languages include Amharic, French, German, Hebrew, Japanese, and Urdu. The majority indicated speaking English for more than 7 years.

3.2. Design

Participants were randomly assigned to one of two feedback conditions. Half of the participants received *only* automated formative strategy feedback (Strategy Condition, n = 60), and half of the participants received *both* formative strategy feedback *and* spelling and grammar feedback (Strategy + SG Condition, n = 59).

3.3. Materials

3.3.1. The Writing Pal (W-Pal)

W-Pal (Authors, xxxx) is an automated writing evaluation (AWE) tool and intelligent tutoring system (ITS) that teaches writing strategies and enables multiple forms of writing practice (e.g., game-based strategy practice and essay writing). W-Pal instruction targets eight different aspects of writing: freewriting, planning, introduction building, body building, conclusion building, unity, paraphrasing, and revision. Importantly, in the current study, we examined *only* the AWE components of W-Pal—participants did not review strategy lessons or play practice games. In the AWE essay writing module, learners are assigned an SAT-style persuasive essay prompt and have about 25 min to compose an initial draft. Once time has elapsed, several natural language processing-driven algorithms provide both a holistic score (on a 6-point scale) and personalized formative feedback. More details about W-Pal have been reported in Authors (xxxx).

3.3.2. Essay prompts

Participants wrote and revised up to six essays in response to prompts adapted from publicly released SAT exam materials. These prompts asked participants to adopt a stance with regard to a central topic, and then to defend that position via evidence, examples, and/or logical reasoning. All prompts were designed to minimize prior knowledge demands such that participants could write from experience rather than constrained educational content or source materials. In each prompt, there was a brief introduction to the topic and then a final prompt question. For example, the prompt about "images and impressions" appears below:

> All around us appearances are mistaken for reality. Clever advertisements create favorable impressions but say little or nothing about the products they promote. In stores, colorful packages are often better than their contents. In the media, how certain entertainers, politicians, and other public figures appear is sometimes considered more important than their abilities. All too often, what we think we see becomes far more important than what really is. Do images and impressions have a positive or negative effect on people?

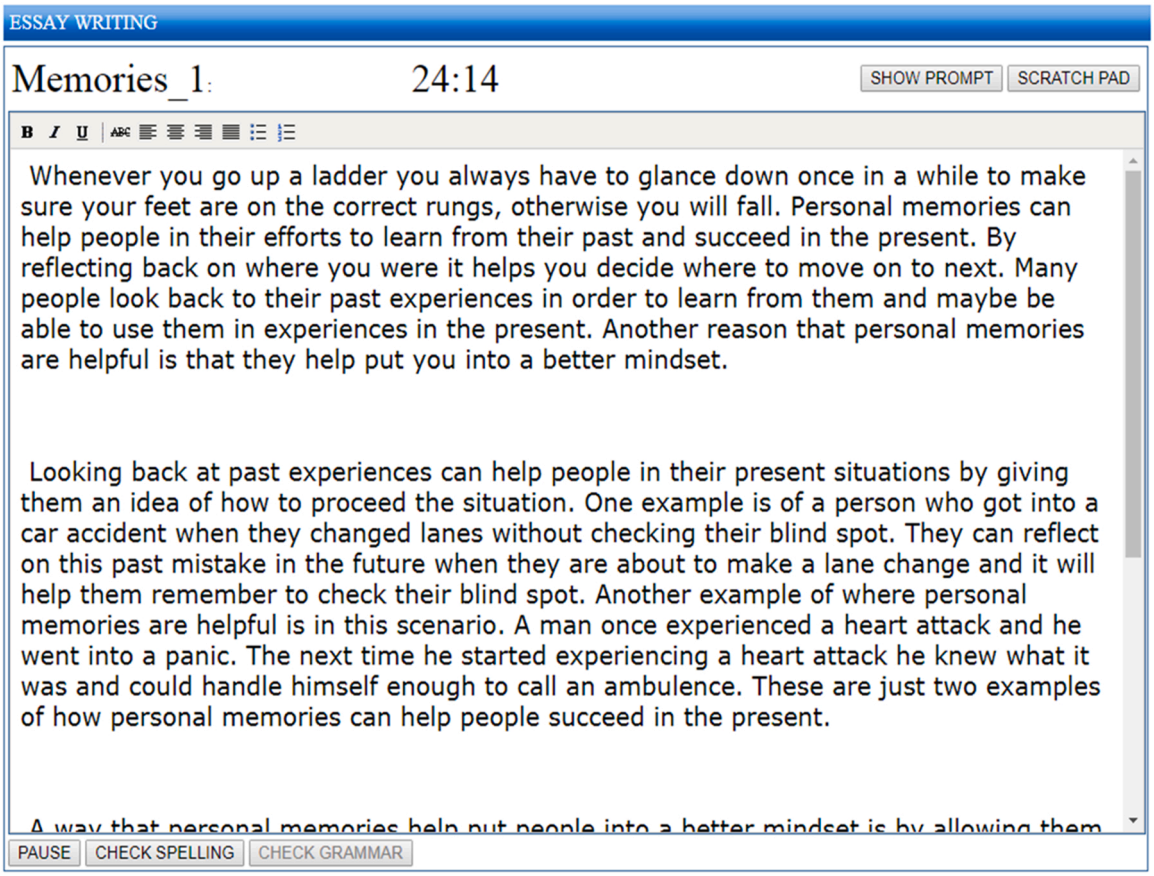


Fig. 1. Writing pal essay writing interface with spelling and grammar check icons.

Table 1 presents the focal question for each prompt, and the complete prompts are reported in Appendix A. For logistical purposes in administering the study, all prompts were presented in the same order for all participants.

3.3.3. Strategy feedback messages

In W-Pal, formative feedback messages recommend actionable steps and strategies for improving an essay via prewriting, drafting, and revising. These messages can address a variety of concerns such as generating and elaboration ideas, organizing ideas, crafting strong introductions and conclusions, providing meaningful examples and evidence, building cohesion, choosing appropriate and precise words, and overall revising. Crucially, participants never receive feedback on all categories for any given essay. Algorithms identify the most salient problem areas first. Such algorithms are based on NLP tools that evaluate linguistic, syntactic, and semantic features relevant to common challenges and weaknesses in participant writing (e.g., poorly developed ideas and low cohesion). Table B2 briefly provides example feedback messages participants could receive regarding *structure*, *conclusion building*, or *cohesion*.

3.3.4. Spelling and grammar feedback

Participants in the Strategy + SG Condition were given access to "Check Spelling" and "Check Grammar" buttons at the bottom of the essay window (Fig. 1). Participants could access either function at any time during writing or revising.

When participants used the checkers, relevant errors were underlined within the text, similar to features common in word processing software. Specifically, errors were detected using the open-source *API LanguageTool* (Miłkowski, 2010; Naber, 2003). Clicking on the error opened a small pop-up window containing potential corrections. To ensure that participants did not overlook the checking tools, a reminder about the checkers appeared when there were five minutes remaining in the writing session. However, participants were not forced to use the tools.

3.3.5. Gates-MacGinitie Reading Test

The Gates-MacGinitie Reading Test (GMRT; MacGinitie, MacGinitie, Cooter, & Curry, 1989) was included to assess reading skill, which is correlated with writing ability (Authors, xxxx). The GMRT is a standardized test in which participants read passages and then answer multiple-choice questions about them. The test contains 48-item multiple-choice items and is a reliable and well-established measure of reading comprehension ($\alpha = 0.85$–0.92; Phillips, Norris, Osmond, & Maynard, 2002).

3.4. Procedure

This project was reviewed and approved by the university's Institutional Review Board prior to all data collection. The study took place over four sessions. In Session 1, participants completed a demographic questionnaire and the GMRT. In Sessions 2 through 4, participants authored essays in response to six persuasive prompts (i.e., two prompts per session) using W-Pal. For each prompt, participants were allotted 25 min to compose an initial draft. After the 25 min elapsed, a pop-up window appeared with an algorithm-generated holistic rating from "Poor" to "Great" (with six levels) along with formative strategy feedback. The window displayed only one feedback message that targeted a critical writing strategy for the current essay. Participants could voluntarily request to receive and view additional feedback—up to a maximum of 10 feedback messages. After participants reviewed their feedback, they were allotted 10 min to revise.

3.5. Essay scoring

W-Pal automatically assigns holistic ratings of essay quality using algorithms validated based on human raters (e.g., Authors, xxxx). However, in the present study, we relied on expert human ratings of essay quality to evaluate additional subscores. The W-Pal system score was correlated ($r = 0.65$) with the holistic scores generated by the human raters. This correlation is consistent with extant work comparing human and automated scores (Authors, xxxx).

Human ratings enabled us to assess not only holistic quality, but performance based on a variety of fine-grained subscales. Specifically, trained raters assigned a *holistic essay quality score* (see Table B3 for rubric) to each essay as well as ratings on nine separate subscales: (1) *grammar, style, and mechanics,* (2) *word choice*, (3) *voice*, (4) *sentence structure,* (5) *introduction paragraph quality*, (6) *body paragraph quality*, (7) *conclusion paragraph quality*, (8) *organization*, and (9) *unity* (see Table B4). Note that the holistic score is a separate scale (1–6) and not a sum of the subscale scores. However, it was provided by the same rater who assigned the subscores. It is also of note that the raters have not only scored these essays, but hundreds, if not thousands of other essays using this same rubric for a number of related research projects.

The trained human raters included four graduate students of English. Rater pairs were trained to a high level of reliability (i.e., all kappas > 0.80) on all metrics on practice essays. Each essay ($N = 1428$) was then scored by two raters. Across raters, the reliability of ratings for unadjudicated scores (ICC) ranged from .79 to .90. If the two raters scored differed by more than 2, the scores were adjudicated by a third party. The final scores for each essay reflect the average score of the two raters.

Descriptive analyses of essay scores, collapsed across feedback condition, are provided in Table B5. Unsurprisingly, holistic scores were strongly correlated with all subscales (i.e., concurrent validity). These correlations are similar to those reported in other studies using expert ratings of holistic and subscores (Crossley et al., 2014; Crossley & McNamara, 2010; Roscoe et al., 2014b). In addition, and consistent with prior research (e.g., Allen et al., 2014), reading skill was positively correlated with holistic score and subscales ($r = 0.47$ to .68). These moderate to strong correlations provide both convergent validity for our scoring rubric as well as evidence for the need to consider the effect of reading skill on essay scores.

4. Results

4.1. Preliminary analyses

Exploratory t-tests indicated no difference in performance between L1 and L2 participants on holistic score and no differences on most of the subscores. The exception was for Grammar, Spelling and Mechanics in which L1 participants scored higher ($M = 3.59$, $SD = 0.73$) than the L2 participants ($M = 3.45$, $SD = 0.80$), $t(223) = 2.11$, $p = .04$. Due to the small sample and even division of L2 participants across conditions, we do not further examine the effects of the experimental manipulation as a function of speaker status.

4.1.1. Use of the spelling and grammar feedback tools

An important first step was to confirm that participants used the spelling and grammar tools provided by W-Pal. Participants in the Strategy + SG Condition authored more than 700 essays in total (i.e., 59 students × 6 essays × 2 drafts). Log data revealed that spelling and/or grammar checking tools were accessed for 622 of these essays; 149 essays exhibited no evidence of using the checkers. Among these 622 cases, the spelling checker was accessed for 592 essays (95.2%) and the grammar checker was accessed for 378 essays (60.8%). Thus, most participants used the spelling and grammar checker on multiple essays, and every participant used the checkers on more than one essay.

When participants accessed a checker, all errors of that type (i.e., spelling or grammar) were visually underlined. Table B6 displays the number of errors of each type logged by the system for (a) initial drafts and (b) revised drafts. As an example, Table B7 shows the frequency of each type of error identified in the *Competition and Cooperation* prompt essays. The majority of errors were spelling errors, including both typographical errors ("teh") and true misspellings ("Instrincly"). Notably, there were far more spelling errors than grammar errors.

4.1.2. Effects of spelling and grammar feedback alone

The primary focus of the study was the combination of automated strategy feedback with spelling and grammar feedback. We were neither theoretically nor practically interested in the effects of spelling and grammar checking in isolation. Nonetheless, we were able to investigate the potential impact of spelling and grammar alone by comparing essay quality across the two conditions for

participants' *initial* draft of their *first* essay (on the topic of *Images and Impressions*). For this essay and draft, participants had not yet received any strategy feedback on any essay—the only source of support (in the SG condition) was the spelling and grammar checkers. Thus, this specific instance allowed us to assess benefits of spelling and grammar feedback versus a "no feedback" control case. A multivariate analysis of variance (MANOVA) tested the effect of condition on all of the scores (holistic and the nine subscores). This analyzes revealed no effect of condition (and thus spelling and grammar feedback) on any of the expert holistic or subscale ratings (all $Fs < 1.00$; Table B8). This outcome suggests that spelling and grammar feedback on its own has little effect on essay quality, even when examining subscales that focus on surface-level features of the essays.

4.2. Effects of strategy feedback + spelling and grammar feedback

Based on prior research, there are plausible reasons to hypothesize that combining spelling and grammar support with formative strategy feedback could be detrimental (H1), beneficial (H2), or have no effect (H3) on writing quality. Thus, we conducted analyzes to examine the effects of combining feedback approaches, along with potential influences of the essay prompts and individual differences.

To assess the effects of combining feedback approaches, along with potential effects of different prompts or individual differences in reading skill, we implemented a series of linear mixed effects models using the *lme4* package in R (Bates, Maechler, Bolker, & Walker, 2015). The models allow us to enter between-subjects and within-subjects fixed factors along with the individual student as a random factor to account for person-level variance. Specifically, we conducted separate analyzes for the holistic score and each subscale essay score as the dependent variable. The models also included (a) condition (strategy feedback, strategy + SG) and (b) essay draft (i.e., initial and revised), while controlling for (c) prompt and (d) reading ability (GMRT; see model structures in Appendix B). The *Reghelper* package (Hughes & R Core Team, 2017) was used to estimate simple slopes.

Unsurprisingly, reading skill was correlated with all essay scores (r's = 0.33–0.49). For each essay score, a baseline model (M0) was created, including the two covariates, GMRT and essay prompt, in order to control for the influence of reading skill ($M = 0.58$, $SD = 0.20$) on writing proficiency as well as the differences that emerge as a function of prompt (Authors, xxxx; Huot, 1990). Model 1 (M1) added 'draft' (initial draft and revision) as a fixed effect to examine the effects of revision. Model 2 (M2) added the fixed effect of feedback condition (Strategy and Strategy + SG), as well as 'draft by condition' and 'GMRT by condition' interaction terms.

Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. For each of the models listed below, significance was determined using likelihood ratio tests between each model and a reduced model. Significant chi-square ($\chi 2$) tests indicate that adding the additional variable(s) improved fit as compared to the previous model. In the following analyses, we report the best fitting model for each score. Thus, if main effects or interactions are not reported, they can be assumed to be nonsignificant.

Table B9 reports average essays scores, as a function of draft and condition, along with the likelihood ratio tests indicating model fit. Overall, Table B9 reveals that students' essay scores improved from initial draft to revision in terms of the holistic score and all but one of the subscales (unity). This result substantiates the value of strategy feedback and opportunities to revise. By contrast, the effects of spelling and grammar feedback were limited to improvements in terms of mechanics (as expected) and three of the other subscores (word choice, voice, and conclusion). These results are discussed in more detail in the following sections. The full models for each subscale are provided in Appendix B.

4.2.1. Holistic essay quality

To evaluate the influence of different types of feedback on overall essay score, we first inspected the holistic score. Holistic essay quality increased significantly when participants revised (i.e., from initial to revised draft), supporting assumptions that strategy feedback and the opportunity to revise improves essay quality. There was not a significant effect of condition (Table B9). Thus, participants in both the Strategy and Strategy + SG conditions were able to use the automated formative feedback to improve their essays. However, the availability of spelling and grammar checking tools conferred no additional benefits. These findings support the hypothesis (H3) that adding spelling and grammar feedback to an AWE system has no discernable effects on overall essay quality.

4.2.2. Grammar, style, and mechanics

Spelling and grammar checkers directly target spelling and grammatic errors, and thus it was unsurprising to observe a significant positive benefit for checking tools on the mechanics subscale score. Specifically, a simple slopes analysis showed that Strategy + SG Condition participants improved significantly from initial draft to revision; Estimate = 0.26, SE = 0.04, $t(1283) = 4.86$, $p < .001$; whereas Strategy Condition participants did not; Estimate = 0.02, SE = 0.04, $t(1283) = 0.50$, $p = .62$. Notably, Strategy + SG condition participants had access to spelling and grammar feedback during *both* stages of writing. The significant interaction suggests that the effects of spelling and grammar feedback primarily affected the quality of the revision. This finding supports the hypothesis (H2) that spelling and grammar feedback contribute positively to evaluations of participants' use of grammar, style, and mechanics.

4.2.3. Writing dimensions that benefitted from spelling and grammar feedback

Although students' essays did not improve holistically with the availability of checking tools, several writing dimensions did appear to benefit. The addition of condition and the interaction terms improved model fit for *word choice, voice,* and *conclusion*, subscales (*organization* subscores showed a similar trend, but did not reach conventional levels of statistical significance). For all three of these subscales, analyzes of the best fit models (M3) revealed significant main effects of draft (participants score improved from initial draft to revision), but no main effect of feedback condition. This was qualified by significant draft by condition interaction indicating that the effect of draft depended on feedback condition. Simple slope analyses, with draft as the focal predictor and condition as the

moderator, indicated that those who had access to spelling and grammar checks tended to increase their subscales at revision (*word Choice*: Estimate = 0.22, SE = 0.04, $t(1283) = 5.54$, $p < .01$; *voice*: Estimate = 0.29, SE = 0.05, $t(1283) = 5.97$, $p < .001$; *conclusion*: Estimate = 0.38 SE = 0.06, $t(1283) = 6.07$, $p < .001$). In contrast, there were no statistically significant improvements for participants in the Strategy Condition (all Estimates < .09; $t < 2$). These findings further support the hypothesis (H2) that spelling and grammar feedback contribute positively to some aspects of writing quality. Notably, word choice and voice are primarily at the word level. Thus, it makes intuitive sense that these subscales would be influenced by the availability of spelling and grammar feedback.

4.2.4. Writing dimensions that did not benefit from spelling and grammar feedback

The availability of spelling and grammar feedback did not appear to influence the *Introduction*, *Body*, and *Sentence Structure* subscales. Students improved on these traits from initial to revised draft, but there was no significant effect of feedback condition. Although these specific findings support H3 (no effect), the broader implication of the mixed results for spelling and grammar feedback supports the conclusion that spelling and grammar feedback does not have uniform impact on essay writing. That is, spelling and grammar feedback supports some aspects of writing, but not all.

4.3. Examples of participant revising

The above analyses suggest that availability of spelling and grammar checking tools did not improve students' essay quality holistically or across all dimensions, but several traits appeared to benefit from these tools: word choice, voice, the quality of conclusion paragraphs, and essay organization. To more concretely illustrate how spelling and grammar feedback was (and was not) used by student writers, we provide below four authentic example essays from participants who received Strategy + SG feedback. These example essays were extracted from participants who demonstrated consistent improvements (e.g., larger mean gains across essays) on target dimensions during revising. Specifically, we targeted writers whose revisions tended to successfully improve word choice quality, voice, conclusion, or organization. These successful revisers were the most likely to exhibit use of the spelling and grammar tools (if at all). Importantly, these examples are not intended to provide a qualitative or mixed-method analysis (e.g., Creswell, 2013; Creswell & Clark, 2017; Saldaña, 2015) that derives generalizable themes or patterns across the dataset. Rather, our purpose was to provide meaningful examples of revising that complement the depersonalized, aggregate quantitative data.

All the example essays were written on the *Loyalty* prompt (i.e., "Should people always maintain their loyalties, or is it sometimes necessary to switch sides?") (see Appendix C). The *Loyalty* prompt was the fourth essay during the study. We selected this essay because participants were far enough along in the study to be sufficiently familiar with the system and tools. For each of the selected example essays ($n = 4$), we identified (a) changes from initial draft to revised draft and (b) use of spelling and grammar feedback for essays.

4.3.1. Participant A

Across all essays, Participant A tended to demonstrate meaningful average gains (i.e., increases from initial to revised draft) on the dimensions of *conclusion* (from *M*initial = 2.75 to *M*revised = 4.37) and *organization* (from *M*initial = 3.58 to *M*revised = 4.62). That is, when revising, Participant A tended to consistently improve upon the quality of concluding paragraphs and essay structure. On the example *Loyalty* essay, Participant A increased from a score of 1.5–4.5 on *conclusion* and from a score of 3.0–5.0 on *organization*. Appendix C provides the full text of the revised essay, with deletions indicated via strikethrough and additions indicated in bold.

The most substantive revisions were the elaboration of "the final reason" (i.e., additional support) and an entirely new concluding paragraph that summarized main ideas. Inspection of log data for the target essay revealed that Participant A accessed feedback on five potential spelling issues (i.e., "bihind," "happns," "strat," "hagout," and "br") along with suggested corrections (i.e., "behind," "happens," "start," "hangout," and "be"). The correction of "hagout" addressed a misspelled word that was retained from the initial draft. All other suggestions appeared in the new content added by the participant; and all corrections were accepted. The participant also accessed feedback on one possible grammatical issue (i.e., "Who cares if they laugh at you" was flagged as a possible interrogatory statement), which they disregarded. The most substantive contributions—the revisions that influenced human ratings of conclusion and organization—likely stemmed from the rhetorical changes rather than edits to spelling. However, spelling and grammar feedback enabled the participant to better communicate (i.e., fewer typos) these added ideas. Overall, this student seemed to leverage spelling and grammar feedback fairly minimally, and instead allocated revisions to larger structural revisions.

4.3.2. Participant B

Across essays, Participant B tended to consistently improve *word choice* (from *M*initial = 3.67 to *M*revised = 4.87) and essay *organization* (from *M*initial = 3.83 to *M*revised = 4.87) from initial draft to revision. On the target *Loyalty* essay, Participant B increased from a score of 4.0–5.0 on *word choice* and from a score of 4.0–4.5 on *organization*. Appendix C provides the full and annotated text of the revised essay.

Participant B implemented several revisions that potentially improved the precision or sophistication of wording within the essay. For instance, the phrase "a little crazy" was replaced with "very haughty," the vague term "they" was replaced with "the person on the receiving end," "unjust way" was replaced with "oppressed manner," and the colloquial word "ok" was replaced with "valid." When revising, log data showed that Participant B accessed feedback for only one spelling issue (i.e., "recieving") and one potential grammatical issue ("in an oppressed way"), both of which were addressed. These changes improved the wording of the essay but represent only two of the participants' many revisions. Thus, for Participant B, it was unclear whether the availability of spelling and grammar feedback had much impact on word choice or organization.

4.3.3. Participant C

Across essays, Participant C tended to improve essay *conclusion* (from *M*initial = 2.17 to *M*revised = 3.75) and *organization* (from *M*initial = 3.50 to *M*revised = 4.91) via revising. On the target *Loyalty* essay, Participant C increased from a score of 1.0–4.5 on *conclusion* and from a score of 3.0–5.5 on *organization*. Appendix C provides the full and annotated text of the revised essay.

Participant C made substantive revisions via elaboration (e.g., "This betrayal caused her mother to realize…" and "For example, they could know the location of…"), restructuring (e.g., repositioning sentences such as "This sense of nationalism…"), and the addition of a new concluding paragraph. Such revisions likely contributed positively to the improved subscale scores for essay organization and conclusion quality. On this essay, Participant C accessed no feedback on grammatical issues and only one issue for spelling (i.e., "brokeness"), which was resolved (i.e., "brokenness"). Thus, spelling and grammar feedback during revising appeared to contribute minimally to improvement of this essay for Participant C.

4.3.4. Participant D

Across essays, Participant D tended to improve essay *voice* (from *M*initial = 2.87 to *M*revised = 4.17) and *conclusion* (from *M*initial = 2.25 to *M*revised = 3.50) via revising. On the target *Loyalty* essay, Participant C increased from a score of 2.5–4.5 on *voice* and from a score of 2.0–4.5 on *conclusion*. Appendix C provides the full and annotated text of the revised essay.

This participant implemented a variety of relatively minor revisions that improve clarity, such as correcting "No matter to who is loyalty you or not you or not you should always be loyal" to "No matter who is loyal to you or not you should always be loyal." The participant also refined wording, such as replacing "will see it in the future" to "will affect your future," and replacing "right decisions" with "correct decisions." Other revisions were mechanical, such as correcting capitalization (e.g., from "i" to "I") or repairing sentence fragments (e.g., from "Doesn't have to be…" to "It doesn't have to be…"). In the conclusion, the participant added a restated thesis with an exclamation point. Notably, the participant never accessed grammar feedback during revision and accessed spelling feedback only once to correct "loyaly" to "loyalty." That is, this participant made additional spelling and grammar corrections that were not directly offered by the system. Thus, although Participant D revised the essay for spelling, grammar, and other conventions, these changes did not appear to be directly elicited by the spelling and grammar checker.

4.3.5. Summary

Overall, these essays exemplify how spelling and grammar tools were used infrequently and contributed to incremental improvements (e.g., clarity and mechanical correctness), yet were not associated with substantive gains in holistic writing quality. These illustrative examples help to visualize how the availability of spelling and grammar feedback was neither harmful nor strongly useful.

5. Discussion

Building upon the utility of word processing programs (e.g., Bangert-Drowns, 1993; Morphy & Graham, 2012), modern automated writing evaluation (AWE) systems provide computer-based support for multiple aspects of writing. Using NLP-based algorithms, AWEs can assess numerous lexical, syntactic, organization, semantic, and rhetorical features of essays to evaluate overall quality and detect potential problems (Allen et al., 2016; Strobl et al., 2019). In turn, these assessments can yield actionable recommendations and strategies that learners can use to improve their writing (Authors, xxxx). This is important, given that prior research has established that formative strategy feedback is one of the most powerful components of writing instruction and writer development (e.g., Graham & Perin, 2007; Parr & Timperley, 2010).

However, one assumption inherited from word processors is that AWE tools can and should assess and provide feedback on spelling and grammar. Spelling and grammar checkers are popular, and many educators and participants expect this functionality from AWE. Often, instructors and designers assume that any opportunity to deliver feedback should be leveraged. However, it is also possible that additional feedback can result in suboptimal effects (e.g., Authors, xxxx). Thus, experimental evaluations of new features represent an important aspect of effective design and development of AWEs and other computer-based learning environments (Authors, xxxx). Although automated spelling and grammar error detection is relatively easy to implement, it was important for us to demonstrate that adding this feedback would not dampen or nullify the benefits of the existing AWE system.

This study explored the effects of incorporating spelling and grammar checking tools within an existing AWE (i.e., W-Pal) that provides formative strategy feedback (see Authors, xxxx). Due to the importance of strategy feedback for good writing, all conditions in our study included strategy feedback. The lack of a "no feedback" control limits the ability to discuss the magnitude of the effect of strategy feedback on its own, but such effects have been well-documented in prior literature. Prior writing research shaped three competing hypotheses. First, spelling and grammar feedback might be *detrimental* (H1) by guiding participants' attention away from conceptual aspects of writing and reinforcing their tendency to focus on superficial edits (Crawford et al., 2008; Fitzgerald, 1987). Second, spelling and grammar feedback could be *beneficial* (H2) by helping participants correct superficial errors quickly, thus enabling or motivating them to expend more effort on deeper concerns (see Graham & Santangelo, 2014; Morphy & Graham, 2012). Finally, spelling and grammar feedback might have *no effect*—contributing little to ratings of writing quality (see Crossley et al., 2014). Importantly, we explored these possibilities across not only the holistic scores, but also in terms of the subscores representing various aspects of essay quality. Overall, our results suggest some modest benefits of spelling and grammar feedback (H2), but largely no effect (H3) of spelling and grammar feedback.

Notably, there was no evidence that spelling and grammar feedback decreased essay quality (H1). More specifically, the availability of spelling and grammar checking tools did not significantly improve holistic essay quality nor did it improve several more fine-grained dimensions of essay quality (i.e., introduction paragraph quality, body paragraph quality, and sentence structure). However, analyses

observed that spelling and grammar feedback appeared to modestly improve writing with respect to word choice, voice, and conclusion paragraphs, but only on the second draft.

Our results indicate that correcting a few spelling and grammar errors had little impact on overall essay quality. This null result is unsurprising but emphasizes the need to focus on strategy feedback rather than spelling and grammar correction. However, our study also revealed that the addition of spelling and grammar feedback did not detract from essay quality. Students' overall essay quality improved in both conditions. Thus, potential concerns that students might be overwhelmed and reject the feedback, or that they might become hyper-focused on less impactful revisions, were unfounded.

It is likely that the effects of spelling and grammar feedback on subscores (e.g., conclusion and organization) was indirect. Given that students tend to focus their revisions on mechanical errors, by quickly identifying and resolving these issues, students had more time to respond to other aspects of the feedback. For example, less skilled writers tend to skip past an extended planning phase and instead engage in *knowledge telling* in which they transcribe their ideas as they go in relatively linear fashion (Bereiter, Burtis, & Scardamalia, 1988). This would suggest that conclusion scores were weaker than introduction and body paragraph scores because they had less time to craft this section in this timed task. In the 10-minute revision time, students with spelling and grammar feedback were able to easily resolve spelling and grammar issues, leaving them more time to increase the length and quality of their concluding paragraph. Indeed, this assumption is supported by the qualitative inspections.

As observed across a few example essays, limited use of spelling and grammar checking tools helped participants produce more technically correct writing (i.e., fewer typos), but the most substantive revisions (e.g., added elaboration and arguments) seemed independent of the spelling and grammar support. Notably, actionable recommendations for improving essay structure, communication of ideas, cohesion, word choice, and so on are commonly provided by W-Pal formative feedback messages (see Table B2). These example essays suggest that the participants were using the formative strategy feedback to make more substantive changes above and beyond the recommendations made by the spelling and grammar checker.

Overall, such findings are most consistent with existing work demonstrating that spelling and grammar feedback are not detrimental, but that this mechanical feedback influences only a few aspects of writing (e.g., Kellogg, Whiteford, & Quinlan, 2010; Rock, 2007).

5.1. Implications

Our study revealed that spelling and grammar feedback on its own had no significant effects on essay quality (i.e., on the first draft; see Section 4.1.2). Although this finding is not surprising given the extant literature on writing, it runs counter to many instructors' intuitions that quality writing hinges on mechanics. Our findings provide further evidence in support of formative strategy feedback.

Our log data suggested that students used the spelling and grammar tools, but not to great effect. It may be the case that students know how to navigate to use these tools, but they may not have the knowledge or skills to be able to use the feedback effectively. Thus, one consideration may be for AWE systems to provide instruction on how to best leverage spelling and grammar feedback along with their instruction on more sophisticated writing strategies. It is also likely that writers differentially seek out and leverage different types of feedback. For example, Hazelton et al. (2021) found that less confident writers tended to rely on grammar tools more than their more confident peers. Thus, the extent to which feedback is sought out and used is likely related to individual differences in skills, motivations, and attitudes. A more comprehensive understanding of the impact of various types of feedback requires considering both prior skills as well as how the feedback is perceived and implemented.

Notably, the findings do not discount the potential benefits of spelling and grammar feedback. The availability of spelling and grammar feedback *in addition to* writing strategy feedback showed no negative effects and, indeed, some modest benefits. Indeed, our findings highlight that quality writing involves mastery of both *content* and the *language* through which that content is conveyed. Developing writers are likely to need assistance with both. While the need to develop more foundational spelling and grammar knowledge and skills as well as development of knowledge of structure and content in tandem is likely apparent to writing researchers, it is not always obvious to instructors and students. The current works represents part of a growing body of research aimed at a deeper understanding of how developing writers use feedback. These studies can help instructors (and AWEs) to deliver better just-in-time support that can help students to interweave quality content with good mechanics.

5.2. Future work

The current study was aligned with prior W-Pal research (e.g., Authors, xxxx; Proske et al., 2014). Thus, the findings are constrained by the functionalities of W-Pal, such as how it permits writing and revising and how it delivers feedback. For example, W-Pal's default setting, used in this study, is to give 25 min to author an initial draft and 10 min to revise. Participants were unable to submit their essays before this time had elapsed. These durations and restrictions were implemented to control time-on-task (e.g., prevent rushing to finish). However, given the impact of self-regulation on writing (Kellogg, 2008; Kellogg & Raulerson, 2007; Santangelo et al., 2016), it is plausible that writing or revision behaviors might differ if participants governed their own writing time. Indeed, a long history of research indicates that students do not often take full advantage of revision opportunities (Attali, 2004; Faigley & Witte, 1981), but students can and do make meaningful additions when encouraged to revise (Authors, xxxx). W-Pal also delivers strategy

feedback only once the participant has submitted their essay, rather than during composition—a feature that appears in several other AWE systems (see Strobl et al., 2019). Thus, spelling and grammar feedback was deployed immediately, whereas strategy feedback was delayed. It would be of value to explore how changing the temporality of these two types of feedback, such as waiting to deliver spelling and grammar until after writing is complete, deploying strategy feedback in-the-moment, or waiting to provide spelling and grammar feedback until after larger structural and content issues have already been addressed (e.g., Koltovskaia, 2020) might impact writing quality.

It is also worth noting that W-Pal feedback targets writing strategies that are communicated in the tutorial video lessons and practice games. Participants in this study did not receive this writing strategy *tutorial instruction* that is part of the larger W-Pal tutoring system. Future work will examine how spelling and grammar feedback is employed in the context of strategy training.

It is also important to note that the strategies highlighted in W-Pal reflect only a subset of the possible types of writing feedback. W-Pal is one of many AWE systems. As outlined previously, different AWEs provide feedback on a variety of aspects of writing, some of which are captured in W-Pal and some of which are not. Thus, one clear direction for future comparative research is to evaluate whether spelling and grammar checking tools perform differently in AWE systems that employ different styles of feedback. Examining these effects in other AWEs will help to assess the generalizability of these findings as well as potentially important boundary conditions. Indeed, the sample in the current study also limits the degree to which broad generalizations can be made. Our study includes adolescents who are predominantly native English speakers. Preliminary analyses suggested that the non-native English speakers produced essays of similar quality, save for the grammar, spelling, and mechanics subscore. However, our sample is too small to investigate this more deeply. The finding suggests that further work should be done to examine the combination of both types of feedback for those who have more difficulty with more foundational writing skills including L2 writers and younger, developing writers.

The present study examined how spelling and grammar feedback influenced essay writing and revision in the context of a lab-based setting wherein students were provided automated feedback. Subsequently, evaluations were provided by expert raters using an established rubric. It will be of value in future work to explore these effects across other forms of essay evaluation. As mentioned previously, non-expert raters appear to value different features of essays than their expert counterparts (e.g., Crossley et al., 2014; Johnson et al., 2017). Students also attend to different features than teachers when evaluating essay quality (Authors, xxxx). Although AWEs were built to emulate teacher feedback, there is evidence that teacher evaluation and feedback can be both qualitatively and quantitatively different from feedback provided by AWEs (e.g., Dikli & Bleyle, 2014). Thus, the effects of automated spelling and grammar feedback may differentially influence essay scores depending on the nature of the evaluation. Future work should examine these effects in the context of the classroom and with ecologically-appropriate evaluations.

Finally, our sample was predominantly native English-speaking. Although there is an abundance of work on spelling and grammar feedback in the context of L2/ESL writers, there is much less work specifically examining the combination of spelling and grammar feedback with strategy feedback on L2/ESL writers' essay quality. Additional studies with larger L2/ESL samples must be conducted to examine this more directly. Such comparisons will contribute to our theoretical understanding of writing, and also help researchers and instructors to better tailor feedback to individual writers.

5.3. Conclusion

Although spelling and grammar feedback tools are a ubiquitous part of our writing experiences, prior research on writing instruction and assessment offered conflicting hypotheses regarding the potential effects of providing spelling and grammar feedback in concert with formative strategy feedback. Our findings suggest modest or minimal benefits of spelling and grammar feedback—formative writing strategy feedback remains the important feedback support that we should provide to participant writers.

The rapid growth of AI-driven feedback shows that computers can do much more than simply detect typographical errors and spelling mistakes. However, it may be unwise to take this mechanical feedback for granted. Given the potential benefits of combining spelling and grammar feedback in concert with strategy feedback, we suggest possible benefits of including instruction in how to be more strategic and mindful of feedback on these errors in conjunction with instruction about deeper aspects of writing.

Author Note

This research was supported by grants from the U.S. DoEd Institute of Education Sciences (R305A120707 and R305A180261) and the U.S. DoD Office of Naval Research (N000141712300). Opinions, findings, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding sources.

Acknowledgements

This research was supported by grants from the [blind] and the [blind]. Opinions, findings, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding sources.

Appendix A. Prompt instructions and essay prompts

Prompt Instructions: For each prompt, participants were given the following instructions: "You will now have 25 min to write an essay on the prompt below. The essay gives you an opportunity to show how effectively you can develop and express ideas. You should,

therefore, take care to develop your point of view, present your ideas logically and clearly, and use language precisely. Think carefully about the issue presented in the following excerpt and the assignment below. Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations."

Essay order	Prompt title	Prompt question
1	Images and Impressions	All around us appearances are mistaken for reality. Clever advertisements create favorable impressions but say little or nothing about the products they promote. In stores, colorful packages are often better than their contents. In the media, how certain entertainers, politicians, and other public figures appear is sometimes considered more important than their abilities. All too often, what we think we see becomes far more important than what really is. Do images and impressions have a positive or negative effect on people?
2	Competition and Cooperation	While some people promote competition as the only way to achieve success, others emphasize the power of cooperation. Intense rivalry at work or play or engaging in competition involving ideas or skills may indeed drive people either to avoid failure or to achieve important victories. In a complex world, however, cooperation is much more likely to produce significant, lasting accomplishments. Do people achieve more success by cooperation or by competition?
3	Winning	From talent contests to the Olympics to the Nobel and Pulitzer prizes, we constantly seek to reward those who are "number one." This emphasis on recognizing the winner creates the impression that other competitors, despite working hard and well, have lost. In many cases, however, the difference between the winner and the losers is slight. The wrong person may even be selected as the winner. Awards and prizes merely distract us from valuable qualities possessed by others besides the winners. Do people place too much emphasis on winning?
4	Loyalty	Loyalty is one of the essential attributes a person must have and must demand of others. Being loyal, faithful, or dedicated to someone or something, is not always easy. People often have conflicting loyalties, and there are no guidelines that help them decide to what or to whom they should be loyal. Moreover, people may be loyal to something harmful or bad. Should people always maintain their loyalties, or is it sometimes necessary to switch sides?
5	Patience	When we are young, we learn from parents and teachers that we should wait patiently for what we want. Few people would dispute the wisdom or truth of this teaching. Our society, however, with its mad rush and hurry and its insistence on instant gratification and quick responses, encourages and rewards impatience. Experience teaches us that we should not and do not have to wait. Is it better for people to act quickly and expect quick responses from others rather than to wait patiently for what they want?
6	Memories	Many persons believe that to move up the ladder of success and achievement, they must forget their past, repress it, and let it go. But others have just the opposite view. They see their old memories as a chance to reckon with their past and integrate past and present. Do personal memories hinder or help people in their effort to learn from their past and succeed in the present?

Appendix B. Complete models for holistic scores and analytic sub-scores

See Tables B1–B9.

Table B1
Linear mixed effect model predicting introduction score.

	Holistic score			
	M0		M1	
	B	CI	B	CI
Fixed Parts				
(Intercept)	2.33***	2.04–2.61	2.26***	1.97 to 2.54
Images	-0.16**	-0.26 to − 0.06	-0.16**	-0.26 to − 0.06
Loyalty	-0.10*	-0.20 to − 0.00	-0.10*	-0.20 to − 0.00
Memories	-0.04	-0.14 to 0.06	-0.04	-0.14 to 0.06
Patience	-0.05	-0.15 to 0.05	-0.05	-0.15 to 0.04
Winning	-0.03	-0.13 to 0.07	-0.03	-0.13 to 0.07
Reading Skill (GMRT)	2.13***	1.68–2.58	2.13***	1.68–2.58
Draft			0.14***	0.09–0.20
Random Parts				
NID	119		119	
ICCID	0.427		0.432	
Observations	1410		1410	
R2/Ω02	.608/.607		.616/.614	

Notes GMRT = Gates-MacGinitie Reading Test; B = beta weight; CI = confidence interval; * $p < .05$ ** $p < .01$ *** $p < .001$.

Table B2
Linear mixed effect model predicting introduction score.

	Introduction score			
	M0		M1	
	B	CI	B	CI
Fixed Parts				
(Intercept)	2.66***	2.42–2.91	2.62***	2.37–2.87
Images	-0.21***	-0.32 to − 0.10	-0.21***	-0.32 to − 0.10
Loyalty	0.04	-0.07 to 0.15	0.04	-0.07 to 0.15
Memories	0.15**	0.04–0.26	0.15**	0.04–0.26
Patience	0.02	-0.09 to 0.13	0.02	-0.09 to 0.13
Winning	0.03	-0.08 to 0.14	0.03	-0.08 to 0.14
Reading Skill (GMRT)	1.93***	1.54–2.32	1.93***	1.54–2.31
Draft			0.10**	0.03–0.16
Random Parts				
N_{ID}	119		119	
ICC_{ID}	0.290		0.292	
Observations	1410		1410	
$R^2/\Omega_0{}^2$	.498/.494		.501/.498	

Notes GMRT = Gates-MacGinitie Reading Test; B = beta weight; CI = confidence interval; * p < .05 ** p < .01 *** p < .001.

Table B3
Linear mixed effect model predicting body score.

	Body score			
	M0		M1	
	B	CI	B	CI
Fixed Parts				
(Intercept)	2.78***	2.52–3.04	2.74***	2.48–3.00
Images	-0.10	-0.20 to 0.00	-0.10	-0.20 to 0.00
Loyalty	-0.22***	-0.32 to − 0.11	-0.22***	-0.32 to − 0.11
Memories	-0.18***	-0.29 to − 0.08	-0.18***	-0.29 to − 0.08
Patience	-0.08	-0.18 to 0.03	-0.08	-0.18 to 0.03
Winning	-0.11*	-0.21 to − 0.00	-0.11*	-0.21 to − 0.00
Reading Skill (GMRT)	1.84***	1.44–2.25	1.84***	1.43–2.24
Draft			0.08*	0.02–0.14
Random Parts				
NID	119		119	
ICCID	0.346		0.348	
Observations	1410		1410	
R2/Ω02	.527/.524		.529/.526	

Notes GMRT = Gates-MacGinitie Reading Test; B = beta weight; CI = confidence interval; * p < .05 ** p < .01 *** p < .001.

Table B4
Linear mixed effect model predicting conclusion score.

	Conclusion score					
	M0		M1		M2	
	B	CI	B	CI	B	CI
Fixed Parts						
(Intercept)	2.16***	1.80–2.52	2.05***	1.69–2.41	2.02***	1.55–2.50
Images	-0.31***	-0.46 to − 0.16	-0.31***	-0.46 to − 0.16	-0.31***	-0.46 to − 0.16
Loyalty	-0.00	-0.15 to 0.15	-0.00	-0.15 to 0.15	-0.00	-0.15 to 0.15
Memories	0.13	-0.03 to 0.28	0.13	-0.02 to 0.27	0.12	-0.02 to 0.27
Patience	0.02	-0.13 to 0.17	0.02	-0.13 to 0.17	0.02	-0.13 to 0.17
Winning	0.09	-0.06 to 0.24	0.09	-0.06 to 0.24	0.09	-0.06 to 0.24
Reading Skill (GMRT)	1.99***	1.42–2.55	1.98***	1.42–2.55	1.74***	0.94–2.53
Draft			0.23***	0.14–0.31	0.38***	0.26–0.50
Cond					0.12	-0.59 to 0.84
Draft * Cond					-0.30***	-0.47 to − 0.13
GMRT * Cond					0.34	-0.82 to 1.49
Random Parts						
NID	119		119		119	
ICCID	0.325		0.330		0.331	
Observations	1410		1410		1410	
R2/Ω02	.467/.463		.478/.474		.483/.479	

Notes GMRT = Gates-MacGinitie Reading Test; B = beta weight; CI = confidence interval; * p < .05 ** p < .01 *** p < .001.

Table B5
Linear mixed effect model predicting grammar, style, & mechanics score.

	Grammar score					
	M0		M1		M2	
	B	CI	B	CI	B	CI
Fixed Parts						
(Intercept)	2.66***	2.41–2.92	2.61***	2.36–2.87	2.67***	2.35–3.00
Images	-0.13**	-0.22 to − 0.04	-0.13**	-0.22 to − 0.04	-0.13**	-0.22 to − 0.04
Loyalty	-0.09*	-0.18 to − 0.00	-0.09*	-0.18 to − 0.00	-0.09*	-0.18 to − 0.00
Memories	-0.06	-0.14 to 0.03	-0.06	-0.14 to 0.03	-0.06	-0.14 to 0.03
Patience	-0.13**	-0.21 to − 0.04	-0.13**	-0.21 to − 0.04	-0.13**	-0.21 to − 0.04
Winning	-0.13**	-0.21 to − 0.04	-0.13**	-0.21 to − 0.04	-0.13**	-0.21 to − 0.04
Reading Skill (GMRT)	1.72***	1.31–2.12	1.71***	1.31–2.12	1.70***	1.14–2.26
Draft			0.10***	0.05–0.15	0.18***	0.11–0.25
Cond					-0.23	-0.74 to 0.27
Draft * Cond					-0.16**	-0.26 to − 0.06
GMRT * Cond					0.22	-0.59 to 1.03
Random Parts						
NID	119		119		119	
ICCID	0.427		0.429		0.423	
Observations	1410		1410		1410	
R2/Ω02	.589/.587		.594/.592		.597/.595	

Notes GMRT = Gates-MacGinitie Reading Test; B = beta weight; CI = confidence interval; * p < .05 ** p < .01 *** p < .001.

Table B6
Linear mixed effect model predicting organization score.

	Organization score					
	M0		M1		M2	
	B	CI	B	CI	B	CI
Fixed Parts						
(Intercept)	2.85***	2.58–3.13	2.78***	2.51–3.05	2.75***	2.39–3.11
Images	-0.22***	-0.33 to − 0.10	-0.22***	-0.33 to − 0.10	-0.22***	-0.33 to − 0.10
Loyalty	-0.04	-0.15 to 0.08	-0.04	-0.15 to 0.08	-0.04	-0.15 to 0.08
Memories	-0.07	-0.19 to 0.04	-0.07	-0.19 to 0.04	-0.07	-0.19 to 0.04
Patience	-0.04	-0.15 to 0.08	-0.04	-0.15 to 0.07	-0.04	-0.15 to 0.07
Winning	-0.00	-0.12 to 0.11	-0.00	-0.12 to 0.11	-0.00	-0.11 to 0.11
Reading Skill (GMRT)	1.67***	1.25–2.10	1.67***	1.24–2.10	1.55***	0.95–2.15
Draft			0.15***	0.09–0.22	0.23***	0.14–0.32
Cond					0.12	-0.42 to 0.66
Draft * Cond					-0.16*	-0.29 to − 0.03
GMRT * Cond					0.13	-0.74 to 1.01
Random Parts						
NID	119		119		119	
ICCID	0.324		0.328		0.330	
Observations	1410		1410		1410	
R2/Ω02	.480/.476		.488/.485		.491/.487	

Notes GMRT = Gates-MacGinitie Reading Test; B = beta weight; CI = confidence interval; * p < .05 ** p < .01 *** p < .001.

Table B7
Linear mixed effect model predicting sentence structure score.

	Sentence structure score					
	M0		M1		M2	
	B	CI	B	CI	B	CI
Fixed Parts						
(Intercept)	2.58***	2.36–2.81	2.54***	2.31–2.77	2.59***	2.29–2.90
Images	-0.04	-0.14 to 0.05	-0.04	-0.13 to 0.05	-0.04	-0.13 to 0.05
Loyalty	-0.07	-0.16 to 0.03	-0.07	-0.16 to 0.03	-0.07	-0.16 to 0.03
Memories	0.04	-0.06 to 0.13	0.04	-0.06 to 0.13	0.04	-0.06 to 0.13
Patience	0.00	-0.09 to 0.10	0.00	-0.09 to 0.10	0.00	-0.09 to 0.10
Winning	-0.01	-0.11 to 0.08	-0.01	-0.11 to 0.08	-0.01	-0.11 to 0.08
Reading Skill (GMRT)	1.78***	1.42–2.14	1.78***	1.42–2.14	1.76***	1.25–2.27
Draft			0.09**	0.03–0.14	0.04	-0.03 to 0.12
Cond					-0.13	-0.59 to 0.33
Draft * Cond					0.09	-0.02 to 0.20
GMRT * Cond					0.08	-0.66 to 0.82
Random Parts						
NID	119		119		119	
ICCID	0.329		0.331		0.335	
Observations	1410		1410		1410	
R2/Ω02	.527/.524		.530/.527		.531/.528	

Notes GMRT = Gates-MacGinitie Reading Test; B = beta weight; CI = confidence interval; * $p < .05$ ** $p < .01$ *** $p < .001$.

Table B8
Linear mixed effect model predicting voice score.

	Voice					
	M0		M1		M2	
	B	CI	B	CI	B	CI
Fixed Parts						
(Intercept)	2.88***	2.65–3.11	2.79***	2.56–3.02	2.81***	2.51–3.11
Images	-0.08	-0.20 to 0.04	-0.08	-0.20 to 0.04	-0.08	-0.19 to 0.04
Loyalty	-0.05	-0.17 to 0.06	-0.05	-0.17 to 0.06	-0.05	-0.17 to 0.06
Memories	0.10	-0.02 to 0.22	0.10	-0.02 to 0.22	0.10	-0.02 to 0.21
Patience	0.08	-0.04 to 0.20	0.08	-0.04 to 0.19	0.08	-0.04 to 0.19
Winning	0.05	-0.06 to 0.17	0.05	-0.06 to 0.17	0.05	-0.06 to 0.17
Reading Skill (GMRT)	1.52***	1.16–1.88	1.52***	1.16–1.88	1.44***	0.94–1.95
Draft			0.18***	0.11–0.25	0.29***	0.19–0.38
Cond					-0.09	-0.54 to 0.37
Draft * Cond					-0.22**	-0.35 to − 0.09
GMRT * Cond					0.23	-0.50 to 0.96
Random Parts						
NID	119		119		119	
ICCID	0.224		0.228		0.231	
Observations	1410		1410		1410	
R2/Ω02	.389/.383		.402/.396		.407/.402	

Notes GMRT = Gates-MacGinitie Reading Test; B = beta weight; CI = confidence interval; * $p < .05$ ** $p < .01$ *** $p < .001$.

Table B9
Linear mixed effect model predicting word choice.

	Word choice score					
	M0		M1		M2	
	B	CI	B	CI	B	CI
Fixed Parts						
(Intercept)	2.79***	2.58–3.00	2.72***	2.51–2.94	2.77***	2.49–3.05
Images	-0.02	-0.12 to 0.07	-0.02	-0.11 to 0.07	-0.02	-0.11 to 0.07
Loyalty	-0.07	-0.16 to 0.02	-0.07	-0.16 to 0.02	-0.07	-0.16 to 0.02
Memories	0.01	-0.09 to 0.10	0.01	-0.08 to 0.10	0.01	-0.08 to 0.10
Patience	-0.07	-0.17 to 0.02	-0.07	-0.17 to 0.02	-0.07	-0.17 to 0.02
Winning	-0.02	-0.12 to 0.07	-0.02	-0.11 to 0.07	-0.02	-0.11 to 0.07
Reading Skill (GMRT)	1.80***	1.47–2.13	1.80***	1.46–2.13	1.70***	1.22–2.17
Draft			0.14***	0.09–0.19	0.21***	0.14–0.29
Cond					-0.16	-0.58 to 0.26
Draft * Cond					-0.15**	-0.25 to − 0.04
GMRT * Cond					0.28	-0.40 to 0.97
Random Parts						
NID	119		119		119	
ICCID	0.304		0.309		0.311	
Observations	1410		1410		1410	
R2/Ω02	.518/.515		.528/.525		.531/.528	

Notes GMRT = Gates-MacGinitie Reading Test; B = beta weight; CI = confidence interval; * $p < .05$ ** $p < .01$ *** $p < .001$.

Appendix C. Revised essays from example participants

Participant a revised essay on "loyalty"

Loyalty is something that everyone should have right? But sometimes you get betrayed by your best friend and you think the best thing to do is to not be loyal to them anymore. However that is not true; sometimes betrayal is the right thing to do because you might find something that they might like to do things that they force you get dragged into, they could be forcing you to do something, and you might find out that they are talking behind your back and leave them.

To start off, your buddy might like to do things that you get dragged into. For example, they might like to go on roller coasters and you don't but you have to because you are scared that they will make fun of you. I used to be scared that my friends would make fun of me because I didn't really like to do the things they did or didn't like the same things I liked. Most of the time I would just stay quiet while my friends talked their lives away. Then I saw some kids bring toys and I saw that I had some of them and I was no longer afraid to show what I liked to do.

The next reason is that they might force you to do things you don't want to do. Sometimes friends may be abusive and this is when you turn the other way. Just say no and if they keep pushing then don't **hangout** with them anymore. There was a time when somebody told me to play with them and then they told me to do something bad and I told them no. They kept asking me whenever I came close to them so I just stopped going towards them and I have never seen them again.

The final reason is that they could be talking behind your back**. Sometimes after you reject someone, they might start talking behind you back or telling people bad things about you. If this ever happens, don't do the same thing. Just ignore and if someone comes up to you with something bad about you, prove them wrong and move on with your own life.**

In conclusion, don't be pushed around by your peers. If you don't like something that they do and you don't, just tell them and leave. Who cares if they laugh at you. It is your own life to live and nobody's business. Loyalty is great but sometimes you have to go the other way because you may not like something they do, they could be forcing you into things, and they might be talking behind your back.

Participant B revised essay on "loyalty"

Loyalty is one of the main character traits people look for in a companion. It is what makes friendships and relationships last so long. It is what makes a person feel comfortable with others, and with loyalty usually comes trust as well. But, in some instances loyalty is hard to maintain. When a friend has wronged their other friend, when **a person makes** choices others do not agree with and when people drift away from each other. In those cases, it is easy to lose sight of being loyal.

People have fought with their friends or significant others since the dawn of time. But sometimes these disputes ~~just~~ go too far. If a person has back stabbed their friend and still expects complete forgiveness, they must be a**very haughty**. If a person has continuously **told lies** to their significant other, the odds are **the person on the receiving** will leave the picture. People do not enjoy being treated poorly or in an **oppressed manner. People** will not stand for that at all. These types of events create ripples in relationships that are sometimes more deep than trying to stay loyal to someone.

In other cases, the loyalty dispute is one sided. If **someone** has a friend that starts to pick up habits that they believe are wrong, it is hard to continue to be in their lives. For example, if a friend started abusing drugs it would be hard to stay with them especially if they

are refusing help. Just like in dating, there are certain deal breakers people have that make them question their loyalty **in friendships.** If the action a friend is doing puts someone in a compromising situation or in any sort of danger, it is **valid** to switch sides and leave them. Sometimes it is about looking after one self rather than somebody else.

A lot of the time people lose interest in each other.They do not start relationships with that thought in mind, but it does happen. It is human nature. Whether it is because the duo went to different schools, got new jobs, or just made new friends it does happen. When this occurs, it is easy to **forget** the friendships **- and relationships and move on. People get more concerned** with themselves and **put others on** the **back burner causing them to** drift away. In turn, they lose the feelings and thoughts they once had for the other person, and it is easy to not stay loyal.

Loyalty is a touchy subject for most people. It is a trait that someone either has or does not **acquire**. Although, it is important to stay loyal to the ones that **are** close to someone, it is understandable to switch sides occasionally. If someone has created problems for them, back stabbed them or just grown apart, it is valid to not be loyal anymore.

Participant C revised essay on "loyalty"

Whether it is being loyal to a specific coffee shop or to the deepest secrets of one's company, loyalty can be a difficult attribute to maintain. The aspects of loyalty are complex; therefore, many individuals have a difficult time staying loyal. Sometimes it is necessary to switch sides, so people should not be held accountable for not being loyal.

Exclusivity is a prominent component of the majority of successful, long-term marriages. Often, a spouse's interests and personality alters. When this change is so significant that one spouse feels unfamiliar with the other, divorce occurs and the mutual loyalty that once was is shattered. For example, when I was in eighth grade, my best friend's parents began arguing. The disagreements escalated until they ultimately divorced. In the end, I was informed that the reason they separated was because her father was having an affair. **This betrayal caused her mother to realize the brokenness in her husband.** When loyalty is not earned or deserved, it is acceptable and necessary to switch sides.

During the time of kings and queens, loyalty to a certain religion meant life or death. Unlike America's limitless freedoms, people could not freely worship what they pleased. In Spain, Catholicism was prominent, as apposed to paganism and Christianity. Groups formed secretly to pray and worship against orders of the royal court. They could be named traitors of the king **and be beheaded**, if exposed. If someone truly believes that their spiritual loyalty lies elsewhere, switching sides is important to their moral being.

The United States Military, FBI, and CIA are known for their stellar commitment and **unbreakable loyalty. This sense of nationalism can cost them their lives.** If a soldier or agent is kidnaped as a prisoner of war, it is critical for them to maintain loyalty for the sake of their country. **For example, they could know the location** of **- the commander in chief or the codes to a sealed file.** It is necessary for them to break their loyalty to improve the quality of their well-being **and prevent their demise.**

The complexities of loyalty goes beyond ";right"; or ";wrong";. People can have conflicting loyalties and switch sides if the circumstances call for it. Faith for a religion, country, or person can change, because nothing is truthfully constant.

Participant D revised essay on "loyalty"

People should always be loyal. That is how you build trust. I wouldn't want to be friends with someone that isn't loyal. No matter who is **loyal to** you or not you should always be loyal. Loyal can be something huge **and** hard to keep. But if you do keep loyalty trust me that will cause huge positive impacts in your life. As well as saying good things about you. People will always look at you as an admiring person as well. Who doesn't want to be admired. I believe everyone at some point wants to be admired. This includes me I would want to be admired maybe not now but in the future. But it all depends with my loyalty now and in the present. What you do now everyone will**affect your** future. The past reflects in the future a lot. **It doesn't** have to be right away but trust me it **will**.

I always admire loyal people. Loyal people are smart because they choose to be loyal and that is a great thing to do. I **have** to accept **i I** think **i I** am pretty loyal personally. Even if **I** wasn't loyal **I** would probably think of the consequences. I am pretty good at choosing things like being loyal. That is why I'd rather stay loyal. Plus good consequences come from **correct** decisions. **Even** if you choose to be loyal **you will** have a positive future. No matter what **I** will always stay loyal. That is like the best decision to make.

People always wants to be surrounded by loyal people. Nobody wants someone fake near them including me **i I** want someone positive in my life. The not loyal people can cause harm and **I** don't want that. I want the positive things by my side only which is loyalty **known as loyal people**. The not loyal people can cause bad things. Plus it isn't that hard to be loyal to everyone. The hard thing in being loyal is commitment. Don't take the risk on people and surround yourself with loyal understanding people. **People should maintain loyalty!**.

References

Allen, L. K., Snow, E. L., Crossley, S. A., Jackson, G. T., & McNamara, D. S. (2014). Reading comprehension components and their relation to the writing process. *L'année psychologique/Topics in Cognitive Psychology, 114*, 663–691.

Allen, L. K., Likens, A. D., & McNamara, D. S. (2019). Writing flexibility in argumentative essays: a multidimensional analysis. *Reading and Writing, 32*(6), 1607–1634.

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 316–329). New York: The Guilford Press.

Allen, L. K., & Perret, C. A. (2016). Commercialized writing systems. In Adaptive educational technologies for literacy instruction (pp. 145-162). Routledge.

Attali, Y. (2004). Exploring the feedback and revision features of criterion. April *Paper presented at the national council on measurement in education in San Diego, CA*.
Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3).
Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology, 37*(1), 67–81. https://doi.org/10.1080/01443410.2016.1223275
Bangert-Drowns, R. L. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational Research, 63*(1), 69–93. https://doi.org/10.3102/00346543063001069
Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7. ⟨https://doi.org/10.18637/jss.v067.i01⟩.
Bereiter, C., Burtis, P. J., & Scardamalia, M. (1988). Cognitive operations in constructing main points in written composition. *Journal of Memory and Language, 27*(3), 261–278.
Boland, J. E., & Queen, R. (2016). If you're house is still available, send me an email: Personality influences reactions to written errors in email messages. *PLoS One, 11*(3), Article e0149885. https://doi.org/10.1371/journal.pone.0149885
Butler, J. A., & Britt, M. A. (2011). Investigating instruction for improving revision of argumentative essays. *Written Communication, 28*(1), 70–96. https://doi.org/10.1177/0741088310387891
Cavaleri, M. R., & Dianati, S. (2016). You want me to check your grammar again? The usefulness of an online grammar checker as perceived by participants. *Journal of Academic Language and Learning, 10*(1), A223–A236.
Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing, 27*(3), 419–436. https://doi.org/10.1177/0265532210364391
Crawford, L., Lloyd, S., & Knoth, K. (2008). Analysis of participant revisions on a state writing test. *Assessment for Effective Intervention, 33*(2), 108–119. https://doi.org/10.1177/1534508407311403
Creswell, J. (2013). *Qualitative inquiry and research design: Five different approaches*. Los Angeles, CA: Sage.
Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research* (3rd ed.). Los Angeles, CA: Sage.
Crossley, S., & McNamara, D. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. *In Proceedings of the Annual Meeting of the Cognitive Science Society, 32*, 32.
Crossley, S. A., Kyle, K., Allen, L. K., & McNamara, D. S. (2014). The importance of grammar and mechanics in writing assessment and instruction: Evidence from data mining. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 300–303). London: UK.
Dikli, S., & Bleyle, S. (2014). Automated Essay Scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006
Elliott, S. 2003. Intellimetric: From Here to Validity. In Automated Essay Scoring: A Cross-Disciplinary Per- spective, ed. M. Shermis and J. Burstein. Hillsdale, N. J.: Lawrence Erlbaum Associates.
El Ebyary, K., & Windeatt, S. (2010). The impact of computer-based feedback on students' written work. *International Journal of English Studies, 10*(2), 121–142. https://doi.org/10.6018/ijes/2010/2/119231
Elliott, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis, & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Erlbaum.
Faigley, L., & Witte, S. (1981). Analyzing revision. *College Composition and Communication, 32*(4), 400–414. https://doi.org/10.2307/356602
Figueredo, L., & Varnhagen, C. K. (2005). Didn't you run the spell checker? Effects of type of spelling error and use of a spell checker on perceptions of the author. *Reading Psychology, 26*(4–5), 441–458. https://doi.org/10.1080/02702710500400495
Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57*(4), 481–506. https://doi.org/10.3102/00346543057004481
Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387. https://doi.org/10.2307/356600
Gillespie, A., & Graham, S. (2014). A meta-analysis of writing interventions for students with learning disabilities. *Exceptional Children, 80*, 454–473. https://doi.org/10.1177/0014402914527238
Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 187–207). New York: Guilford.
Graham, S. (2021). A walk through the landscape of writing: Insights from a program of writing research. *Educational Psychologist*. https://doi.org/10.1080/00461520.2021.1951734
Graham, S., Aitken, A., Hebert, M., Santangelo, T., Camping, A., Harris, K. R., … Ng, C. (2020). Do children with reading difficulties experience writing difficulties? A meta-analysis. *Journal of Educational Psychology, 2020*, 643.
Graham, S., Bañales, G., Ahumada, S., Muñoz, P., Alvarez, P., & Harris, K. R. (2020). Writing strategies interventions. *Handbook of Strategies and Strategic Processing*, 141–151.
Graham, S., Bollinger, A., Booth Olson, C., D'Aoust, C., MacArthur, C., McCutchen, D., & Olinghouse, N. (2012). *Teaching elementary school students to be effective writers: A practice guide (NCEE 2012-4058)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
Graham, S., Bruch, J., Fitzgerald, J., Friedrich, L., Furgeson, J., Greene, K.… Smither Wulsin, C. (2016). *Teaching secondary students to write effectively (NCEE 2017-4002)*. Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education.
Graham, S., Capizzi, A., Harris, K. R., Hebert, M., & Morphy, P. (2014). Teaching writing to middle school participants: A national survey. *Reading and Writing, 27*(6), 1015–1042. https://doi.org/10.1007/s11145-013-9495-7
Graham, S., & Harris, K. R. (2016). A path to better writing: Evidence-based practices in the classroom. *The Reading Teacher, 69*(4), 359–365. https://doi.org/10.1002/trtr.1432
Graham, S., Harris, K. R., & Chambers, A. B. (2015). Evidence-based practice and writing instruction: A review of reviews. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed.). New York, NY: Guilford.
Graham, S., MacArthur, C. A., & Fitzgerald, J. (2013). *Best practices in writing instruction*. Guilford Press.
Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent participants. *Journal of Educational Psychology, 99*(3), 445–476. https://doi.org/10.1037/0022-0663.99.3.445
Graham, S., & Santangelo, T. (2014). Does spelling instruction make participants better spellers, readers, and writers? A meta-analytic review. *Reading and Writing, 27*(9), 1703–1743. https://doi.org/10.1007/s11145-014-9517-0
Grammarly (2019). *Retrieved Nov 20, 2019 from* ⟨https://www.grammarly.com/⟩.
Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment, 8*(6), 4–43.
Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*(3), 369–388. https://doi.org/10.1177/0741088312451260
Harris, K. R., Graham, S., MacArthur, C., Reid, R., & Mason, L. H. (2011). Self-regulated learning processes and children's writing. *Handbook of self-regulation of learning and performance*, 187–202.
Hazelton, L., Nastal, J., Elliot, N., Burstein, J., & McCaffrey, D. F. (2021). Formative automated writing evaluation: A standpoint theory of action. *Journal of Response to Writing, 7*(1), 3.
Heift, T., & Rimrott, A. (2008). Learner responses to corrective feedback for spelling errors in CALL. *System, 36*(2), 196–213. https://doi.org/10.1016/j.system.2007.09.007
Hillocks, G., Jr. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies. *American Journal of Education, 93*(1), 133–170. https://doi.org/10.1086/443789

Huang, S., & Renandya, W. A. (2020). Exploring the integration of automated feedback among lower-proficiency EFL learners. *Innovation in Language Learning and Teaching, 14*(1), 15–26. https://doi.org/10.1080/17501229.2018.1471083
Hughes, J., & R Core Team (2017). *Reghelper: Helper functions for regression analysis*. R package version 0.3, 3.
Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237–263. https://doi.org/10.3102/00346543060002237
Johnson, A. C., Wilson, J., & Roscoe, R. D. (2017). College student perceptions of writing errors, text quality, and author characteristics. *Assessing Writing, 34*, 72–87.
Kellogg, R. T. (2008). Training writing skills: A cognitive development perspective. *Journal of Writing Research, 1*(1), 1–26. https://doi.org/10.17239/jowr-2008.01.01.1
Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college participants. *Psychonomic Bulletin & Review, 14*(2), 237–242. https://doi.org/10.3758/BF03194058
Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research, 42*(2), 173–196. https://doi.org/10.2190/EC.42.2.c
Kiuhara, S. A., Graham, S., & Hawken, L. S. (2009). Teaching writing to high school participants: A national survey. *Journal of Educational Psychology, 101*(1), 136–160. https://doi.org/10.1037/a0013097
Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing.* , Article 100450. https://doi.org/10.1016/j.asw.2020.100450
Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing, 27*, 1–18. https://doi.org/10.1016/j.jslw.2014.10.004
Li, Z., Feng, H., & Saricaoglu, A. (2017). The short-term and long-term effects of AWE feedback on ESL students' development of grammatical accuracy. *CALICO Journal, 34*(3), 355–375.
Lin, P. H., Liu, T. C., & Paas, F. (2017). Effects of spell checkers on English as a second language participants' incidental spelling learning: A cognitive load perspective. *Reading and Writing, 30*(7), 1501–1525. https://doi.org/10.1007/s11145-017-9734-4
Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, 1–30. https://doi.org/10.1080/09588221.2020.1743323
MacArthur, C. A., Philippakos, Z. A., & Graham, S. (2016). A multicomponent measure of writing motivation with basic college writers. *Learning Disability Quarterly, 39*(1), 31–43. https://doi.org/10.1177/0731948715583115
McNamara, D. S., & Allen, L. K. (2017). Toward an integrated perspective of writing as a discourse process. In M. Schober, A. Britt, & D. N. Rapp (Eds.), *Handbook of discourse processes* ((2nd ed.)). New York: Routledge.
MacGinitie, W. H., MacGinitie, R. K., Cooter, R. B., & Curry, S. (1989). Assessment: Gates-MacGinitie reading tests. *The Reading Teacher, 43*(3), 256–258.
McCarthy, K. S., Likens, A. D., Johnson, A. M., Guerrero, T. A., & McNamara, D. S. (2018). Metacognitive overload!: Positive and negative effects of metacognitive prompts in an intelligent tutoring system. *International Journal of Artificial Intelligence in Education, 28*(3), 420–438.
McCarthy, K. S., Roscoe, R. D., Likens, A. D., & McNamara, D. S. (2019, June). Checking It Twice: Does Adding Spelling and Grammar Checkers Improve Essay Quality in an Automated Writing Tutor?. In International Conference on Artificial Intelligence in Education (pp. 270-282). Springer, Cham.
Marshall, J. C. (1967). *Composition errors and essay examination grades re-examined, 4* pp. 375–385). American Educational Research Association. https://doi.org/10.3102/00028312004004375
Midgette, E., Haria, P., & MacArthur, C. (2008). The effects of content and audience awareness goals for revision on the persuasive essays of fifth- and eighth-grade participants. *Reading and Writing, 21*(1–2), 131–151. https://doi.org/10.1007/s11145-007-9067-9
Miłkowski, M. (2010). Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience, 40*(7), 543–566. https://doi.org/10.1002/spe.971
Morphy, P., & Graham, S. (2012). Word processing programs and weaker writers/readers: A meta-analysis of research findings. *Reading and Writing, 25*(3), 641–678. https://doi.org/10.1007/s11145-010-9292-5
Naber, D. (2003). *A rule-based style and grammar checker* (Unpublished Master's thesis). Germany: Universität Bielefeld.
O'Neill, R., & Russell, A. M. T. (2019). Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. *Australasian Journal of Educational Technology, 35*(1), 42–56. https://doi.org/10.14742/ajet.3795
Otnes, H., & Solheim, R. (2019). Acts of responding. Teachers' written comments and students' text revisions. *Assessment in Education: Principles, Policy & Practice, 26*(6), 700–720.
Palermo, C., & Wilson, J. (2020). Implementing automated writing evaluation in different instructional contexts: A mixed-methods study. *Journal of Writing Research, 12*(1), 63–108. https://doi.org/10.17239/jowr-2020.12.01.04
Page, E. B. (2020). Project Essay Grade: PEG. In M. D. Shermis, & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Erlbaum.
Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and participant progress. *Assessing Writing, 15*(2), 68–85. https://doi.org/10.1016/j.asw.2010.05.004
Phillips, L. M., Norris, S. P., Osmond, W. C., & Maynard, A. M. (2002). Relative reading achievement: A longitudinal study of 187 children from first through sixth grades. *Journal of Educational Psychology, 94*(1), 3–13. https://doi.org/10.1037/0022-0663.94.1.3
Potter, R., & Fuller, D. (2008). My new teaching partner? Using the grammar checker in writing instruction. *English Journal, 98*(1), 36–41.
Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research, 26*(4), 407–425. https://doi.org/10.2190/CX92-7WKV-N7WC-JL0A
Proske, A., Narciss, S., & McNamara, D. S. (2012). Computer-based scaffolding to facilitate students' development of expertise in academic writing. *Journal of Research in Reading, 35*, 136–152.
Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning, 31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994
Rock, J. L. (2007). The impact of short-term use of CRITERIONS on writing skills in ninth grade. *ETS Research Report Series, 2007*(1). i-24.
Roscoe, R. D., Crossley, S. A., Snow, E. L., Varner, L. K., & McNamara, D. S. (2014). Writing quality, knowledge, and comprehension correlates of human and automated essay scoring. In W. Eberle, & C. Boonthum-Denecke (Eds.), *Proceedings of the 27th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 393–398). Palo Alto: CA: AAAI Press.
Saldaña, J. (2015). *The coding manual for qualitative researchers* (3rd ed.). Los Angeles, CA: Sage.
Santangelo, T., Harris, K. R., & Graham, S. (2016). Self-regulation and writing: Metaanalysis of the selfregulation processes in Zimmerman and Risemberg's model. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 174–193). New York, NY: Guilford Press.
Scardamalia, M., & Bereiter, C. (1986). Research on written composition. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 778–803). New York: Macmillan.
Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007
Stevenson, M., & Phakiti, A. (2019). Automated feedback and second language writing. In K. Hyland, & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (2nd ed., pp. 125–142). Cambridge University Press.
Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education, 131*, 33–48. https://doi.org/10.1016/j.compedu.2018.12.005
Underwood, J. S., & Tregidgo, A. P. (2006). Improving student writing through effective feedback: Best practices and recommendations. *Journal of Teaching Writing, 22*(2), 73–98.
Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 157–180. https://doi.org/10.1191/1362168806lr190oa

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100*, 94–109.
Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research, 58*(1), 87–125.
Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing, 36*, 90–102. https://doi.org/10.1016/j.asw.2018.02.004

Kathryn S. McCarthy is Assistant Professor of Educational Psychology at Georgia State University. Her research explores the cognitive processes involved in discipline-specific literacy and the extent to which these processes vary across learners and contexts.

Rod D. Roscoe is an Associate Professor of Human Systems Engineering in the Polytechnic School of the Ira A. Fulton Schools of Engineering. His research investigates the intersection of learning science, computer science, and user science to inform effective educational technologies.

Laura K. Allen is an Assistant Professor in the Department of Psychology at University of New Hampshire. Her research theoretically and empirically investigates the higher-level cognitive skills that are required for successful text comprehension and production, as well as the ways in which performance in these domains can be enhanced through strategy instruction and training.

Aaron D. Likens is an Assistant Professor in the Department of Biomechanics at the University of Nebraska at Omaha. His research focuses on human variability and developing new technologies that improve learning, performance, and health.

Danielle S. McNamara is a Professor of Psychology at Arizona State University. She focuses on educational technologies and discovering new methods to improve students' ability to understand challenging text and convey their thoughts and ideas in writing. Her work (see http://soletlab.com) integrates various approaches including the development of game-based tutoring systems (e.g., iSTART, Writing Pal), the development of natural language processing tools, and the use of learning analytics across multiple contexts.