Computers & Education 100 (2016) 94–109

Contents lists available at ScienceDirect

Computers & Education

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

Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality

Joshua Wilson*, Amanda Czik

University of Delaware, United States

ARTICLE INFO

Article history:
Received 30 October 2015
Received in revised form 7 May 2016
Accepted 9 May 2016
Available online 11 May 2016

Keywords:
Automated essay evaluation
Interactive learning environments
Writing
English Language Arts

ABSTRACT

Automated Essay Evaluation (AEE) systems are being increasingly adopted in the United States to support writing instruction. AEE systems are expected to assist teachers in providing increased higher-level feedback and expediting the feedback process, while supporting gains in students' writing motivation and writing quality. The current study explored these claims using a quasi-experimental study. Four eighth-grade English Language Arts (ELA) classes were assigned to a combined feedback condition in which they received feedback on their writing from their teacher and from an automated essay evaluation (AEE) system called PEG Writing®. Four other eighth-grade ELA classes were assigned to a teacher feedback-only condition, in which they received feedback from their teacher via GoogleDocs. Results indicated that teachers gave the same median amount feedback to students in both condition, but gave proportionately more feedback on higher-level writing skills to students in the combined PEG + Teacher Feedback condition. Teachers also agreed that PEG assisted them in saving one-third to half the time it took to provide feedback when they were the sole source of feedback (i.e., the GoogleDocs condition). At the conclusion of the study, students in the combined feedback condition demonstrated increases in writing persistence, though there were no differences between groups with regard to final-draft writing quality.

1. Introduction

A commonly used method for teaching writing is to provide instructional feedback (Biber, Nekrasova, & Horn, 2011; Black & William, 2009; Hattie & Timperley, 2007; Kellogg & Whiteford, 2009). Instructional feedback is information provided by an agent—such as a teacher, peer, or computer—that indicates both correctness/incorrectness and ways to improve performance or understanding (Hattie & Timperley, 2007; Parr & Timperley, 2010). Struggling writers, in particular, need targeted instructional feedback because they tend to produce shorter, less-developed, and more error-filled and problem-laden texts than their peers (Troia, 2006).

However, the role of instructional feedback in the teaching of writing is not without controversy. Proponents advocate its role in supporting motivation and writing quality by (a) indicating to the author his/her position relative to a desired level of

* Corresponding author. University of Delaware, School of Education, 213E Willard Hall Education Building, Newark, DE 19716, United States.
E-mail address: joshwils@udel.edu (J. Wilson).

http://dx.doi.org/10.1016/j.compedu.2016.05.004
0360-1315/

quality; (b) identifying areas in need of improvement related to low-level writing skills (spelling, word choice, mechanics, grammar) or high-level skills (idea development and elaboration, organization, rhetoric); and (c) prompting additional practice attempts in which the author incorporates and eventually internalizes the feedback (Ferster, Hammond, Alexander, & Lyman, 2012; Kellogg, Whiteford, & Quinlan, 2010; Parr & Timperley, 2010). In contrast, others argue that providing instructional feedback is (a) too time consuming and leads to teacher burnout (Anson, 2000; Baker, 2014; Lee, 2014); (b) too difficult for teachers to provide given the complexity of writing ability (Marshall & Drummond, 2006; Parr & Timperley, 2010); and (c) ineffective or incapable of achieving substantial, generalizable gains in students' writing performance (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Biber et al., 2011; Kluger & DeNisi, 1998). Nevertheless, instructional feedback continues to be recommended as a method for teaching writing (APA, 2015; Graham, Harris, & Santangelo, 2015; Graham, Hebert, & Harris, 2015; Graham, McKeown, Kiuhara, & Harris, 2012; Graham & Perin, 2007).

In the U.S., an increasingly common form of instructional feedback for writing is feedback provided by automated essay evaluation systems (AEE) (Warschauer & Grimes, 2008). AEE are web-based formative writing assessment software programs which provide students with immediate automated feedback in the form of essay ratings and individualized suggestions for improvement when revising (Shermis & Burstein, 2013). Some of the principal benefits of AEE are efficiency and flexibility. While there is no consensus regarding the optimal timing of feedback (see Shute (2008) for review), immediate feedback is often preferred (Chan, Konrad, Gonzalez, Peters, & Ressa, 2014; Ferster et al., 2012) especially in classroom settings (Hattie & Timperley, 2007; Shute, 2008). In addition, unlike teacher or peer feedback, automated feedback allows students to control feedback timing. Students receive feedback when they request it, either in the middle of, or after having completed, an essay draft. This enables feedback to be immediately actionable, accelerating the practice-feedback loop (Foltz, Streeter, Lochbaum, & Landauer, 2013; Kellogg et al., 2010).

While automated feedback addresses some of the barriers faced by teachers when providing instructional feedback, the intended use of AEE systems is to complement and not replace teacher feedback (Foltz, 2014; Foltz et al., 2013; Kellogg et al., 2010). Indeed, AEE is thought to free up instructional time and allow teachers to be more selective in the type of feedback they provide, thereby improving students' writing motivation and writing performance (Grimes & Warschauer, 2010). For instance, after implementing AEE in her high school, a school administrator reported that, "[AEE] has helped motivate our students to write while making it easier for educators to provide the feedback needed to ensure growth in writing" (Schmelzer, 2004, p.34).

Yet, the growing adoption of AEE in the U.S. has been accompanied by a number of concerns and fears. For instance, despite its intended role as a complement to teacher feedback, some fear that AEE will come to replace the teacher as primary feedback agent (Ericcson & Haswell, 2006; Herrington & Moran, 2001), and thereby negate the social communicative function of writing (National Council of Teachers of English [NCTE], 2013). Others are concerned that AEE can be easily fooled to assign high scores to essays which are long, syntactically complex, and replete with complex vocabulary (Bejar, Flor, Futagi, & Ramineni, 2014; Higgins & Heilman, 2014). Concerns such as these have led some groups to summarily reject the use of AEE (Conference on College Composition and Communication, 2014; NCTE, 2013).

This debate over AEE's virtues and ills is compounded by two related issues. First, there is a dearth of research on AEE used for the purpose of formative assessment—i.e., assessment for, rather than of learning (Black & William, 2009). By far, the majority of research has focused on the psychometric properties of the automated scoring engine, rather than documenting evidence that automated feedback is associated with desired changes in teacher feedback practices or students' writing motivation or writing quality (Stevenson & Phakiti, 2014). Indeed, a recent chapter on the formative use of AEE in the Handbook of Writing Research still primarily discusses the features of the AEE scoring systems and the reliability and agreement of those systems with human essay ratings. The chapter authors acknowledge that research "still needs to be conducted to gain a more comprehensive understanding of the impact of [automated] feedback than can guide best-use practice" (Shermis, Burstein, Elliot, Miel, & Foltz, 2015, p.406). Second, previous research has most often examined the effects of automated feedback in isolation of teacher feedback (Stevenson & Phakiti, 2014). Such designs lack ecological validity and may inadvertently bolster fears that adoption of AEE will replace teachers as feedback agents.

Given the controversies surrounding the use of instructional feedback and AEE, as well as the dearth of prior research focusing on the intended usage of AEE, the current study was designed to explore the implications for instruction and student performance when teacher feedback on writing was combined with automated feedback. Specific outcomes of interest included the amount, type, and level of teacher feedback; students' writing motivation; and final-draft writing quality. To further provide context for this study, three areas of prior research will be discussed: (1) categorization of teacher feedback on writing, (2) effects of teacher and automated feedback on writing motivation, and (3) effects of teacher and automated feedback on writing quality.

2. Background

2.1. Categorizing teacher feedback on writing

Teacher feedback on writing is commonly categorized as having at least two components: feedback type and feedback level. *Feedback type* relates to the manner in which feedback is presented to the student. A common distinction is between direct and indirect feedback (Biber et al., 2011; Black & William, 1998; Cho, Schunn, & Charney, 2006; DeGroff, 1992; Shute, 2008). Direct feedback (i.e., directives) involves teachers making a correction or directly telling students what needs to be

revised. Indirect feedback involves teachers guiding students to form their own concept. Types of indirect feedback include queries and informatives (DeGroff, 1992; Shute, 2008). Queries take the form of questions to the author, and request a clarification or response. Informatives are designed to transmit ideas, opinions, and information, but do not explicitly require students to perform a specific edit or revision. A final type of teacher feedback is praise. Praise expresses approval regarding aspects of the student's effort, performance, or writing quality (Cho et al., 2006; Nelson & Schunn, 2009; Peterson, Childs, & Kennedy, 2004).

Feedback level refers to the specific writing subskills that are the target of the teacher's feedback message. Biber et al. (2011) refer to this as the "focus" of the feedback, for which they distinguished five broad foci for instructional feedback on writing: lexis, grammar, mechanics, organization, and content. Prior research has differed with regard to the grain size of analysis within each of these levels. For example, AbuSeileek (2013) specified several components of feedback on grammar and mechanics including: capitalization, fragments and run-ons, misused words, negation, noun phrases, possessives and plurals, punctuation, questions, relative clauses, subject-verb agreement, and verb phrase. Matsumara, Patthey-Chavez, Valdés, and Garnier (2002) grouped teacher feedback across categories into surface-level or content-level feedback. Surface-level feedback targeted lower-level writing skills such as mechanics, usage, grammar, spelling, sentence structure and formatting. Content-level feedback targeted higher-level writing skills such as concepts and structures in the student's writing. Likewise, Peterson et al. (2004) categorized teacher feedback as editive or revisional. Editive feedback was given with the intent of encouraging lower-order revisions such as mechanical, grammatical, lexical, or syntactic changes; whereas, revisional feedback was given with the intent of encouraging higher-order revisions like informational, organizational, or holistic changes. Similarly, Parr and Timperley (2010) distinguished between four deep features constituting higher-level feedback (audience, structure, content, and language resources) and three surface-level features constituting lower-level feedback (grammar, spelling, punctuation).

Based on this prior research, the current study categorizes teacher feedback according to both type and level. Feedback type was coded as directive, query, informative, or praise. Feedback level was coded as one of eleven components, organized into the following two broad categories:

- Lower-level writing skills (7 components): spelling, capitalization, punctuation, sentence structure, grammar, formatting, and word choice
- Higher-level writing skills (4 components): ideas and elaboration, organization, style, and self (i.e., feedback directed at the author's writing process or experience)

A further description of this study's coding framework, along with definitions and examples, is presented in Appendix A.

2.2. Instructional feedback and writing motivation

2.2.1. Effects of teacher feedback on writing motivation

Theoretical models of writing ability (Hayes, 2012) and empirical research (Graham, Berninger, & Fan, 2007) underscore the importance of students' writing motivation and writing dispositions for promoting writing achievement. Teacher feedback is a key source of students' perceived self-efficacy and writing motivation (Pajares, 2003), though students vary in what kind of feedback they want. Some value positive comments very highly (Cho et al., 2006) while others discount praise as a mitigation device (Hyland & Hyland, 2001). Indeed, motivational beliefs about writing have been shown to mediate the relationship between students' writing activity (e.g., the frequency with which they engage in specific writing tasks in and out of school) and writing quality (Troia, Harbaugh, Shankland, Wolbers, & Lawrence, 2013).

Prior research on the effects of teacher feedback on student writing motivation has found that students' self-efficacy changes even after a single feedback episode (Dujinhower, Prins, & Stokking, 2010). Indeed, effective writing teachers appear to leverage the motivational aspects of teacher feedback to promote students' persistence and effort (Harward et al., 2014). However, writing motivation may be negatively affected if feedback is misunderstood, inaccurate or misdirected, or processed ineffectively (Zumbrunn, Mars, & Mewborn, 2016). Similarly, feedback may affect students' writing motivation differently based on their initial self-efficacy beliefs. For example, students with lower initial self-efficacy may be concerned that feedback will simply identify their weaknesses in writing, and for these students, feedback may disincline them to write (Dujinhower et al., 2010).

2.2.2. Effects of automated feedback on writing motivation

Though limited, prior research suggests that automated feedback may be associated with positive effects on students' writing motivation. Warschauer and Grimes (2008) reported that teachers and students in four secondary schools in the U.S. attributed gains in students' motivation to write and revise to the use of AEE programs. Similar results were reported in a subsequent study: 30 out of 40 surveyed teachers agreed or strongly agreed that students' motivation was higher with AEE than with basic word processing (Grimes & Warschauer, 2010). Students also indicated statistically significant higher than chance levels of agreement with the statement: "Writing with MyAccess! [an AEE system] has increased my confidence in my writing." Foltz (2014) also found that students using an AEE system called Write-to-Learn spent a substantial amount of time

writing and revising their responses to prompts. While time spent writing and revising are not direct indicators of motivation, they do suggest that adoption of AEE was associated within increased time on task and persistence.

Though these studies provide some initial evidence that AEE has positive effects on students' motivation, clearly more research is needed. Specifically, little is known about the effects on writing motivation when automated feedback and teacher feedback are combined. Given that teacher feedback also affects student motivation, if the use of AEE were to lead to changes in the type and level of teacher feedback, this may have additional attendant increases or decreases in students writing motivation.

2.3. Instructional feedback and writing quality

2.3.1. Effects of teacher feedback on writing quality

Prior research reports highly variable findings on the effects of teacher feedback on writing quality. For instance, a recent meta-analysis reported an average weighted effect size of 0.77 for the effects of teachers providing feedback on writing quality, an effect size much larger than those reported for peer, self, and automated feedback (Graham et al., 2015). However, a meta-analysis by Biber et al. (2011) provided a more nuanced analysis of the effect of written feedback for both English-only (L1) instruction and English as a Second Language. For instance, though instructional feedback was found to have an effect size of 1.20 for L1 studies using pretest/posttest designs, this effect disappeared for L1 studies using treatment/control designs ($D = -0.03$). In addition, not all types of feedback were equally effective: directive feedback in the form of comments was more effective than non-directive feedback. Nor were all levels of feedback: feedback on content and form yielded larger effects than did feedback on either content or form in isolation. Nevertheless, feedback provided as written comments resulted in large gains in grammatical accuracy and overall quality. Thus, while the effects of teacher feedback appear to be positive in aggregate, the relationship between teacher feedback and writing quality is complex, and likely mediated by the type and level of feedback provided to students, among other factors.

2.3.2. Effects of automated feedback on writing quality

A small but growing body of research suggests that automated feedback supports modest gains in students' writing quality (Graham et al., 2015; Morphy & Graham, 2012; Stevenson & Phakiti, 2014). For instance, in their meta-analysis of the effects of various software programs on weaker writers/readers' motivation to write, Morphy and Graham (2012) reported an average weighted effect size of 1.46 for three studies of automated feedback when compared to studies using paper-and-pencil as a control condition. Findings from individual studies not included in their meta-analysis are generally consistent: Automated feedback appears to support improvements in the overall quality of students' essays while concomitantly reducing the frequency of mechanical errors across successive revision to an essay draft (Foltz, 2014; Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005; Kellogg et al., 2010; Shermis, Garvan, & Diao, 2008; Wade-Stein & Kintsch, 2004; Wilson & Andrada, 2016; Wilson, Olinghouse, & Andrada, 2014). However, automated feedback appears limited to scaffolding improvements in a single essay, versus supporting transfer to improved independent performance (Stevenson & Phakiti, 2014; Wilson & Andrada, 2016; Wilson et al., 2014). Yet, these findings must be taken cautiously since the majority of prior research has employed within-subjects designs or have used experimental/quasi-experimental designs with weak counterfactuals, such as a no-feedback control condition.

3. Study purpose

This study examined effects on teacher feedback, and students' writing motivation and final-draft writing quality associated with a combined automated + teacher feedback condition, in which students received feedback from an AEE system called PEG Writing® as well as their teacher, and a teacher-feedback-only condition, in which students received feedback from their teacher via the comment and edit functions of GoogleDocs. To date, no research has evaluated the effects of a combined teacher + automated feedback condition against a teacher-feedback-only condition with respect to these outcomes. Three research questions guided the study:

RQ1: What are the effects associated with feedback condition on the amount, type, and level of teacher feedback on students' writing?
RQ2: What is the effect associated with feedback condition on students' writing motivation?
RQ3: What is the effect associated with feedback condition on students' final-draft writing quality?

4. Methods

4.1. Setting and participants

This study was conducted in a middle school in an urban school district in the mid-Atlantic region of the United States. The district serves approximately 10,000 students in ten elementary schools, three middle schools, and one high school. In this

district, 43% of students are African-American, 20% are Hispanic/Latino, and 33% White. Approximately 9% of students are English Language Learners, and 50% of students come from low income families.

Two eighth-grade English Language Arts (ELA) teachers provided consent to participate in this study. Both teachers were experienced, having taught for a total of 12 and 19 years, respectively. One teacher had earned a Master's degree and the other was in the process of earning it (Bachelor's +21 credits). Each teacher taught a total of four class periods of ELA daily.

Since the school did not use academic tracking, two classes were randomly selected from each teacher's schedule to assign to the teacher + automated feedback condition (hereafter referred to as PEG + Teacher), and two classes to the teacher-feedback-only condition (hereafter referred to as GoogleDocs). Thus, teachers instructed classes assigned to both feedback conditions, allowing instruction to be held relatively constant across conditions.

After receiving consent and assent, a total of 151 students were assigned by classroom to either the PEG + Teacher or GoogleDocs conditions. Though classes were randomly assigned to feedback conditions, the study sampled intact classes, resulting in a quasi-experimental design. While sampling intact classrooms results in a weaker design than randomly assigning students to condition, this design removed threats of compensatory rivalry and resentful demoralization since all students within a classroom utilized the same software (either PEG Writing® or GoogleDocs). Six students moved during the study timeframe, leaving a final sample of 145 students.

Table 1 reports demographics for this sample, organized by feedback condition. No students received special education services. Per district policy, since more than 51% of the school's population qualified as low-income students, all students in the school received free lunch. Chi-Square tests and Analysis of Variance (ANOVA) confirmed that the groups were equal with respect to all demographic variables.

4.2. Description of PEG Writing®

PEG Writing® is a web-based formative writing assessment software program developed by Measurement Incorporated (MI). It is an interactive learning environment designed to promote increased learning of writing skills and increased writing achievement. PEG Writing® supports several types of interactions in the system: (a) learner-system interactions, via students receiving quantitative and qualitative automated feedback after submitting their essay for scoring; (b) teacher-system interactions, via teachers reviewing PEG Writing®'s score reports; (c) teacher-learner interactions, via teachers providing feedback to students to supplement and complement the automated feedback; and, (d) learner-learner interactions, via PEG Writing®'s peer review functions. The current study focused on the first three types of interactions.

PEG Writing® is built around an automated essay scoring engine called Project Essay Grade (PEG). PEG was developed by Ellis Batten Page (Page, 1966, 1994, 2003) and acquired by MI in 2003 who has since redesigned and enhanced the scoring engine (Shermis et al., 2015). PEG uses a combination of techniques such as natural language processing, syntactic analysis, and semantic analysis to measure more than 500 variables that are combined using machine-learning algorithms that predict essay ratings assigned by expert raters (Bunch, Vaughn, & Miel, 2016; Shermis et al., 2015). A number of empirical studies have established the reliability and criterion validity of PEG's scoring system (Keith, 2003; Shermis, 2014; Shermis, Koch, Page, Keith, & Harrington, 2002).

Students and teachers access PEG Writing® by visiting www.pegwriting.com and inputting their username and password. Students then select among system-created or teacher-created writing prompts in multiple genres (narrative, informational, argumentative). Prompts may also have associated stimulus materials, such as readings, videos, artwork, music, or links to websites.

Once a prompt is assigned, students can select from one of several embedded graphic organizers to support prewriting. After prewriting, students have up to 60 min to complete and submit their initial draft for evaluation by PEG. Once submitted, students immediately receive scores for six traits of writing quality: Idea Development, Organization, Style, Sentence Structure, Word Choice, and Conventions. Each of these traits is scored on a 1–5 scale and combined to form an Overall Score

Table 1
Sample demographics.

	PEG + teacher ($n = 72$)	GoogleDocs ($n = 73$)
Gender (n)		
Male	39	34
Female	33	39
Race (n)		
Hispanic/Latino	19	19
African American	25	31
White	26	21
Asian	1	1
Unreported	1	1
ELL (n)	2	0
Age (months)		
M	168.58	169.04
SD	4.92	5.98

$F_{(1, 143)} = 0.26$, $p = 0.609$ comparing groups for age.

ranging from 6 to 30. In addition, students receive feedback on grammar and spelling, as well as trait-specific feedback that encourages students to review and evaluate their text with regard to the features of that specific trait. This is done by providing students with general characteristics of how good writers develop these specific traits in their writing. For example, for the Word Choice trait, students are told that good writers "use precise vocabulary in sophisticated ways; choose words to make the writing interesting and engaging; and use content-specific vocabulary to support topic explanation and information." An example of a feedback statement for Word Choice is "Your use of many different words makes your essay stronger." Thus, PEG's trait-specific feedback tends be rather broad (i.e., nonspecific), as it is tied to the trait scores students receive and the general evaluation criteria underlying those specific scores.

Once students receive their feedback, they may revise and resubmit their essays as many times as they wish (the maximum is 99 times), each time receiving new feedback. Students also receive customized links to multimedia interactive lessons on specific writing skills to help them practice and strengthen their writing. PEG also enables students to receive feedback from their teachers, who can embed comments directly within the students' essays (similar to that of adding comments in Microsoft Word®) or via summary comments located in a text box beneath PEG's score report. Students may also leave comments for their teacher with a similar function. Appendix B provides screen shots of PEG's qualitative and quantitative feedback, and an example of teacher feedback provided via the PEG Writing® environment.

4.3. Study measures

4.3.1. Prior literacy achievement

Study constraints required drawing upon existing district data to establish equivalence of groups at pretest in terms of prior literacy ability. Prior literacy achievement was measured using the district-administered *STAR Reading* assessment, a computer-adaptive reading test developed by Renaissance Learning (range: 0–1400).

4.3.2. Amount, type, and level of teacher feedback

Each feedback message given to students by the teachers was first parsed into feedback units (i.e., idea units) according to procedures outlined in Cho et al. (2006) and Hayes and Berninger (2010): a feedback unit was defined as a stand-alone message addressing a single aspect of the text. To ensure accuracy, all teacher-feedback messages were parsed into feedback units by both the first and second author. Differences were resolved by consensus. This process resulted in identifying a total of 3830 teacher-feedback units across 145 essays.

Then, each teacher-feedback unit was coded using a coding framework containing two dimensions: feedback type and feedback level (see Section 2.1 and Appendix A). The first and second author trained on the coding framework until reaching a minimum inter-rater reliability (IRR) of 80% exact agreement for coding feedback type and feedback level, respectively. Percentage of exact agreement was calculated as: {[Total agreements/(total agreements + disagreements)]*100}. Once reaching this criterion, the second-author coded all teacher-feedback units for the full sample ($n = 145$). A random sample of 30% of students' writing samples ($n = 43$) was double coded by the first author as a reliability check. IRR was high. Percent of exact agreement for Feedback Type was 94% (range: 78–100%), and for Feedback Level it was 84% (range: 70–100%). All disagreements were resolved via consensus prior to data analysis.

Fig. 1 presents an example of how feedback messages were parsed into feedback units and coded using this framework.

4.3.3. Writing motivation

Writing Motivation was assessed using the Writing Disposition Scale (WDS), a scale designed for elementary and middle school students (Piazza & Siebert, 2008). See Appendix C. The WDS asks students to rate their agreement with 11 Likert-scale items, evaluating various dispositions towards writing such as confidence (items 1, 5, 11), passion (items 2, 6, 8, and 9), and persistence (items 3, 4, 7, and 10). The WDS was administered at pretest and posttest. Cronbach's Alpha was reported as 0.89 for the entire instrument in the original study (Piazza & Siebert, 2008), and was 0.83 for the present study.

4.3.4. Measures of final-draft writing quality

Final-draft writing quality was assessed using two measures: (a) the PEG Overall Score and PEG trait scores; and (b) a Holistic Quality score, as measured by the first author and a trained research assistant. Details on the PEG Overall and trait scores are found in Section 4.2. Essays of students in the PEG + Teacher group were automatically scored by PEG upon submission of the essays to its scoring engine. Essays of students in the GoogleDocs group were copied and pasted from GoogleDocs into PEG to obtain the PEG Overall Score and PEG trait scores.

Holistic writing quality was evaluated with the narrative rubric used for the 2011 eighth-grade U.S. National Assessment of Educational Progress (National Assessment Governors Board, 2010). The rubric evaluated students' ability to convey experience, real or imagined, using elaboration and detail, a clear and effective organization, and well-controlled and accurate use of sentence structure, word choice, grammar, usage, and mechanics. The rubric identified the following scoring levels: 0 = Off topic/illegible, 1 = Little or no skill, 2 = Marginal skill, 3 = Developing skill, 4 = Adequate skill, 5 = Competent skill, 6 = Effective skill. Raters trained until reaching a minimum of 80% agreement within 1 point. Then, all writing samples were double-scored. Students' final Holistic Quality score was calculated as the sum of the two raters' individual scores (range: 0–12).

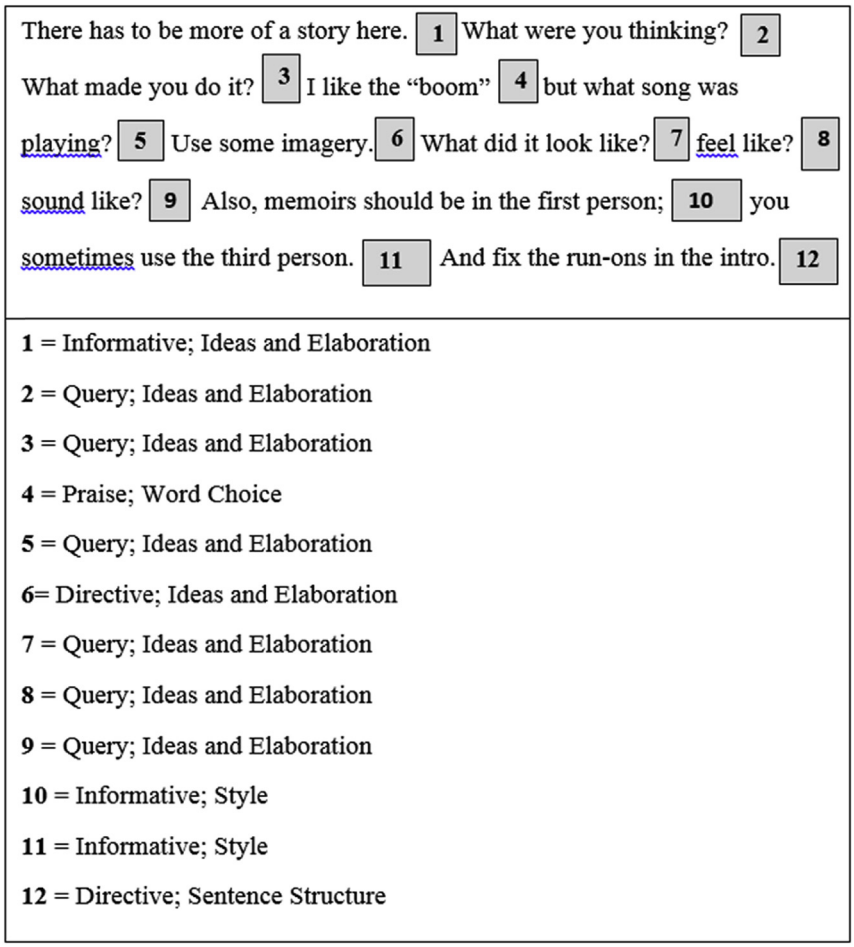

Fig. 1. Example of how a feedback message was parsed and coded.

IRR for the holistic quality score was calculated as quadratic weighted kappa. Unlike kappa which assumes that scoring categories have no ordinality, quadratic weighted kappa takes into account the weighted distance between ratings, applying a greater penalty for ratings that are further apart than to ratings which disagree but are close together (Shermis, 2014). Quadratic weighted kappa was 0.83, indicating a high-level of inter-rater reliability. Spearman's rank-order correlation between the two raters was $\rho = 0.81$.

4.3.5. *Teacher perceptions of feasibility, utility, and desirability*

An eight-item survey was administered to teachers, asking them to share their perceptions of the feasibility, utility, and desirability of providing feedback via the two software systems.

4.4. *Study procedures*

The current study spanned 11 school days (i.e., 11 1-h class periods). For the duration of the study, both teachers kept laptop carts in their classrooms for use with each class period. The laptop carts housed large enough sets of Google ChromeBooks to accommodate each teacher's largest classroom. The ChromeBooks were used by students in both feedback conditions, enabling consistency of hardware across conditions. Teachers were asked to keep instructional logs stating the focus of the teacher's and students' activity throughout the study duration. Instructional logs indicated that teachers presented information on the same topics and provided the same number of class sessions to work on memoir writing for students in both the PEG + Teacher and GoogleDocs conditions.

On Day 1, student assent forms, and pretest literacy and writing motivation data were collected. Pretest literacy data was collected via teacher report. Writing motivation data was collected by having teachers administer and collect the WDS. Then, teachers introduced students to their district-assigned curriculum module on memoir writing, a form of narrative writing, one of the three main genres addressed in the Common Core State Standards (Common Core State Standards Initiative, 2010).

Teachers introduced the key features of memoir writing to all their classes, and introduced the teacher-created writing prompt for the memoir unit, which read:

> We have all had interesting life experiences. Some are good, funny, or exciting, while others are bad, sad, or devastating. Choose one experience from your life and tell the story. Once you have chosen your topic, you may choose to turn it into a scary story, drama, elaborate fiction, science fiction, comedy, or just tell it how it is. Be sure to organize your story and elaborate on your details. Your audience wasn't there so you need to tell them every little detail.

On Day 2, teachers provided additional instruction on the characteristics of memoirs, and taught a mini-lesson titled "blowing up the moment," a lesson about description and style in memoir writing. Then, teachers gave students in the PEG + Teacher condition an opportunity to learn how to use PEG. Earlier in the school year, the first author trained the two teachers on the use PEG during three 30 min training sessions. Then, teachers subsequently trained their students how to use the program in 1 h-long class period. Students were taught how to select and complete the graphic organizers embedded within PEG, input and submit text for evaluation, examine feedback on spelling and grammar, examine trait specific feedback, and complete an essay revision. Students in the GoogleDocs group required no training since they were accustomed to using GoogleDocs in other courses for independent and shared writing tasks. Students were familiar with accessing, developing, and formatting drafts written in GoogleDocs, as well as responding to edits and revisions suggested by teachers who had editing privileges on each student's GoogleDocs account.

On Days 3–9, teachers began each class session with a short mini-lesson on a characteristic of memoir writing, such as dialogue, transitions, figurative language, vivid verbs, and "show don't tell." The mini-lessons lasted approximately 15 min. The remainder of the class periods was devoted to students independently brainstorming ideas, completing graphic organizers, drafting, revising, and editing their individual memoirs. Thus, aside from the teacher-led whole group mini-lessons students completed all work on their memoirs independently.

Between Days 3–8, teachers were instructed to review and provide feedback on their students' writing on at least one, and no more than two, occasions for students in both conditions. Teachers were directed to provide feedback as they normally would, commenting on those aspects of a student's text which they deemed salient. In addition, teachers were allowed to determine which functions of the respective writing software they wished to use for providing feedback. For instance, GoogleDocs enabled teachers to directly edit students' texts in the manner of "track changes," and to provide margin comments. Teacher feedback in the PEG + Teacher condition was delivered in the form of embedded- and summary comments (see Appendix B); teachers could not directly edit students' text as they could when using GoogleDocs. All drafting and revising in the PEG + teacher condition was carried out within PEG Writing®; thus, students had the opportunity to receive as much automated feedback as they wished by revising and resubmitting their memoir to PEG for evaluation. However, the number of instances in which teachers provided feedback was held constant across condition: 1–2 occasions.

On Day 10, students completed their memoirs, submitted their final drafts for grading, and were re-administered the writing motivation survey. On Day 11, teachers completed the brief survey regarding their experiences providing feedback via PEG Writing® and GoogleDocs.

4.5. Data analysis

4.5.1. Writing motivation

To estimate differences between feedback conditions on individual items of the Writing Dispositions Scale (WDS) the non-parametric form of the *t*-test, the Mann-Whiney test (Mann & Whitney, 1947), was used. The Mann-Whitney test compares whether the distribution of a variable is statistically significantly different across independent groups. This difference is calculated not on the means, but on the sum of ranks in each group (Field, 2014). Though there are different schools of thought regarding whether or not Likert scale ratings are best analyzed using parametric or nonparametric analyses (see Allen & Seaman, 2007; de Winter & Dodou, 2010), we elected to use non-parametric analyses for the following reasons: (a) we analyzed each of the Likert items individually, rather than creating composite scores; (b) we do not believe the underlying distribution, or interpretation, of the response categories within each item is interval in nature, believing instead that it is ordinal in nature; thus, means and standard deviations are invalid parameters for such data, as are parametric analyses; (c) seven of the 11 WDS items at pretest, and four of the 11 items at posttest, violated assumptions of normality; and consequently (d) we were interested in whether the response distributions, not the means, were different across groups.

4.5.2. Amount of teacher feedback

Differences in the amount of teacher feedback were calculated based on the median number of feedback units given by teachers across each feedback condition. This variable did not meet the distributional assumptions necessary for parametric analysis, thus the median, and not the mean was used. Accordingly, the Mann-Whitney test was used to estimate differences across groups.

4.5.3. Type and level of teacher feedback

To estimate differences in the type and level of teacher feedback between conditions we first converted raw counts of feedback type and feedback level into proportions by dividing specific counts by the total number of feedback units for each essay. Doing so controlled for the variation in the amount of feedback that students received from their teachers, and for the

variation in essay length which contributed to this variation. For example, to create proportions of queries, we used this equation: (count of queries/total feedback units for essay). The same was done to calculate proportions of feedback given for the remaining three feedback types and eleven feedback levels. The distributional assumptions necessary for parametric analysis were met for only two of the 15 proportion variables. Thus, the Mann-Whitney test was used to estimate differences in the proportion of feedback type and feedback level for all variables across conditions.

4.5.4. Writing quality

Finally, to estimate differences in final-draft writing quality scores, a series of one-way ANOVAs was conducted with feedback condition as the fixed factor—each of these variables met the distributional assumptions necessary for use of ANOVA.

4.5.5. Effect sizes

Effect sizes for Mann Whitney tests are reported as *r*, where $r = \frac{Z}{\sqrt{N}}$. Effect sizes for ANOVA are reported as Cohen's *D*, a standardized measure of the difference between two means expressed in standard deviation units.

4.5.6. Method of controlling Type-I error

Given that multiple comparisons were conducted for the WDS items, and the proportion of lower-level feedback, and the proportion of higher-level feedback, the Benjamini-Hochberg procedure was used to control the false discovery rate (Benjamini & Hochberg, 1995; Thissen, Steinberg, & Kuang, 2002). The false discovery rate is the proportion of significant results that are false positives. The B-H procedure is an acceptable method for controlling Type-I error that has greater power (i.e., less likelihood of Type-II errors) than the more conservative Bonferroni procedure (Thissen et al., 2002; U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2013). Specifically, within each family of multiple comparisons, we compared each of the statistically significant results to a B-H adjusted alpha critical value. When p_{observed} was less than p_{adjusted}, the result was considered statistically significant.

4.5.7. Method of handling missing data

Any missing data from the pretest or posttest administration of the WDS was handled using multiple imputation because it was assumed that missing data was treated as missing at random. This assumption was made because there were no differences in groups on pretest demographic or achievement measures, and the cause of missing data was either due to student absence or due to accidentally skipping a survey item. For the WDS pretest administration, only one student was absent resulting in 0.66% missing data across the sample. For the WDS posttest administration, a combination of skipped items and student absences resulted in levels of missingness ranging from 8.61% to 9.93% across items.

Rather than listwise delete cases with missing data it is preferable to use multiple imputation techniques (Peugh & Enders, 2004; Wilkinson & Task Force on Statistical Inference, 1999). In the current study, multiple imputations were conducted using Mplus V7.1 to estimate values of the missing categorical survey data. Instead of hot deck or cold imputation methods (see Andridge & Little, 2010 for review), Mplus uses Bayesian analysis (Rubin, 1987; Schafer, 1997) via the "TYPE = BASIC" command to generate multiple data sets using Markov chain Monte Carlo (MCMC) algorithms (Muthén & Muthén, 1998–2012)—interested readers are referred to Asparouhov and Muthén (2010) for additional information on the computational processes used by Mplus for multiple imputation. These data sets are independent draws based on the missing data posterior. A total of 10 multiply-imputed data sets were generated. Data analysis was conducted on a final data set using the mode of the imputed values across the 10 multiply-imputed data sets. To confirm that results did not differ as a result of conducting multiple imputation, all analyses were re-estimated using only data from complete cases (i.e., listwise deletion). No differences in results were noted; therefore, results are reported using data that included multiple imputations for missing data.

5. Results

5.1. Pretest analyses

A one-way ANOVA indicated that groups were equivalent with regard to prior literacy achievement: PEG + Teacher ($M = 916.23$, $SD = 256.79$), GoogleDocs ($M = 851.37$, $SD = 252.56$); $F_{(1, 142)} = 2.34$, $p = 0.129$. Non-parametric analyses performed on the individual writing-motivation survey items revealed that the null hypothesis of equal distributions across feedback conditions was retained in all cases. Hence, at pretest, groups were equivalent with respect to prior literacy ability and writing motivation.

5.2. Analysis of the amount, type, and level of feedback across conditions

Table 2 presents descriptive statistics for the amount, type, and level of teacher feedback across feedback conditions. A series of Mann-Whitney *U* tests were conducted. Findings indicated that there was no statistically significant difference in the amount of feedback teachers gave to students across conditions. Neither were there statistically significant differences in the

proportion of feedback types across conditions. Teachers in both conditions primarily relied on directives to convey feedback, followed by informatives and queries. Praise was the least frequently occurring feedback type in either condition.

With regard to lower-level feedback, when compared to B-H adjusted alpha critical values, two contrasts were statistically significant and one contrast was marginally statistically significant. A statistically significant greater proportion of feedback on capitalization was given by teachers in the GoogleDocs condition ($r = 0.33$), while a greater proportion of feedback on sentence structure was given by teachers in the PEG + Teacher condition ($r = 0.20$). Teachers gave a marginally statistically significant greater proportion of feedback on formatting in the GoogleDocs condition ($r = 0.18$; $p_{observed} = 0.033 > p_{adjusted} = 0.029$).

With regard to higher-level feedback, when compared to B-H adjusted alpha critical values, one contrast was statistically significant. Teachers gave a greater proportion of feedback on idea development ($r = 0.19$) in the PEG + Teacher condition.

When aggregating lower-level and higher-level writing skills, teachers gave a statistically significant greater proportion of feedback on higher-level writing skills in the PEG + Teacher condition, though the effect was small ($r = 0.22$).

5.3. Posttest analyses of writing motivation

Non-parametric analyses were performed on the individual posttest survey items to examine whether there were statistically significant differences in the distribution of responses across conditions. When compared to B-H adjusted critical alpha values, all contrasts were non-statistically significant, except for item 10—"I take time to solve problems in my writing." The mean ranks of the PEG + Teacher and GoogleDocs conditions were 65.19 ($n = 72$) and 80.70 ($n = 73$), respectively: $U = 2066.00$, $Z = -2.38$, $p = 0.017$. Examination of the individual frequency data for this item revealed that 68% of students in the PEG + Teacher feedback condition agreed or strongly agreed with this statement (an increase of 7% from pretest), as compared to 49% of students in the GoogleDocs condition (a decrease of 1% from pretest). This resulted in a small but statistically significant effect size: $r = 0.20$.

To further investigate this finding the average number of essay drafts completed by students in each condition was compared using a one-way ANOVA. Completing greater number of essay drafts may be interpreted as a form of persistence, the disposition evaluated in Item 10 of the WDS. Results indicated that students in the PEG + Teacher condition completed a higher average number of essay drafts ($M = 11.28$, $SD = 6.81$) than students in the GoogleDocs condition ($M = 7.18$, $SD = 2.29$): $F_{(1, 138)} = 22.287$, $p < 0.001$, $D = 0.81$. Thus, students' self-report information appeared consistent with their observed behavior in this regard.

5.4. Posttest analyses of writing quality

Table 3 presents results of the series of one-way ANOVAs which examined the effects of feedback condition on the PEG Overall and Trait scores, and Holistic Quality. The null hypothesis of equal means was retained in all cases. However, a small effect size favored the PEG + Teacher group for the Conventions trait ($D = 0.22$), and a small effect size favored the GoogleDocs condition on the Organization trait ($D = 0.15$).

Table 2
Comparison of teacher feedback by condition.

	PEG + teacher	GoogleDocs	Mann Whitney	*r*
	Median (range)	Median (range)		
Total feedback units	19.50 (0–57)	15.50 (1–88)		
Feedback type				
Directives	0.52 (0.00–1.00)	0.58 (0.00–1.00)		
Queries	0.13 (0.00–0.45)	0.10 (0.00–1.00)		
Informatives	0.28 (0.00–1.00)	0.23 (0.00–0.82)		
Praise	0.00 (0.00–0.22)	0.00 (0.00–0.24)		
Lower-level feedback				
Spelling	0.03 (0.00–0.20)	0.02 (0.00–0.30)		
Capitalization	0.00 (0.00–0.17)	0.02 (0.00–1.00)	$p < 0.001$, PEG < Google	0.33
Punctuation	0.04 (0.00–0.17)	0.00 (0.00–0.50)		
Sentence structure	0.09 (0.00–1.00)	0.03 (0.00–0.38)	$p = 0.016$, PEG > Google	0.20
Grammar	0.10 (0.00–0.50)	0.10 (0.00–1.00)		
Formatting	0.00 (0.00–1.00)	0.03 (0.00–0.44)	$p = 0.033$, PEG < Google	0.18
Word choice	0.05 (0.00–0.75)	0.01 (0.00–0.50)		
Higher-level Feedback	**0.52 (0.00–1.00)**	**0.44 (0.00–1.00)**	**$p = 0.011$, PEG > Google**	**0.22**
Organization	0.00 (0.00–0.50)	0.00 (0.00–0.50)		
Idea development	0.41 (0.00–0.92)	0.30 (0.00–0.94)	$p = 0.029$, PEG > Google	0.19
Style	0.00 (0.00–0.14)	0.00 (0.00–0.33)		
Self	0.03 (0.00–0.60)	0.00 (0.00–0.29)		

Note. Effect size *r* reported as the absolute value. Bold font indicates a contrast between the aggregated higher-level feedback types and aggregated lower-level feedback types.

5.5. Teacher survey data

A survey administered to teachers at the conclusion of the study sought to ascertain teacher perceptions of the feasibility and desirability of using PEG Writing® and GoogleDocs to provide students with feedback. Responses are presented in Table 4. When asked to estimate the amount of time spent providing feedback to students in each condition, both teachers agreed that PEG Writing® was more efficient than GoogleDocs. They estimated that providing feedback saved one-third to fifty percent of the time it took to provide feedback via GoogleDocs. However, when asked to select which system was easier to use, opinions were mixed. Yet, both teachers indicated that, in comparison to GoogleDocs, PEG Writing allowed them to devote more energy to commenting on content, was easier and more motivating for students to use, and promoted greater student independence.

6. Discussion

To our knowledge, this was the first study to compare the effects of a combined teacher + automated feedback condition against a teacher-feedback-only condition (GoogleDocs) A recent literature review indicated the absence of such comparisons and the need to utilize a more ecologically valid experimental design than customary automated feedback versus no-feedback control designs (Stevenson & Phakiti, 2014).

6.1. Effects on teacher feedback: amount, type, and level

The use of PEG Writing® did have associated effects on teachers' feedback practices. Results of the current study indicated that teachers gave an equal amount of feedback to students with or without the addition of automated feedback. The finding that teachers independently provided the same median number of feedback units to students in both feedback conditions is encouraging, suggesting that AEE adoption may not signal reduction in the overall amount of teacher feedback given to students, as some have feared it might.

Teachers also did not change the type of feedback they gave across conditions. Teachers primarily relied on directives and informatives to communicate their feedback to students, using few queries and little praise. Teachers' reliance on directive feedback is consistent with prior research (e.g., Clare, Valdés, & Patthey-Chavez, 2000), and is consistent with findings that directive feedback is more effective at improving writing quality than non-directive types of feedback, such as queries and informatives (Biber et al., 2011), and more effective than praise (Hattie & Timperley, 2007).

Teachers did, however, differ in the proportion of feedback provided to students across lower- and higher-level writing skills. Teachers gave proportionately more feedback on capitalization and formatting in the GoogleDocs condition, and proportionately more feedback on sentence structure and idea development in the PEG + teacher condition. This resulted in overall difference in the proportion of higher-level feedback given across conditions, favoring the PEG + teacher condition.

This finding may help explain why, despite giving equal amounts of feedback across conditions, teachers reported that using PEG was more efficient than GoogleDocs for providing students with feedback on their writing. It is possible that providing feedback on capitalization and formatting—which included things like capitalizing proper nouns and beginning words of sentences, capitalizing dialogue, indenting paragraphs, and adding new lines for dialogue—may be more time-consuming than providing feedback on sentence structure and content. Sentence structure and content comprise broader levels of text (i.e., a broader grain size of evaluation) and are perhaps more quickly evaluated and commented on than more fine-grained features such as capitalization and formatting. Indeed, if teachers needed to stop every few words to comment on capitalization or formatting errors in GoogleDocs, their progress through the text was likely hindered, resulting in greater expenditure of time and energy than with PEG, which addressed those aspects of the text to a greater extent. Future experimental research could test this hypothesis by timing how long it takes teachers to provide feedback on texts that have differing proportions of errors in fine-grained features (e.g., capitalization, formatting, punctuation) versus problems at broader features, such as sentence construction and idea development.

However, it is important to note that the effect sizes for differences in feedback proportions across conditions were quite small (range: 0.18–0.33). Thus, the current study finds only partial support for the premise that AEE affords teachers the

Table 3
Comparison of writing quality measures by condition.

	PEG + teacher (*M*, *SD*)	GoogleDocs (*M*, *SD*)	ANOVA	Cohen's *D*
PEG overall score	20.26 (3.37)	20.44 (3.39)	$F_{(1,142)} = 0.10$, $p = 0.749$	−0.05
PEG idea development	3.18 (0.68)	3.18 (0.63)	$F_{(1,142)} = 0.01$, $p = 0.982$	0.00
PEG organization	3.26 (0.73)	3.37 (0.77)	$F_{(1,142)} = 0.72$, $p = 0.398$	−0.15
PEG style	3.47 (0.53)	3.48 (0.63)	$F_{(1,142)} = 0.01$, $p = 0.940$	−0.02
PEG word choice	3.34 (0.63)	3.36 (0.67)	$F_{(1,142)} = 0.01$, $p = 0.934$	−0.03
PEG sentence structure	3.79 (0.79)	3.70 (0.88)	$F_{(1,142)} = 0.45$, $p = 0.502$	0.11
PEG conventions	3.22 (0.59)	3.07 (0.77)	$F_{(1,142)} = 1.83$, $p = 0.179$	0.22
Holistic quality	6.17 (2.47)	6.01 (2.55)	$F_{(1,142)} = 0.13$, $p = 0.715$	0.06

Note. The PEG + teacher condition is the reference group for all effect sizes.

Table 4
Teacher perceptions of PEG Writing® and GoogleDocs.

Survey item	Response	
	Teacher A	Teacher B
Estimate the amount of time it took to provide feedback to students in the GoogleDocs condition.	6 h total	15–20 min each
Estimate the amount of time it took to provide feedback to students in the PEG Writing condition.	3 h total	10–15 min each
Which system was easier for you to use?	GoogleDocs	PEG Writing
Which system was more efficient for you to use?	PEG Writing	PEG Writing
Which system allowed you to devote more energy to commenting on content?	PEG Writing	PEG Writing
Which system seemed easier for students to use?	PEG Writing	PEG Writing
Which system was more motivating for students to use?	PEG Writing	PEG Writing
Which system promoted greater student independence?	PEG Writing	PEG Writing

ability to focus on providing feedback on higher-level writing skills. Teachers did devote proportionately more feedback to higher-level writing skills, but they still provided a substantial amount of feedback on lower-level writing skills, and provided more feedback on sentence structure in the PEG condition than in GoogleDocs. This suggests that PEG may not have adequately addressed the range of students' needs with regard to lower-level writing skills. Similar findings have been reported in studies examining the use of AEE in second-language writing instruction where students typically make high degrees of errors in lower-level writing skills; such studies have reported a lack of consistency in errors identified by AEE and teachers (Chen & Cheng, 2008; Dikli & Bleyle, 2014).

As the fields of natural language processing, machine learning, and computational linguistics continue to advance, it is possible that the accuracy of AEE feedback on lower-level writing skills will improve. In that case, teachers may achieve a more marked division of labor when using AEE to provide feedback on student writing than was found in the current study.

6.2. Effects on students' writing motivation

Consistent with prior research (Grimes & Warschauer, 2010; Warschauer & Grimes, 2008), students using AEE reported increases in their writing motivation. While the groups were equivalent at pretest, at the conclusion of the study students in the PEG + Teacher feedback condition reported stronger agreement with Item 10 of the WDS—"I take time to solve problems in my writing"—than did students in the GoogleDocs condition. This self-report data was further supported by the fact that students in the PEG + Teacher condition completed a statistically significantly greater average number of essay drafts than their peers in the GoogleDocs condition. Differences in the amount of drafts may be explained by the fact that students received quantitative and qualitative feedback each time they submitted their essay to PEG. This was likely motivating for students, reinforcing their persistence to address the feedback they received in order to gain a higher score.

However, effects on other items relating to persistence, such as items 3, 4, and 7, were not found, nor were effects on other writing dispositions, such as confidence, assessed by items 1, 5, and 11. This may have been due, in part, to the short duration of the study (11 class sessions). Changes in the broader construct of writing motivation may require additional exposure to AEE, teacher feedback, and increased opportunities for students to experience success. In addition, that effects were limited to item 10 may have been because that item appears to target a writing disposition most proximal to the processes of revising and editing. Cognitive theories of writing formulate revision as a problem-solving process (Allal, Chanquoy, & Largy, 2004; Hayes, 2012), requiring students to re-read, evaluate, diagnose, and select an appropriate action to repair the diagnosed problem. Sufficient motivation to engage in this process, and to exhibit persistence when independently solving rhetorical problems, is a mark of a skilled writer. Studies have shown that struggling writers typically lack motivation and make few revisions to their initial drafts (Troia, 2006). Thus, using an AEE system, in conjunction with teacher feedback, may afford the possibility of promoting initial gains in persistence which can then be leveraged to address the cognitive demands of revision.

6.3. Effects on students' final-draft writing quality

With respect to final-draft writing quality, results showed no statistically significant differences between conditions for the PEG Overall Score, PEG trait scores, or Holistic Quality. Previous research suggests that automated feedback leads to modest but statistically significant gains in writing quality for students in secondary and undergraduate settings (Stevenson & Phakiti, 2014) and for struggling writers (Morphy & Graham, 2012). Thus, it is surprising that the PEG did not support differential gains in writing quality in contrast to a teacher-feedback only condition. There are, however, several potential explanations for this finding.

First, the length of the intervention (10 sessions) may have been insufficient to register an effect on the PEG trait scores or the Holistic quality measure. Writing is a complex skill which develops slowly, and the intervention was brief, so it is possible that improvements occurred at a grain size finer than our outcome measures were able to detect. Had there been additional fine-grained linguistic measures available for analysis, such as the proxy scores used by PEG in its scoring model, it is possible we would have detected more of an effect. Future research is needed which increases students' exposure to PEG Writing and

its automated feedback via a longer intervention cycle. Such a situation may result in a greater effect on writing quality at the trait or holistic level.

Second, the majority of prior experimental and quasi-experimental research supporting the effectiveness of automated feedback for improving writing quality has focused on contrasting automated feedback against a no-feedback control condition (Stevenson & Phakiti, 2014). The counterfactual used in the current study was stronger, a teacher-feedback condition. Findings from the current study underscore the importance of research that adopts stronger, and more ecologically valid counterfactuals. Such research is necessary for obtaining a better understanding of the true effects of automated feedback when used as intended, effects which are likely smaller than those reported in prior research using no-feedback control conditions.

Third, while teachers in the PEG + teacher condition did provide proportionately more feedback on idea-development and higher-level writing skills, the effect sizes were small ($r = 0.19$ and 0.22, respectively). Teachers still gave a substantial amount of feedback on lower-level writing skills when using PEG. However, teachers received no instruction regarding how to integrate their feedback with PEG; they were only trained how to view PEG's feedback and how to add their own. Thus, it is promising that they intuitively shifted their behavior in the direction of the division of labor expected when using AEE, though this shift was not of sufficient magnitude to affect statistically significant differences in final-draft writing quality. A promising line of future research may be to leverage this intuitive shift in behavior and provide teachers with explicit instruction regarding how they can best complement automated feedback with effective teacher feedback shown to improve writing quality.

Fourth, effects of teacher feedback are highly variable, and not all feedback is of equal efficacy for improving writing quality (Biber et al., 2011). Given that the study collected no measure of the quality of teacher feedback—we simply described the type and level of that feedback—it is possible that teachers were not providing feedback of sufficient quality for improving overall writing quality. That is, even if teachers provided proportionately more feedback on idea-development in the PEG + teacher condition, simply increasing feedback amount does not itself signify increasing feedback quality.

Furthermore, feedback provision is not commensurate with feedback implementation. Feedback implementation is influenced by cognitive and affective features of the feedback, as well as a student's understanding and agreement with the feedback message (Nelson & Schunn, 2009). Thus, it is also possible that even if teachers provided sufficient, high-quality feedback on higher level writing skills to affect an increase in final-draft writing quality, students may not have implemented the feedback. In which case, no additional gains in writing quality would be expected. However, the current study did not analyze feedback implementation by tracing changes in the content and quality of drafts after receipt of feedback. This is a limitation of the current study, and future research should supplement measures of feedback amount, type, and quality with measures of feedback quality and feedback implementation to more carefully ascertain causal explanations for observed findings.

6.4. Additional limitations and future research

Study findings must be interpreted in light of the following limitations. First, the sample was drawn from a middle school serving a diverse, high-poverty population of students. While study findings were generally consistent with prior research, it is important to recognize that a different pattern or magnitude of effects may be observed in samples of different composition.

Second, due to study constraints, no measure of prior writing ability was collected. Nevertheless, given the consistent relationship between reading and writing skills (e.g., Abbott, Berninger, & Fayol, 2010; Ahmed, Wagner, & Lopez, 2014), were groups to have significantly differed in their writing skills, it is likely that statistically significant differences in reading ability would have been observed. Furthermore, existing literature documents a consistent relationship between writing motivation and achievement (Graham et al., 2007), and had one condition exhibited significantly greater writing achievement this would likely also have been reflected in the disposition ratings at pretest.

Third, students composed and received feedback on a single writing assignment, memoir writing, for which students possessed all relevant topic knowledge. It is unclear whether similar results would have been found had a prompt been assigned that placed greater cognitive demands on students. Given the literature on genre and task effects in writing (e.g., Graham, Hebert, Sandbank, & Harris, 2014; Schoonen, 2012), it is important that future research attempt to replicate results across different writing genres and tasks.

Fourth, the finding related to reduction in time spent providing feedback had only one form of data as support, teacher's self-report. Concrete data such as measures of time on task were not collected. As suggested earlier, it will be important for future research to obtain such measures to corroborate teacher self-report data and to examine the effects of different types of writing problems (e.g., lower-level errors or rhetorical problems) on teachers' time-spent-reviewing.

Future research should also explore whether exposing students to a combined teacher + automated feedback condition over several writing assignments would have cumulative effects on writing motivation and writing quality. It may be possible that initial increases in writing persistence reported in this study are leveraged to promote other writing dispositions, or increases in writing motivation more broadly. Similarly, it is possible that there is a cumulative effect on improvements in writing quality with additional exposure to a combined feedback condition (e.g., Ramineni, Calico, & Li, 2015). Alternatively, it is possible that initial gains in motivation may plateau if they are simply an artifact of the novelty effect associated with initial

adoption of technology-based interventions (Cheung & Slavin, 2013). These hypotheses should be tested in future research which utilizes a longer intervention cycle and which requires students to complete multiple writing tasks.

Finally, the current study was focused on writing outcomes, which is only one important facet of the validity argument for or against the use of AEE in instructional contexts. However, another important facet of validity is substantive validity (see Messick, 1996); that is, whether or not students engage in intended cognitive processes. Future research could utilize think-aloud protocols to identify similarities and differences in how students perceive and act on teacher and automated feedback. If these forms of feedback promote different cognitive processes (e.g., rereading and reviewing versus editing), this would have pedagogical implications and substantive implications for teachers and stakeholders concerned about the adoption of AEE in instructional settings.

7. Conclusion

With the increasing adoption of AEE in classroom settings in the U.S., it is important to carefully understand the associated effects on teachers' feedback practices and key student outcomes, such as writing motivation and writing quality, when AEE is used as intended. The current study provides partial support for the claim that AEE will afford teachers the ability to focus on higher-level writing skills, while increasing students' writing motivation and writing quality. Nevertheless, study findings indicate that AEE, even with its limitations, may have benefits for both teachers and students.

Acknowledgements

An earlier version of this work was presented as a paper presented at the 10th Workshop on Innovative Use of NLP for Building Educational Applications in Denver, Colorado, in June of 2015. This research was supported in part by a Delegated Authority contract from Measurement Incorporated® to University of Delaware (EDUC432914150001). The opinions expressed in this paper are those of the authors and do not necessarily reflect the positions or policies of this agency, and no official endorsement by it should be inferred.

Appendix B. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.compedu.2016.05.004.

References

Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of Educational Psychology, 102*, 281–298.

AbuSeileek, A. F. (2013). Using track changes and word processor to provide corrective feedback to learners in writing. *Journal of Computer Assisted Learning, 29*, 319–333.

Ahmed, Y., Wagner, R. K., & Lopez, D. (2014). Developmental relations between reading and writing at the word, sentence and text levels: a latent change score analysis. *Journal of Educational Psychology, 106*, 419–434.

Allal, L., Chanquoy, L., & Largy, P. (2004). *Revision: Cognitive and instructional processes*. Boston, MA: Kluwer Academic Publishers.

Allen, I. E., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality Progress, 40*(7), 64–65.

American Psychological Association, Coalition for Psychology in Schools and Education. (2015). *Top 20 principles from psychology for preK–12 teaching and learning*. Retrieved from http://www.apa.org/ed/schools/cpse/top-twenty-principles.pdf.

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistics Review, 78*, 40–64.

Anson, C. M. (2000). Response and the social construction of error. *Assessing Writing, 7*, 5–21.

Asparouhov, T., & Muthén, B. (2010). *Multiple imputation with Mplus*. Mplus Web Notes. Retrieved from http://ww.statmodel2.com/download/Imputations7.pdf.

Baker, N. L. (2014). "Get it off my stack": teachers' tools for grading papers. *Assessing Writing, 19*, 36–50.

Bangert-Drowns, R. L., Kulik, C. L., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238.

Bejar, I. I., Flor, M., Futagi, Y., & Ramineni, C. (2014). On the vulnerability of automated scoring to construct-irrelevant response strategies (CIRS): an illustration. *Assessing Writing, 22*, 48–59.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*, 289–300.

Biber, D., Nekrasova, T., & Horn, B. (2011). *The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. TOEFL iBT™ research report*. Princeton, NJ: Educational Testing Service.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7–74.

Black, P., & William, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*, 5–31.

Bunch, M. B., Vaughn, D., & Miel, S. (2016). Automated scoring in assessment systems. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 611–626). Hershey, PA: IGI Global.

Chan, P. E., Konrad, M., Gonzalez, V., Peters, M. T., & Ressa, V. A. (2014). The critical role of feedback in formative instructional practices. *Intervention in School and Clinic, 50*(2), 96–104.

Chen, C. E., & Cheng, W. E. (2008). Beyond the design of automated writing evaluation: pedagogical practices and perceived learning effectiveness in EFL writing classes. *Learning and Technology, 12*(2), 94–112.

Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: a meta-analysis. *Educational Research Review, 9*, 88–113.

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on writing: typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication, 23*, 260–294.

Clare, L., Valdés, R., & Patthey-Chavez, G. G. (2000). *Learning to write in urban elementary and middle schools: An investigation of teachers' written feedback on student compositions*. Center for the Study of Evaluation Technical Report No. 526. Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Common Core State Standards Initiative. (2010). *Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf.
Conference on College Composition and Communication. (2014). *Writing assessment: A position statement*. Retrieved from http://www.ncte.org/cccc/resources/positions/writingassessment.
DeGroff, L. (1992). Process-writing teachers' responses to fourth grade writers' first drafts. *The Elementary School Journal, 93*, 131–144.
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.
Dujinhower, H., Prins, F. J., & Stokking, K. M. (2010). Progress feedback effects on students' writing mastery goal, self-efficacy beliefs, and performance. *Educational Research and Evaluation, 16*, 53–74.
Ericcson, P. F., & Haswell, R. J. (2006). In *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.
Ferster, B., Hammond, T. C., Alexander, R. C., & Lyman, H. (2012). Automated formative assessment as a tool to scaffold student documentary writing. *Journal of Interactive Learning Research, 23*, 81–99.
Field, A. (2014). *Discovering statistics using IBM SPSS statistics* (4th ed.). Los Angeles, CA: Sage.
Foltz, P. W. (2014). *Improving student writing through automated formative assessment: Practices and results*. Paper presented at the International Association for Educational Assessment (IAEA), Singapore.
Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the intelligent essay assessor. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 68–88). New York, NY: Routledge.
Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary Street®: computer support for comprehension and writing. *Journal of Educational Computing Research, 33*, 53–80.
Graham, S., Berninger, V., & Fan, W. (2007). The structural relationship between writing attitude and writing achievement in first and third grade students. *Contemporary Educational Psychology, 32*(3), 516–536.
Graham, S., Harris, K. R., & Santangelo, T. (2015). Research-based writing practices and the common Core. *Elementary School Journal, 115*, 498–522.
Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: a meta-analysis. *Elementary School Journal, 115*, 523–547.
Graham, S., Hebert, M., Sandbank, M. P., & Harris, K. R. (2014). Assessing the writing achievement of young struggling writers: application of generalizability theory. *Learning Disability Quarterly*, 1–11.
Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104*, 879–896.
Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools – A report to Carnegie Corporation of New York*. Washington, DC: Alliance for Excellent Education.
Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: a multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6), 1–44. Retrieved December 12, 2014 from http://www.jtla.org.
Harward, S., Peterson, N., Korth, B., Wimmer, J., Wilcox, B., Morrison, T. G., et al. (2014). Writing instruction in elementary classrooms: why teachers engage or do not engage students in writing. *Literacy Research and Instruction, 53*, 205–224.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.
Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*(3), 369–388.
Hayes, J. R., & Berninger, V. W. (2010). Relationships between idea generation and transcription: how the act of writing shapes what children write. In C. Braverman, R. Krut, K. Lundsford, S. McLeod, S. Null, P. Rogers, et al. (Eds.), *Traditions of writing research* (pp. 166–180). New York, NY: Routledge.
Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63*(4), 480–499.
Higgins, D., & Heilman, M. (2014). Managing what we can measure: quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement Issues and Practice, 33*(3), 36–46.
Hyland, F., & Hyland, K. (2001). Sugaring the pill: praise and criticism in written feedback. *Journal of Second Language Writing, 10*(3), 185–212.
Keith, T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis, & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147–167). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Kellogg, R. T., & Whiteford, A. P. (2009). Training advanced writing skills: the case for deliberative practice. *Educational Psychologist, 44*, 250–266.
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.
Kluger, A. N., & DeNisi, A. (1998). Feedback interventions: toward the understanding of a double-edged sword. *Current Directions in Psychological Science, 7*(3), 67–72.
Lee, I. (2014). Feedback in writing: issues and challenges. *Assessing Writing, 19*, 1–5.
Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*, 50–60.
Marshall, B., & Drummond, M. J. (2006). How teachers engage with assessment for learning: lessons from the classroom. *Research Papers in Education, 21*, 133–149.
Matsumara, L. C., Patthey-Chavez, G. G., Valdés, R., & Garnier, G. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *Elementary School Journal, 103*, 3–25.
Messick, S. (1996). Validity of performance assessments. In G. W. Philips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). Washington, DC: National Center for Educational Statistics.
Morphy, P., & Graham, S. (2012). Word processing programs and weaker writers/readers: a meta-analysis of research findings. *Reading and Writing, 25*, 641–678.
Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
National Assessment Governors Board. (2010). *Writing framework for the 2011 National Assessment of Educational Progress*. Washington, DC: U.S. Department of Education.
National Council of Teachers of English. (2013). *NCTE position statement on machine scoring*. Retrieved from http://www.ncte.org/positions/statements/machine_scoring.
Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science, 37*, 375–401.
Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238–243.
Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of Experimental Education, 62*(2), 127–142.
Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis, & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: a review of the literature. *Reading & Writing Quarterly, 19*, 139–158. http://dx.doi.org/10.1080/10573560308222.
Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15*, 68–85.
Peterson, S., Childs, R. M., & Kennedy, K. (2004). Written feedback and scoring of sixth-grade girls' and boys' narrative and persuasive writing. *Assessing Writing, 9*, 160–180.
Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: a review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525–556.
Piazza, C. L., & Siebert, C. F. (2008). Development and validation of a writing dispositions scale for elementary and middle school students. *The Journal of Educational Research, 101*(5), 275–286.

Ramineni, C., Calico, T., & Li, C. (2015). Integrating product and process data in an online automated writing evaluation system. In O. C. Santos (Ed.), *Proceedings of the 8th International Conference on Educational Data Mining*. Madrid, Spain.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
Schmelzer, D. M. (2004). Turning out better writers: practice makes perfect, and computer-based writing makes for more practice. *Multimedia & Internet @ Schools, 11*(3), 34.
Schoonen, R. (2012). The validity and generalizability of writing scores. The effects of rater, task, and language. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van Den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology, and practices* (pp. 1–22). Leiden, The Netherlands: Brill.
Shermis, M. D. (2014). State-of-the-art automated essay scoring: competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76.
Shermis, M. D., & Burstein, J. C. (2013). *Handbook of automated essay evaluation*. New York, NY: Routledge.
Shermis, M. D., Burstein, J. C., Elliot, N., Miel, S., & Foltz, P. W. (2015). Automated writing evaluation: an expanding body of knowledge. In C. A. McArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 395–409). New York, NY: Guilford.
Shermis, M. D., Garvan, C. W., & Diao, Y. (2008, March). *The impact of automated essay scoring on writing outcomes*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.
Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62*, 5–18.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189.
Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65.
Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg Procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*, 77–83.
Troia, G. A. (2006). Writing instruction for students with learning disabilities. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 324–336). New York, NY: Guilford.
Troia, G. A., Harbaugh, A. G., Shankland, R. K., Wolbers, K. A., & Lawrence, A. M. (2013). Relationships between writing motivation, writing activity, and writing performance: effects of grade, sex, and ability. *Reading and Writing, 26*, 17–44.
U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2013, March). *What Works Clearinghouse: Procedures and standards handbook (version 3.0)*. Retrieved from http://whatworks.ed.gov.
Wade-Stein, D., & Kintsch, E. (2004). Summary street: interactive computer support for writing. *Cognition and Instruction, 22*, 333–362.
Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*, 22–36.
Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist, 54*, 594–604.
de Winter, J. C. F., & Dodou, D. (2010). Five-point Likert items: T-test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation, 15*(11), 1–16.
Wilson, J., & Andrada, G. N. (2016). Using automated feedback to improve writing quality: Opportunities and challenges. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 678–703). Hershey, PA: IGI Global.
Wilson, J., Olinghouse, N. G., & Andrada, G. N. (2014). Does automated feedback improve writing quality? *Learning Disabilities: A Contemporary Journal, 12*, 93–118.
Zumbrunn, S., Mars, S., & Mewborn, C. (2016). Toward a better understanding of student perceptions of written feedback: a mixed methods study. *Reading and Writing, 29*, 349–370.