# Automated Writing Instruction and Feedback: Instructional Mode, Attitudes, and Revising

**Rod D. Roscoe**
Human Systems Engineering
Arizona State Univ.
rod.roscoe@asu.edu

**Laura K. Allen**
Psychology
Mississippi State Univ.
laura.allen22@gmail.com

**Adam C. Johnson**
I/O Psychology
Arizona State Univ.
acjohn17@asu.edu

**Danielle S. McNamara**
Psychology
Arizona State Univ.
danielle.mcnamara@asu.edu

This study evaluates high school students' perceptions of automated writing feedback, and the influence of these perceptions on revising, as a function of varying modes of computer-based writing instruction. Findings indicate that students' perceptions of automated feedback accuracy, ease of use, relevance, and understandability were favorable. Immediate perceptions of feedback received on a selected essay were minimally related to how and whether students revised their essays. However, attitudes formed over multiple sessions were significantly related to revising. More importantly, the mode of instruction appeared to influence how feedback perceptions shaped revising behaviors. Students who engaged in traditional writing-based training and practice seemed to focus on their own perceived writing abilities when deciding how to revise. In contrast, students who also received strategy instruction and game-based practice attended more carefully to the perceived quality of the automated feedback.

**Keywords**: automated writing evaluation; user experience; user perceptions; writing instruction

## Introduction

Formative feedback (Shute, 2008) aids writing development by clarifying criteria, revealing discrepancies, and providing strategies for improvement (Kellogg, 2008; Parr & Timperley, 2010). However, feedback efficacy and uptake may be mediated by recipients' perceptions—students prefer feedback that seems detailed, encouraging, fair, and specific (Harks et al., 2014; Kaufman & Schunn, 2011; Nelson & Schunn, 2009). Such perceptions may be highly salient for *automated writing evaluation* (AWE) in which assessment is driven by software algorithms (Shermis & Burstein, 2013). Although feedback from teachers may not be perfect, the ratings originate from humans with assumed abilities for comprehending text and appreciating subjective factors such as humor. In contrast, AWE users may doubt whether a computer can perform these tasks, which may decrease adoption of the feedback (Deane, 2013). To further explore this issue, the current study explores high school students' evaluations of AWE feedback, and how these judgments influence students' revising behavior within the context of different training modes.

### Automated Writing Evaluation and Feedback

AWE systems are software programs that automate writing assessments such as scoring, error detection, and feedback (Shermis & Burstein, 2013). These functions are driven by natural language processing (NLP) tools that extract text data about text structure, wording, syntax, cohesion, and meaning (McNamara, Graesser, McCarthy, & Cai, 2014). Statistical relations between text features and writing quality can then be leveraged to assess writing (Ramineni & Williamson, 2013).

Despite such computational power, AWE excludes aspects of writing that are difficult to detect automatically, such as logic (Deane, 2013). Consequently, students might view such feedback as less valid. In fact, evaluations of AWE systems have found that many students do not revise after receiving automated feedback (Attali & Burstein, 2006; Stevenson & Phakiti, 2013; Wilson, Olinghouse, & Andrada, 2014).

Fortunately, usability studies suggest that students perceive at least some utility for AWE (Grimes & Warschauer, 2010; Roscoe, Allen, Weston, Crossley, & McNamara, 2014). One study (Roscoe, Wilson, Johnson, & Mayra, 2017) explored how college students' perceptions of automated feedback influenced writing quality, revising, and future intentions. Overall, students' perceptions had minimal impact on their "in the moment" use of the software to write and revise. Immediate perceptions were not correlated with revising. However, positive perceptions predicted willingness to use the software again or to recommend it to others. Thus, although immediate perceptions may not have a strong influence, more holistic or cumulative attitudes could have a greater impact.

Another source of students' attitudes toward AWE are their subjective user experiences (Roscoe, Branaghan, Cooke, & Craig, 2017). For instance, beliefs about one's own writing ability (Bruning, Dempsey, Kauffman, McKim, & Zumbrunn, 2013) may shape how and whether they respond to feedback. Another potential factor is the mode of training. A *traditional* approach assigns students to write, receive feedback, and then revise. This iterative and intensive writing practice may inspire strong attitudes as a function of cumulative experiences. If the feedback is consistently perceived as accurate or helpful, students' may develop greater trust in the feedback. If the feedback is consistently viewed as wrong or useless, then students may develop aggravation and resistance.

One alternative is to *enhance* AWE with tutorials on writing strategies (Roscoe & McNamara, 2013; Wilson & Czik, 2016) and practice games (Allen, Crossley, Snow, & McNamara, 2014; Roscoe et al., 2014). This mode changes the user experience as less time is generally spent on writing and revising, but new opportunities for learning are introduced through strategy instruction (Graham & Perin, 2007; MacArthur, Phillipakos, & Ianetta, 2015) and games (Connolly et al., 2012; Warren, Dondlinger, & Barab, 2008). Such training may help students understand or implement the feedback, rendering the feedback more relevant and usable.

### Research Questions and Hypotheses

Prior research suggests that perceptions of AWE feedback influence students' use of the software. Unanswered questions pertain to the effects of immediate perceptions of feedback received on a specific essay versus attitudes formed over time, and the effects of varying modes of AWE-based training. The current study addresses three research questions:

- How do students perceive the quality of feedback from an AWE system?
- How do students' perceptions and attitudes toward AWE feedback influence their revising?
- How do different modes of AWE training influence students' perceptions, attitudes, or response to feedback?

## Method

### Writing Pal (*W-Pal*)

The current study uses W-Pal, a tutoring system that trains writing strategies for adolescent students (Crossley, Allen, & McNamara, 2016; Roscoe & McNamara, 2013; Roscoe et al., 2014). W-Pal includes multimedia strategy lessons and a suite of mini-games for strategy practice. Students also practice writing and revising prompt-based, argument essays with automated feedback. Holistic essay scores are generated via NLP algorithms that evaluate lexical, syntactic, semantic, and rhetorical features (McNamara, Crossley, & Roscoe, 2013; McNamara, Crossley, Roscoe, Allen, & Dai, 2015). These NLP algorithms generate a score on a 6-point scale (from "Poor" to "Great") and inform formative assessments and feedback. The feedback focuses on actionable suggestions related to elaboration, structure, connecting with readers, making and supporting claims, and cohesion.

### Participants and Conditions

High school students ($N$ = 85) from the southwest United States enrolled in a 10-session (~3 weeks) workshop with W-Pal. Students were paid for their participation. Ethnically, 5.7% of students identified as African-American, 12.5% Asian, 19.3% Caucasian, and 54.5% Hispanic. Average age was 15.6 years with 62.1% female. Average grade level was 10.4 with 40.2% of students reporting a GPA of $\leq 3.0$.

Students in the *Traditional Mode* condition ($n$ = 45) interacted with *only* the essay and feedback tools. These students wrote (~25 minutes) and revised (~10 minutes) one training essay and then one additional practice essay per session with automated feedback. The durations of writing and revising activities were designed to mimic typical timed essay writing in standardized testing settings. Across all sessions, these students wrote and revised 16 essays.

Students in the *Enhanced Mode* condition ($n$ = 40) practiced writing and also received lessons and games. In each session, these students wrote (~25 minutes) and revised (~10 minutes) one training essay with feedback. The essay scoring and feedback tools were identical across both conditions. Students then completed one lesson module including lessons and games. Total time-on-task was designed to be equivalent to the traditional condition. Students in this condition wrote and revised 8 essays (i.e., half as much writing practice as the traditional condition).

### Essay Corpus

Across all sessions, students collectively wrote and revised over 1000 essays on a variety of prompts. For this study, we extracted essays written in the *final training session*. Students were asked to respond to a prompt about "fame" (i.e., *Are people motivated to achieve by personal satisfaction rather than by money or fame?*) and defend their point of view. The rationale for focusing on the final training essay is that students were, by that point, comfortable with the software and procedures. Thus, their immediate perceptions of the feedback received would not be influenced by factors such as novelty or confusion about navigating the system.

### Self-Assessment and Feedback Assessment Ratings

Students *self-assessed* the quality of their original draft holistically using a 6-point scale from 1 (Poor) to 6 (Great). *Feedback perceptions* were assessed after students had written an original draft, received feedback, and revised. Students rated *accuracy*, *ease of use*, *relevance*, and *understandability* on a scale from 1 (most negative) to 4 (most positive).

Students' self-assessments and feedback ratings for the selected essay represent *immediate perceptions* of their writing and feedback for that single essay. To estimate students' overarching *attitudes* toward their writing and automated feedback, we averaged ratings across *all essays*. Average ratings stem from multiple writing and revising experiences throughout the study, and thus we interpret higher mean ratings as indicative of more positive attitudes.

### Revision Coding

Coding of revisions entailed identifying each edit and then coding the (a) action taken to change the text and (b) whether revisions preserved or changed the meaning of the text (adapted from Roscoe et al., 2017).

*Revision Actions. Additions* are revisions in which new text is inserted in the essay, and *deletions* occur when text is removed without replacement. *Substitutions* occur when existing text is replaced with new text. *Reorganizations* reorder or move text from one section to another. To assess coding reliability, two raters independently categorized revisions in a subset of 30 essays. Inter-rater reliability was high ($\kappa$ = .92) and a single researcher completed the coding.

*Coding of Revision Impact. Superficial edits* preserve the meaning of the surrounding text. For example, writers might reorder sentences without changing the concepts discussed. *Substantive edits* alter the ideas in the essay. For instance, authors might substitute text that changes the interpretation of events (e.g., "the man *went* to the store" versus "the man *hurried* to the store"). As above, inter-rater reliability was strong ($\kappa$ = .81) and a single researcher completed the coding.

## Results

### Self-Assessments and Perceptions of Automated Feedback

Two one-way MANOVAs assessed the effect of condition on students' immediate perceptions and attitudes regarding their own writing and automated feedback (Table 1). There were no significant differences in immediate perceptions, Wilks' $\lambda$ = .94, $F(5,79) = 1.04$, $p = .40$, or attitudes, Wilks' $\lambda$ = .94, $F(5,79) < 1.00$, $p = .45$. Both modes were associated with positive self-evaluations and feedback evaluations.

Table 1. *Mean Ratings by Condition*

| | Enhanced | | Traditional | |
|---|---|---|---|---|
| Rating | *M* | *SD* | *M* | *SD* |
| Immediate Perceptions | | | | |
| Self-Assessment | 3.4 | 1.3 | 3.6 | 1.1 |
| Accuracy | 3.2 | 0.7 | 3.3 | 0.8 |
| Ease of Use | 3.3 | 0.8 | 3.0 | 0.9 |
| Relevance | 3.2 | 0.8 | 3.2 | 0.9 |
| Understandability | 3.2 | 0.7 | 3.3 | 0.8 |
| Attitudes | | | | |
| Self-Assessment | 3.7 | 1.0 | 3.6 | 0.8 |
| Accuracy | 3.2 | 0.5 | 3.3 | 0.5 |
| Ease of Use | 3.2 | 0.6 | 3.1 | 0.7 |
| Relevance | 3.1 | 0.6 | 3.2 | 0.6 |
| Understandability | 3.3 | 0.5 | 3.3 | 0.6 |

**Essay Scores and Revising**

*Essay Scores*. A 2 (draft) by 2 (condition) mixed ANOVA tested W-Pal-assigned essay scores as a function of revising and condition. Across conditions, students slightly improved their essay scores from original drafts ($M$ = 2.7 $SD$ = 1.0) to revised drafts ($M$ = 2.9, $SD$ = 1.1), $F(1,83) = 13.23$, $p < .001$, $d = .19$. There was no difference by condition.

*Revising*. A one-way MANOVA tested the frequency of revisions by condition and found no difference, Wilks' $\lambda = .928$, $F(8, 76) < 1.00$, $p = .66$ (Table 2).

Across conditions, all revision types occurred but varied in frequency. Additions were most common (47.5%). W-Pal often suggests elaborative strategies (e.g., generating ideas), and students may have responded to such feedback by adding detail. The next most frequent revisions were substitutions (33.6%). W-Pal feedback offers paraphrasing and cohesion-building strategies to help students improve wording, structure, and flow. Thus, the substitutions may reflect students' attempts to replace unclear or incorrect text with alternative text. Deletions (15.4%) and reorganizations (2.5%) occurred less often. Most revisions were superficial (61.8%) but substantive revisions were also observed (38.2%).

Table 2. *Mean Frequency of Revisions*

| | Enhanced | | Traditional | |
|---|---|---|---|---|
| Revisions | *M* | *SD* | *M* | *SD* |
| Total | 10.4 | 10.2 | 11.5 | 9.7 |
| Addition | 5.3 | 5.1 | 5.2 | 4.7 |
| Deletion | 1.6 | 2.7 | 1.8 | 3.2 |
| Substitution | 3.1 | 3.8 | 4.3 | 4.8 |
| Reorganization | 0.4 | 0.9 | 0.2 | 0.6 |
| Superficial | 6.3 | 9.1 | 7.2 | 7.7 |
| Substantive | 4.1 | 4.3 | 4.3 | 5.1 |

Students' revision actions and types were not significantly associated with higher quality final drafts (range of $r$ = -.03 to .12) nor to quality gains from original to revised drafts (range of $r$ = -.19 to .07). These correlations suggest that students' revisions were haphazard—they neither systematically improved nor harmed their essays. For example, some substantive additions might have improved quality by incorporating a new example, but other additions may have added inappropriate content that weakened their argument.

**Feedback Perceptions and Revising**

Correlations were conducted between perceptions, attitudes, and revising. Separate analyses were conducted for immediate perceptions and attitudes within each condition. For brevity, broad patterns are summarized rather than reporting multiple large correlation matrices.

*Immediate Perceptions: Traditional Mode.* No links were observed between immediate perceptions and revising behaviors in the traditional condition. Immediate self-assessments and feedback perceptions were not related to the total number of revision nor the frequency of any specific type of revision (all $r$s > .29, all $p$s > .05). Students revised regardless of how they perceived the feedback, and regardless of whether they self-assessed their original essay.

*Immediate Perceptions: Enhanced Mode.* A similar lack of correlations was observed between immediate perceptions and revising in the enhanced condition. Self-assessments and immediate perceptions were unrelated to the frequency of revisions of any type (all $r$s < .30, all $p$s > .05).

*Attitudes: Traditional Mode.* Attitudes about feedback accuracy, ease of use, relevance, and understandability were generally unrelated to revising behaviors (most $r$s < .30, $p$s > .05). There was one exception: attitudes about *ease of use* were related to *substitution revisions* ($r$ = .30, $p$ = .045).

An interesting and more consistent pattern was observed for students' *self-assessments*. Self-assessment attitudes were significantly and negatively correlated with *total revisions* ($r$ = -.38, $p$ = .01), *sentence revisions* ($r$ = -.36, $p$ = .014), *additions* ($r$ = -.44, $p$ = .002), *deletions* ($r$ = -.37, $p$ = .011), and *substantive revisions* ($r$ = -.38, $p$ = .01). Thus, in the traditional mode, students who tended to view their writing as higher quality also revised significantly less.

*Attitudes: Enhanced Mode.* Self-assessment attitudes were unrelated to revising behaviors ($r$s < .21, $p$s > .19). However, attitudes regarding *feedback accuracy* were significantly and positively related to *total revisions* ($r$ = .32, $p$ = .041) and *additions* ($r$ = .34, $p$ = .031), and the same trend was observed for *deletions* ($r$ = .31, $p$ = .053) and *substantive revisions* ($r$ = .31, $p$ = .051). Students who tended to view the feedback as more accurate tended to revise significantly more.

**Discussion and Conclusion**

AWE is increasingly used to support writing instruction, yet automated feedback that is doubted or disliked may be rejected. Such perceptions are highly salient for AWE because the software is expected to enact nuanced, ill-defined, and subjective writing assessment tasks that are difficult even for experienced human raters (Hamp-Lyons, 2002; Huot, 1996).

This study examined high school students' perceptions and attitudes toward automated feedback provided by the W-Pal tutoring system. Analyses contrasted two modes of computer-based training: a *traditional mode* in which students wrote and revised numerous essays with automated formative feedback, and an *enhanced mode* wherein students wrote half as much but also received direct strategy instruction and educational practice games. All students rated the quality of their own original drafts and the quality of the feedback received. For a given essay, these ratings represented *immediate perceptions*. When averaged across all sessions and essays, these ratings

approximated cumulative *attitudes*. Students' revisions were extensively coded for one target essay extracted from the final training session. Correlational analyses tested links between perceptions, attitudes, and revising within each mode.

**Findings and Implications for AWE**

*Feedback Perceptions, and Attitudes.* Students' ratings of automated feedback accuracy, ease of use, relevance, and understandability were favorable. This approval was observed in both immediate perceptions and attitudes, and corroborates prior work on the feasibility of AWE (Grimes & Warschauer, 2010; Roscoe et al., 2017). Students in both conditions also seemed able to use W-Pal feedback to revise and somewhat improve their essays. These revisions appeared to align with W-Pal formative feedback and strategies: they attempted to incorporate new content, substitute old material with new material, and make both superficial (e.g., word choice) and substantive (e.g., meaning and logic) revisions.

*Instructional Modes.* Students in both groups wrote essays of similar quality and exhibited comparable levels of revising. These equivalencies suggest that both training modes can be viable for computer-supported writing instruction. However, differences were observed for the effects of perceptions and attitudes on revising within each mode.

Immediate perceptions and self-ratings were unrelated to the frequency of revisions in either condition. In accord with prior work (Lipnevich & Smith, 2009; Roscoe et al, 2017), students' "in the moment" evaluations of automated feedback did not strongly influence whether or how they revised.

In contrast, students' attitudes—assessed by averaging their cumulative feedback perceptions and self-assessments over all sessions rather than a single session—demonstrated significant correlations with revising behaviors. As suggested by Roscoe and colleagues (2017), cumulative attitudes formed over multiple experiences may have a greater impact than fleeting perceptions. Importantly, the pattern of correlations also differed as a function of instructional mode. Students who received traditional, essay-based practice revised more when they viewed *themselves* to be less skilled writers (i.e., negative correlations between frequency of revisions and average self-assessments). In contrast, students who received a blend of strategy instruction, game-based practice, and essay-based practice were primarily influenced by attitudes *toward the feedback.* Instead of focusing on their self-assessed writing quality or abilities, these students revised more when they considered the W-Pal feedback to be accurate.

One interpretation is that the strategy tutorials and games influenced the ways in which students prioritized sources of information about their writing. When navigating the revising stage of the writing process, students might ask, "Do I need to revise?" and, if so, "What should I do?" One way to answer these questions is to consult *internal, metacognitive metrics*—self-assessments and beliefs about one's own writing ability (Bruning et al., 2013; Harris, Santangelo, & Graham, 2010). For example, students who believe themselves to be generally decent writers may assume their initial drafts are "good enough" and require few edits. Similarly, a lack of self-efficacy may lead students to avoid revising because it seems unlikely to be successful. Another source of information are *external evaluations and recommendations*, such as feedback from peers (Patchan & Schunn, 2015; Patchan et al., 2016) or teachers (Ferris, 2014; Parr & Timperley, 2010). Praise or critique from a trusted reviewer can be used to determine whether and what type of revisions are necessary.

Students who interacted solely with W-Pal's AWE tools seemed to focus on metacognitive self-assessments. They revised more if they felt that their writing was not typically very good (i.e., valid self-regulation). One possibility is that students' trusted their own self-judgments more than the software. This aligns with prior findings that students trust human evaluations more than automated feedback (Dikli & Bleyle, 2014; Lipnevich & Smith, 2009).

Students who also interacted with W-Pal's strategy tutorials and practice games seemed more attuned to the external, automated feedback. These students appeared to pay more attention to W-Pal feedback and, if it seemed to be accurate, revised their essays. One interpretation is that the W-Pal tutorials and games made the feedback more concrete or otherwise improved students' trust in the system.

Research on trust in automation (e.g., Hoff & Bashir, 2015; Schaefer, Chen, Szalma, & Hancock, 2016) describes how features of the users, automation, and environment can influence trust. For example, trust is generally higher when users feel that they understand how the automated system operates, perceive the automation as reliable, and when tasks are cognitively supported (Schaefer et al., 2016). In this study, all students interacted with the same automated feedback system. However, the additional tutorials and games may have seemed to reveal the "inner workings" of W-Pal—what the software was "looking for" when assessing writing—which contributed to a sense of better understanding of how W-Pal operates. The lessons and games provided added cognitive support for interacting with the automated feedback, perhaps inspiring more feedback awareness or trust.

In future work on AWE, an expanded analysis of students' beliefs and expectations of human-automation interactions (e.g., their mental models; see Endsley, 2017) may be fruitful. This wider lens may shed light on how and when students choose to use automated feedback, and may guide new hypotheses about the effective design of AWE tools. For instance, do students believe that the system is merely "counting words" (see Perelman, 2014)? Do they believe the system utilizes spelling and grammar-checking akin to word processing? If students understand that NLP algorithms typically model myriad linguistic features, what degree of weighting or importance do they believe is placed on each feature? Future research on students' conceptions of automation may further reveal how and whether doubts about AWE feedback arise and influence their behavior.

**Acknowledgments**

This research was supported by the Institute of Education Sciences, US Department of Education, through Grant R305A120707. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

## References


Allen, L. K., Crossley, S. A., Snow, E. L., & McNamara, D. S. (2014). Game-based writing strategy tutoring for second language learners: game enjoyment as a key to engagement. *Language Learning and Technology*, *18*, 124-150.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v. 2. *Journal of Technology, Learning, and Assessment*, *4*, 123-212.

Bruning, R., Dempsey, M., Kauffman, D. F., McKim, C., & Zumbrunn, S. (2013). Examining dimensions of self-efficacy for writing. *Journal of Educational Psychology*, *105*, 25-38.

Crossley, S. A., Allen, L. K., & McNamara, D. S. (2016). The Writing Pal: A writing strategy tutor. In S. A. Crossley & D. S. McNamara (Eds.), *Adaptive educational technologies for literacy* (pp. 204-224). New York, NY: Routledge.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*, 7-24.

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.

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors*, *59*, 5-27.

Ferris, D. R. (2014). Responding to student writing: Teachers' philosophies and practices, *19*, 6-23.

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

Grimes, D., & Warschauer, M. (2010). Utility in a fallible took: a multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, *8*, 4-43.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, *8*, 5-16

Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2014). The effects of feedback on achievement, interest, and self-evaluation: the role of feedback's perceived usefulness. *Educational Psychology*, *34*, 269-290.

Harris, K. R., Santangelo, T., & Graham, S. (2010). Metacognition and strategies instruction in writing. In H. S. Waters & W. Schneider (Eds.), *Metacognition, strategy use, and instruction* (pp. 226-256). New York, NY: Guildford Press.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*, 407-434.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication*, *47*, 549–566.

Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: their origin and impact on revision work. *Instructional Science*, *39*, 387-406.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, *1*(1), 1-26.

Lipnevich, A. A., & Smith, J. K. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*, *5*, 319-333.

MacArthur, C. A., Philippakos, Z. A., & Ianetta, M. (2015). Self-regulated strategy instruction in college developmental writing. *Journal of Educational Psychology*, *107*, 855-867.

McNamara, D. S., Crossley, S. A., & Roscoe, R. D. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*, 499-515.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35-59.

McNamara, D. S., Graesser, A. C., McCarthy, P., Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

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.

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing*, *15*, 68-85.

Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: How students respond to peers' texts of varying quality. *Instructional Science*, *43*, 591-614.

Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016). The nature of feedback: How peer feedback features affect students' implementation of rate and quality of revisions. *Journal of Educational Psychology*, *108*, 1098-1120.

Perelman, L. (2014). When the "state of the art" is counting words. *Assessing Writing*, *21*, 104-111.

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*, 25-39.

Roscoe, R. D., Allen, L., Weston, J, Crossley, S., & McNamara, D.S. (2014). The Writing Pal Intelligent Tutoring System: usability testing and development. *Computers and Composition*, *34*, 39-59.

Roscoe, R. D., Branaghan, R. J., Cooke, N. J., & Craig, S. D. (2017). Human systems engineering and educational technology. In R. D. Roscoe, Scotty D. Craig, & I. Douglas (Eds.), *End-user considerations in educational technology design*. Hershey, PA: IGI Global.

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, *105*, 1010-1025.

Roscoe, R. D., Wilson, J., Johnson, A. C., & Mayra, C. R. (2017). Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior*, *70*, 207-221.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy of future systems. *Human Factors*, *58*, 377-400.

Shermis, M. D., & Burstein, J. C. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*, 153-189.

Stevenson, M., & Phakiti, A. (2013). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*, 51-65.

Warren, S. J., Dondlinger, M. J., & Barab, S. A. (2008). A MUVE towards PBL writing: Effects of a digital learning environment designed to improve elementary student writing. *Journal of Research on Technology in Education*, *41*, 113-140.

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., Olinghouse, N. G., & Andrada, G. N. (2014). Does automated feedback improve writing quality? *Learning Disabilities: A Contemporary Journal*, *12*, 93-118.