Learning Disabilities: A Contemporary Journal 12(1), 93-118, 2014

Does Automated Feedback Improve Writing Quality?

Joshua Wilson*
University of Delaware

Natalie G. Olinghouse
University of Connecticut

Gilbert N. Andrada
Connecticut State Department of Education

The current study examines data from students in grades 4-8 who participated in a statewide computer-based benchmark writing assessment that featured automated essay scoring and automated feedback. We examined whether the use of automated feedback was associated with gains in writing quality across revisions to an essay, and with transfer effects to improved first-draft performance or accelerated growth when composing and revising a follow-up prompt. Three-level hierarchical linear modeling revealed that writing quality improved across revisions, though growth decelerated over time. Females and students with higher prior writing achievement scored higher for first-draft performance, but students receiving free and reduced lunch grew at a slower rate than their peers. No significant transfer effects were observed, but the effect of socioeconomic status on growth in writing quality was no longer significant in models describing performance on a follow-up prompt.

Keywords: Automated feedback, writing quality, writing assessment.

Providing instructional feedback is an effective method for improving students' writing quality (Graham, Harris, & Hebert, 2011; Graham, McKeown, Kiuhara, & Harris, 2012). Specifically, a recent meta-analysis summarizing 27 studies involving feedback from various sources (e.g., adult, peer, self, or computer) found an average weighted effect size of 0.61 on measures of writing quality (Graham, Hebert, & Harris, under review). Struggling writers, in particular, need targeted instructional feedback given the range of difficulties they have. For example, struggling writers tend to produce shorter, less-developed, and more error-filled texts than their peers, and tend to spend very little time planning and revising their text (Troia, 2006).

However, though instructional feedback may be effective at addressing these deficits, feedback is often difficult and time-consuming for teachers to provide. Consequently, educators are increasingly relying on automated essay-scoring systems (AES) to provide students with immediate quantitative and qualitative feedback regarding their writing. Indeed, computer-based benchmark and formative writing assessments that incorporate AES are currently under-development by two national assessment consortia: the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for the Assessment of Readiness for College and Careers (PARCC) (U.S. Department of Education, 2012a, 2012b). As AES and automated feedback become more widely adopted and utilized in classroom settings, further research is needed to

*Please send correspondence to: Joshua Wilson, School of Education, Willard Hall, Newark, DE 19716.
Email: joshwils@udel.edu.

explore whether automated feedback is useful for scaffolding improvements in writing quality, especially for struggling writers.

Instructional Feedback

Hattie and Timperley (2007) define feedback as information provided by an "agent" regarding specific aspects of a student's task performance or conceptual understanding. *Instructional feedback* goes a step further, in that the provided information not only indicates correctness/incorrectness, but clearly indicates ways to improve performance or understanding (Hattie & Timperley, 2007). This definition of instructional feedback applies to feedback provided by any type of agent, such as an adult, peer, or computer (i.e., automated feedback).

One theory that explains the effects of instructional feedback on writing quality is Vygotsky's (1978) notion of the zone of proximal development (ZPD). By providing feedback, the agent—be it a teacher, peer, or computer—creates a space for development, called the zone of proximal development, in which the student gradually internalizes the instructional feedback and achieves a new level of independent performance. Based on this theory, the knowledge and experience of both the agent and the student are key components of effective instructional feedback.

The agent. To provide effective feedback, the agent must accurately evaluate a text and diagnose key areas for improvement. To do so, requires knowledge of what characterizes high-quality writing; however, such knowledge may be hard to acquire in terms of writing. First, there is no universally agreed-upon definition of writing quality (Elliott, 2005). Indeed, what characterizes high-quality writing has been shown to vary across raters and tasks (Baker, Abedi, Linn, & Niemi, 1995; Chen, Niemi, Wang, Wang, & Mirocha, 2007), as well as writing purposes and genres (Olinghouse, Santangelo, & Wilson, 2012; Olinghouse & Wilson, 2013). Second, writing is a complex, multicomponential skill (Berninger & Swanson, 1994; Flower & Hayes, 1980; Hayes, 2012; Hayes & Berninger, in preparation) comprised of numerous subskills such as spelling, grammar, word choice, sentence structure, organization, and idea development. This complexity complicates diagnosis of writing difficulties—students may exhibit a wide array of writing strengths and weaknesses (Berninger, Abbott, Whitaker, Sylvester, & Nolan, 1995; Scott, 2009). The complexity of writing also may make it more difficult for the agent to provide feedback that falls within students' ZPD. Third, the provided feedback may vary across adults and peers, or even within the same individual from one time period to another.

The student. In order to leverage feedback to improve writing quality, students must be able to understand and apply the feedback targeted at their ZPD (Beach & Friedrich, 2006). While research shows that variables such as grade-level, gender, prior writing achievement, and socio-economic status predict significant variance in: (a) writing performance, (b) writing knowledge, and (c) writing motivation and self-regulation skills (Berninger & Swanson, 1994; Graham & Harris, 2000; Loban, 1976; Olinghouse & Graham, 2009; Troia, Shankland, & Wolbers, 2012), additional research is needed to explore whether variables such as these also predict the ability to benefit from instructional feedback in writing.

Automated Feedback

Compared to adult- or peer-feedback, automated feedback has a number of advantages. Automated feedback removes the knowledge-barrier humans face with providing effective feedback. AES is able to instantly evaluate hundreds of text markers and identify writing errors. Based on this evaluation, AES systems generate either: (a) a single score representing the overall quality of the student's text (akin to holistic scoring), (b) separate scores across specific traits of writing (akin to analytic scoring), or (c) trait scores and a measure of overall quality which is formed by either averaging or summing the separate trait scores. Based on their specific writing errors, students receive automated feedback suggesting ways to improve their writing. Unlike human raters, AES represents a 100% consistent scoring system. AES is able to consistently identify error patterns in students' texts and no matter how many times the same text is submitted for evaluation it will always receive the same score. Finally, because automated feedback provides instant feedback, it can help accelerate the feedback-practice loop essential for developing writing skills (Kellogg & Whiteford, 2009).

Automated feedback and writing performance. Only a small number of studies have examined the effect of automated feedback on writing quality. These studies have focused on two automated feedback systems, ETS's e-rater® and Criterion® system and Pearson's Summary Street system. No previous studies have investigated the effects of automated feedback from Project Essay Grade (PEG; Page, 1966, 1994), the AES system which is the subject of the present study.

Kellogg, Whiteford, and Quinlan (2010) assigned undergraduate students into three groups which varied exposure to automated feedback provided by e-rater and Criterion: no feedback, intermittent feedback, or continuous feedback. Students composed and revised essays based on automated feedback regarding grammar, usage, mechanics, style, organization, and development. In addition, students also received an e-rater overall quality score (range: 1-6). Results indicated that undergraduate students who received continuous automated feedback successfully reduced the number of errors in grammar, usage, mechanics and style in final drafts of three essays, and did so to a greater degree than students receiving either intermittent or no feedback. However, there were no reliable differences in the e-rater overall quality score between feedback-conditions across drafts.

A second study examining the effects of automated feedback from e-rater and Criterion (Shermis, Wilson Garvan, & Diao, 2008) used hierarchical linear modeling to examine whether continuous practice and exposure to automated feedback was associated with growth in the writing quality of final drafts of seven subsequent essays written by students in grades 6-8 and 10. Analyses examined growth in several measures, including: the e-rater overall quality score; text length; number of unique words; and errors in grammar, usage, mechanics, and style. Students demonstrated improvements in each of the dependent measures across the seven final drafts, with eighth graders outperforming other students. Although results indicated growth in the e-rater overall quality score, this effect may be masking practice effects and general growth effects. Since the timeframe of the study was long (11 months) and there was no control for practice effects or classroom instruction, it is possible that there may have been very little improvement from first draft to final draft and that practice

and classroom instruction, not automated feedback, explained the growth in e-rater overall quality across essays.

Two studies examined the effects of automated feedback provided by Summary Street on the quality of summaries written by upper-elementary and middle-school students. The first, a study by Wade-Stein and Kintsch (2004), used a repeated measures design to analyze the effect of Summary Street feedback on sixth-graders' summary writing. Half the students composed and revised a summary with feedback from Summary Street, while half the students received feedback only on the length of their summary. Conditions were counterbalanced so all students were exposed to both feedback conditions. Dependent measures were time on task, human-scored content ratings, and human-scored holistic quality ratings (1-5 scale). Students exposed to Summary Street feedback spent more time on task and wrote summaries rated higher for both content and holistic quality. However, the effect on overall quality disappeared when controlling for students' improvements in content ratings, the aspect of the text for which they received automated feedback.

A second study (Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005) used within-classroom random assignment to assign 111 eighth grade students to either an experimental condition that utilized feedback from Summary Street or a control condition that utilized a word processor which provided feedback regarding spelling and text length. Students in each condition practiced composing summaries twice a week for four weeks. Results indicated that students using Summary Street composed summaries judged to be superior to those composed in the control condition in terms of holistic quality, organization, content, sparing use of detail, and style. There were no effects of condition on mechanics, though this is unsurprising given that control students had access to the word processor's spell check function. The authors did not conduct a similar analysis as in the previous study (Wade-Stein & Kintsch, 2004) to determine if these effects were explained by improvements in the content of the summaries.

Transfer effects of automated feedback. Vygotsky's (1978) theory of the zone of proximal development suggests that sufficient exposures to instructional feedback will result in gradual internalization and subsequent improvements in independent performance. Three of the previous studies explicitly looked at maintenance and transfer; however, evidence of transfer effects appears limited to improving performance in specific traits, rather than improving overall quality. Two weeks after the conclusion of a practice phase, Kellogg et al. (2010) asked students to write and revise a maintenance prompt independent of automated feedback. Results showed a transfer effect for students in the continuous feedback condition on reducing the number of errors in grammar, usage, mechanics, and style, but no effect on the e-rater quality score. Wade-Stein and Kintsch (2004) counterbalanced students' exposure to automated feedback from Summary Street: half of the sample used Summary Street in week one and half of the sample used it in week two. Students who received Summary Street during week-one maintained strong content scores when creating summaries independent of automated feedback in week-two, but no significant improvements in holistic quality were noted. Franzke et al. (2005) examined transfer effects in two ways: (a) unaided transfer in completing a reading comprehension post-test which included a variety of items, including summary items; and (b) aided transfer in com-

posing summaries of increasingly difficult texts over a four-week period. While there was no difference between conditions on overall post-test comprehension scores, students in the Summary Street condition outperformed control students on summary items. Students in the Summary Street condition were also able to maintain their level of performance across time despite being exposed to increasingly more difficult texts to summarize. These findings were confirmed in a later study (Caccamise, Franzke, Eckhoff, Kintsch, & Kintsch, 2007 [study 1]) of students in grades seven through nine which showed that, in contrast to a control group, the use of Summary Street was associated with transfer effects when composing a post-test summary without the aid of the feedback program. However, in this study, students from the control group were taken from multiple classes across the curriculum which may or may not have been exposed to instruction or practice composing summaries.

In summary. Previous research suggests positive effects of automated feedback on writing ability. Indeed, in a recent meta-analysis (Graham, Hebert, & Harris, in press), the overall weighted effect size of automated feedback across four studies was reported to be 0.38. Though this is a moderate positive effect, additional research is needed given that the use of automated essay scoring and automated feedback is poised to assume larger roles in classroom learning. Indeed, previous studies are inconclusive with regards to whether the effects of automated feedback are isolated to improving specific traits of writings—those that are targeted by the feedback program—or whether they generalize to improvements in the overall quality of the text across several traits. In addition, mixed evidence of transfer effects further supports the need for additional research to examine whether automated feedback appears to target students' zone of proximal development.

The Present Study

The current study extends previous research in several ways. First, this is the first study to explore automated feedback provided by PEG. Second, we examine growth across multiple revisions to a writing prompt (range: 3-18 revisions). Third, we explore whether student characteristics predict the ability to apply automated feedback to improve writing quality. Finally, we explore transfer effects in two ways: (a) by assessing whether students demonstrate improved first-draft performance on a subsequent prompt (i.e., unaided transfer), and (b) by assessing whether students utilize automated feedback more efficiently and effectively to revise a second prompt (i.e., aided transfer). Accordingly, we asked the following research questions:

RQ1: Is the use of automated feedback associated with growth in writing quality across successive revisions to a writing prompt? If so, what is the shape of this growth?

RQ2: Is there a differential effect of grade-level, gender, prior writing achievement, or socio-economic status on first-draft performance or growth in writing quality across revisions?

RQ3: Is the use of automated feedback associated with improved first-draft performance or more rapid growth when composing and revising a follow-up prompt?

Method

Data Source

The current study analyzes data from a sample of fourth- through eighth-grade students who participated in a statewide classroom benchmark writing assessment administered in academic year 2012. In accord with The Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99), both individual participant data as well as the identity of the state were de-identified.

The statewide benchmark writing assessment is a freely available, voluntary (non-compulsory), non-stakes assessment designed to support classroom formative and summative writing assessment. It is a computer-based assessment offered to students in grades 3-12 and available continuously throughout the entire school year. Students access the assessment by logging-in to a web-based application via the state department of education and inputting their unique ID and password. System-created writing prompts are provided in multiple genres (e.g., narrative, persuasive, expository, informational), and teachers also may create their own prompts. Examples of system-created prompts are: "What Makes for Success?" (persuasive), "An Important Person" (informative/descriptive), and "Magical Remote Control" (narrative). Examples of teacher-created prompts are: "Is the Cafeteria Ruining Your Life?" (persuasive), "Adventure to Antarctica" (informative), and "Personal Narrative" (narrative). Figure 1 displays a screen shot of the prompt-selection screen. Once selected, students have up to 60 minutes to type their response (see Figure 2).

Once completed, student responses are immediately scored via an AES system called Project Essay Grade (PEG; Page, 1966, 1994). PEG reports scores on a 1-6 scale for six traits of writing ability: overall development, organization, support, sentence structure, word choice, and mechanics (see Appendix A). Also reported is an Overall Score (range: 6-36), which is formed as the sum of these six trait scores. PEG also provides students with individualized automated feedback which students use to revise their response (see Appendix B)[1]. Students may revise and resubmit their response as many times as they like. Based on results, customized links to online writing instruction resources are provided to teachers and students. Finally, the assessment includes a data-management function, enabling teachers to monitor student and classroom performance (see Figure 3).

Project Essay Grade. PEG was developed by Ellis Page and colleagues (Page, 1966; Page, 1994) and was acquired by Measurement Incorporated in 2003. PEG uses a syntactic text parser to measure "proxes" (i.e., syntactic indicators) of intrinsic characteristics of writing quality called "trins" (i.e., traits). These proxes are combined in a regression-based algorithm to model human holistic scores and analytic scores across six traits of writing ability. Numerous studies have shown PEG to demonstrate high reliability and validity both for predicting pairs of human raters' holistic scores (e.g., Keith, 2003; Shermis, Mzumara, Olson, & Harrington, 2001) as well as analytic trait scores (Page, Poggio, & Keith, 1997; Shermis, Koch, Page, Keith,

1 The examples of feedback provided in Appendix B represents the feedback received by students participating in the statewide benchmark writing assessment in AY 2012. As of July 12, 2013 Measurement Incorporated updated its automated feedback both in terms of its alignment with measured traits and with students' grade-levels.

& Harrington, 2002). Indeed, in a study comparing holistic and trait ratings of six human raters with PEG (Shermis et al., 2002), confirmatory factor analysis was used to estimate the latent true score of an essay from all possible pairs of human raters and PEG. PEG demonstrated the highest standardized coefficient on the latent essay true score (.89), superior to all but one pair of human raters.

Figure 1. **Student View of Prompt Selection Screen**

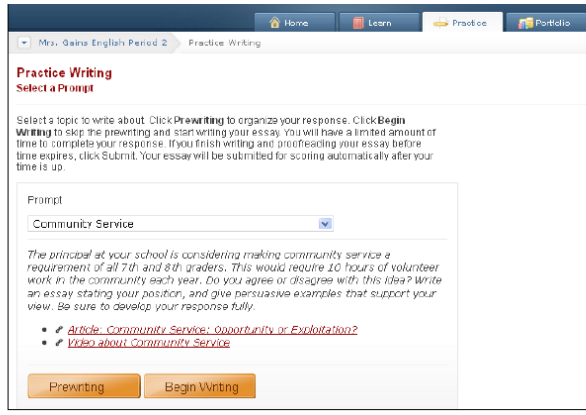


Figure 2. **Student View of Prompt Screen**

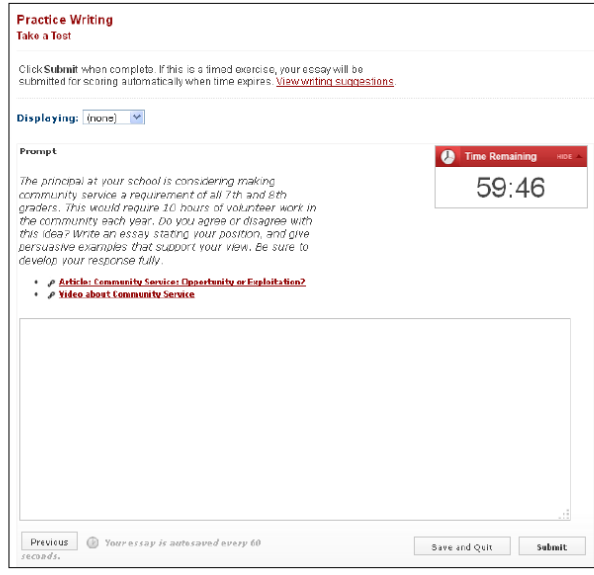
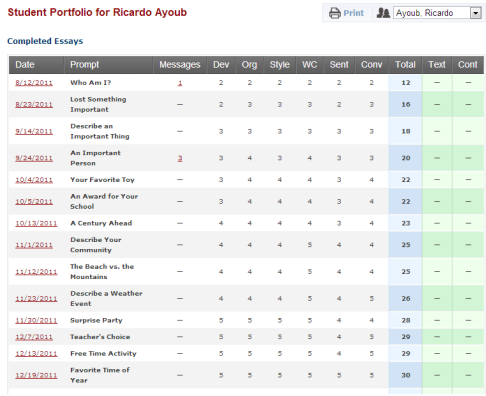


Figure 3. **View of a Student's Portfolio of Completed Essays**



Student Portfolio for Ricardo Ayoub

Print | Ayoub, Ricardo

Completed Essays

Date	Prompt	Messages	Dev	Org	Style	WC	Sent	Conv	Total	Text	Cont
8/12/2011	Who Am I?	1	2	2	2	2	2	2	12	–	–
8/23/2011	Lost Something Important	–	2	3	3	3	2	3	16	–	–
9/14/2011	Describe an Important Thing	–	3	3	3	3	3	3	18	–	–
9/24/2011	An Important Person	3	3	4	3	4	3	3	20	–	–
10/4/2011	Your Favorite Toy	–	3	4	4	4	3	4	22	–	–
10/5/2011	An Award for Your School	–	3	4	4	4	3	4	22	–	–
10/13/2011	A Century Ahead	–	4	4	4	4	3	4	23	–	–
11/1/2011	Describe Your Community	–	4	4	4	5	4	4	25	–	–
11/12/2011	The Beach vs. the Mountains	–	4	4	4	5	4	4	25	–	–
11/23/2011	Describe a Weather Event	–	4	4	4	5	4	5	26	–	–
11/30/2011	Surprise Party	–	5	5	5	5	4	4	28	–	–
12/7/2011	Teacher's Choice	–	5	5	5	5	4	5	29	–	–
12/12/2011	Free Time Activity	–	5	5	5	5	4	5	29	–	–
12/19/2011	Favorite Time of Year	–	5	5	5	5	5	5	30	–	–

Recently, the Hewlett Foundation sponsored a competition among nine AES vendors to compare performance of AES engines and pairs of human raters for reliably scoring extended essay responses. Results indicated that PEG was the highest performing AES system for scoring essay-length extended response items (Shermis & Hammer, 2012; Automated Student Assessment Prize [ASAP], 2012). PEG also won a follow-up competition for scoring short constructed response items (ASAP, 2013). Currently, PEG is one of three AES systems, along with ETS's Criterion and Pearson's WriteToLearn®, selected for a nationwide classroom trial in seventh-grade classrooms across the United States (https://classroomtrials.rampit.com/).

Despite evidence of the reliability and validity of scores generated from PEG, no previous study has examined whether PEG has utility as an automated feedback system. Given the literature supporting the use of instructional feedback for promoting learning and achievement (Hattie & Timperley, 2007), it is logical that the instructional feedback provided by PEG may assist students in scaffolding improvements in writing quality. However, there has been no previous research to empirically assess whether this logic holds. This gap in the literature served as further impetus for the current study.

Participants

The current study analyzes data from fourth- through eighth-grade students who participated in the statewide benchmark writing assessment between September 2012 and January 2013. A total of 4,162 students in grades 4-8 from 28 schools completed a minimum of one essay during this time. To estimate a growth-model, a minimum of three observations per student was necessary; thus, for our sample we selected all students who completed two or more revisions to their first draft. This yielded a total of 955 students from 14 different schools to answer research questions 1 and 2. To examine transfer effects (research question 3), we retained students who completed: (a) a first draft to a subsequent writing prompt (i.e. unaided transfer), and (b) a minimum of two revisions to this follow-up prompt using PEG feedback (i.e., aided transfer). Respectively, this yielded: (a) 739 students from 12 schools,

and (b) 486 students from 12 schools. Table 1 provides demographic data for each successive sample.

Chi-square tests of independence were used to assess statistically significant differences in proportions of demographic variables across the three different samples. For each comparison test, the null hypothesis of equal proportions was retained. A one-way ANOVA was used to assess statistically significant differences in mean prior writing achievement across samples. Again, the null hypothesis of equal means was retained [F (2, 2,177) = 2.748, p = 0.064]. Thus, the samples did not statistically significantly differ on any of the observed demographic or prior achievement variables.

Table 1. ***Demographic and Descriptive Information for Successive Samples***

Variable	Sample A	Sample B	Sample C
Total Students (*n*)	955	739	486
Total Schools (*n*)	14	12	12
Grade-band (%)			
4-5	8.5	4.3	3.1
6-8	91.5	95.7	96.9
Gender (%)			
Male	45.0	43.6	42.8
Female	55.0	56.4	57.2
FRL (%)	34.3	32.6	33.3
Race (%)			
White	59.4	61.7	65.8
Hispanic	22.5	20.0	18.9
Black	12.4	12.3	10.9
Other ethnicities	5.7	6.0	4.4
SPED[a] (%)	6.1	6.5	5.8
ELL[b] (%)	3.1	2.3	2.3
Prior Writing Achievement[c] (*M*, *SD*)	257.62 (38.67)	262.77 (61.13)	263.20 (61.83)

Note. Sample A was used to answer research questions 1 and 2. Sample B was used to answer research question 3 regarding improved initial performance on a successive prompt. Sample C was used to answer research question 3 regarding accelerated growth revising a successive prompt.
[a]SPED = Students who receive special education services with an individualized education plan (IEP). [b]ELL = English language learners. [c]Prior writing achievement based on students' scale score (range:100-400) from the prior year's state writing achievement test.

Design and Data Analysis

Hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002) was used to examine whether using automated feedback to revise a writing prompt was associated

with gains in writing quality. A three-level HLM longitudinal model was specified in which repeated observations of the outcome variable (level-1) were nested within individual students (level-2) who were nested within different schools (level-3). Compared to analyses which ignore the presence of such clustering, HLM yields more precise parameter estimates and standard errors which thereby reduces Type I error rates (McCoach, 2010). A further advantage of HLM for analyzing longitudinal data is the ability to analyze unbalanced datasets, in which the number of measurement occasions varies across subjects (Singer & Willett, 2003), as it does in this study.

All analyses were completed using HLM 7 (Raudenbush, Bryk, & Congdon, 1988) using Full Information Maximum Likelihood (FIML) estimation. All estimated models iterated and converged without errors.

Prompts and Revisions

The post-hoc nature of the study design meant that several variables were not controlled, such as: (a) the topic of the prompts students responded to, (b) the genre of those prompts, (c) whether prompts were teacher- or system-created, (d) the elapsed time between revisions of a prompt, and (e) the elapsed time between students completing their first and second prompt. Across the sample, students responded to one or more of 110 different writing prompts. Of these, 42.8% were persuasive, 37.1% were informative/descriptive, 9% were expository, and 3.8% were narrative. The majority of these prompts were teacher-created (64.4%). Students also exhibited a wide range in the number of revisions they completed (range: 2-74). For data analysis we elected to use a trimmed dataset which included students who revised up to a maximum of 18 revisions, a range which represented the number of revisions completed by 95% of the sample (range: 2-18 prompts). The elapsed time between revisions and between the first and second essay varied across students. Some students completed several revisions within a single one-hour session, while others completed their revisions over several days.

Measures

A 3-level HLM growth model allows inclusion of predictor variables at each level of the model. However, the present study posed substantive questions pertaining only to level-1 (the shape of individual growth trajectories) and level-2 (individual predictors of performance and growth). Thus, models included predictor variables at these levels, but not at level-3.

Outcome variable. The PEG Overall Score was used as the outcome variable measuring writing quality. It is the sum of each student's individual trait scores, and ranged from 6-36. We selected this outcome variable because individual traits were highly correlated (range: $r = .732$-$.918$), it had greater variability than the individual trait scores, and we were interested in the effect of AES feedback on overall writing quality. The PEG Overall Score provides a highly reliable measure of writing quality across time—PEG is absent of human-rater error and displays intra-rater reliability of 1.00. Thus, growth in writing quality was operationalized as growth in PEG Overall Score across successive revisions to a writing prompt.

Level-1 variables. Initial inspection of students' individual growth curves revealed a gradual rise in Overall Score followed by a plateau after several revisions.

Hence, a polynomial growth model was estimated using two variables to measure time: (a) *Time*, a variable which simply counted each successive revision; and (b) $Time^2$, a variable formed by raising *Time* to the second-power. Together these variables were used to estimate the instantaneous rate of change and deceleration of students' growth curves. Both variables were centered such that 0 represented students' performance on their first draft prior to completing any revisions.

Level-2 variables. Several level-2 variables were used to determine whether certain student characteristics were associated with the ability to apply automated feedback to improve writing quality. First, a dummy-coded variable called *Middle-School* represented whether a student was in grades 6-8 (Middle School = 1) or grades 4-5 (Middle School = 0). We hypothesized that grade-level effects, if any, would be restricted to the difference between middle-school and upper-elementary students rather than successive grade-level differences. Second, a dummy-coded variable called *Female* (1 = females, 0 = males) was included because of documented gender-effects in the development of writing ability favoring females (see Berninger & Swanson, 1994). Third, we measured students' prior writing achievement using performance on the spring 2012 state writing assessment. This variable was reported as a scale-score (range: 100-400) and was included in models using grand-mean centering. Finally, we used a dummy-coded variable called *Free or Reduced Lunch* (1 = FRL, 0 = not FRL) as a measure of socio-economic status. Table 1 includes descriptive data for these predictors.

Results

Research Question 1

Table 2 reports results of the unconditional means model, unconditional quadratic growth model, and final conditional model for data from students' first prompt. The variance components of the unconditional means model were used to calculate intra-class correlations. The proportion of variance in the PEG Overall Score that was within students (level-1) was 12.64%, while 83.48% fell between students within schools (level-2), and 3.88% fell between schools (level-3). These findings support the choice to evaluate a three-level model but to focus on explaining within- and between-student variance, rather than between-school variance. We tested several level-1 growth models: linear, logarithmic, exponential (raised to 0.5 power), and quadratic. The quadratic model provided the best fit to the data and yielded a 68% reduction in within-student variance compared to the unconditional means model.

The final results showed that, when averaging across students within schools, the PEG Overall Score for a first-draft was approximately 21.5 points. With each successive revision, the PEG Overall Score increased 0.5 points and decelerated at the rate of -0.02 points multiplied by the respective values for each time variable. The nonlinear and concave shape of growth (Figure 4) suggests that a plateau occurs as students maximize the benefits they derive from the automated feedback. The point at which growth slows to 0 is called the saturation point, and it can be estimated using the equation $[(-1*\pi_{1i})/(2*\pi_{2i})]$ (Singer & Willett, 2003). In this sample, the saturation point appears with the completion of 11 revisions, suggesting that after 11 revisions

students are predicted to have maximized their growth, from 21.5 points to 24.51 points.

Table 2. ***Results of Models Estimating Initial Performance and Growth Across Revisions for Prompt 1.***

		Unconditional Means Model		Unconditional Quadratic Growth Model		Final Conditional Model	
		Coefficient (S.E.)	*t (df)*	Coefficient (S.E.)	*t (df)*	Coefficient (S.E.)	*t (df)*
Fixed Effects							
Initial Status, π_{0i}	Intercept	22.42*** (0.38)	59.63 (13)	21.51*** (0.37)	58.04 (13)	20.78*** (0.44)	47.81 (13)
	Gender					0.98*** (0.29)	3.38 (922)
	Prior Writing Achievement					0.06*** (0.004)	14.01 (922)
	FRL					-	-
Rate of Change, π_{1i}	Intercept			0.52*** (0.06)	8.89 (13)	0.57*** (0.06)	10.28 (13)
	FRL					-0.19** (0.07)	-2.68 (922)
Deceleration, π_{2i}	Intercept			-0.02*** (0.003)	-7.43 (13)	-0.03*** (0.003)	-7.71 (922)
	FRL					0.010~ (0.01)	1.93 (922)
Variance Components							
Level-1 (time)	Within-person	3.56		1.15		1.14	
Level-2 (students)	In initial status	23.52***		27.10***		22.287***	
	In rate of change	-		0.75***		0.75***	
	In deceleration	-		0.003***		0.002***	
Level-3 (schools)	In initial status	1.09***		1.05***		1.45***	
	In rate of change	-		0.02~		0.01**	
	In deceleration	-		0.00		-	

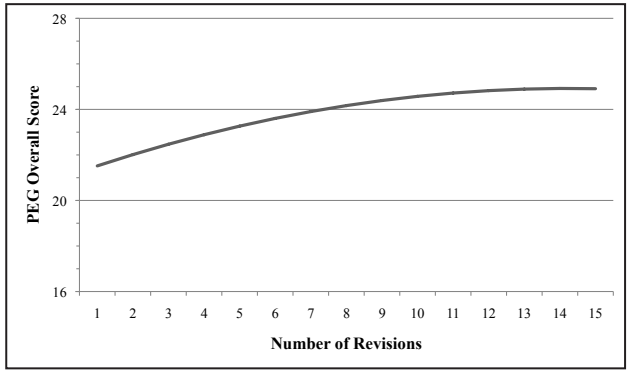
Goodness-of-fit Statistics							
	Deviance	26480.44	$df = 4$	22938.80	$df = 16$	22719.97	$df = 17$
	AIC	26488.44		22970.80		22753.97	
	BIC	26507.88		23048.59		22836.62	
	SBIC	26495.18		22997.77		22782.63	
^{2} Difference Test		-		3541.646^{***}, $df = 12$		218.83^{***}, $df = 1$	

Note. $^{\sim}p$ .10; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Research Question 2

We were next interested in determining whether certain students experienced a differential benefit from automated feedback. Variables of interest were Upper-Elementary, Female, Prior Writing Achievement, and FRL. Several successive models were evaluated using these variables as predictors of the intercept and each of the growth slopes. The final random coefficients model is presented in the last columns of Table 2. For parsimony, all non-significant parameters were excluded from the model. Middle-School was a non-significant predictor for each of the fixed effects. Female and Prior Writing Achievement predicted variation between students' first-draft performance but not rate of improvement (Figure 5); females and students with higher prior writing achievement received higher PEG Overall Scores on their first drafts. These predictors yielded a 17.75% reduction in between-student variance in the intercept. The only predictor of slope was FRL, which predicted both slope parameters. After controlling for Female and Prior Writing Achievement, students who received free or reduced lunch grew slightly slower than peers with low-SES (Figure 6). Adding FRL to the model yielded a 12.99% reduction in between-student variance in the linear growth slope (γ_{100}), but no reduction in variance in the quadratic growth slope (γ_{200}), likely because this variance was negligible to begin with ($\tau_{\beta 20} = 0.003$, from the unconditional growth model).

Figure 4. **Unconditional Quadratic Growth Model (Prompt 1). Full range of PEG Overall Score (6-36) restricted in Y axis to better illustrate effects.**



Research Question 3

We then examined whether the experience of using automated feedback to repeatedly revise an essay transferred to improved initial performance (i.e., unaided transfer) or accelerated growth on a subsequent prompt when again using automated feedback (i.e., aided transfer). Table 3 reports results of the unconditional means model, unconditional quadratic growth model, and final conditional model for students' second prompt. The variance components of the unconditional means model were used to calculate the intra-class correlations between different levels of the model. In this sample, the proportion of variance in the PEG Overall Score that was within students (level-1) was almost double that of the first sample: 24.21%. In

addition, there was slightly less between-student variance (67.95%), but slightly more between-school variance (7.85%). Again, these findings support the choice to evaluate a three-level model, while suggesting that the majority of variance lies within and between students.

Figure 5. **The Effect of Gender and Prior Writing Achievement on Growth in Writing Quality, for Students not Receiving Subsidized Lunch (Prompt 1). Full range of PEG Overall Score (6-36) restricted in Y axis to better illustrate effects.**

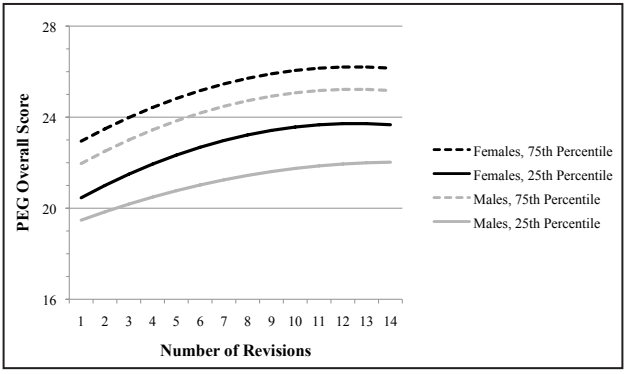
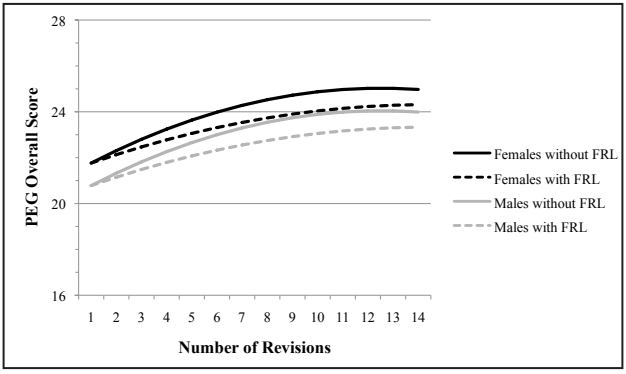


Figure 6. **The Effect of FRL on Males and Females of Average Prior Writing Achievement (Prompt 1). Full range of PEG Overall Score (6-36) restricted in Y axis to better illustrate effects.**



The quadratic model was again the best-fitting model to explain students' growth in writing quality. Compared to the unconditional means model, it explained 71.13% of the within-student variance in growth in writing quality. For their second essay, the average first draft score for this sample was 21.14, with an instantaneous rate of chance of 0.6 points and deceleration of 0.02 points. The saturation point for this model was reached with 12 revisions, at which point the PEG Overall Score grew from 21.14 to 24.70 points.

Table 3. ***Results of Models Estimating Initial Performance and Growth Across Revisions for Prompt 2***

		Unconditional Means Model		Unconditional Quadratic Growth Model		Final Conditional Model	
		Coefficient (S.E.)	*t (df)*	Coefficient (S.E.)	t (df)	Coefficient (S.E.)	*t (df)*
Fixed Effects							
Initial Status, π_{0i}	Intercept	22.07*** (0.52)	42.50 (11)	21.14*** (0.62)	34.19 (11)	20.60*** (0.64)	32.12 (11)
	Gender					1.77*** (0.33)	5.30 (711)
	Prior Writing Achievement					0.02_{***} (0.003)	5.50 (711)
	Free Reduced Lunch					-1.68*** (0.37)	4.48 (711)
Rate of Change, π_{1i}	Intercept			0.58*** (0.09)	6.28 (11)	0.64*** (0.05)	11.84 (11)
	Free Reduced Lunch					-	-
Deceleration, π_{2i}	Intercept			-0.02*** (0.004)	-6.378 (11)	-0.03*** (0.003)	-7.96 (711)
	Free Reduced Lunch					-	-
Variance Components							
Level-1 (time)	Within-person	6.18		1.96		1.96	
Level-2 (students)	In initial status	19.01***		25.00***		24.48***	
	In rate of change	-		0.64***		0.66***	
	In deceleration	-		0.002***		0.002***	
Level-3 (schools)	In initial status	2.20***		3.28***		3.22***	
	In rate of change	-		0.05**		0.01**	
	In deceleration	-		0.00		-	

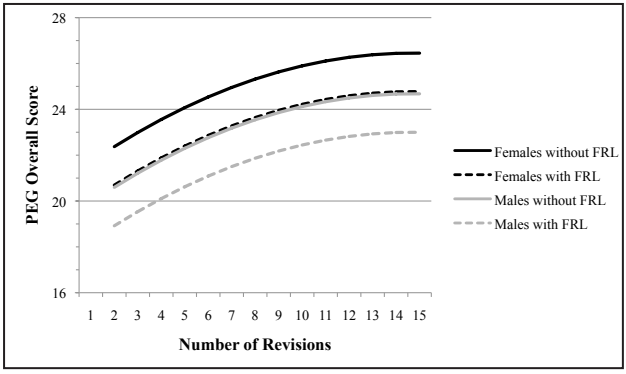
Goodness-of-fit Statistics							
	Deviance	20575.52	$df = 4$	17756.22	$df = 16$	17673.95	$df = 16$
	AIC	20583.52		17788.22		17705.95	
	BIC	20602.96		17861.90		17779.64	
	SBIC	20590.25		17811.10		17728.83	
^{2} Difference Test		-		2819.30^{***}, $df = 12$		n/a	

Note. ˜p .10; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

T-tests comparing the fixed-effects parameters of the unconditional quadratic growth model of students' first and second prompts suggest consistency in intercept (t = 0.52, pooled df = 22, p = 0.61, CI = [-1.13, 1.86]), in linear growth slope (t = -0.59, pooled df = 22, p = 0.56, CI = [-0.29, 0.16]), and in quadratic growth slope (t = -0.22, pooled df = 22, p = 0.83, CI = [-0.001, 0.009]). Thus, there is insufficient evidence to suggest that engaging in a single cycle of revision aided by automated feedback is associated with significant unaided transfer to improved first-draft performance or accelerated growth in writing quality when revising a subsequent prompt with automated feedback (i.e., aided transfer).

The final random coefficients model describing growth in writing quality for students' second prompt revealed that Middle-School was again a non-significant predictor of the fixed effects, and that Female and Prior Writing Achievement were significant predictors of the intercept. These predictors yielded a 10.08% reduction in between-student variance in the intercept. However, unlike the final model for prompt one, FRL was no longer a significant predictor of either growth slope, but was a significant predictor of the intercept (γ_{020} = -1.68, t = -4.48, df = 711, $p < 0.001$; see Figure 7).

Figure 7. **The Effect of Receiving Free or Reduced Lunch on Males and Females of Average Prior Writing Achievement (Prompt 2). Full range of PEG Overall Score (6-36) restricted in Y axis to better illustrate effects.**



Discussion

The current study examined whether the use of automated feedback provided by an AES system called Project Essay Grade was associated with growth in writing quality across successive revisions to a writing prompt. In addition, we examined whether student characteristics were associated with differential effects of automated feedback. Finally, we assessed whether repeatedly engaging in automated feedback to revise a single essay was associated with either improved first draft performance (unaided transfer) on a second prompt, or to more rapid growth when using automated feedback to revise that prompt (aided transfer).

Results of research question one indicate that automated feedback is associated with improvements in writing quality, as measured by the PEG Overall Score.

Although the per revision gain was about a 0.5 points, a very minor gain, this finding is significant considering that previous studies have had difficulty documenting effects of automated feedback on measures of overall writing quality (e.g., Kellogg et al., 2010; Wade-Stein & Kintsch, 2004).

While a minor gain, it is one that adds up across revisions. Indeed, one of the benefits of AES and automated feedback is the ability to continually engage in the cycle of feedback and practice critical for improving writing skills (Kellogg & Whiteford, 2009). This may partially explain the results observed in the present study. Compared to previous studies, students in our sample experienced many exposures to automated feedback, completing a minimum of two and upwards of 18 revisions to a single writing prompt. Given the complexity of writing ability, it is reasonable to assume that multiple and repeated exposures of automated feedback are necessary before observing significant improvements in measures of overall quality.

However, results also indicate automated feedback should not be used ad infinitum to revise an essay. The shape of students' growth was quadratic and concave, revealing that growth slowed to zero once 11-12 revisions had been completed. Thus, despite its positive effects, there is a limit to how much one can benefit from applying similar feedback to the same essay.

With research question two, we evaluated easily observable student characteristics that may be related to the ability to apply and benefit from automated feedback. Results indicated that grade-band was not significantly related to first draft performance or growth in writing quality. This finding is different from that of Shermis et al. (2008), who observed eighth graders outperforming students from grades 6, 7, and 10, on the e-rater quality score, text length, number of unique words, and error correction. Differences in study findings may simply be due to sample size and estimation techniques. The Shermis et al. (2008) study was over weighted with eighth graders (35% of the study) and grade-level was estimated as single grade-level differences. In contrast, our sample had very few students from grades 4-5 (8.5% of the study) and we examined grade-level differences in terms of differences between students in middle-school and upper-elementary students. Additional research using more equal sample weighting is needed to determine whether automated feedback is of equal benefit to students of different ages.

Prior writing achievement and Female were both significant predictors of students' first draft performance, though neither variable was related to growth in writing quality. This finding is important for teachers of struggling writers. Our results suggest that struggling writers benefitted from automated feedback as much as higher performing writers; however, automated feedback did not close the gap between the two groups of students.

In contrast, students who received free and reduced lunch grew slower than their peers. This finding, however, should not be over-interpreted as FRL likely acts as a proxy for many underlying variables, such as reading ability, language comprehension, attention, motivation, and self-regulation skills. Instead, this finding is taken to confirm our hypothesis that different students may derive differential benefit from automated feedback. Future research should continue to explore student characteristics that may mediate the effect of automated feedback on writing quality. Such

research should specifically examine differences between struggling writers, students with disabilities, and typically-achieving peers in order to determine what prerequisite skills and knowledge are needed to derive maximum benefits from automated feedback.

While repeated exposure to automated feedback was associated with gains in writing quality across revisions to an essay, there was no evidence of transfer effects when comparing HLM models for prompt one and prompt two. Kellogg et al. (2010) and Wade-Stein and Kintsch (2004) both reported unaided transfer effects for reducing errors in grammar, usage, mechanics, and style, but not for overall quality. While it is certainly important for students to improve their error detection skills and reduce spelling and grammar errors, these gains did not result in concomitant increases in measures of overall quality. Ideally, an automated feedback system would support growth in multiple areas of writing ability.

One limitation facing automated feedback from supporting such growth is that each AES system analyzes text differently. For example, PEG uses a syntactic text parser to evaluate word- and sentence-level indicators of text quality, while Summary Street uses latent semantic analysis to examine content-match between student texts and source documents. Consequently, a single AES system may be limited to supporting improvements in those aspects of writing for which it evaluates. For struggling writers and students with disabilities it is likely necessary to pair automated feedback with additional high quality writing instruction to address the full range of their writing deficits.

A second limitation is that AES systems typically provide feedback about writing performance—i.e., task-level performance—rather than about writing process, writing strategies, or writing metacognitions. Hattie and Timperley (2007) caution that receiving only task-level feedback may result in students ignoring important procedural or strategic aspects that promote achievement. Given the robust effects of strategy instruction on writing quality (Graham et al., 2012; Graham & Perin, 2007), perhaps multiple types of feedback addressing multiple aspects of writing ability will promote stronger and more transferable gains in writing ability.

Finally, because AES systems and computational linguistics are continually evolving, conclusions regarding the efficacy of PEG automated feedback must be contextualized within the time period from which the data was gathered (AY 2012). Indeed, in July 2013, Measurement Incorporated updated PEG's automated feedback by adding new feedback statements and more closely aligning feedback to students' grade level and the writing trait.

Limitations

The study exhibited the following limitations. First, the nature of the study design required collapsing across prompt variables such as topic, genre, and prompt creator. Each of these variables has been shown to explain significant variance in student achievement (Chen et al., 2007; Matsumura, Patthey-Chavez, Valdez, & Garnier, 2002; Olinghouse, Santangelo, & Wilson, 2012). Thus, it is likely that some of the unexplained variance in first-draft performance and growth in writing quality may be attributed to unmeasured aspects of the prompts or tasks. Second, the study did not control for the number of revisions a student completed or the time frame within

which they were completed. Unmeasured qualitative differences among the students within the sample may explain this variability. While we attempted to identify characteristics that might explain some of these differences, the limitations of the source database meant that we were constrained to measuring easily observable characteristics such as grade-band, gender, prior writing achievement, and free/reduced lunch status. Other variables such as writing-related knowledge, attention, motivation, and self-regulation may simultaneously explain some of the differences in the range of revisions students completed, as well as the unexplained between-student variance in first draft performance and growth in writing quality. Finally, our analysis of transfer effects did not control for the possibility that students received additional instruction between the time they completed their first essay and started their second essay. This issue would be more problematic had the study found evidence of transfer effects, but it is important to note here.

Future Research

Though this study noted gains in writing quality, they were minor. It is unclear whether this was due to aspects relating to the feedback students received or the ability of the students to understand and apply the feedback. Thus, future research should more closely assess the content of the feedback students receive, as well as students' knowledge of revising behaviors. It is possible that provided feedback only targets certain aspects of writing and not others, or that the feedback is not targeted within students' zone of proximal development—perhaps the feedback was too simple for some students and too difficult for others. Also, research should consider whether providing multiple types of feedback (e.g., task, strategic, and procedural) is more effective than providing a single type.

Second, research also should consider the possibility that students need additional instruction on how best to utilize automated feedback to improve their writing. Indeed, previous research has found that, particularly among struggling writers, knowledge of procedures for revision is lacking (MacArthur, Graham, & Schwartz, 1991; Graham, MacArthur, & Schwartz, 1995). These students typically make surface-level changes more akin to the process of editing than the substantive changes implied by the process revising. Research also should continue to explore student characteristics that may mediate the effect of automated instructional feedback. Potential variables include attention, motivation, writing-related knowledge, and self-regulation skills.

Finally, future studies should directly examine the effect of automated feedback for struggling writers and students with disabilities. Given that AES and automated feedback can accelerate the feedback-practice cycle necessary for growth in writing ability (Kellogg et al., 2010), it would be beneficial to understand whether automated feedback could be used as part of an intervention for those most at-risk.

Conclusion

There is still a lot to be learned about automated feedback and its effect on writing quality. While automated feedback may not currently address all the complexities of writing ability, or provide students with multiple types of feedback (e.g., task-focused, procedural, strategic, metacognitive), the field of AES continues to

evolve, and the limits of computational linguistics have yet to be reached. The positive effects of automated feedback noted in this study suggest that AES may aid teachers in the difficult and time consuming work of evaluating, diagnosing, and responding to student text. In turn, this may afford teachers the ability to devote time and energy to providing feedback regarding other important aspects of writing ability.

References

Automated Student Assessment Prize (2012, April 12). *Automated essay scoring systems demonstrated to be as effective as human graders in big trial.* Retrieved from http://www.measurementinc.com/news/mis-peg-software-leads-automated-essay-scoring-competition.

Automated Student Assessment Prize (2013, October 4). *Winners of competition announced.* Retrieved from http://www.measurementinc.com/news/mis-automated-essay-scoring-system-takes-first-prize-national-competition.

Baker, E. L., Abedi, J., Linn, R. L., & Niemi, D. (1995). Dimensionality and generalizability of domain independent performance assessments. *Journal of Educational Research, 89*, 197-205.

Beach, R., & Friedrich, T. (2006). Response to writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 222-234). New York, NY: Guilford.

Berninger, V. W., & Swanson, H. L. (1994). Modifying Hayes and Flower's model of skilled writing to explain beginning and developing writing. In E. Butterfield (Ed.), *Children's writing: Toward a process theory of the development of skilled writing* (pp. 57-81). Greenwich, CT: JAI Press.

Berninger, V. W., Abbott, R. D., Whitaker, D., Sylvester, L., & Nolen, S. B. (1995). Integrating low- and high-level skills in instructional protocols for writing disabilities. *Learning Disability Quarterly, 18*, 293-309.

Caccamise, D., Franzke, M., Eckhoff, A., Kintsch, E., & Kintsch, W. (2007). Guided practice in technology-based summary writing. In D. S. McNamara, *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 375-396). Mahwah, NJ: Erlbaum.

Chen, E., Niemie, D., Wang, J., Wang, H., & Mirocha, J. (2007). *Examining the generalizability of direct writing assessment tasks.* CSE Technical Report 718. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Elliott, N. (2005). *On a scale: A social history of writing assessment in America.* New York, NY: Peter Lang Publishing.

Flower, L. S., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L.W. Gregg, & E.R. Sternberg (Eds.), *Cognitive processes in writing* (pp. 3-29). Hillsdale, NJ: Lawrence Erlbaum Associates.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., and Dooley, S. (2005). Summary Street®: Computer support for comprehension and writing. *Journal of Educational Computing Research, 33*, 53–80

Graham, S., & Harris, K. R. (2000). The role of self-regulation and transcription skills in writing and writing development. *Educational Psychologist, 35*, 3-12.

Graham, S., Harris, K., & Hebert, M. A. (2011). *Informing writing: The benefits of formative assessment. A Carnegie Corporation Time to Act report.* Washington, DC: Alliance for Excellent Education.

Graham, S., Hebert, M. A., & Harris, K. *Formative Assessment and Writing: A Meta-analysis with Implications for Common Core.* Manuscript submitted for publication.

Graham, S., MacArthur, C., & Schwartz, S. (1995). Goal setting and procedural facilitation on the revising behavior and writing performance of students with writing and learning problems. *Journal of Educational Psychology, 87*, 230-240.

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). Meta-analysis of writing instruction for students in elementary grades. *Journal of Educational Psychology, 104*, 879-896.

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing adolescents in middle and high schools – A report to Carnegie Corporation of New York*. Washington, DC: Alliance of Excellent Education.

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*, 369-388.

Hayes, J. R., & Berninger, V. W. (in preparation). Cognitive processes in writing: A framework. In, B. Arfe, J. Dockrell, & V. Berninger (Eds.), *Writing development and instruction in children with hearing, speech, and language disorders.* New York, NY: Oxford University Press.

Keith, T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147-168). Mahwah, NJ: Erlbaum.

Kellog, R. T., & Whiteford, A. P. (2009). Training advanced writing skills: The case for deliberate practice. *Educational Psychologist, 44*, 250-266.

Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students to write? *Journal of Educational Computing Research, 42*, 173-196.

Loban, W. (1976). *Language development: Kindergarten through grade twelve.* Research report no. 18. National Council of Teachers of English, Champaign, IL.

MacArthur, C. A., Graham, S., & Schwartz, S. (1991). Knowledge of revision and revising behavior among students with learning disabilities. *Learning Disability Quarterly, 14*, 61-73.

Matsumara, L. C., Patthey-Chavez, G. G., Valdez, R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *The Elementary School Journal, 103*, 3-25.

McCoach, D. B. (2010). Hierarchical linear Modeling. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 123-140). New York, NY: Routledge.

Olinghouse, N. G., & Graham, S. (2009). The relationship between the discourse knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology, 10*, 37-50.

Olinghouse, N. G., Santangelo, T., & Wilson, J. (2012). Examining the validity of single-occasion, single-genre, holistically-scored writing assessments. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, and H. Van den Bergh (Eds.), *Measuring writing. Recent insights into theory, methodology and practices* (pp.55-82). Leiden, The Netherlands: Brill.

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing, 26*, 45-65.

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*, 127-142.

Page, E. B., Poggio, J. P., Keith, T. Z. (1997, March). *Computer analysis of student essays: Finding trait differences in student profile.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. T. (1988). *HLM: Hierarchical linear modeling.* Chicago: Scientific Software International, Inc.

Scott, C. M. (2009). Language-based assessment of written expression. In G. A. Troia (Ed.), *Instruction and assessment for struggling writers: Evidence-based practices* (pp. 358-385). New York, NY: Guilford.

Shermis, M. D., & Hammer, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. Downloaded from National Council on Measurement in Education (NCME).

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62*, 5-18.

Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington, S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment and Evaluation in Higher Education, 26*, 247-259.

Shermis, M. D., Wilson Garvan, C., & 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.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford.

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., Shankland, R. K., & Wolbers, K. A. (2012). Motivation research in writing: Theoretical and empirical considerations. *Reading and Writing Quarterly, 28*, 5-28.

U. S. Department of Education, Office of the Deputy Secretary, Implementation and Support Unit. (2012a). *Race to the Top assessment: Partnership for Assessment of Readiness for College and Careers year one report.* Washington, DC: Author. Downloaded from: http://www2.ed.gov/programs/racetothetop-assessment/performance.htmlU. S. Department of Education, Office of the Deputy Secretary, Implementation and Support Unit. (2012b). *Race to the Top assessment: Smarter Balanced Assessment Consortium year one report.* Washington, DC: Author. Downloaded from: http://www2.ed.gov/programs/racetothetop-assessment/performance.htmlVygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*, 333-362.

Author's Note

Joshua Wilson, School of Education, University of Delaware; Natalie G. Olinghouse, Department of Educational Psychology, University of Connecticut; Gilbert N. Andrada, Psychometrics and Applied Research, Bureau of Student Assessment, Connecticut State Department of Education.

The authors would like to thank Trish Martin, Shayne Miel, Fran Brown, Tiwanna Bazemore, and Andrew O'Neill of Measurement Incorporated for their assistance in preparing the datasets used for analysis.

Correspondence concerning this article should be addressed to Joshua Wilson, University of Delaware School of Education, Willard Hall, Newark, DE 19716. Email: joshwils@udel.edu.

Appendix A: PEG Six Trait Measures

Quantitative feedback is provided to students via an automated essay scoring (AES) engine called Project Essay Grade[2] (PEG; Page, 1966, 1994) which analyzes text for syntactic features associated with writing quality and uses a regression-based algorithm to generate six analytic trait scores used to summarize the student's writing performance on a particular draft. These trait scores are similar to those measured on commonly-used analytic rubrics and each is scored on a 1-6 scale (1 = low, 6 = high).

PEG Writing Scores

Overall Development: Measures the degree to which the writer demonstrates awareness of purpose, task, and audience.

Organization: Measures the degree to which the writing shows a logical organization and focus.

Support: Measures the writer's inclusion of elaboration, reasons, and examples that develop the writing.

Sentence Structure: Measures the correctness and variety of sentences in the text.

Word Choice: Measures the extent to which the text includes vivid and varied vocabulary.

Mechanics: Measures the correct use of capitalization, punctuation, and spelling throughout the text.

Overall Score: Sum of the individual six trait scores (range: 6-36).

2 PEG is proprietary to Measurement Incorporated (MI), an assessment company based out of Durham, North Carolina. MI uses PEG as part of its benchmark writing assessments and other assessment products in which student writing is evaluated.

Appendix B: Examples of PEG Qualitative Feedback

- Be sure to use "a" before a consonant sound and "an" before a vowel sound.
- Break up sentences to make them easier to understand.
- Check carefully for spelling errors.
- Don't overuse "and" and "then" to connect sentences.
- Don't use "of" as a verb. For example, "could of" should be "could have".
- Don't use two helping verbs together: Write: "I might go," instead of "I might could go."
- Keep it active; where possible avoid the passive voice: Instead of: "The law was passed by congress" write: "Congress passed the law."
- Know your comparatives and superlatives: Write: "better" not "more better", write: "funniest", not "most funniest."
- Make sure every sentence has a verb.
- Try to use more advanced words.
- Vocabulary is too simple for grade level.