Read Writ
DOI 10.1007/s11145-016-9695-z

Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities

Joshua Wilson[1]

Abstract The present study examined growth in writing quality associated with feedback provided by an automated essay evaluation system called PEG Writing. Equal numbers of students with disabilities (SWD) and typically-developing students (TD) matched on prior writing achievement were sampled ($n = 1196$ total). Data from a subsample of students ($n = 655$) was used to investigate evidence of transfer to improved first-draft performance on a follow up writing prompt. Three-level hierarchical linear modeling was used. Findings indicated that SWD produced first drafts of lesser quality than TD students, but grew at a faster rate and were able to close the gap in writing quality after five revisions. However, these effects were moderated by school quality and the availability of internet-connected devices in schools. There was no evidence of transfer for either group of students. Results document a positive association between the use of PEG Writing and growth in writing quality for SWD, and underscore the importance of having sufficient technology resources for maximizing this growth.

Keywords Automated essay evaluation · Project Essay Grade (PEG) Writing · Students with disabilities · Writing quality

Introduction

Though an essential twenty-first century skill, writing outcomes for the nation's youth are concerning. Results of recent National Assessments of Educational Progress [National Center for Educational Statistics (NCES), 2012; Persky, Daane, & Jin, 2002; Salahu-Din, Persky, & Miller, 2008] indicate that only one third of

✉ Joshua Wilson
joshwils@udel.edu

[1] University of Delaware, School of Education, 213E Willard Hall Education Building, Newark, DE 19716, USA

Published online: 03 October 2016

students in Grades 4, 8, and 12 meet or exceed grade-level proficiency in writing. For students with disabilities (SWD), this percentage shrinks to an alarming 5 %.

One recommendation for improving these outcomes is to increase the use of technology during writing instruction (Graham & Harris, 2013; National Commission on Writing, 2003). This recommendation is echoed in the Common Core State Standards (2010) which expects students to use technology during independent and collaborative writing. Indeed, research has shown that the use of computers, educational technology, and assistive technology can improve students' writing performance (Graham & Perin, 2007; Graham, McKeown, Kiuhara, & Harris, 2012b), particularly for struggling writers and SWD (MacArthur, 2009; Morphy & Graham, 2012).

One increasingly prevalent form of educational technology is automated essay evaluation (AEE) software. AEE is web-based software which provides students with immediate feedback on their writing in the form of essay ratings and individualized suggestions for improvement. The current study examines the effects associated with the use of an AEE program called PEG Writing on writing achievement for SWD as compared to typically-developing (TD) students.

Writing and students with disabilities

Writing is complex. It requires students to regulate and coordinate multiple component skills across sub-lexical (handwriting), lexical (spelling, word choice), syntactic (grammar, sentence structure, punctuation), and discourse levels of language (organization, content) (Abbott, Berninger, & Fayol, 2010). In addition, students must integrate topic and discourse knowledge from long-term memory (Bereiter & Scardamalia, 1987; Hayes, 2012) while executing and regulating the cognitive processes of proposing, evaluating, translating, transcribing, and reviewing (Hayes, 2012). Such coordination and regulation places great strain on limited cognitive resources, including attention and working-memory (McCutchen, 2011). To free up cognitive resources, writers must acquire automaticity in lower-level transcription skills (handwriting/keyboarding and spelling) and utilize strategies for planning, drafting, and revising (Harris & Graham, 2013).

The complexity of writing makes it a difficult skill for SWD to acquire. SWD demonstrate skill deficits across each level of language; lack well-developed strategies for planning, drafting, and revising (i.e., they lack procedural knowledge); and exhibit deficits in working memory, self-regulation, motivation, and persistence (Dockrell, Lindsay, & Connelly, 2009; Graham & Harris, 2013; Troia, 2006). Consequently, instructional approaches aimed at increasing automaticity of transcription skills, genre and procedural knowledge, and metacognitive and self-regulatory control are effective for improving the writing performance of SWD (Berninger, Nagy, Tanimoto, Thompson, & Abbott, 2015; Englert et al., 2009; Gillespie & Graham, 2014; Graham & Harris, 2003; Graham & Perin, 2007).

Linking all three of these major instructional components—automaticity of writing skills, writing-related knowledge (including strategies), and self-regulatory control—is the need for students to engage in substantial amounts of practice and to receive feedback on their performance (Biber, Nekrasova, Horn, 2011; Graham &

Perin, 2007; Graham et al., 2012a, b; Kellogg & Whiteford, 2009; McCutchen, 2011). Unfortunately, results of recent surveys suggest that this need is unmet for the majority of U.S. students. Elementary and middle-school teachers report instructing writing for 15 min or less per day, and cite the time cost of evaluating writing as a substantial barrier to increasing writing practice and opportunities for feedback (Applebee & Langer, 2009; Gilbert & Graham, 2010; Graham, Capizzi, Harris, Hebert, & Morphy, 2013).

Given the complexity of writing and the need to increase writing practice and feedback to maximize instructional benefits, educators in the U.S. are increasingly adopting AEE software to support the teaching and learning of writing.

Automated essay evaluation (AEE) software

AEE is the formative application of automated essay scoring. Automated essay scoring is the application of natural language processing, latent semantic analysis, and machine learning for the purpose of generating essay quality ratings that are highly predictive of those assigned by expert raters (Shermis, 2014; Shermis & Burstein, 2003; Shermis & Hammer, 2013). When automated essay scoring is paired with automated feedback—computer-generated suggestions for improving essay quality (Shermis, 2014; Shermis & Burstein, 2003; Shermis & Hammer, 2013)—the application is referred to as automated essay evaluation or AEE (Shermis & Burstein, 2013). An obvious benefit of AEE is efficiency. Essay ratings and automated feedback are returned to students immediately. This vastly accelerates the practice-feedback loop essential for writing proficiency (Kellogg, Whiteford, & Quinlan, 2010).

AEE programs have been developed by most major education technology and research companies, including Education Testing Service (*Criterion*), Pearson (*Write-to-Learn*), Measurement Incorporated (*PEG Writing*), Pacific Metrics (*Writing Power*), Vantage Learning (*My Access*), and Turnitin (*Revision Assistant*). AEE is also widely used. PEG Writing—the program focused on in the current study—is being used to support writing instruction and formative assessment for approximately 650,000 students in the United States and several other countries (C. Dobbs, personal communication, August 8, 2016).

How AEE may support the writing performance of students with disabilities

AEE may support the writing performance of SWD because it combines word processing and feedback. Both of these components have separately been shown to be effective for improving writing outcomes (Graham & Perin, 2007; Graham, Hebert, & Harris, 2015; Graham et al., 2012a, b), and when combined in an AEE program, may support the writing performance of SWD. Indeed, word processing software supports students' transcription skills and facilitates revision by enabling easy insertion, deletion, and rearranging of text (MacArthur, 2009). These capabilities are beneficial for weaker writers. In their meta-analysis of the effects of word processing software on writing outcomes for weaker writers, Morphy and Graham (2012) reported positive, moderate to large effects sizes (*d* range .48–1.42)

for the effect of word processing on writing quality, text length, development and organization of text, mechanical correctness, and motivation. The effect size for measures of writing quality for word processing programs that provided feedback or that encouraged process writing was quite large ($d = 1.46$), though this was limited to a review of three studies, only one of which was with a modern AEE system and that study did not sample SWD (Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005).

AEE may also be beneficial for SWD because it accelerates the practice-feedback loop by providing *immediate*, corrective feedback (Kellogg et al., 2010). This in turn appears to increase the amount of writing and revising students do (Grimes & Warschauer, 2010; Moore & MacArthur, 2016; Wilson & Czik, 2016). Consequently, AEE may be particularly beneficial for SWD because increasing opportunities for practice and feedback are strongly recommended pedagogical approaches for this population of students (Gersten & Baker, 2001; Vaughn, Gersten, & Chard, 2000).

Furthermore, AEE may also help support deficits in procedural knowledge and metacognition commonly exhibited by SWD (see Lin, Monroe, & Troia, 2007). SWD typically lack knowledge of what and how to revise, and instead make surface level edits that do little to improve their overall writing quality (MacArthur, Graham, & Harris, 2009; MacArthur, Graham, & Schwartz, 1991). Thus, AEE may benefit SWD because AEE externalizes part of the reviewing and revising process that proficient writers generally internalize (see Hayes, 2004). Automated feedback is designed to detect and diagnose multiple areas of improvement such as organization, idea development, sentence structure, word choice, and conventions.

Finally, AEE may support writing motivation and self-efficacy, key components of theories of writing ability (Hayes, 2012; Pajares, 2003) and areas of deficit commonly exhibited by SWD (Troia, 2006). Use of the word processor itself appears to increase student motivation: Morphy and Graham (2012) reported an average weighted effect size of $d = 1.42$ for use of word processing programs on measures of motivation. AEE goes a step further than basic word processing by providing instantaneous feedback. As indicated by self-reports of teachers and students, AEE appears to increase students' writing motivating (Grimes & Warschauer, 2010; Wilson & Czik, 2016). Submitting writing to an AEE system may also help students avoid social stigma that may result from sharing weak writing with teachers or classmates (Nobles & Paganucci, 2015).

Thus, for these reasons—supporting students' transcription skills, increasing the amount of practice and feedback students receive, supporting deficits in procedural knowledge, and supporting affective dimensions of writing skill—AEE may help support improvements in the writing quality of SWD.

Effects of AEE on writing quality

While no previous studies have evaluated the effects of AEE on writing quality for SWD, a recent review of research (Stevenson & Phakiti, 2014) attempted to summarize the small but growing body of research on the effects of AEE on writing quality in typically-developing students (TD). The authors concluded that automated feedback has a modest effect on measures of writing quality, but does

not appear to support transfer to improved independent performance—by 'transfer' it is meant the ability to acquire skills in one context and effectively apply them in another (National Research Council, 2000). That is, the experience of receiving, interpreting, and acting upon automated feedback supports immediate improvements to the essay being composed, but does not appear to develop students' knowledge and skills to the point where they produce improved first drafts when responding to future writing prompts.

This conclusion is consistent with prior research by Wilson and colleagues (Wilson & Andrada, 2016; Wilson, Olinghouse, & Andrada, 2014) on the effects of of automated feedback for scaffolding growth in students' writing quality across essay drafts and transfer to improved independent performance. TD in Grades 4–8—approximately 1000 students in each study—improved their overall quality score after multiple revisions, but when composing a subsequent essay, the quality of their first-drafts was not statistically significantly different than those of their prior essay. This finding appears to hold for different subgroups, too. There were no differences in growth rates or transfer as a function of gender, grade level, or prior reading and writing achievement levels. The effect of socio-economic status, measured as free or reduced lunch (FRL) status, was inconclusive. FRL negatively influenced growth in one study (Wilson, Olinghouse, & Andrada, 2014), but had no effect on growth in a subsequent study (Wilson & Andrada, 2016). No previous research, however, has examined whether these findings hold for SWD.

In addition, questions remain about the contextual factors that may influence the effects of AEE on writing quality for different subgroups of students, such as SWD. The benefits of AEE are manifested when students repeatedly engage in a process of drafting, reviewing feedback, and revising (Wilson & Czik, 2016). Yet, AEE implementation and usage is quite variable. In some cases, students complete multiple rounds of revision (Foltz, Lochbaum, & Rosenstein, 2011; Wilson & Andrada, 2016; Wilson et al., 2014) while in other cases, students complete only a single draft (Attali, 2004; Warschauer & Grimes, 2008).

This variability may be due to how schools implement AEE and how teachers incorporate it into their curriculum. For instance, in their study of four school districts, Warschauer and Grimes (2008) reported that a major impediment to AEE usage was teachers' inability to create and assign their own prompts within the AEE system. Teacher-created prompts tend to be more closely tied to the curriculum, which meant that teachers could not use AEE to support content-area learning and assessment. Today, several AEE systems, including PEG Writing, allow teachers to create their own prompts, though it remains unclear whether automated feedback is equally beneficial (i.e., students improve at equal rates) when receiving feedback on teacher-created prompts for which the system was not trained to score (see Wilson & Andrada, 2016).

Variability in AEE implementation may also be due to the availability of computers within a school: as computer availability rises, so too might AEE usage and its concomitant effects on students' writing. It is reasonable to assume that students in schools with one-to-one laptops will be able to use AEE more frequently, and thus derive greater benefit than students in schools that only have 2–3 desktop computers per classroom or require teachers to schedule time to visit a computer lab. Warschauer and Grimes (2008) explored the association between

technology availability and AEE usage, but their results are inconclusive because, regardless of technology availability, their study reported minimal AEE usage due to the inability of teachers to create their own prompts. Thus, to more carefully ascertain the effects of contextual factors on the effects of AEE on writing quality, the current study separately examined the effect of prompt source (i.e., teacher- or system-created prompts) and the availability of computers within a school. By doing so, it was possible to more fully parse out the influence of contextual factors on the comparative effects of AEE for SWD and TD.

Study purpose

This study was the first of its kind to explore the effects of AEE with SWD. Specifically, the current study examined the influence of disability status and other student- and school-level variables on growth in students' writing quality across successive revisions to an essay. In addition, the question of transfer was also explored by comparing students' first draft performance on a subsequent writing prompt to that of a previous prompt. The following research questions guided the study: (1) What is the shape and rate of growth in students' writing quality in response to automated feedback from PEG Writing across successive revisions to a writing prompt? (2) Do students with and without disabilities grow at equal rates? (3) Do other student-level variables (grade, gender, socio-economic status, English Language Learner status, race, prompt source, and prompt genre) and school-level variables (school quality and computer availability) moderate the associated effects of AEE on initial performance and growth? (4) Do students with and without disabilities demonstrate improved performance on a follow-up writing prompt (i.e., is there evidence of transfer)?

Methods

This article reports results of hierarchical growth models of a sample of students in one state who, in the spring of 2014, used an AEE system called PEG Writing. During this time, the state Department of Education provided PEG Writing to districts, schools, and teachers as a resource to support classroom formative writing assessment. In accord with the Family Educational Rights and Privacy Act (FERPA), both individual participant data, as well as the identity of the state are de-identified for reporting here. Data used for the analyses were obtained from a state database which compiled students' PEG Writing data together with student demographic and state-test achievement data.

Participants

Primary sample

The primary study consisted of 1196 students in grades 4–8 from 26 districts and 46 schools. The sample was formed in the following manner. First, all SWD

who utilized PEG Writing and who had complete demographic and achievement data were selected ($n = 598$). In the current study, SWD were defined as those students identified as having an Individualized Education Plan, and excluded those who had a 504 Plan. Information regarding SWD's specific disability categories was unavailable for reporting; however, the sample was indicative of the broader SWD population of the state during the 2013–2014 school year (G. Andrada, personal communication, July 29, 2016) which included, in order of prevalence, Specific Learning Disability (33.7 %), Other Health Impairment (20.3 %), Speech Language Impairment (15.5 %), Autism (11 %), Emotional Disturbance (7.9 %), Intellectual Disability (3.5 %), and all other disabilities (8 %).

A comparison sample of TD students was selected using stratified random sampling. The sampling variable was performance on the prior year's state writing achievement test used for accountability purposes (described below). Equal numbers of TD students were selected to match the number of SWD at each of the five achievement bands for which writing performance was reported: Band 1 = Below Basic ($n = 130$), Band 2 = Basic ($n = 336$), Band 3 = Proficient ($n = 388$), Band 4 = Goal ($n = 260$), and Band 5 = Advanced. ($n = 82$). It is important to note that these achievement bands were not used by the state when making disability determinations. Thus, equal numbers of SWD and TD students were sampled ($n = 598$), with the full sample totaling 1196 students.

Demographic data for the primary sample is reported in Table 1. Chi square tests indicated that the groups were equivalent with regard to the proportions of males and English Language Learners (ELLs). However, proportionately more SWD were educated in schools that the state classified as "excelling", while more TD students were educated in schools classified as "transitioning" (see below regarding School Classification). Additionally, proportionately more eighth graders were in the TD group. The TD group also had more Hispanic and African-American students, and students who received free or reduced lunch (FRL).

Transfer sample

To explore transfer effects, all students from the primary sample who completed a follow-up writing prompt were retained for analysis ($n = 655$). Approximately equal numbers of SWD and TD students were retained ($n = 330$ and 325, respectively). Demographics of the transfer sample are also reported in Table 1. Again, the groups were equivalent with regard to the proportion of males and ELLs. As in the primary sample, a greater proportion of SWD were educated in schools classified as Excelling. The SWD group included a greater amount of sixth graders. As before, the TD group included a greater amount of Hispanic and African-American students, and students receiving FRL.

PEG Writing

PEG Writing is web-based formative assessment and feedback system developed by Measurement Incorporated and designed for students in Grades 3–12. It is based on

Table 1 Demographics of primary and transfer samples

	Primary sample		Chi square	Transfer sample		Chi square
	SWD ($n = 598$)	TD ($n = 598$)		SWD ($n = 330$)	TD ($n = 325$)	
Districts	24	24		17	16	
Schools	45	41		32	30	
School classification			40.11, $df = 2$, $p < .001$			13.62, $df = 2$, $p = .001$
1 = Excelling	103	53		65	43	
2 = Progressing	263	211		148	122	
3 = Transitioning	232	334		117	160	
Grade			13.17, $df = 4$, $p = .010$			9.44, $df = 4$, $p = .051$
4	63	78		33	58	
5	95	74		53	47	
6	167	141		122	100	
7	157	147		84	81	
8	116	158		38	39	
Race			33.93, $df = 4$, $p < .001$			13.03, $df = 4$, $p = .011$
Hispanic	92	145		51	71	
African American	59	84		22	37	
Asian	11	25		6	9	
White	424	337		245	206	
Multiracial	12	7		6	2	
Males	372	358	0.69, $df = 1$, $p = .407$	208	190	1.43, $df = 1$, $p = .231$
ELL	47	60	1.74, $df = 1$, $p = .188$	24	24	0.00, $df = 1$, $p = .956$
Free or reduced lunch	196	257	13.22, $df = 1$, $p < .001$	98	124	5.23, $df = 1$, $p = .022$

SWD students with disabilities, *TD* typically-developing students, *ELL* English-language learners

the pioneering work of Ellis Page and colleagues (Page, 1966, 2003), who developed the first automated essay scoring system called Project Essay Grade, hence the acronym PEG. Students access the system by visiting www.pegwriting.com and inputting their username and password. Students select from a number of prompts in a variety of genres aligned to those targeted in the Common Core State Standards: narrative, informative/explanatory, and argumentative/persuasive. Teachers have the option of assigning system-created prompts or creating their own prompts to align with their particular curriculum. Both system-created and teacher-created prompts may have stimulus materials, such as text files, videos, images, or links to URLs.

Once students select their prompt, they are given the option to pre-write with the aid of built-in electronic graphic organizers. Students then move to a "practice" screen where they create their initial draft, either from scratch or from the support of their completed graphic organizer. Students have up to 60 min to complete their initial draft. Once completed, their draft is submitted to PEG for automated scoring and feedback. Students receive immediate scores for six traits of writing, each on a 1–5 scale: Ideas and Development, Organization, Style, Sentence Structure, Word Choice, and Conventions. An overall score, formed as the sum of the six traits (range 6–30), is also assigned. Students also receive automated feedback in the form of spelling and grammar feedback, trait-specific suggestions for improvement, and customized links to PEG Writing's multimedia interactive skill-based tutorial lessons. Students can revise and resubmit their essays as many times as they like, receiving new feedback each time.

PEG Writing also includes a number of instructional management tools for teachers, including the ability to create and share prompts, create groupings for peer feedback (accessed by students within the PEG Writing site), view students' writing histories and portfolios, and view and comment on students' drafts. Teachers can also provide feedback to students in the form of comments embedded in the body of the draft or as summary comments. Students may also leave comments for their teachers using a similar function.

Prompts and revisions

Data from the primary sample ($n = 1196$) comes from the first prompt to which they responded in PEG Writing, and data from the transfer sample ($n = 655$) comes from the subsequent prompt to which students responded. The primary sample responded to one of 167 prompts; the transfer sample responded to one of 153 prompts. All prompts were on different topics. Thus, results should be interpreted as the associated effect of automated feedback from PEG Writing when averaging across prompt topic. On each occasion, students responded to argumentative prompts and informative/explanatory prompts with greater frequency (approximately 40 %, respectively) than narrative prompts (approximately 20 %). Given these differences, the effect of prompt genre was examined in prediction models.

Measures

Prior writing achievement

Prior writing achievement was measured by performance on the spring 2013 state writing assessment used for accountability purposes. The state assessment was administered to students annually in Grades 3–8, and included two subtests: the Direct Assessment of Writing (DAW) and the Editing and Revising test (ER). The DAW consisted of one 45-min writing prompt, with narrative texts assigned in primary grades, and informative and argumentative tasks assigned in upper-elementary and middle grades. Texts were scored holistically by pairs of trained raters who each assigned a score (range 1–6) based on consideration of the following features: elaboration and detail, logical organization, and proficiency of language usage. Students' final scores were the sum of the two raters' scores (range 2–12). The ER test consisted of 32–40 multiple-choice items testing students' ability to identify errors in written language. Raw scores on both subtests were combined to form a weighted composite score which was converted into a scale score (range 100–400). Grade-specific cutpoints were established to delineate five achievement bands based on the scale score range: Below Basic, Basic, Proficient, Goal, and Advanced. SWD and TD students were matched based on their achievement band using these grade-specific cutpoints.

PEG overall score

The primary dependent variable was the PEG overall score, which was formed as the sum of the six trait scores assigned to students' drafts by PEG: Ideas and Development, Organization, Style, Sentence Structure, Word Choice, and Conventions. PEG uses a combination of techniques in its scoring algorithm including natural language processing, syntactic analysis, and semantic analysis to measure thousands of variables that are combined in a regression-based algorithm that accurately predicts holistic and analytic scores assigned by trained raters (see Bunch, Vaughn, & Miel, 2016 for in depth description of PEG's scoring algorithms). Several empirical studies have established the reliability and criterion validity of PEG's essay ratings (Keith, 2003; Shermis, 2014; Shermis, Koch, Page, Keith, & Harrington, 2002). For example, PEG yielded quadratic weighted kappas ranging from .70 to .84 across eight test sets of human-scored essays (Shermis, 2014). In addition, unlike humans, PEG is free of intra-rater variability. These features make the PEG overall score an ideal outcome variable for growth modeling since its scale, accuracy, and reliability remains consistent over time (Singer & Willet, 2003).

While the PEG overall score was the main dependent variable of interest, additional analyses were conducted to examine performance on each of the six PEG trait scores which comprised the PEG overall score. In this way, it was possible to provide a more nuanced look into the comparative effects of AEE for SWD and TD.

Level 1 variables

The primary Level 1 variable was Time, operationalized as the number of essay drafts (i.e., revisions) completed by each student. Three different Level 1 models were tested: a linear model in which Time counted each successive essay draft, a logarithmic model which was formed by taking the natural log of Time, and a quadratic model which added Time raised to the second power to the linear model. In each case, Time was centered such that 0 represented students' initial draft.

Level 2 variables

Level 2 variables included measures designed to detect inter-individual variability in performance and growth in response to PEG Writing. Eight Level 2 variables were analyzed. The primary Level 2 variable of interest was Special Education status (SPED), coded 1 for SWD and 0 for TD. Other Level 2 variables of interest included (a) Grade level, centered such that 0 represented Grade 4; (b) Gender (i.e., Male), coded 1 for males and 0 for females; (c) Free or reduced lunch status (FRL), a measure of students' socio-economic status, coded 1 for FRL and 0 for not-FRL; (d) English Language Learner (ELL) status, coded 1 for ELL and 0 for not-ELL; (e) Race/Ethnicity, for which four dummy variables were created to represent five ethnicities; (f) Prompt source, a dummy-coded variable indicating whether students responded to a teacher- or system-created prompt, coded 1 and 0 respectively; and, (g) Prompt genre, dummy-coded to represent whether the prompts were argumentative, informative, or narrative.

Level 3 variables

Level 3 variables were those measured at the school level and which may have affected AEE implementation and effectiveness. Two variables were included in the growth models: School classification and Students per internet-connected computer (SPIC). Data for both of these variables was gathered from the state Department of Education website from the most recent year available (2012–2013).

The first variable, School classification, was a measure of school performance developed by the State Department of Education as part of its accountability system. In this system, schools were classified as "Excelling," "Progressing," "Transitioning," "In Review," or "Turnaround" based on a composite of the following factors: School Performance Index (SPI; a measure based on the average performance of all students on state assessments), percentage of students scoring in the Advanced level (Band 5) of state assessments, subgroup performance, 4-year cohort graduation rate, and participation rates in state assessments. Excelling schools had the highest average SPI, greatest percentage of students scoring in the Advanced level on state achievement tests, lowest gaps in subgroup performance, highest 4-year cohort graduation rate, and highest participation rates in state assessments. Turnaround schools were those where the converse was true. In the 2012–2013 year, the state classified 15 % of its schools as Excelling, 29 % as Progressing, 40 % as Transitioning, 10 % as In Review, 4 % as Focus, and 2 % as Turnaround. As shown

in Table 1, no schools in the present sample were classified in the latter three categories. For data analysis, this variable was centered such that 0 represented an Excelling school.

The second variable, Students per internet-connected computer (SPIC), was used as an indicator of availability of computers within each school relative to the size of the school population. This metric was compiled by the state department of education. An SPIC of 1.0 indicated equal numbers of students and computers, while an SPIC of 3.0 indicated three students for every one computer. The grand mean of SPIC for the primary sample was 2.99 ($SD = 1.18$, range 1.00–5.90). There were no statistically significant differences in SPIC for either SWD or TD students; both groups' mean SPIC values were equivalent to the grand mean. For data analysis, SPIC was centered by subtracting 1.0 from every SPIC value; in this way, 0 represented an SPIC ratio equal to 1.0.

Design and data analysis

The present study used three-level hierarchical linear growth modeling (HLM) to evaluate (a) the shape and rate of growth of students' writing quality across successive revisions to a writing prompt, and (b) the effects of student- and school-level variables on performance and growth. Repeated observations of the outcome variable, the PEG overall score, constituted the Level 1 model, which were nested within students (Level 2) who were nested within schools (Level 3). HLM is ideal for analyzing unbalanced datasets in which the number of measurement occasions differs across participants (Singer & Willett, 2003). Indeed, there was considerable variability with regard to the number of essay drafts students completed (range 1–37). HLM allows estimation of fixed effects using the complete dataset and estimation of variance components with available data.

All analyses were completed using HLM version 7.0 (Raudenbush, Bryk, & Congdon, 1988) and estimated using Full Information Maximum Likelihood. All estimated models iterated and converged without errors. To ensure against violations of the assumption of normality, robust standard errors are reported for all fixed effects.

Results

RQ1: shape and rate of growth across drafts

Table 2 reports results of the unconditional means model, unconditional logarithmic growth model, and the conditional model testing the effect of special education status (SPED) for data from the primary sample's performance on their first writing prompt. Intra-class correlations (ICC) calculated from the unconditional means model showed that 11.25 % of the variance in the PEG overall score was at the level of within-student change (Level 1), 62.56 % of the variance was at the student level (Level 2), and 26.19 % of the variance was at the school level (Level 3). Findings

Table 2 HLM models predicting PEG overall score for primary sample

	Unconditional means model		Unconditional logarithmic growth model		Conditional model with SPED as predictor	
	Coefficient (SE)	*t* (*df*)	Coefficient (SE)	*t* (*df*)	Coefficient (SE)	*t* (*df*)
Fixed effects						
Initial status, π_{0ij}						
γ_{000}	17.57 (0.39)	45.08*** (46)	17.20 (0.37)	46.36*** (46)	17.90 (0.41)	44.15*** (46)
SPED					−1.27 (0.24)	−5.35*** (1100)
Growth, π_{1ij}						
γ_{100}			1.03 (0.12)	8.54*** (46)	0.85 (0.15)	5.8*** (46)
SPED					0.35 (0.16)	2.15* (1100)
Variance components						
Level-1 (time)						
σ^2	2.35***		0.76***		0.76***	
Level-2 (students)						
$\tau\pi_{00}$	13.09***		15.05***		14.63***	
$\tau\pi_{11}$			2.54***		2.56***	
Level-3 (schools)						
$\tau\beta_{00}$	5.48***		4.79***		5.24***	
$\tau\beta_{11}$			0.18*		0.18*	
Goodness of fit statistics						
D	13,366.82	*df* = 4	12,270.91	*df* = 9	12,242.08	*df* = 11
AIC	13,374.82		12,288.91		12,264.08	
BIC	13,395.17		12,334.69		12,320.03	
SBIC	13,368.00		12,273.56		12,245.33	
χ^2 difference test	–		1095.91***, *df* = 4		33.65***, *df* = 2	

PEG overall score range = 6–30

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

support the choice to estimate a three-level model and explore predictors at each of these levels.

Though accounting for the least variance in the model, Level 1 variance was statistically significantly different from zero: $\sigma^2 = 2.35$, $SE = 0.08$, $Z = 28.53$, $p < 0.001$. Several Level l models were estimated including linear, logarithmic, and quadratic growth models. All models demonstrated a statistically significant improvement in model fit over the unconditional growth model. However, the logarithmic model demonstrated the best fit to the data, and was superior to the other models. The unconditional logarithmic growth model explained 67.7 % of the Level 1 variance in the unconditional means model. Figure 1 displays growth in students' PEG overall score based on the best fitting Level 1 model. Student growth occurred

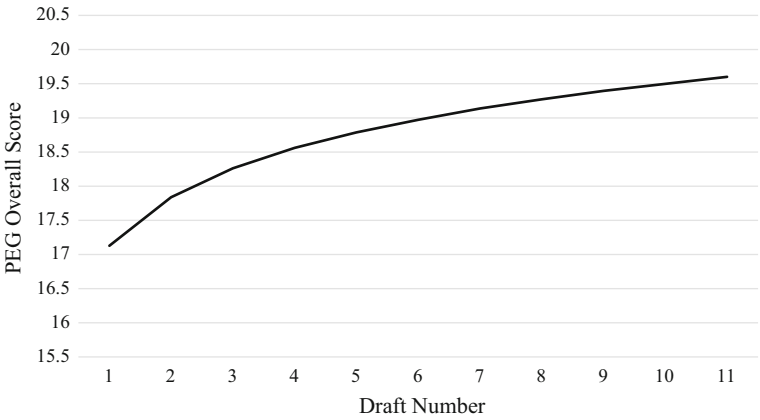

Fig. 1 Unconditional logarithmic growth model. PEG overall score range = 6–30

most frequently between the first few drafts and progressively slowed with increasing revisions.

RQ2: differential effects for students with and without disabilities

Prior to exploring the effect of special education status in the HLM growth model, basic usage statistics were calculated for SWD and TD students to investigate whether there were systematic differences in PEG Writing usage across groups. First, the number of essay drafts for the first prompt were calculated and compared for each group. While the range of drafts completed by SWD was larger (1–37) than that for TD students (1–22), there were no statistically significant differences in the average number of drafts completed between groups, as indicated by both parametric and non-parametric tests: $F(1, 1194) = 0.42$, $p = 0.519$; Mann–Whitney U = 180,167.50; Wilcoxon W = 359,268.50, Standardized Test Statistic = 0.249, Asymptotic Sig. (two-sided test) = 0.803. Next, groups were compared with regard to the frequency with which they responded to teacher-created or system-created prompts. There were no statistically significant differences between groups in relation to prompt source: $\chi^2 = 0.16$, $df = 1$, $p = 0.686$. SWD responded to 298 teacher-created prompts and 300 system-created prompts while TD students responded to 291 teacher-created prompts and 307 system-created prompts. Thus, there were no group differences PEG Writing usage with regard to the amount of drafts completed or the frequency with which students responded to teacher- or system-created prompts.

HLM was then used to analyze the effect of disability status on initial performance and growth. Correspondingly, SPED was entered as a predictor of both the intercept and slope of the logarithmic growth model. SPED was a statistically significant predictor of both parameters. As shown in the last column of Table 2, SWD tended to perform, on average, 1.27 points lower than TD students on their initial drafts. However, they grew at a statistically significantly faster rate than their TD peers matched on prior writing achievement: $\beta = 0.35$, $t = 2.15$, $p = 0.032$.

Table 3 Estimates of growth in PEG trait scores for TD and SWD

PEG trait score Parameter	Unconditional means model Estimate (SE)	Unconditional log growth model Estimate (SE)	Conditional growth model Estimate (SE)
Overall development			
Intercept	2.69*** (0.07)	2.91** (0.06)	3.03*** (0.07)
			*−0.22*** (0.04)*
Slope		0.16*** (0.02)	0.13*** (0.03)
			0.05^{ns} (0.03)
Organization			
Intercept	3.13*** (0.08)	3.07*** (0.07)	3.18*** (0.08)
			*−0.19*** (0.04)*
Slope		0.17*** (0.03)	0.12*** (0.03)
			*0.08** (0.03)*
Support			
Intercept	2.87*** (0.08)	2.81*** (0.07)	2.94*** (0.08)
			*−0.23*** (0.04)*
Slope		0.16*** (0.03)	0.14*** (0.03)
			0.04^{ns} (0.03)
Sentence structure			
Intercept	2.70*** (0.06)	2.65*** (0.05)	2.75*** (0.06)
			*−0.19*** (0.04)*
Slope		0.16*** (0.02)	0.15*** (0.03)
			0.02 ns (0.04)
Word choice			
Intercept	3.08*** (0.06)	3.02*** (0.06)	3.13*** (0.07)
			*−0.20*** (0.04)*
Slope		0.16*** (0.02)	0.12*** (0.02)
			0.07 (0.03)*
Conventions			
Intercept	2.84*** (0.07)	2.75*** (0.07)	2.87*** (0.07)
			*−0.22*** (0.04)*
Slope		0.23*** (0.02)	0.19*** (0.02)
			*0.08** (0.03)*

Conditional growth model tests the effect of special education status (SPED) on the intercept and slope of the unconditional logarithmic growth model. Estimates of the coefficients of the SPED terms are italicized beneath their associated parameter. PEG trait scores range from 1 to 5

ns not significant

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Follow-up analysis: PEG trait scores

To provide a more nuanced analysis of the effect of disability status on initial performance and growth in writing quality across revisions, six additional HLM

analyses were conducted in which each of the six PEG trait scores was used as a dependent variable. These follow-up analyses estimated an unconditional means model, an unconditional logarithmic growth model, and a conditional growth model in which disability status was the sole predictor—estimation of the full conditional model with all student and school-level variables was reserved for the model predicting the PEG overall score (see results for RQ3). Table 3 reports results of these analyses.

The unconditional logarithmic growth models for each of the traits showed that students improved their scores on each of the six traits, suggesting that students were able to utilize feedback from PEG to make improvements on each of these dimensions of their writing. However, first-draft responses produced by SWD scored approximately 0.20 points lower than TD students for each of the PEG traits. Investigation of the effect of disability status on the slope parameters for each model indicated that, compared to TD students, SWD demonstrated accelerated growth for the traits of organization, word choice, and conventions, but not for the other three traits. Thus, the accelerated growth that SWD demonstrated relative to TD students for the PEG overall score (Table 2) can be understood as the result of accelerated improvements in those three traits of writing quality.

RQ3: student and school-level predictors of initial performance and growth

Exploratory analyses were used to assess the contribution of additional student-level (Level 2) and school-level (Level 3) variables to the model predicting growth in the PEG overall score. The final conditional model, trimmed to exclude non-statistically significant predictors, is presented in Table 4.

Non-statistically significant predictors included race and grade level. Race was not a statistically significant predictor of the intercept or slope. This was also true when race was entered in the model as a dichotomized variable (0 = White, 1 = non-White). Grade level did not predict the growth slope, but it was a statistically significant predictor of the intercept. However, after accounting for the effect school classification on the intercept, grade level no longer predicted variance and was thus dropped from the final model reported in Table 4.

With respect to the intercept, several Level 2 variables were statistically significant predictors. In addition to special education status (SPED), the effects of Male, FRL, and ELL on the intercept were all statistically significant and in the negative direction. The effect of prompt source was positive and indicated that students scored approximately 1.8 points higher on their initial drafts when responding to teacher-created prompts versus system-created prompts. One of the two dummy variables for prompt genre was statistically significant: students scored approximately 1.2 points higher when responding to narrative prompts as compared to argumentative or informative prompts. In addition, one Level 3 variable, school classification, was a statistically significant predictor of the intercept. Students in Excelling schools produced initial drafts of higher quality than students in Progressing schools who performed higher than students in Transitioning schools. Thus, the intercept of the final conditional model can be interpreted as the PEG overall score in Excelling schools for female, non-ELL, TD students who did not

Table 4 Final conditional HLM growth model predicting PEG overall score for primary sample

	Coefficient (SE)	*t* (*df*)
Fixed effects		
Initial status, π_{0ij}		
Intercept γ_{000}	19.03 (0.73)	26.23*** (45)
School classification	−1.03 (0.41)	−2.49** (45)
SPED	−1.33 (0.21)	−6.28*** (1093)
Male	−1.30 (0.25)	−5.28*** (1093)
FRL	−1.23 (0.18)	−7.00*** (1093)
ELL	−1.25 (0.46)	−2.70*** (1093)
Prompt source	1.77 (0.63)	2.81** (1093)
Prompt genre[a]	1.18 (0.31)	3.77*** (1093)
Growth, π_{1ij}		
Intercept, γ_{100}	0.89 (0.13)	6.86*** (46)
SPED	1.03 (0.35)	2.98** (1093)
School classification	−0.37 (0.19)	−2.00* (1093)
SPIC	−0.15 (0.07)	−2.15* (1093)
Variance components		
Level-1 (time)		
σ^2	0.76***	
Level-2 (students)		
$\tau\pi_{00}$	13.08***	
$\tau\pi_{11}$	2.51***	
Level-3 (schools)		
$\tau\beta_{00}$	2.86***	
$\tau\beta_{11}$	0.15 ns	
Goodness of fit statistics		
D	12,070.75	$df = 19$
AIC	12,108.75	
BIC	12,205.40	
SBIC	1208.37	

PEG overall score range = 6–30

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

[a] Prompt genre tests the effect of responding to narrative versus informative and argumentative prompts

receive FRL and who responded to system-created prompts in either argumentative or informative genres.

With respect to the growth slope, the only student-level variable to predict variance in individual growth was SPED: SWD grew at a faster rate than TD students. Both school-level variables were tested as predictors of the growth slope and the effect of SPED on the growth slope. Neither variable predicted variance in the growth slope, but both variables predicted variance in the effect of SPED on growth in the PEG overall score across drafts. Thus, based on the final conditional model, the growth slope can be interpreted as the rate of improvement in PEG overall score for TD students in Excelling schools where there was one student per internet-connected computer (SPIC = 1.0). The addition of the remaining parameters can be interpreted to mean that SWD grew at a faster rate than TD students,

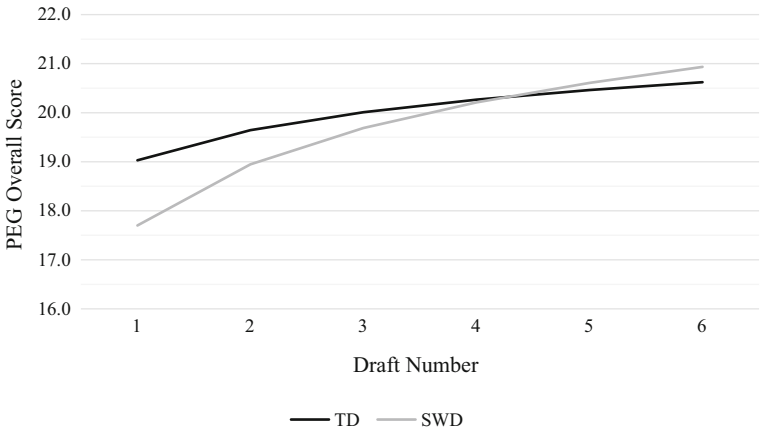

Fig. 2 Effect of special education status on growth in PEG overall score (range 6–30) for students with and without disabilities holding all other variables constant in the model. *TD* typically-developing students, *SWD* students with disabilities

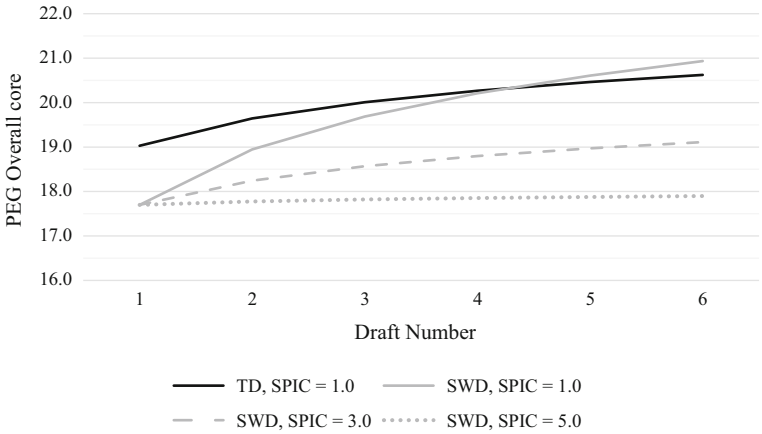

Fig. 3 Effect of increasing proportions of students per internet-connected computer (SPIC) on the effect of special education status on growth in writing quality across essay drafts holding all other variables constant in the model. *TD* typically-developing students, *SWD* students with disabilities. PEG overall score range = 6–30

but this growth rate slowed as school quality decreased ($\gamma_{111} = -0.37$, $t = -2.00$, $df = 1093$, $p = 0.046$) and as the availability of computers in schools decreased ($\gamma_{112} = -0.15$, $t = -2.15$, $df = 1093$, $p = 0.032$).

Representative growth trajectories are presented in Figs. 2 and 3. Figure 2 presents growth trajectories for students with and without disabilities while holding all other predictors constant at their intercept. As shown on the graph, though SWD produced initial drafts of lower quality, they were able to close the gap. By the fifth draft SWD had improved their PEG overall score to the point where it was equivalent to that of TD students.

Figure 3 illustrates the effect of SPIC on the growth in writing quality of SWD across successive drafts. The figure illustrates that in schools where there were 1:1 devices, SWD were able to close the performance gap within five drafts, and by the sixth draft even slightly exceed the writing quality of their TD peers. However, when the SPIC increased to 3:1—illustrated in the dashed line of the graph—SWD needed to complete 6 drafts simply to score at the level where TD students scored on their first draft. When the SPIC increased to 5:1—illustrated in the dotted line of the graph—growth of SWD slowed to near zero.

RQ4: transfer effects

Using data from the transfer sample ($n = 655$), it was possible to explore whether students demonstrated improved independent performance on a subsequent prompt they completed within PEG Writing. A growth model for the transfer sample was calculated using SPED as a predictor of the intercept.

Table 5 presents results of the Unconditional Means Model, Unconditional Logarithmic Growth Model, and Conditional Model with SPED as a predictor of the intercept. Based on the variance components of the Unconditional Means Model, 12.13 % of the variance in growth in writing quality was at Level 1. Students varied very little in their growth across revisions; however, this variability was statistically significantly different from zero: $\sigma^2 = 2.56$, $SE = 0.10$, $Z = 24.55$, $p < 0.001$. The majority of the variance, 61.80 %, was between individuals (Level 2), and 26.07 % of the variance was at Level 3.

As before, the logarithmic model proved the best fit to the data. This model explained 63.29 % of the Level 1 variance in the Unconditional Means Model. SPED was again a statistically significant predictor of the intercept. SWD scored, on average, 0.89 points below TD students. Results of a t test comparing the coefficients of SPED from the first and second prompt indicated failure to reject the null hypothesis of equal intercepts: $t = 0.94$, $df_{\text{pooled}} = 926$, $p = 0.348$. No improvement was found for TD students, either: $t = 0.19$, $df_{\text{pooled}} = 921$, $p = 0.849$.

Discussion

This study examined whether students with and without disabilities experienced differential benefit from automated feedback provided by an AEE system called PEG Writing. Equal samples of SWD and TD, matched on prior writing achievement, were selected to form a primary sample of 1196 students in grades 4–8 who used PEG Writing in the winter and spring of 2014. Hierarchical linear growth modeling was used to analyze the shape and rate of growth, as well as identify predictors of students' growth in writing quality across essay drafts scaffolded by automated feedback from PEG. A subsequent analysis examined evidence of transfer to improved first draft performance using a subsample of students who completed a second writing prompt in PEG. Study findings are discussed in turn.

Table 5 Hierarchical linear growth models of transfer sample's performance on the second prompt

	Unconditional means model		Unconditional logarithmic growth model		Conditional model with SPED predicting intercept	
	Coefficient (SE)	*t* (*df*)	Coefficient (SE)	*t* (*df*)	Coefficient (SE)	*t* (*df*)
Fixed effects						
Initial status, π_{0ij}						
γ_{000}	17.81 (0.46)	39.12*** (34)	17.33 (0.46)	37.83*** (34)	17.77 (0.52)	34.25*** (34)
SPED					−0.89 (0.33)	−2.65** (584)
Growth, π_{1ij}						
γ_{100}			1.13 (0.13)	8.83*** (34)	1.14 (0.13)	8.85*** (34)
Variance components						
Level-1 (time)						
σ^2	2.56***	0.94***		0.94***		
Level-2 (students)						
$\tau\pi_{00}$	13.03***	14.94***		14.69***		
$\tau\pi_{11}$		1.75***		5.67***		
Level-3 (schools)						
$\tau\beta_{00}$	5.50***	5.50***		1.76***		
$\tau\beta_{11}$				0.16**		
Goodness of fit statistics						
D	8662.39	$df = 4$	7818.72	$df = 9$	7810.49	$df = 10$
AIC	8670.39		7836.72		7830.49	
BIC	8688.33		7877.09		7875.34	
SBIC	8663.47		7821.16		7813.19	
χ^2 difference test	–	843.66***, $df = 5$		8.23**, $df = 1$		

PEG overall score range = 6–30

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Shape and rate of growth

A logarithmic growth model proved to be the best fit to the data. Previous studies of PEG Writing (Wilson & Andrada, 2016; Wilson et al., 2014) reported that a quadratic model displayed the best fit to the data. However, both logarithmic and quadratic models are consistent in illustrating that the majority of students' growth in writing quality occurred within the first few drafts, and gradually slowed with subsequent drafts.

Why might a pattern of decelerating growth occur? One possibility is that during the first few drafts students rely on PEG's trait feedback to make substantive changes to their essay, adding and deleting content, adjusting organization, and improving cohesion. Then, after these substantive changes are made, students make edits to correct their spelling and grammar, and improve their word choice. Revisions of a substantive nature tend to have a larger effect on measures of writing

quality, while spelling and grammar edits tend to have a much smaller effect (Clare, Valdes, & Pathey-Chavez, 2000; MacArthur et al., 2009; Matsumara, Patthey-Chavez, Valdes, & Garnier, 2002). This pattern of first revising and then editing is also consistent with common instructional approaches, such as the Writing Process model. Additional research is needed, however, to test this hypothesis by tracking specific changes in students' text across drafts.

Comparisons of SWD and TD students

Findings for SWD are encouraging both in terms of usage patterns and outcomes. SWD engaged in the same average amount of revision, and responded to teacher-created prompts with equal frequency as TD students. The latter finding is important because it is suggestive of equal access to the general education curriculum, since teacher-created prompts are typically curriculum specific.

It was also encouraging that, while SWD tended to perform on average 1.27 points lower than TD students on their initial drafts, their overall writing quality grew at a statistically significantly faster rate. SWD were able to close this performance gap in their PEG overall score after five drafts.

Follow-up analyses of the six PEG trait scores provided a more nuanced understanding of the associated effects of automated feedback on the growth in writing of SWD. SWD demonstrated more rapid growth with respect to the traits of organization, word choice and conventions. This is an encouraging finding, suggesting that AEE is able to help SWD make higher-level revisions (i.e., improving the organization and coherence of a text) and not just make improvements in lower-level characteristics like word choice and conventions, the type of changes SWD typically make without substantial intervention (Graham, MacArthur, & Schwartz, 1995; MacArthur et al., 1991, 2009).

Additional predictors of initial performance and growth

Student and prompt variables

Aside from special education status, no other student or prompt variable explained variance in the growth slope. However, males, students receiving free or reduced lunch, and ELLs all scored lower on their initial drafts than their counterparts. These findings are consistent with those of prior studies of AEE (Wilson & Andrada, 2016; Wilson et al., 2014) and with reports of subgroup performance on NAEP writing assessment (NCES, 2012; Persky et al., 2002; Salahu-Din et al., 2008). Consistent with prior studies (Wilson & Andrada, 2016; Wilson et al., 2014), these demographic characteristics did not predict growth, and thus were not determinants of a student's ability to improve their writing quality when using PEG Writing. Given the persistent achievement gaps faced by ELLs and students from low-SES households, ideally these students would grow at a faster rate in order to close the gap. Future research that intentionally focuses on these subgroups is needed. Other demographic variables of interest were race and grade level, but neither were statistically significant predictors of initial performance and growth.

With respect to prompt genre and prompt source, each was a statistically significant predictor of initial performance but not growth. Students composed higher quality first drafts for narrative prompts than they did for argumentative or informative prompts. Findings are consistent with research on genre differences in writing performance (Graham, Harris, & Hebert, 2011; Graham, Hebert, Sandbank, & Harris, 2016; Olinghouse & Wilson, 2013) and may possibly reflect differences in students' familiarity with these genres (Gillespie, Olinghouse, & Graham, 2013; Lin et al., 2007).

Students who responded to teacher-created prompts scored slightly lower on their first drafts, but grew at an equivalent rate regardless of the prompt source. In a prior study, students scored slightly higher when they responded to teacher-created prompts (Wilson & Andrada, 2016), but this difference is likely a function of sample specific differences in the features of the teacher-created prompts. The relative difficulty of teacher-created prompts likely varies by teacher, school, district, and academic year. A more important finding is that PEG's feedback appears equally beneficial regardless of the prompt source. This is encouraging since the inability to create prompts has proven to be prohibitive to AEE usage (Warschauer & Grimes, 2008). Further, this opens up possibilities for utilizing PEG Writing to support writing in the content-areas, or to support individual learning objectives as outlined in the individualized education plans of SWD.

School-level variables

School classification was a statistically significant predictor of initial performance but not of growth. Students in Excelling schools produced higher quality first drafts than those in Progressing schools who outperformed students in Transitioning schools. School classification had no effect on growth, but did explain variance in the effect of special education status on growth. SWD improved their text quality at the most rapid rate in Excelling schools with growth slowing as school quality decreased. The growth rate of SWD was also affected by the proportion of students per internet-connect computer (SPIC) within a school. As shown in Fig. 3, even in Excelling schools, growth of SWD slowed to a near halt when SPIC ratios increased to 5:1 (i.e., five students for every one computer). This is an important finding, and one that underscores Graham and Harris's (2013) admonishment that "No student with LD should head to school in the morning without a computer to write on" (p. 35). In the current study, SWD were able to improve their writing quality at a faster rate than their typically-developing peers, but only when the school had sufficient numbers of internet-connected devices available. Thus, schools and districts considering adopting AEE systems must ensure the availability of sufficient technological resources (computers, servers, fast internet-connection speeds) to maximize the potential of these systems for supporting growth and improvement of all students, and particularly for SWD.

Transfer

The current study was also interested in the issue of transfer, the ability to generalize knowledge and skills from one situation to the next (National Research Council, 2000). In the current study this was operationalized as growth in first-draft writing quality from one prompt to the next. However, no such growth was found. This is consistent with prior research on PEG Writing (Wilson & Andrada, 2016; Wilson et al., 2014), and findings from a recent literature review of AEE: students commonly improve their writing quality in response to the feedback they receive, but rarely show evidence of transfer (Stevenson & Phakiti, 2014). The issue of transfer is an important one, and should continue to be explored, particularly in studies that examine AEE usage over extended timeframes. Given the complexity of writing, and the numerous skills which must be developed and coordinated, growth is typically slow (Berninger, 2000), underscoring the importance of successive practice over extended timeframes for detecting evidence of transfer. One prior study (Shermis, Garvan, & Diao, 2008) looked at growth across the school year, but high rates of attrition and lack of a comparison group, make it difficult to attribute transfer effects to AEE and to draw generalizable conclusions. Thus, further longitudinal research is needed to assess whether AEE is able to support transfer of learning.

Limitations and future research

Study findings must be interpreted in light of study limitations. First, no causal inferences can be drawn from the study design. The study was drawn from available data provided by the state department of education, and no experimental manipulation occurred. It was not possible to control for several factors such as the amount of time elapsed between essay drafts, whether or not students used PEG's electronic graphic organizers, the instructional context within which AEE was introduced and utilized, use of multimedia tutorials, or whether additional teacher feedback was provided. Efforts were made to control one known source of variance, students' prior writing ability, by using a stratified random sample that had equal amounts of SWD and TD students from each of the five achievement bands of the 2013 state writing assessment. In addition, multiple other demographic, prompt, and school variables were adjusted for statistically through the HLM models. Still, findings only document an association between automated feedback and improvements in writing quality. Future research employing rigorous experimental and quasi-experimental designs is needed to establish a causal link between use of AEE and benefits for SWD. Furthermore, as described in the introduction, there are several ways in which AEE *may* support the writing outcomes of SWD. It remains unclear exactly *how* AEE benefits SWD. For instance, without a comparison group it was not possible to ascertain the degree to which the benefits of AEE may have differed from that of basic word processing software. Nevertheless, documenting a positive association between the use of PEG Writing and growth in writing quality of SWD is an important first step in this line of research.

Second, without a measure of writing ability external to PEG Writing, it is unclear whether the ability to benefit from PEG's feedback as indicated by improvement in the PEG overall score would also signify improvement on external measures of writing skill (i.e., whether human raters would also detect improvements in overall writing quality from first draft to final draft). While this is a possibility, given the body of research indicating high degrees of consistency between PEG scores and those assigned by trained raters (Shermis, 2014; Shermis & Hammer, 2013), it is likely that growth in the PEG overall score would parallel growth in scores assigned by trained raters.

Next, SWD sampled in this study were those identified in a state database as receiving special education services via the provision of an IEP. Information regarding students' specific disability categories was not available for reporting. However, as described earlier, the prevalence of students served under the different IDEA categories in this state broadly mirrors that of the nation, with the majority of students receiving services in high-incidence areas (Specific Learning Disabilities, Other Health Impairment, Speech and Language Impairment). Thus, similar to NAEP analyses which report results for a heterogeneous group of SWD, current findings are likely to generalize to the population of SWD more broadly, but may not be replicable for specific subgroups within that population.

In sum, additional research should adopt rigorous experimental and quasi-experimental designs, should adopt multiple outcome measures, and measure these over extended timeframes. Furthermore, research should explore potential variability in the effects of AEE systems on specific well-defined subgroups of SWD such as students with specific learning disability, attention deficit hyperactivity disorder, or students with emotional/behavioral disorders.

Educational implications

Study findings cautiously point to the potential for AEE to support intervention and remediation efforts for SWD. Word processing software is already recommended for struggling writers and SWD (MacArthur, 2009; Morphy & Graham, 2012). AEE goes a step further by combining word processing and instantaneous feedback to accelerate the practice-feedback cycle (Kellogg et al., 2010) so critical for SWD (Vaughn et al., 2000). Though additional research is clearly needed, the current study documents a promising association between use of AEE and growth in writing quality for SWD, albeit one that is moderated by the availability of sufficient computing resources within a school.

If AEE is adopted to support intervention efforts, educators may wish to consider the findings that maximal student growth occurred within the first few drafts and transfer was not noted. This suggests that relying on excessive revising/editing of a single piece of writing is unlikely to make lasting gains for students. This is not to minimize the value and importance of iterative practice and revision within a single essay: SWD closed initial gaps in writing quality only after five drafts. Instead, students should iteratively improve multiple essays across the school year. Different prompts draw upon different sources of topic and genre knowledge (Ruth & Murphy, 1988), and thereby elicit new challenges. Thus, while cognitive theories of

writing emphasize the importance of practice for developing automaticity and proficiency (Alexander, 2003; Kellogg & Whiteford, 2009; McCutchen, 2011), educators should recognize that students may require multiple forms of practice to improve.

Indeed, AEE might make such a recommendation feasible. For instance, with PEG Writing, teachers can assign personalized or system-created prompts to students throughout the school year, and across content areas, and students can complete as many revisions as they wish, each time receiving instantaneous feedback to scaffold their next revision. In doing so, AEE presents a feasible solution to addressing the time costs of evaluating writing that most educators cite as the barrier for intensifying writing instruction and intervention.

Acknowledgments This research was supported in part by a Delegated Authority contract from Measurement Incorporated® to University of Delaware (EDUC432914150001). The opinions expressed in this paper are those of the author and do not necessarily reflect the positions or policies of this agency, and no official endorsement by it should be inferred.

References

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

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher, 32*(8), 10–14.

Applebee, A. N., & Langer, J. A. (2009). What is happening in the teaching of writing? *English Journal, 98*(5), 18–28.

Attali, Y. (2004). *Exploring the feedback and revision features of criterion.* Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Berninger, V. W. (2000). Development of language by hand and its connections with language by ear, mouth, and eye. *Topics in Language Disorders*, *20*(4), 65–84.

Berninger, V. W., Nagy, W., Tanimoto, S., Thompson, R., & Abbott, R. D. (2015). Computer instruction in handwriting, spelling, and composing for students with specific learning disabilities in grades 4–9. *Computers & Education, 81*, 154–168.

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

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

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

Common Core State Standards Initiative. (2010). *Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/ELA-Literacy/.

Dockrell, J. E., Lindsay, G., & Connelly, V. (2009). The impact of specific language impairment on adolescents' written text. *Exceptional Children, 75*, 427–446.

Englert, C. S., Mariage, T. V., Okolo, C. M., Shankland, R. K., Moxley, K. D., Courtad, C. A., et al. (2009). The learning-to-learn strategies of adolescent students with disabilities: Highlighting, note taking, planning, and writing expository texts. *Assessment for Effective Intervention, 34*, 147–161.

Foltz, P. W., Lochbaum, K. E. & Rosenstein, M. B. (2011). *Analysis of student ELA writing performance for a large scale implementation of formative assessment.* Paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans, LA.

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

Gersten, R., & Baker, S. (2001). Teaching expressive writing to students with learning disabilities: A meta-analysis. *The Elementary School Journal, 101*, 251–272. doi:10.1086/499668.

Gilbert, J., & Graham, S. (2010). Teaching writing to elementary students in grades 4–6: A national survey. *Elementary School Journal, 110*, 494–518.

Gillespie, A., & Graham, S. (2014). A meta-analysis of writing interventions for students with learning disabilities. *Exceptional Children, 80*, 454–473.

Gillespie, A., Olinghouse, N. G., & Graham, S. (2013). Fifth-grade students' knowledge about writing process and writing genres. *Elementary School Journal, 113*, 565–588.

Graham, S., Bollinger, A., Booth Olson, C., D'Aoust, C., MacArthur, C., McCutchen, D., et al. (2012a). *Teaching elementary school students to be effective writers: A practice guide (NCEE 2012-4058).* Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Graham, S., Capizzi, A., Harris, K. R., Hebert, M., & Morphy, P. (2013). Teaching writing to middle school students: A national survey. *Reading and Writing, 27*, 1015–1042.

Graham, S., & Harris, K. R. (2003). Students with learning disabilities and the process of writing: A meta-analysis of SRSD studies. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 323–344). New York, NY: Guilford.

Graham, S., & Harris, K. R. (2013). Common core state standards, writing, and students with LD: Recommendations. *Learning Disabilities Research and Practice, 28*, 28–37.

Graham, S., Harris, K. R., & 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., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *Elementary School Journal, 115*, 523–547.

Graham, S., Hebert, M., Sandbank, M. P., & Harris, K. R. (2016). Assessing the writing achievement of young struggling writers: Application of generalizability theory. *Learning Disability Quarterly, 39*, 72–82.

Graham, S., MacArthur, C., & Schwartz, S. (1995). Effects of 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. (2012b). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104*, 879–896.

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

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6), 1–44.

Harris, K. R., & Graham, S. (2013). "An adjective is a word hanging down from a noun": Learning to write and students with learning disabilities. *Annals of Dyslexia, 63*, 65–79.

Hayes, J. R. (2004). What triggers revision? In L. Chanquoy, P. Largy, & L. Allal (Eds.), *Revision: Cognitive and instructional processes* (pp. 9–20). Boston, MA: Kluwer Academic.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*(3), 369–388.

Keith, T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147–167). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

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

Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research, 42*(2), 173–196.

Lin, C. S., Monroe, B. W., & Troia, G. A. (2007). Development of writing knowledge in grades 2–8: A comparison of typically developing writers and their struggling peers. *Reading and Writing Quarterly: Overcoming Learning Difficulties, 23*, 207–230.

MacArthur, C. A. (2009). Reflections on research on writing and technology for struggling writers. *Learning Disabilities Research and Practice, 24*, 93–103.

MacArthur, C. A., Graham, S., & Harris, K. R. (2009). Insights from instructional research on revision with struggling writers. In L. Allal, L. Chanquoy, & P. Largy (Eds.), *Revision cognitive and instructional processes* (pp. 125–139). Boston, MA: Kluwer Academic.

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., Valdés, R., & Garnier, G. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *Elementary School Journal, 103*, 3–25.

McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Writing Research, 3*, 51–68.

Moore, N. S., & MacArthur, C. A. (2016). Student use of automated essay evaluation technology during revision. *Journal of Writing Research, 8*, 149–175. doi:10.17239/jowr-2016.08.01.05.

Morphy, P., & Graham, S. (2012). Word processing programs and weaker writers/readers: A meta-analysis of research findings. *Reading and Writing, 25*, 641–678.

National Center for Education Statistics. (2012). *The Nation's Report Card: Writing 2011 (NCES 2012-470)*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Commission on Writing for America's Families, Schools, and Colleges. (2003). *The neglected "R": The need for a writing revolution*. Iowa City, IA: The College Board.

National Research Council. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. Washington, DC: The National Academies Press. doi:10.17226/9853.

Nobles, S., & Paganucci, L. (2015). Do digital writing tools deliver? Student perceptions of writing quality using digital tools and online writing environments. *Computers and Composition, 38*, 16–31.

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. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading and Writing Quarterly, 19*, 139–158. doi:10.1080/10573560308222.

Persky, H. R., Daane, M. C., & Jin, Y. (2002). *The Nation's Report Card: Writing 2002 (NCES 2003-529)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department for Education.

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

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex.

Salahu-Din, D., Persky, H., & Miller, J. (2008). *The Nation's Report Card: Writing 2007 (NCES 2008-468)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Shermis, M. D., Garvan, C. W., & Diao, Y. (2008). *The impact of automated essay scoring on writing outcomes.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Shermis, M. D., & Hammer, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions*. New York, NY: Routledge.

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

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

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65.

Troia, G. A. (2006). *Instruction and assessment for struggling writers: Evidence-based practices*. New York, NY: Guilford.

Vaughn, S., Gersten, R., & Chard, D. J. (2000). The underlying message in LD intervention research: Findings from research syntheses. *Exceptional Children, 67*, 99–114.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*, 22–36.

Wilson, J., & Andrada, G. N. (2016). Using automated feedback to improve writing quality: Opportunities and challenges. In Y. Rosen, S. Ferrara & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 678–703). Hershey, PA: IGI Global.

Wilson, J., & 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.