Research Article
Coordinated Symposium
NCME 2018 Annual Meeting

Writing Mentor™: Writing Progress Using Self-Regulated Writing Support

Jill Burstein, *Educational Testing Service*

Norbert Elliot, *University of South Florida*

Beata Beigman Klebanov, *Educational Testing Service*

Nitin Madnani, *Educational Testing Service*

Diane Napolitano, *Educational Testing Service*

Maxwell Schwartz, *Educational Testing Service*

Patrick Houghton, *Educational Testing Service*

Hillary Molloy, *Educational Testing Service*

Abstract

The Writing Mentor™ (WM) application is a Google Docs add-on designed to help students improve their writing in a principled manner and to promote their writing success in postsecondary settings. WM provides automated writing evaluation (AWE) feedback using natural language processing (NLP) methods and linguistic resources. AWE features in WM have been informed by research about postsecondary student writers often classified as developmental (Burstein et al., 2016b), and these features address a breadth of writing sub-constructs (including use of sources, claims, and evidence; topic development; coherence; and knowledge of English conventions). Through an *optional* entry survey, WM collects self-efficacy data about writing and

DOI: 10.37514/JWA-J.2018.2.1.12

English language status from users. Tool perceptions are collected from users through an *optional* exit survey. Informed by language arts models consistent with the Common Core State Standards Initiative and valued by the writing studies community, WM takes initial steps to integrate the reading and writing process by offering a range of textual features, including vocabulary support, intended to help users to understand unfamiliar vocabulary in coursework reading texts. This paper describes WM and provides discussion of descriptive evaluations from an Amazon Mechanical Turk (AMT) usability task situated in WM and from *users-in-the-wild* data. The paper concludes with a framework for developing writing feedback and analytics technology.

Keywords: automated writing evaluation, feedback, natural language processing, self-efficacy, self-regulated writing, writing analytics, Writing Mentor

1.0 Introduction: Literacy and Natural Language Processing Solutions

Low literacy is a social challenge that affects citizens on a global level. This challenge has implications for critical and practical aspects of social participation, such as employability, and self-esteem and self-confidence (EU High Level Group, 2012). In terms of the global impact of the literacy issue, the Organisation for Economic Co-operation and Development (OECD) reports that, on average, about 20% of students in OECD countries do not attain the baseline level of proficiency in reading (PISA Results in Focus, 2016). In the United States (US), we find literacy challenges in K–12 and postsecondary levels. The National Center for Education Statistics (NCES) reports that average National Assessment for Educational Progress (NAEP) reading assessment scores are marginally proficient for 12th graders in the United States (US) (Musu-Gillette et al., 2017). While factors contributing to the US literacy challenge overall may be complex, the large number of English language learners (ELL) enrolled in US K–12 schools is one factor to the literacy challenge. In 2014–15, it was reported that 4.8 million ELLs were enrolled in K–12, and about 9.6% were participating in ELL programs.[1,2] It is reported that ELL students in cities made up an average of 14.2% of total public school enrollment, and in suburban areas, ELL students made up an average of 8.9% of public school enrollment (Musu-Gillette et al., 2017). Another factor appears in postsecondary contexts. In Fall 2017, it was reported that approximately 20.4 million students were expected to be enrolled in two- and four-year institutions, and this number is expected to rise over the next several years[3]. Millions of these enrolled students reportedly lack the prerequisite skills to succeed (NCES, 2016). Further, it is reported that more than 50% of students entering two-year colleges, and nearly 20% of students enrolled in four-year post-secondary institutions are placed in math, reading, and writing

[1] https://www2.ed.gov/datastory/el-characteristics/index.html
[2] https://ncela.ed.gov/files/fast_facts/05-19-2017/ProfilesOfELs_FastFacts.pdf
[3] https://nces.ed.gov/programs/digest/d16/tables/dt16_303.10.asp

developmental courses (Complete College America, 2012). Nearly 40% of two-year college students do not complete their developmental courses, and in four-year institutions, one-third or fewer of students placed in remediation graduate in six years; a number of reasons for low course and college completion rates are noted, including lack of preparation in reading and writing (Complete College America, 2012).

This article describes the Writing Mentor™ (WM) app—a technology solution to the literacy challenge—designed to help student writers progress and improve their writing in postsecondary settings (Madnani et al., 2018). Building on previous automated writing evaluation (AWE) research (Attali & Burstein, 2006; Burstein et al., 1998; Burstein et al., 2004; Shermis & Burstein, 2013), the app can support writers globally. This paper provides background information and a description of the app, a discussion of a crowd-sourcing (Amazon Mechanical Turk (AMT)) usability task situated in WM, and a qualitative evaluation of real-world use based on event log data collected through WM from submissions from *users-in-the-wild*.

Personalized learning is a goal of WM. The paper therefore concludes with a proposed framework for further developing writing feedback and analytics technology. To that end, the concluding discussion leverages the Institute of Education Sciences (IES) practice guidelines for effective writing (Graham et al., 2016), general attention to personalized learning (Pane, 2017), and identification of variables associated with low proficiency writers (Perin & Lauterbach, 2018).

2.0 The Writing Mentor™ Application[4]— A Google Docs Add-On

Writing Mentor (WM)[5] is a Google Docs add-on designed to provide instructional writing support. The app is intended to provide student writers with 24/7 support for academic writing, especially those in postsecondary settings. To that end, the app draws the user's attention to four key writing subconstructs expected in academic writing—specifically, credibility of claims, topic development, coherence, and editing.

2.1 Motivation

AWE systems have typically supported the measurement of pertinent writing skills for automated scoring of large-volume, high-stakes assessments (Attali & Burstein, 2006; Shermis et al., 2015) and online instruction (Burstein et al., 2004; Foltz et al., 2013; Roscoe et al., 2014). AWE has been used primarily for on-demand essay writing on standardized assessments. However, the real-time, dynamic nature of natural language processing (NLP)-based AWE affords the ability to generate linguistic analyses for a range of writing genres in postsecondary education, such as on-demand essay writing tasks, argumentative essays from the social sciences, and lab reports in STEM courses (Burstein et al., 2016a). Burstein et al. (2017) conducted an exploratory secondary data analysis that examined relationships between NLP-derived linguistic features extracted from on-demand writing samples from college students, and broader success

[4] https://mentormywriting.org
[5] The app was in beta at the time that this manuscript was written.

indicators (such as SAT and ACT composite and subject scores). Findings from Burstein et al. (2017) suggested that writing can provide meaningful information about student knowledge related to broader outcomes (college success indicators and learning outcomes measures). These findings also suggest that AWE may have greater potential for educational analytics beyond prevalent uses for writing assessment and instruction. AWE analyses can be used to generate *feedback* (to provide students with meaningful information to support their writing), and *educational analytics* (to inform various stakeholders, including students, instructors, parents, administrators, and policy-makers).

The WM app targets struggling writers and English learner (EL) populations enrolled in 2- and 4-year colleges. The app is intended to provide *one-stop-shopping* for writers who are looking for some writing help. Other apps that provide access to individual users, such as Grammarly, typically focus on English conventions only. Applications such as ETS's *Criterion*® (Burstein et al., 2004) and Turnitin's Revision Assistant[6] provide feedback above and beyond English conventions, but are not currently consumer-based and require institutional subscriptions, which limits accessibility.

A key motivation of WM development was to conduct research that serves to inform personalized learning with regard to improving writing. Therefore, WM collects user *event logs* that can be used to better understand the types of feedback that users seek and how writing feedback promotes document revision. The app contains an *optional* entry survey that asks users how confident they are as writers, and if English is the first language they learned to speak. Responses to that survey allow us to examine how self-reported self-efficacy about writing and English language proficiency may be related to feedback preferences. The app also includes an *optional* user perception exit survey of the application.

2.2 Application

The app can be installed from the Google Docs add-on store. The app provides users with actionable feedback related to the writing being convincing (e.g., *claims and sources*), well-developed (*topic development*), coherent (e.g., *flow of ideas*), and well-edited (e.g., *knowledge of English conventions*). Figures 1 and 2 illustrate the design of the app in its attention to feedback and provide a snapshot of the user experience. Users can select from any feedback type that is available. The idea is for the users to determine on their own what types of feedback are relevant to their submission and review and reflect on those feedback types to determine what revisions might improve the submission. Feedback is presented by a friendly, *non-binary*[7] persona named "Sam" (a gender-neutral name in English). Features selected in WM were informed by previous research with university faculty (Burstein et al., 2016a and Burstein et al., 2016b), literature related to English learners' language development that informed a language activity-generation app, the Language Muse™ Activity Palette (Madnani et al., 2016), as well as collaboration with writing research subject matter experts and classroom practitioners.

[6] See http://turnitin.com/en_us/2487-revision-assistant.
[7] *Non-binary* in this context indicates that the gender of the character is not exclusively masculine or feminine.

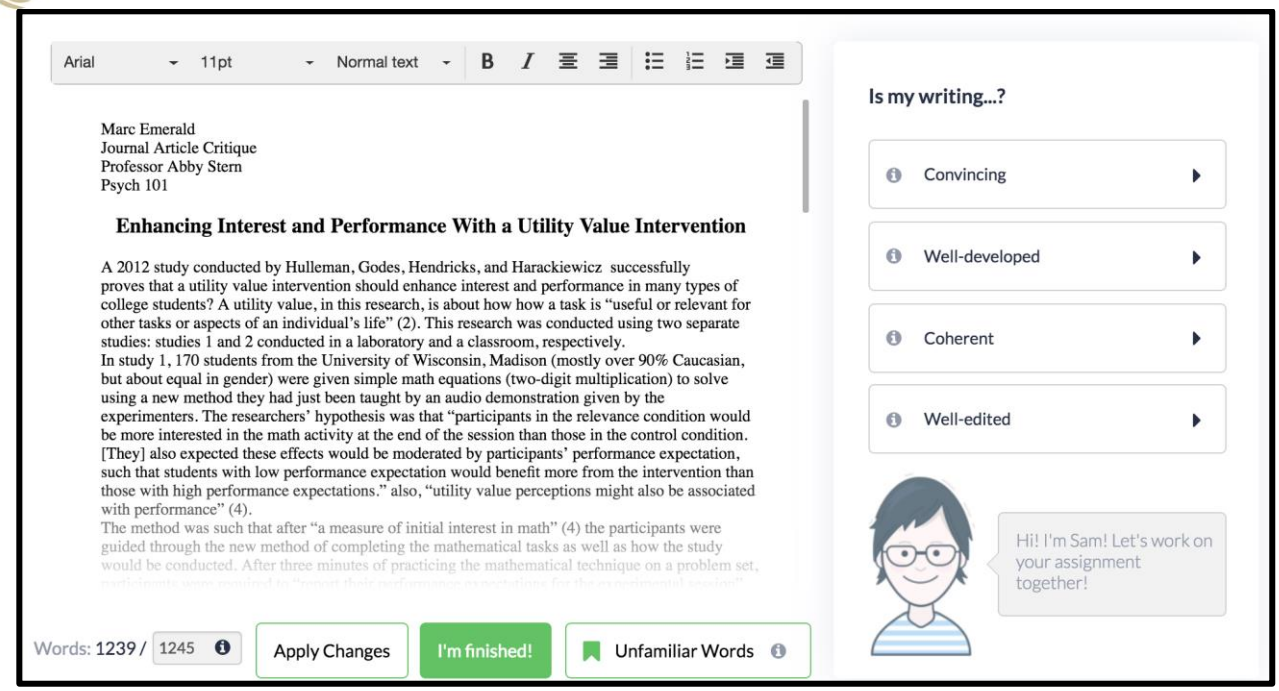

Figure 1. Main page illustrating full set of feature types.

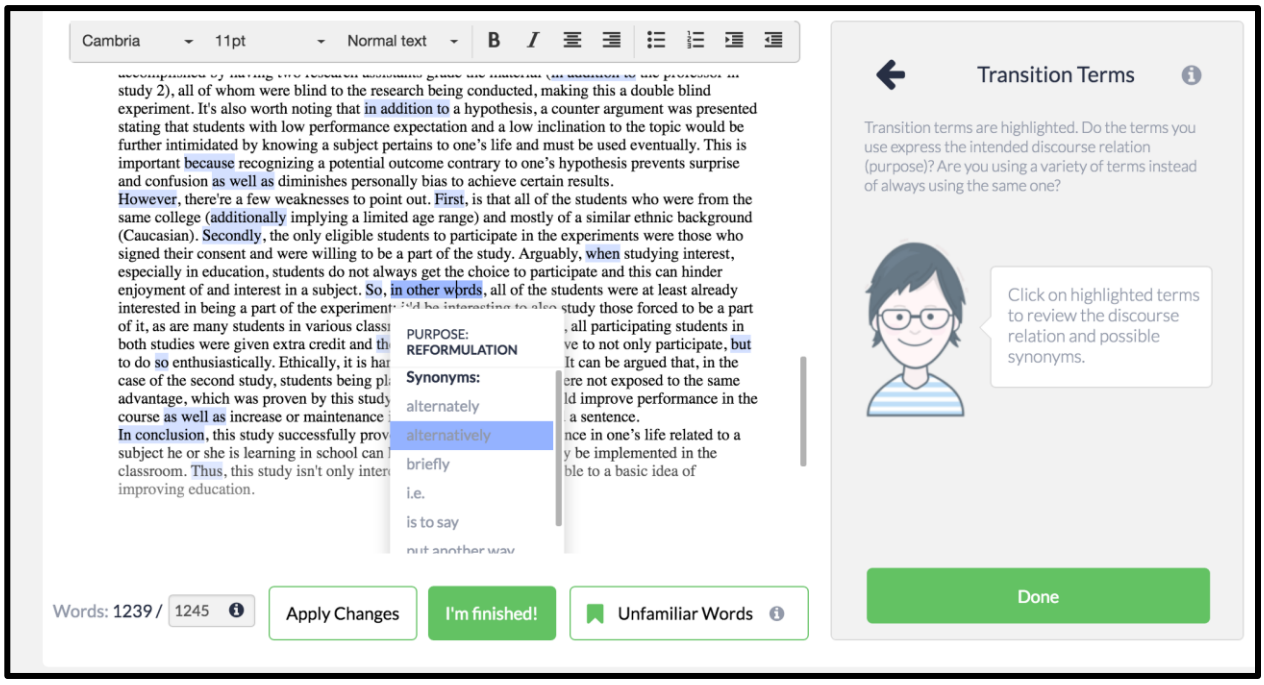

Figure 2. Screenshot to illustrate user experience viewing feedback for *Transition Terms*. Users see highlighted *Transition Terms* and can review potential synonyms to vary term use.

As illustrated in Table 1, feedback leverages ETS's NLP capabilities and lexical resources, and currently uses the Wordnik API[8] to provide synonyms and lexical resources to help users understand unfamiliar words that they may encounter while reading external sources.

Table 1

Feature Types, Subconstructs, and Related NLP Features

Feature Name	*Writing Mentor* Subconstruct	NLP-Based Feature / Resource Description
CLAIMS	Convincing	Arguing expressions from a discourse cue and argument expression lexicon that contains sets of discourse cue terms and relations (e.g., *contrast*, *parallel*, *summary*), and arguing expressions, classified by *stance* (i.e., for/against), and *hedge* and *booster* status. This is an extension of the *cluelex* from Burstein et al. (1998).
SOURCES	Convincing	Rule-based scripts detect in-text formal citations consistent with MLA, APA, and Chicago style citation formats.
TOPIC DEVELOPMENT	Well-developed	Detection of main topics and their related word sets (Beigman Klebanov et al., 2013; Burstein et al., 2016a)
FLOW OF IDEAS	Coherent	Leverages terms in a document generated from Topic Development (above) main topics and their related word sets (Beigman Klebanov et al., 2013; Burstein et al., 2016a)
TRANSITION TERMS	Coherent	Identifying the same lexicon as in the claims above.
LONG SENTENCES	Coherent	Sentences identified with a syntactic parser that contain one independent clause and >= one dependent clause.
TITLE & SECTION HEADERS	Coherent	Rule-based scripts detect titles and section headers.
PRONOUN USE	Coherent	Pronouns identified from a syntactic parser.

[8] See http://developer.wordnik.com/.

Feature Name	*Writing Mentor* Subconstruct	NLP-Based Feature / Resource Description
ERRORS IN GRAMMAR, USAGE, & MECHANICS	Well-edited	Nine automatically-detected *grammar* error feature types, 12 automatically-detected *mechanics* error feature types, and 10 automatically-detected *word usage* error feature types (Attali & Burstein, 2006)
CLAIM VERBS	Well-edited	Verbs from a discourse cue and argument expression lexicon that contains sets of discourse cue terms and relations (e.g., contrast, parallel, summary), and arguing expressions, classified by *stance* (i.e., for/against), and *hedge* and *booster* status. This is an extension of the *cluelex* from Burstein et al. (1998).
WORD CHOICE	Well-edited	Rule-based script that detects words and expressions related to a set of 13 "unnecessary" words and terms, such as *very*, *literally*, *a total of*.
CONTRACTIONS	Well-edited	*Contractions* are identified from a syntactic parser.

To support research, WM includes a brief, *optional* 3-question *entry survey* that asks users to let us know about 1) self-reported confidence about writing, 2) why they are using the app, and 3) English language status (i.e., if English was the first language they learned to speak). The app also includes an *optional exit perception survey* containing 11 items, which is adapted from the System Usability Survey (SUS) (Brooke, 1996). In addition to feedback, the app provides a report illustrating the different feedback types that the user viewed. The report can be saved as a PDF file that can be shared with others. For instance, if shared with an instructor, the report can provide the instructor with a sense of student writing support needs based on which features their students engage with.

WM captures users' *event log data* for research purposes to collect data that we can analyze to understand more about our users and their writing support needs. *Event logs* capture information such as time stamps, feature use, and document revisions. In addition, survey response data is collected and stored in the event logs and easily accessible for research.

3.0 Writing Mentor Evaluations

3.1 Amazon Mechanical Turk[9] Usability Study

Amazon Mechanical Turk (AMT) is a platform for hiring on-demand workers. Organizations post *human intelligence tasks* (HITs) to the AMT site. AMT workers respond to HITs and are selected based on eligibility for specific HIT criteria. This section describes an AMT study designed to help evaluate usability of the WM app.

3.1.1 Participants. One hundred and eight participants successfully completed essay submissions with varying participation on the other components of the HIT (described below). Results are reported based on these 108 participants. We originally posted 110 HITs: 20 which required "*masters*"[10]; 25 which required a 100% HIT acceptance rate; and 65 which required a 95% HIT acceptance rate, meaning that their HIT was successfully completed 95% of the time.

3.1.2 Methods: Human Intelligence Task (HIT). The HIT required users to complete the following tasks, and compensation for each task is in parentheses next to the task: 1) complete a 350–500 word essay response to the prompt provided ($10.00); 2) (a) per instructions provided, install WM and try out all app components with the intention of providing an evaluation of the feedback, and (b) respond to an 11-question exit perception survey built into the app ($5.00); and 3) submit open-ended feedback comments in a Google Form accessed at a link that was provided ($5.00). We indicated that the full set of tasks would take approximately two hours. AMT workers were paid once their HIT was evaluated as satisfactorily completed according to the instructions and within two weeks of completing the HIT.

3.1.3 Results. Note that of the 108 participants, only 90 included their unique AMT identification number on their essay. Findings reported from the event logs include only those 90 participants. As discussed earlier, WM event logs contain a record of the different events from a user session. In this section, we discuss the subset of event log data related to entry and exit surveys. We also provide a summary of AMT worker usability feedback provided in a Google Form outside of the WM app. (See *Human Intelligence Task* section 3.1.2 above.) These data support a qualitative analysis to allow us to see who our AMT workers are with regard to their self-efficacy and English language status, and their impressions of the tool.

3.1.3.1 Self-efficacy. The WM entry survey includes an *optional* question about self-efficacy. Figure 3 shows self-reported self-efficacy for 90 AMT study participants who responded to the survey. Most AMT study participants (98%) self-report that English was the first language they learned to speak. Figure 3 shows that the majority identify as *pretty confident* writers (64%), as compared to *not very confident* (30%) or *very confident* (6%) writers.

[9] https://www.mturk.com/
[10] "*Masters*" have completed 1,000 HITs and maintain a minimum 99% approval rating.

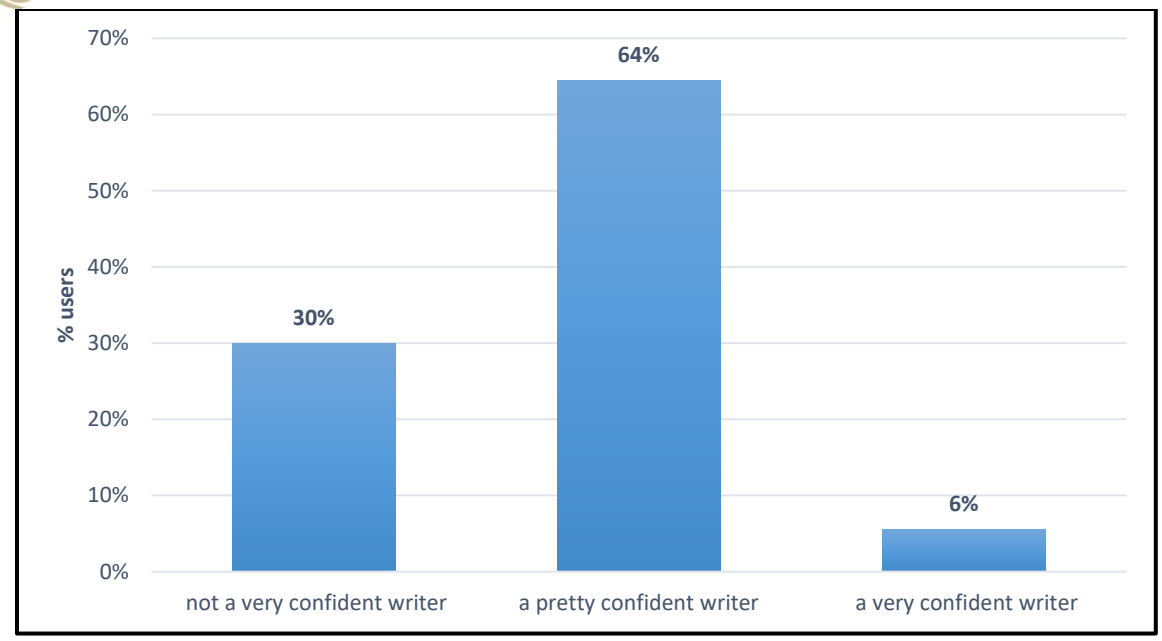

Figure 3. Self-reported *self-efficacy* from AMT workers (N=90).

3.1.3.2 Tool perception. The WM exit survey contains 11 items. The first ten item statements are adapted from the SUS survey (Brooke, 1996). The final question is open-ended.[11] Figure 4 shows the average ranking across the 86 AMT study participants who responded to the perception survey. To provide a ranking to a survey item, a user could select one star (*) to five stars (*****). Orange bars indicate a negative statement (-), where a lower ranking indicates a more positive impression (i.e., disagreement with a negative statement); green bars indicate a positive statement (+) where a higher ranking indicates a more positive impression (i.e., agreement with a positive statement).

[11] The full set of questions are as follows:

1. I think that I would like to use the Writing Mentor frequently.
2. I found the Writing Mentor to be unnecessarily complex.
3. I thought the Writing Mentor was easy to navigate.
4. I needed to learn a lot of things before I could get going with the Writing Mentor.
5. I found the various functions in the Writing Mentor were well-integrated.
6. I thought there was too much inconsistency in the Writing Mentor.
7. I would imagine that most people would learn to use an application like this one very quickly.
8. I found the Writing Mentor very cumbersome to follow.
9. I felt very confident navigating the Writing Mentor.
10. I would need to learn a lot about Writing Mentor before I would recommend it to others.
11. I wish Writing Mentor had...[…..]

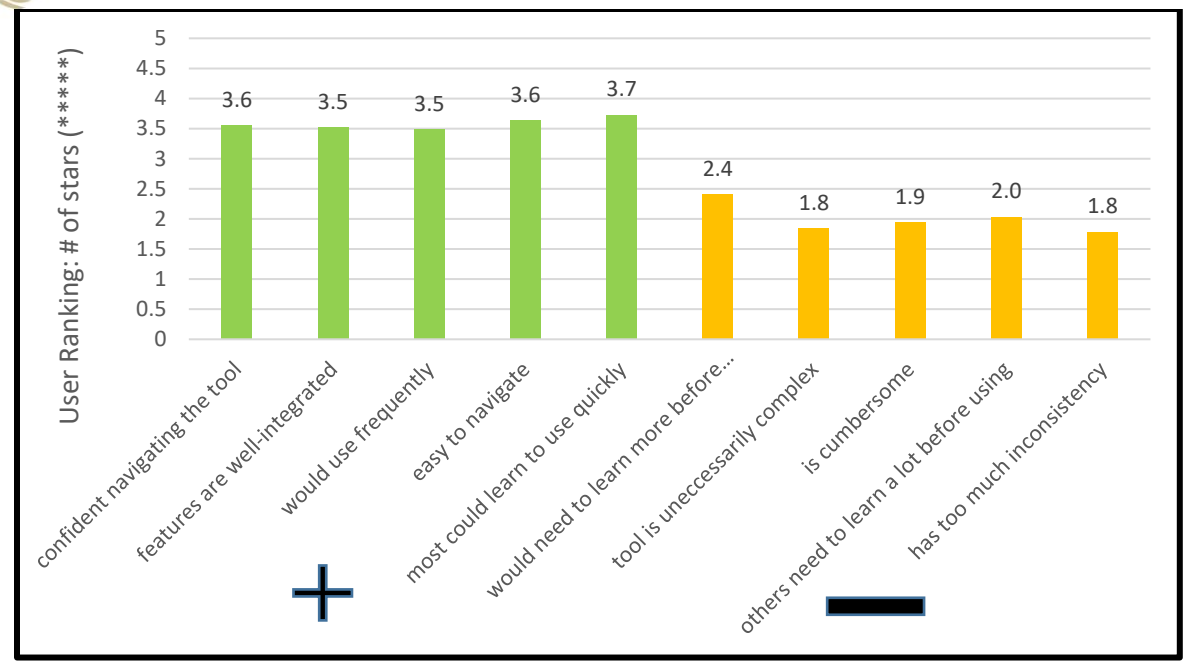

Figure 4. Average rankings for 10 SUS statements from AMT workers. Orange indicates a negative statement so we anticipate a lower ranking; green indicates a positive statement so we anticipate a higher ranking. (N=86)

Figure 4 illustrates that the direction of the rankings suggests users had a positive impression of the app. Specifically, the rankings for the orange categories tend to be lower than the mid-point (between 2.5–3.0), indicating most responses are in disagreement with the negative statements (e.g., Users disagree with the statement: *I found the Writing Mentor to be unnecessarily complex*.); averages for all negative statements are at or lower than 2.5. The average rankings for the green categories (e.g., *I think I would like to use the Writing Mentor frequently*.) are above the midpoint; all rankings for green categories are at or above 3.5.

3.1.3.3 Feedback. One of the authors used NVivo to code the open-ended feedback provided by the AMT study participants. The feedback was organized into positive and negative comments. The comments were further placed into descriptive categories that reflected what the comment was about. Table 2 illustrates the different positive and negative categories and provides example comments for each category.

Table 2

AMT Study Participant Open-ended Feedback Examples.

Category	Example comments		
P: Use cases	*I think it is helpful in understanding what the structure of a good evidence-based, persuasive essay should be.*	*I could see how this would be really useful to people who wrote frequent papers and needed help looking for errors and suggestions.*	*Overall I would love to let my 9th grader use this when she is writing her essays for school*
P: Layout	*The layout is wonderful and easy to follow.*	*I found it really easy to navigate and really intuitive to use. I would think that it is ready for launch.*	*Overall I liked the organization of the add-on.*
P: Reuse	*I feel like this is a great app already and I look forward to using it much more.*	*I really liked it and am going to keep it installed.*	*I think I would use it on a continuing basis to edit papers.*
P: Instructional value	*I think it is helpful in understanding what the structure of a good evidence-based, persuasive essay should be.*	*I found the Mentor App to be very useful. I did not find any bugs, but the tips provided were very insightful and useful.*	*I really liked that I had choices on how the Writing Mentor would work. Being able to receive feedback about the writing as a whole was very helpful.*
P: Feature compliments	*I liked that keywords were chosen from my work and put into a list. From that list I could tell that there was a lot of unity within my writing.*	*I find the synonyms feature in "Transition Terms" to be very helpful. I do use a thesaurus in writing to help keep from being so redundant with certain words but I still end up sounding that way sometimes.*	*Going through each step definitely made my argument stronger, and more cited for claims I wanted to present.*
N: Design	*Sam's text is light gray on a white background, which not only makes it difficult to see but also gives the impression that it isn't important.*	*I think a standard beginner tutorial would be extremely useful when it comes to using this application. I figured out how to use everything pretty quickly, but I could see others struggling with it at first.*	*It was extraordinarily awkward having to scroll back and forth in the pop-up window in order to be able to use the app. Would definitely consider changing the frame size to make it less cumbersome to use.*

Category	Example comments		
N: Feature enhancements	*I think this is great and would love to use it with my middle-school students. But before I could do that, it would need to be able to help them with conclusions and transitions between paragraphs.*	*It would be helpful if sentence structure examples could be added.*	*I did put that I would like to see a way to correct long sentences,*

Note. Positive feedback categories are preceded by "P:" and negative feedback categories are preceded by "N:".

3.2 Users-in-the-wild

WM was first released at the end of November 2017. Eight months later, by July 30, 2018, about 2,693 *users-in-the-wild* had been recorded in the WM event logs.[12,13] Note that we do not intentionally collect personally identifiable information, so we do not know who users are and the purpose for tool use. Potential tool use scenarios may include the following: Users may have come to the tool in good faith to work on their writing. However, some users, such as writing instructors, may be using the tool just to test its instructional utility. Writing instructors who have been interested in using the tool for students have informed us of their use in this capacity. Others may be using the tool for the purpose of identifying tool idiosyncrasies, which has been a common scenario over the years with regard to evaluations of AWE systems (Winerip, 2012).

We now turn to a snapshot of information derived from the event logs from users-in-the-wild approximately eight months after the release of the app in late November 2017. Note that the analyses provided below are qualitative in nature; these analyses help us to understand emerging trends that can inform future system design and the development of personalized learning feedback.

3.2.1 Writing Mentor app user profiles. As of July 31, 2018, 90% (2,430/2,693) of users had responded to the entry survey. Figure 5 shows the percentage of users who self-reported at the different levels of self-efficacy with regard to writing. Users identified as follows: 40% were *not very confident*, 51% were *pretty confident*, and 8% were *very confident* writers. Eighty-one percent reported that English was the first language that they learned to speak. These findings reflect similar trends in the AMT study participants (See Figure 3).

[12] An additional approximately 113 users had participated in an Amazon Mechanical Turk (AMT) study conducted as a way to collect more formal evaluation of the tool. These users are excluded from this analysis, and we report only on the non-AMT users-in-the-wild.

[13] For this analysis, we excluded AMT worker logs and logs that the project team generated for testing purposes.

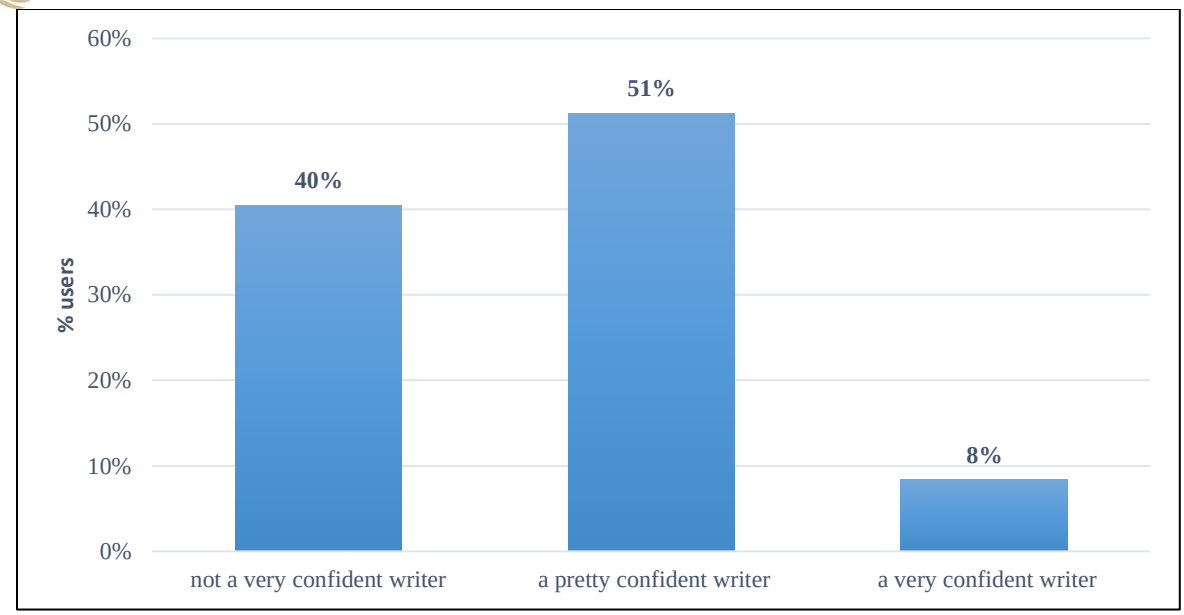

Figure 5. Self-reported writing self-efficacy for users-in-the-wild. Percentage of users who reported being *not very confident, pretty confident,* or *very confident writers* (N= 2,430).

3.2.2 User perceptions. As mentioned earlier, the exit perception survey is adapted from the SUS survey (Brooke, 1996) and contains 11 questions. (See footnote 11.) Figure 6 illustrates that perceptions from these users show similar trends to those from AMT participants (See Figure 4). Orange bars indicate rankings for a *negative statement*, where a lower ranking indicates a more positive impression (i.e., disagreement with a negative statement); green bars indicate a *positive statement*, where a higher ranking indicates a more positive impression (i.e., agreement with a positive statement). Consistent with the AMT participant survey findings, user-in-the-wild average rankings for negative statements are below 3.0, and all average rankings for positive statements are at 3.5 or higher.

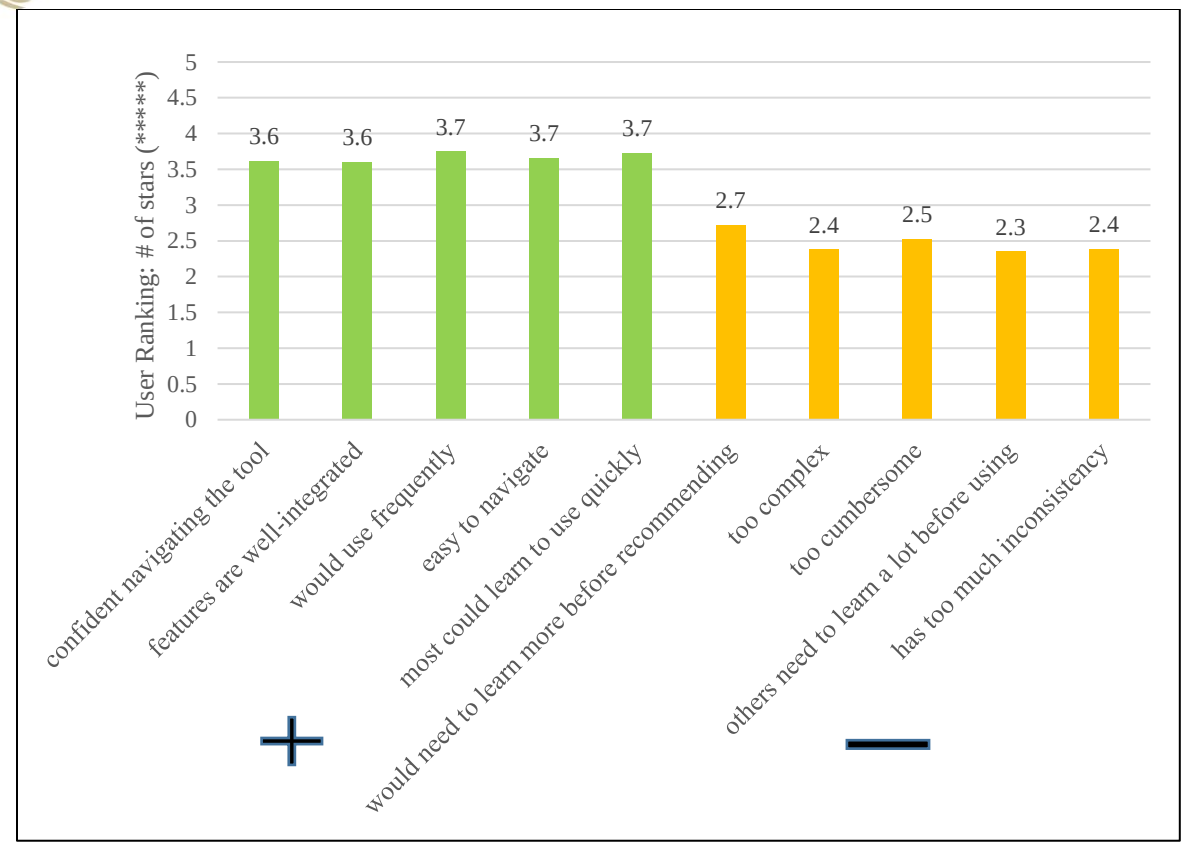

Figure 6. Average rankings from users-in-the-wild for 10 SUS statements. Orange indicates a *negative statement*, so we anticipate a lower ranking; green indicates a *positive statement*, so we anticipate a higher ranking. (N=355)

Figure 7 shows the number of users who return to use Writing Mentor for another document. Repeated use suggests that a user perceives a benefit from tool use. Figure 7 shows that 28% of users (750/2,693) are returning and submitting multiple documents. This finding is consistent with the overall modestly positive ranking in Figure 6 for the statement *I think that I would use the Writing Mentor frequently*.

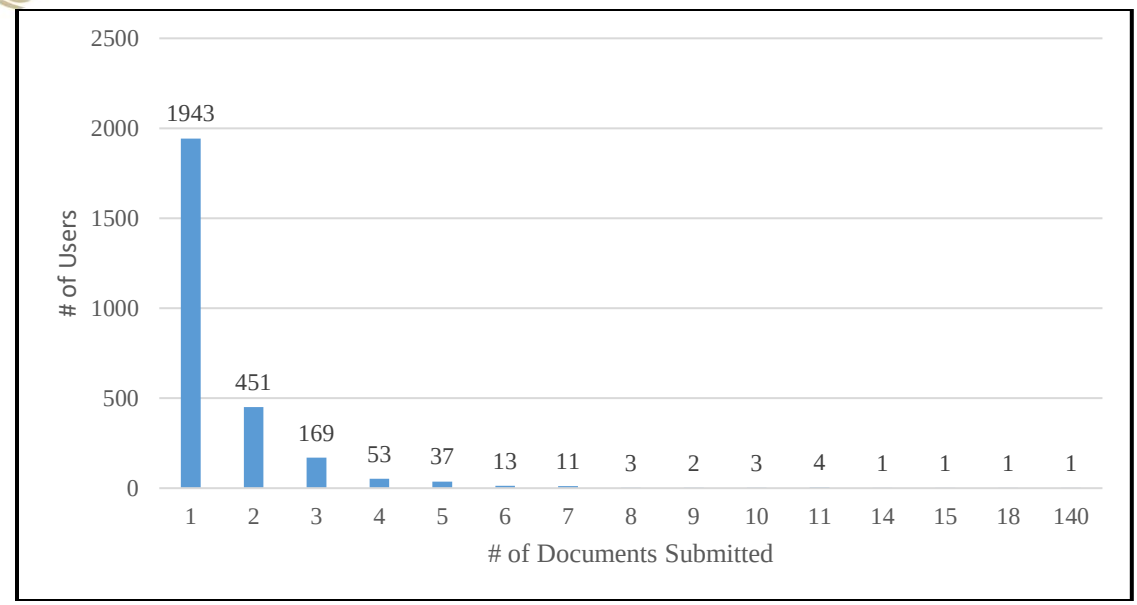

Figure 7. Descriptive statistics based on 2,693 users showing for the number of users how many documents were submitted.

3.2.3 Self-efficacy, English language proficiency, & tool engagement. In this section, we provide a snapshot of relationships between self-efficacy and tool engagement. We used the event logs to extract responses (N=2,430) to the entry survey related to self-efficacy and English language status, and tool use preferences. As discussed earlier, the event logs provide information about how users interact with the Writing Mentor app. Figure 8 illustrates *preferred feature type given self-efficacy*.

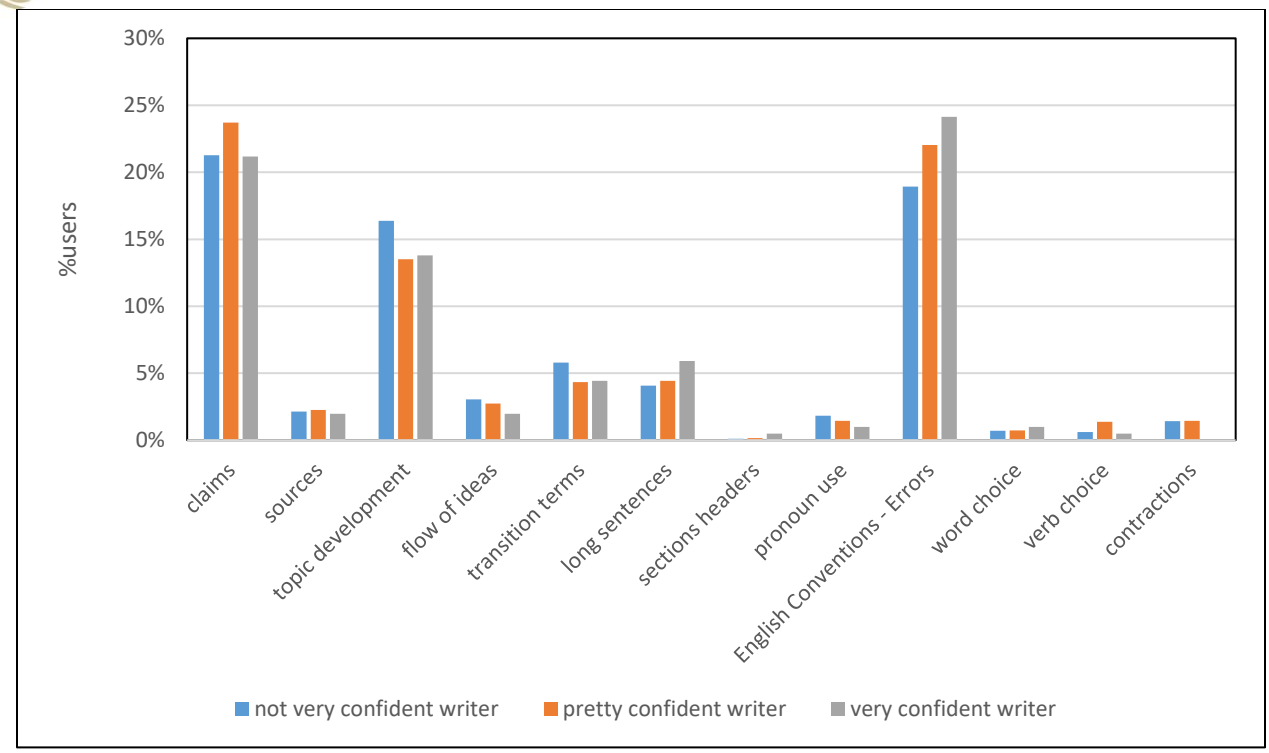

Figure 8. Percentage of preferred feature types given self-efficacy (N=2,430).

To generate the information in Figure 8, for each user, we determined the single feature which a user preferred by keeping track of how much time a user spent with different features. The feature for which they spent the most time was considered the *preferred feature*. Figure 8 shows relative feature preferences, given self-efficacy. For each self-efficacy category, we computed a sum across all users from that category and then averaged across the total number of users in a self-efficacy category.

Figure 8 illustrates that across self-efficacy categories, the most preferred features across all groups appear to be the *grammar errors* feature, followed by the *claims* feature, and then the *topic development* feature. The chart suggests that *not very confident* and *pretty confident* writers' most preferred feature is *claims*, while *very confident* writers' most preferred feature is *grammar errors*. Overall, users appear to be experimenting with the majority of features.

Figure 9 shows users' average rankings for the perception survey statement *I think I would use Writing Mentor frequently*, given their preferred feature. Overall, Figure 9 suggests that users who are returning to use the app and revisiting a feature type are also ranking WM positively (i.e., ranking is greater than the midpoint of 3) across preferred feature types.

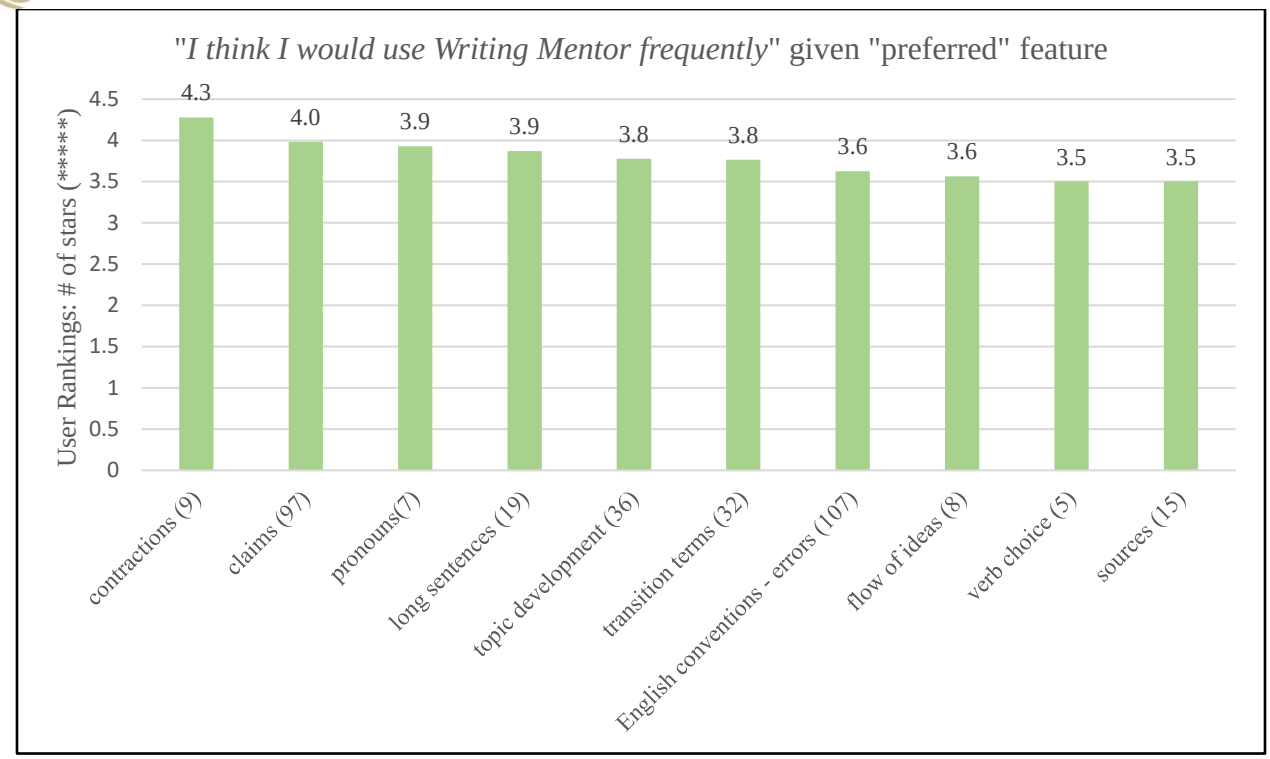

Figure 9. Average rankings to perception survey statement *I think I would use Writing Mentor frequently*, given their preferred (most used) feature. Features were included only if they were associated with at least 5 users. N=337. The 337 users a) responded to the exit survey, and b) the event logs captured evidence of a preferred feature.

Figure 10 shows the *average number of minutes spent on preferred features* for 2,051 users who returned to the system for multiple sessions. One session indicates that a user worked on a document only once. Multiple session use indicates that a user returned to work on a document multiple times. For users who worked on a document for multiple sessions, we can identify the features where they spent the most time. There does not appear to be a direct relationship between the features where users spent the most time (i.e., *claims*, *topic development*, and *English conventions – errors*) and the actual number of minutes spent with those features. While a top preferred feature *claims* (12.7 minutes) is also one for which many users spent relatively more time, for the other top two preferred features, *grammar errors* (9.4 minutes) and *topic development* (3.8 minutes), average actual time spent with those features was lower. While many users are spending relatively more time on these features, on average, users may spend less actual time on those features because they require less time (e.g., quick correction of a verb error). While more research would be needed to better understand this relationship, the good news is that users appear to be spending, on average, more than three minutes on the majority of preferred feature types. This information on use of preferred features suggests that users are engaging with most feedback feature types.

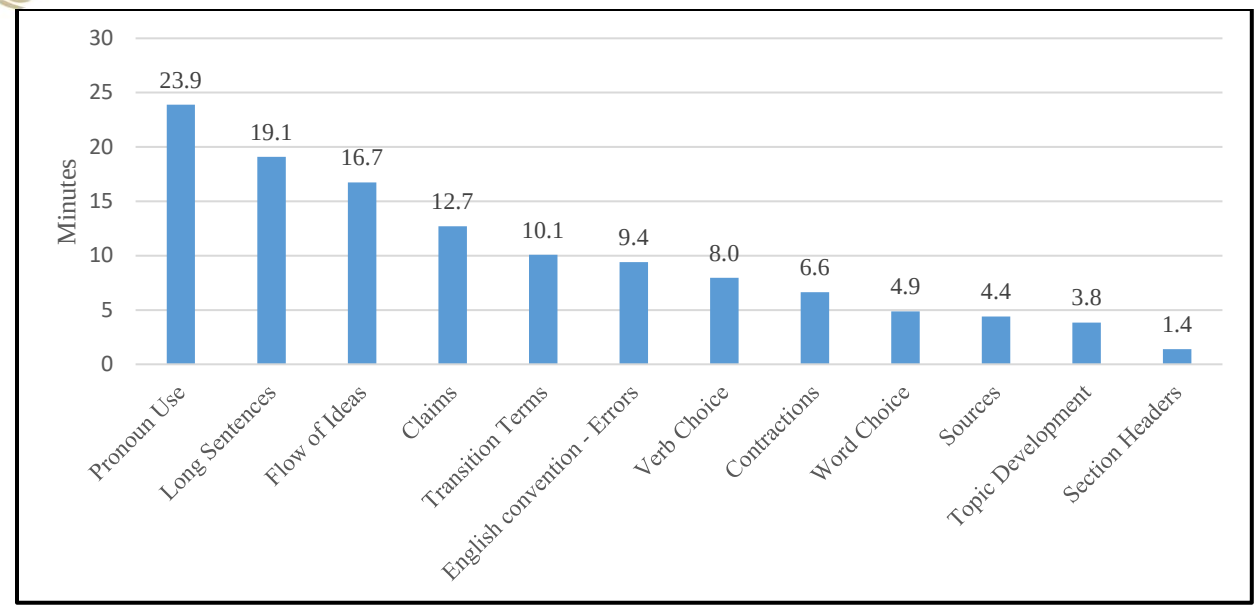

Figure 10. Average time (in minutes) users spent with their *preferred* (most visited) feature. The figure shows the average time spent across the 84% of users (2,051/2,430) who returned to the system, allowing us to access a preferred feature through the event logs.

3.2.4 Evaluation of changes in user submissions. The section above discussed event log data that shows which features users visit and how much time they spend with these features. In this section, we focus on the kinds of revisions users are making to their texts while they are working in WM. We provide evaluations of changes observed in user writing across multiple submissions for a single text.

3.2.4.1 Data. We used a subset of 1,951 texts across our 2,693 users reported in this paper. For these 1,951 texts, the user had submitted multiple versions of a single document.

3.2.4.2 Method. We evaluated changes in users' writing between the *first* and *last*[14] submission for a single document. For all features from WM, we generated feature counts for the *first* and *last* submission of a text. We then subtracted the counts from the *first* submission from the counts from the *last* submission to report the *difference*, where *difference = last count – first count*. Note that positive values indicate an increase in a feature (i.e., a higher feature value in the last text [revision] and a lower feature value in the first text [draft]).

3.2.4.3 Results. Table 3 illustrates that approximately 50% (967/1951) of the documents evaluated had non-zero word count differences between the first and last submissions; this indicates that text was revised across WM *sessions* for the document—specifically the first and

[14] The *last* submission represents the last one submitted at the time we did the data extract. It is possible that the user would continue to work on that document, and at a later point in time, additional changes may be made.

last sessions working on the document.[15] Thirty-six percent (710/1951) showed increased word count, suggesting increased productivity; the remaining 13% (257/1951) showed reduced word count, suggesting text "clean-up." About 50% percent (N=984) of texts contained no word count differences. Table 3 shows the *mode, mean* and *standard deviation* for the word count *difference* for the approximately 50% (N=967) of submissions that contained a non-zero word count difference between the first and last submission of the same text. While the *mode* was +/-1, indicating only small change for most users, overall, there is evidence of revision when using the tool.

Table 3

Mean and Standard Deviation (SD) for Word Count Differences.

WC class	Mode	Mean	SD
WC pos (N=710)	1	181.4	558.9
WC neg (N=257)	-1	-491.4	391.2

Note. WC pos (positive) shows the values for texts where word count increased; *WC neg* (negative) show the values for text where word counts decreased.

Figure 11 suggests that feature changes are occurring between the first and last submissions of a text. The figure illustrates the *mean feature difference* value for a subset of features reported as either *most preferred* (Figure 7) or for which users *spent the most time* (Figure 9). The data suggests that larger differences in word count between the first and last submission were indicative of changes associated with WM feature types. Therefore, for each feature, we show the mean feature difference in relation to relative changes in word count. Figure 11 illustrates the mean feature differences based on sets of essays in three conditions, where the user changed: 1) >= 1 word, 2) >= 5 words, and 3) >= 10 words. Note that the change could be in either direction (i.e., the user added or removed words). Generally speaking, the larger the *threshold value* (i.e., >=1, >=5, or >=10), the larger the mean feature difference. Though a deeper analysis will need to be conducted to explore the mapping between feedback and actual revision, this finding suggests that users may be revising their texts relative to the WM feedback.

[15] Note that a *session* is a single "visit" to the app. Each time a user returns to work on ("visit") the app to work on the same document is considered a session. Multiple sessions suggest a user is continuing to work on (revise) a document.

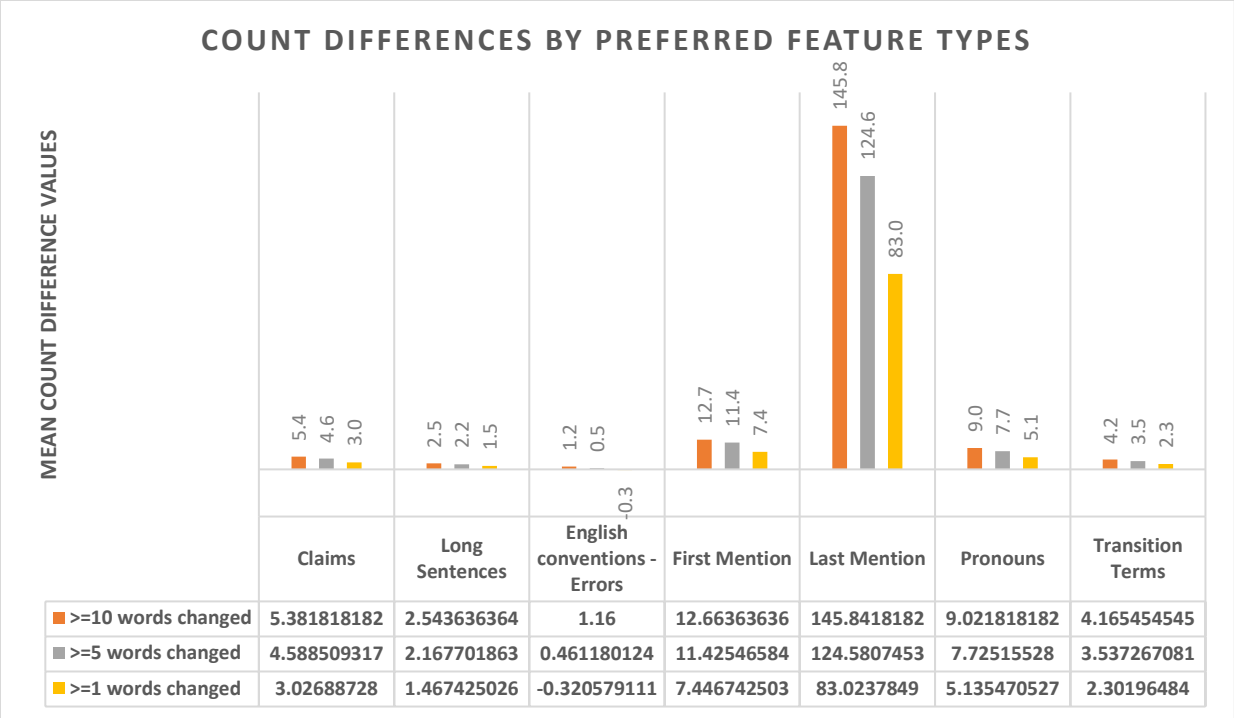

Figure 11. Difference means illustrate the means for differences in feature counts computed between the first and last submission for a single text where >= 1 word was changed, >=5 words were changed, and >=10 words were changed. *First Mention* = position of the word of the first mention of a main topic keyword; *Last Mention* = position of the word of the last mention of a main topic keyword.

We see negative values for *English Conventions – Errors* with *>=1 word changed*, indicating a reduction of error types; however, as the user introduces additional changes, there is an increase in errors. This finding might be expected since as the user writes more, they may introduce more errors. The reader should be reminded that the "last" submission is not necessarily the user's final submission, and they may return to continue revising at a later time. While we need to conduct qualitative analysis to better understand this, it is also possible that as the user writes more, they are introducing proper nouns (e.g., surnames) that the system might not recognize and falsely identify as spelling errors. These do not need to be corrected, so these remain in the longer text. For *claims*, *long sentences*, *pronouns*, and *transition terms*, we see an increased positive value with increased word count differences. This indicates that users are *adding* claims, long sentences, pronouns, and transition terms as they revise. A more targeted analysis will need to be conducted to understand the relevance of these changes, e.g., if changes specific to a feature were made while the user was working in a feature category of the tool or if the changes may be a more general result of writing more. The *first mention* and *last mention* features are related to the *topic development* and *flow of ideas* features. (See Figure 12, which illustrates the *topic development* feedback feature.) *First mention* tells us the initial position of a main topic word (e.g., 5th word, 8th word, etc.) in the text that is identified by the system. *Last*

mention tells us the final position of a main topic word in the text that is identified by the system. Note that the main topic is a single word that is associated with a set of related words in which the mentions of that word are highlighted. Referring back to Figure 10, we see that the mean difference values for *first mention* and *last mention* suggest that there are shifts in text position for main topic words. Further, we observe that the longer the text, the greater the shift. The *first mention* shift suggests that the initial mention of a topic word is slightly later in the text—about 7–13 words. This suggests the word is moved to a later point in a sentence or possibly to the following sentence; the *last mention* is also shifted, but the shift is much greater—about 83–145 words later. This suggests the *last mention* appears several sentences later in the text. This last observation also suggests that the writer may be revising based on system feedback related to the *flow of ideas* category. The *flow of ideas* section of WM leverages the main topic (and keywords) from the *topic development* section of the system. (See Figure 13). The system advises the writer to review the distribution of the main topic words to ensure that the topic is discussed throughout the text (i.e., mentioned early on and later in the text).

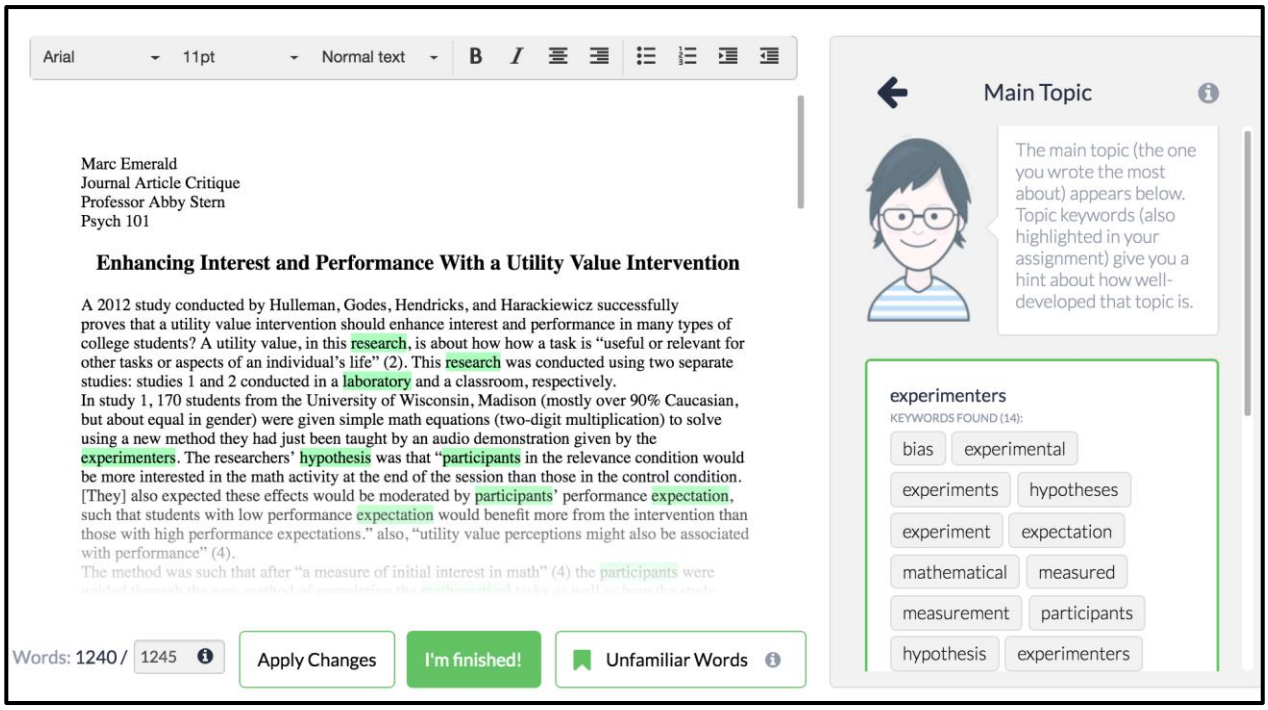

Figure 12. Screenshot of Writing Mentor's main topic and keyword identification.

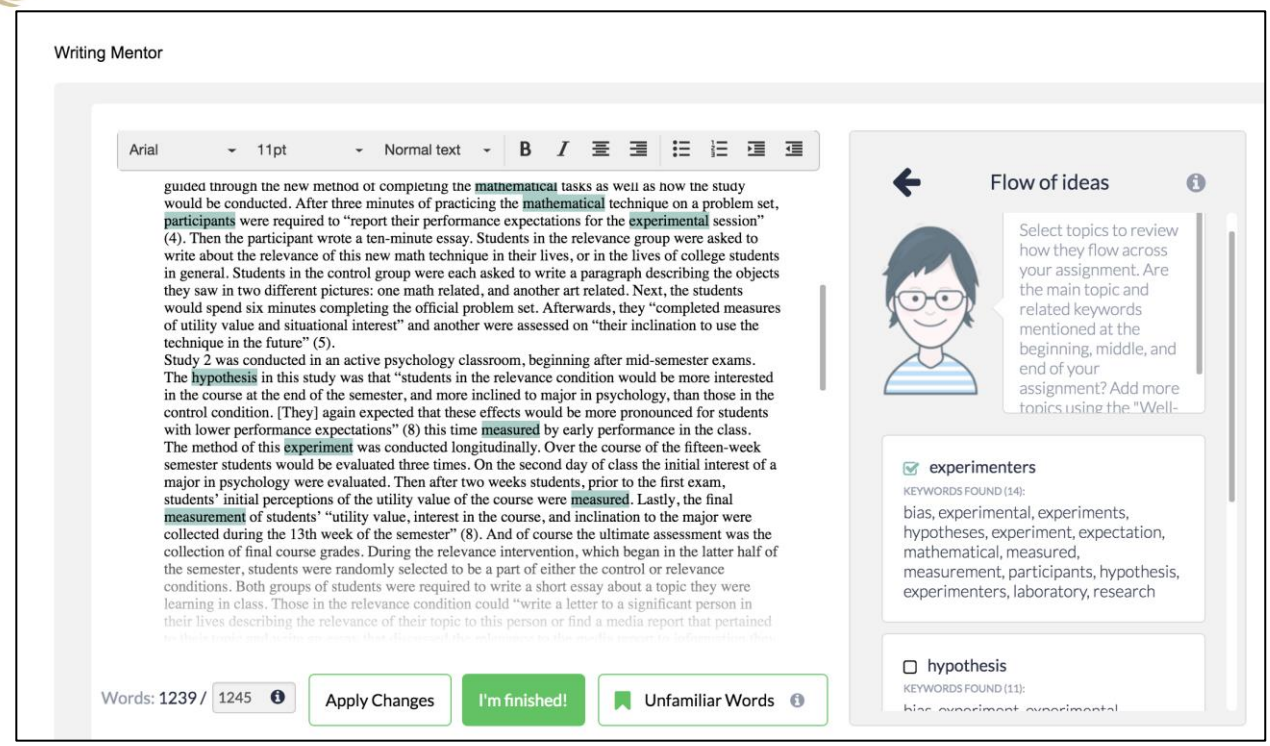

Figure 13. Screenshot of Writing Mentor's *flow of ideas* feedback where the user can review how main topic keywords are distributed in the text.

Findings from this analysis suggest that writers are making feedback-related changes to texts as they are revising. A finer-grained, qualitative analysis will need to be conducted to study the relationship between the changes and the system feedback.

3.2.5 User suggestions. The exit perception survey contained an open-ended question: *I wish that Writing Mentor had ….* Users could fill in a dialog box with suggestions. Example suggestions quoted from users are shown in Table 4 below. Suggestions are categorized by type: interface, usability, feature enhancements, and new features.

Table 4

User Suggestions in Response to the Exit Survey Question: I wish that Writing Mentor had ….

Interface	Usability	Feature Enhancements	New Features
a better proportioned interface. The add-on scales a bit oddly on my monitor. I had to use sliders to position the interface properly.	*I loved the writing mentor I though it worked well I just wish I had more experience in writing to actually use the advice better.*	*a better understanding of run ons. There were lines in my writing that I dont [sic] believe to be run ons.*	*more active walkthroughs to help make changes instead of just highlighting problems.*
A clear word count and easy high contrast text.	*Easier navigation - full screen instead of a smaller window/frame, and funtion [sic]/menu tabs across the top, as in a word processing program rather than on the right side.*	*recommended words to use for the sentences as a correction*	*more interactive comments.*
A little larger screen area		*More suggestions for pronouns*	*could write my paper for me*
A slightly easier-to-follow flow. The "Done" button generally brings you back to the next session, except under "Review Topic Development."		*correction of citations*	*Shown me examples*
simple animations for the avatar. the suggestions / tips are straightforward text. it would help if the avatar has simple animations.		*spelling helping*	*I wish it had personalized comments and feedback.*
		more humor	*A way to check for a conclusion*

4.0 Discussion

As discussed earlier, NLP solutions for writing instruction and assessment have a relatively long history and a substantial body of work that is now commercially available. Writing Mentor is one example of how NLP technology can create literacy solutions that are globally accessible 24/7. As NLP researchers continue to consider solutions to address literacy struggles, they should carefully consider the needs of different populations of learners and scalability of these solutions to ensure that they are accessible to a broad range of learners.

In terms of future development of the WM app, three sources of research may prove especially useful: attention to personalized learning (Pane et al., 2017); the Institute of Education Sciences (IES) practice guidelines for effective writing (Graham et al., 2016); and identification of variables associated with low-proficiency writers (Perin & Lauterbach, 2018).

4.1 Personalized Learning

The concept of Personalized Learning (PL) has a rich history in US and UK education. Combined with advances in computing in the 1980s, PL aimed to provide programmed instruction according to specified domain models of constructs in language acquisition and mathematical proficiency. Current examples in writing studies research include personalized grading contracts (Inoue, 2014). Pane et al. (2017) examined PL in 40 schools dedicated to PL-based instruction. The report concludes that there is evidence that implementation of PL practices may be related to more positive effects on achievement. Among the challenges, barriers to PL implementation included "poor integration of data systems, tensions between competency-based practices and meeting grade-level standards, and the time needed to develop personalized lessons" (p. 40). Among the findings relevant to the present study, one stands out: "Finding high-quality standalone technology-based materials was a challenge" (p. 12).

In light of emphasis on PL, WM provides an integrated system by which self-assessment and performance information relevant to writing construct may be obtained. As this article illustrates, cognitive and intrapersonal domain knowledge may be obtained from the four feedback feature categories and the user entry survey, respectively. As a form of one-stop shopping, WM holds the potential to play a valuable role in PL-based curricula.

4.2 IES Effective Writing Instruction

In the US, the Institute of Education Sciences (IES) publishes practice guides to provide educators with evidence-based guidelines to address pedagogical challenges. Operationally, IES develops and publishes practice guides in conjunction with expert panels according to a levels of evidence model. *Teaching Secondary Students to Write Effectively* (Graham et al., 2016) is one such report.

Table 5 demonstrates the ways that WM is aligned with the practice guide offered by Graham and his colleagues. Three observations may be made regarding the alignment. First, WM offers an explicit instructional model in which writing processes are modeled according to a principled fashion. As such, students are explicitly engaged in targeted writing strategies in which a model-practice-reflect orientation is provided (Graham et al., 2017, p. 2). WM provides explicit strategies for planning and goal setting, drafting, evaluating, revising, and editing—key features of the Model-Practice-Reflect instructional cycle. Put simply, students are not left on their own to interpret a score based on a generic rubric. Rather, a process model is enacted with each use of WM. Second, WM advances a language arts model in which writing and reading are seen as complementary activities. Users are iteratively invited to return to the text at hand, re-read what they have written, and expand their capabilities through targeted feedback. Third, WM

enacts formative assessment—commonly known as assessment for the sake of learning. As Bennett (2011) has observed in his review of formative assessment, in order to realize demonstrable benefit from formative assessment, new development should focus on conceptualizing defined approaches built around process and methodology grounded within specific content domains. WM achieves such goals in its attention to modeling the writing process and its capability of providing instruction according to a defined model of the writing construct.

Table 5

Alignment of WM with IES Effective Writing Strategies

IES standard	Writing Mentor feedback	Alignment
Recommendation 1. Explicitly teach appropriate writing strategies using a Model-Practice-Reflect instructional cycle.	Convincing, well-developed; coherent, well-edited	Adopts a targeted, cognitive approach to composing and revision Writing strategy is modeled, students practice using WM, and students are invited to reflect at each stage of the process.
Recommendation 2. Integrate writing and reading to emphasize key writing features.	Feedback is provided in terms of definitions, writing strategies, and invitations to explore more about these concepts.	Language arts approach fosters reading and writing connections.
Recommendation 3. Use assessments of student writing to inform instruction and feedback.	Rather than scores, a formative, heuristic approach is used for each writing strategy.	A formative assessment cycle is followed based on machine analysis, targeted feedback, and identified revision goals.

4.3 Low-skilled College Students

In their study of postsecondary students demonstrating poor writing skills, Perin and Lauterbach (2018) used persuasive essays and written summaries from low-skilled, college developmental education students to identify the diverse ways that developmental writers address these two genres. This finding does not align with current educational practice in which developmental education in writing groups students into a single classroom based on standardized placement scores that are not designed to be diagnostic of specific literacy needs. As Perin and Lauterbach observe, "differing pattern of skills seems to require more individualized approaches" (p. 70). As the authors conclude, here is a possibility for systems such as WM and other automated systems to support basic writing students at the college level.

> Given the heterogeneity of the writing skills of students who will inevitably be grouped for instruction under current practices, perhaps an automated scoring system could be leveraged by instructors to identify specific writing problems.

> Focusing only on difficulties identified by the automated scoring engine rather than a wider range of skills, some of which students may already have mastered, may lead to more efficient use of class time. Such diagnostic use of automated scores could potentially free up instructors to focus more on content and meaning. However, further research would be needed to test the usefulness of this idea and identify the scoring system and indices that would best inform such differentiated instruction. (p. 70)

While automated systems have been used to provide students with additional help in order to improve their writing without remediation (Klobucar et al., 2013), WM offers new ways to provide individualized help to students based on feedback related to the writing being convincing, well-developed, coherent, and well-edited. Additionally, student self-assessment of efficacy and proficiency can be combined with such feedback in order to personalize instruction. Our present work suggests that WM holds the potential to meet the challenges of personalized learning in terms of improving student writing.

Author Biographies

Jill Burstein is a Director of Research for the Natural Language Processing Group in the Research Division at Educational Testing Service in Princeton, New Jersey.

Norbert Elliot is a Research Professor at the University of South Florida and Managing Editor of *Journal of Writing Analytic*s.

Beata Beigman Klebanov is a Senior Research Scientist at Educational Testing Service.

Nitin Madnani is a Senior Research Scientist for the Natural Language Processing Group in the Research Division at Educational Testing Service in Princeton, New Jersey.

Diane Napolitano is a Research Engineer at Educational Testing Service in Princeton, New Jersey.

Maxwell Schwartz is an Assistant Research Engineer at Educational Testing Service in Princeton, New Jersey.

Patrick Houghton is a Research Associate at Educational Testing Service in Princeton, New Jersey.

Hillary Molloy is a Senior Research Assistant at Educational Testing Service in Princeton, New Jersey.

Acknowledgments

Research presented in this paper was supported by ETS. Any opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily reflect the views of ETS. An earlier version of this paper was presented at the 2018 Annual Meeting of the National Council for Educational Measurement (NCME) at the Coordinated Symposium organized by the first two

authors, entitled *What Writing Analytics Can Tell Us About Broader Success Outcomes*. We thank Jiangang Hao for contributions to event log design. We would also like to acknowledge Dolores Perin and Mya Poe for their expert advice related to Writing Mentor research. We thank our colleagues at 10clouds.com for outstanding front-end web development work.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2.0. *Journal of Technology, Learning, and Assessment, 4*(3).

Beigman Klebanov, B., & Flor, M. (2013). Word association profiles and their use for automated scoring of essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 1148–1158.

Beigman Klebanov, B., Madnani, N., Burstein, J., & Somasundaran, S. (2014). Content importance models for scoring writing from sources. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, Baltimore, MD.

Beigman Klebanov, B., Stab, C., Burstein, J., Song, Y., Gyawali, B., & Gurevych, I. (2016). Argumentation: Content, structure, and relationship with essay quality, In *Proceedings of the 3rd Workshop on Argument Mining, ACL 2016*, Berlin, Germany.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*, 5–25.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 4–7.

Burstein, J., Beigman Klebanov, B., Elliot, N., & Molloy, H. (2016a). A left turn: Automated feedback & activity generation for student writers. *Proceedings of the 3rd Language Teaching, Language & Technology Workshop*, co-located with Interspeech, San Francisco, CA.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online service. *AI Magazine*, *25*(3), 27–36.

Burstein, J., Elliot, N., & Molloy, H. (2016b). Informing automated writing evaluation using the lens of genre: Two studies. *CALICO Journal 33*(1), 117–141.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998, August). Automated scoring using a hybrid feature identification technique. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1* (pp. 206-210). Association for Computational Linguistics.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Advances in Natural Language Processing, 18*(1), 32–39.

Burstein, J., McCaffrey, D., Beigman Klebanov, B., & Ling, G. (2017). Exploring relationships between writing and broader outcomes with automated writing evaluation. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, EMNLP 2017, Copenhagen, Denmark.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In M. D. Shermis, & J. Burstein (Eds.), *Handbook for automated essay scoring* (pp. 55–67). New York, NY: Routledge,.

CCSSO. (2010). Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects. Appendix A: Research supporting key elements of the Standards. Washington, DC.

Cohen, K. B., & Demner-Fushman, D. (2014). *Biomedical natural language processing* (Vol. 11). John Benjamins Publishing Company.

Coleman, R., & Goldenberg, C. (2012). The Common Core challenge: English language learners. *Principal Leadership*, 46-51.

Collins-Thompson, K and Callan, J. (2004). *A language modeling approach to predicting reading difficulty*. In *Proceedings of the HLT/NAACL*.

Complete College America. (2012). *Remediation: Higher education's bridge to nowhere*. Retrieved from http://completecollege.org/docs/CCA-Remediation-final.pdf

EU High Level Group of Experts on Literacy. (2012). *Final Report*. Retrieved from http://ec.europa.eu/dgs/education_culture/repository/education/policy/school/doc/literacy-report_en.pdf

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K (2013). Implementation and applications of the Intelligent Essay Assessor. In M. Shermis & J. Burstein, (Eds.), *Handbook of automated essay evaluation* (pp. 68–88). New York: Routledge.

Graesser, A.C., McNamara, D.S., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*, 223–234.

Graham, S., Bruch, J., Fitzgerald, J., Friedrich, L., Furgeson, J., Greene, K., Kim, J., Lyskawa, J., Olson, C.B., & Smither Wulsin, C. (2016). *Teaching secondary students to write effectively* (NCEE 2017-4002). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Retrieved from the NCEE website: http://whatworks.ed.gov.

Heilman, M., & Smith, N.A. (2010). Good question! Statistical ranking for question generation. In *Proceedings of NAACL*.

Inoue, A. B (2014). Theorizing failure in US writing assessments. *Research in the Teaching of English*, *48*, 330–352.

Klobucar, A., Elliot, N., Deess, P., Rudniy, O., & Joshi, K. (2013). Automated scoring in context: Rapid assessment for placed students. *Assessing Writing*, *18*, 62–84

Madnani, N., Burstein, J., Sabatini, J., Biggers, K., & Andreyev, S. (2016). Language Muse: Automated linguistic activity generation for English language learners. *Proceedings of ACL-2016 System Demonstratio*ns, 79-84.

Madnani, N., Burstein, J., Elliot, N., Klebanov, B. B., Napolitano, D., Andreyev, S., & Schwartz, M. (2018). Writing Mentor: Self-Regulated Writing Feedback for Struggling Writers. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (pp. 113-117).

Musu-Gillette, L., de Brey, C., McFarland, J., Hussar, W., Sonnenberg, W., & Wilkinson-Flicker, S. (2017). *Status and trends in the education of racial and ethnic groups 2017* (NCES 2017-051). U.S. Department of Education, National Center for Education Statistics. Washington, DC. Retrieved from http://nces.ed.gov/pubsearch.

NCES (2016). *Remedial coursetaking at U.S. public 2- and 4-year institutions: scope, experience, and outcomes*. (NCES 2016-405). Retrieved from https://nces.ed.gov/pubs2016/2016405.pdf

Pane, J. F., Steiner, E. D., Baird, M. D., Hamilton, L. S. & Pane, J. D. (2017). *Informing progress: Insights on personalized learning implementation and effects*. RAND. Retrieved from https://www.rand.org/pubs/research_reports/RR2042.html

Perin, D., & Lauterbach, M. (2018). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education*, *28*, 56–78.

PISA (2016). *PISA 2015 results in focus*. Retrieved from https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Roscoe, R., Varner, L., Weston, J., Crossley, S., & McNamara, D. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition*, *34*, 39–59.

Sabatini, J., Bruce, K., Steinberg, J., & Weeks, J. (2015). *SARA reading components tests, RISE forms: Technical adequacy and test design, 2nd edition*. (ETS-RR-15-32). Princeton, NJ: Educational Testing Service.

Shermis, M., Burstein, J., Elliot, N., Miel, S., & Foltz, P. (2015). Automated writing evaluation: An expanding body of knowledge. In C. A. McArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research, 2nd ed.* (pp. 395–409). New York, NY: Guilford.

Somasundaran, S., Burstein, J., & Chodorow, M. (2014). *Lexical chaining for measuring discourse coherence quality in test-taker essays*. COLING 2014, Dublin, Ireland.

Winerip, M. (2012, April). Facing a robo-grader? Just keep obfuscating mellifluously. *New York Times*, pp A11. Retrieved from http://www.nytimes.com/2012/04/23/education/robo-readers-used-to-grade-test-essays.html