A preliminary look at the impact of spell checker use during an L2 English writing assessment

Ikkyu Choi & Yeonsuk Cho
Educational Testing Service

Spell checkers are popular among writers but are seldom used or studied in second language (L2) writing assessment contexts. Little is known about how L2 test takers use spell checkers or how their responses are impacted by spell checker use. Making a first step towards addressing this knowledge gap was the goal of this study. We aimed to gain a preliminary understanding of L2 test takers' spell checker use and its impact with a total of 61 adult English learners who responded to two computer-delivered English writing tasks. Half were randomly selected to be provided with a built-in spell checker. The resulting responses and keystroke logs from the test takers were analyzed to examine their spell checker use and its impact on the product and process of their writing. The test takers who chose to use the spell checker were highly comparable to those who did not use it in terms of their general English proficiency. A series of regression analyses indicated that the spell checker users, on average, wrote fewer words and made fewer spelling errors than those who did not use it. The spell checker users also tended to write at a slower pace for one of the two tasks.

Key words: L2 writing assessment, spell checker, spelling error, writing process

Introduction

Spell checkers need little introduction. They have been one of the most popular and frequently used tools for writers in the digital environment (MacArthur, 1999) and are provided as a built-in tool in many word processing software packages, such as Microsoft Word and Google Docs. They have also attracted interest as a potentially useful tool for second language (L2) writing pedagogy, especially for learners at the beginning level (e.g., Green & Youngs, 2001; Ndiaye & Vandevemter Faltin, 2003; Warschauer, 2010). However, spell checkers are seldom used in L2 writing assessment contexts, as spelling accuracy, unless it interferes with meaning, is not considered as a key construct of well-known L2

Email address for correspondence: ichoi001@ets.org

writing assessments such as IELTS (Cullen et al., 2014) and TOEFL iBT® (ETS, 2012). On the other hand, there is evidence suggesting that spelling accuracy is one of the anxieties test takers can have when they take a writing assessment (Pearson, 2012). If spelling accuracy is not a part of the core construct, but may cause test anxiety, it may seem obvious to allow test takers to use a spell checker during an L2 writing assessment. However, such a decision should be made based on a thorough consideration of the impact of spell checker use on test taker performance.

Most spell checkers make suggestions for misspelled words. These suggestions, even if they are not always accurate, seldom become an issue in a real-world writing context because writers often have time and resources to properly evaluate and verify the suggestions. However, suggestions from a spell checker can complicate the construct of a writing assessment. Spell checkers often make multiple suggestions for a misspelled word. In an assessment context, this essentially amounts to introducing the ability to select the most appropriate suggestion to the construct of writing. It is also possible that none of the suggestions are appropriate, which may confuse test takers. Moreover, the quality of suggestions depends on how close a misspelled word is to the correct one. Therefore, test takers who make trivial spelling errors would be more likely to receive more relevant suggestions than test takers who make more serious spelling errors. This adaptive nature of spell checker suggestions can become a fairness issue because it inherently favors test takers who do not make serious spelling errors to begin with.

From a practical point of view, test takers may use a spell checker inefficiently. Going back to a spelling error and correcting it takes time. Again, this is hardly problematic in a real-world writing context but becomes an entirely different matter in an assessment context in which test takers need to finish writing within a set time limit. When test takers use a spell checker to detect and correct spelling errors, they are, at least implicitly, choosing to spend time on spelling correction over other possible writing activities, such as writing more or correcting other types of errors. This is not necessarily a problem. All activities test takers choose to do, such as preparing an outline instead of starting writing right away, take time. However, given that spelling accuracy is seldom regarded as a key construct in L2 writing assessments, it may be undesirable for test takers to spend much time correcting trivial spelling errors, leaving less time to focus on more important writing activities. It is also possible for test takers to prefer shorter and simpler words (e.g., “say” instead of “announce”) and/or to end up with a correctly spelled but compromised word choice.[2] These potential concerns should be investigated such that assessment developers can make an informed decision about whether to provide a spell checker for a given assessment. However, the lack of spell checkers in L2 assessment

[2] In the remainder of this paper, we call these inefficient and mistaken spell checker uses as misuses.

contexts resulted in the lack of empirical studies on the use of spell checkers in such contexts. Little is known about how test takers would use spell checkers during an L2 writing assessment, or what the impact of spell checker use would be on test taker responses.

Making a first step towards addressing this knowledge gap is the goal of this study. We aim to gain a preliminary understanding of whether and how test takers use a spell checker during an L2 writing assessment, and whether the use of spell checker would have a substantive impact on how they respond to writing tasks. These research objectives are addressed through a case study involving adult English learners and experimental English writing tasks with a built-in spell checker. Both the product and process of writing are examined to provide a thorough account of the pattern and impact of their spellchecker use. We expect this study to form a solid empirical basis for large scale investigations on, and active discussions about, the introduction of spell checkers for L2 writing assessments.

Review of Literature

Spelling Errors in L1 and L2

Spelling errors have been studied intensively in terms of their frequency. Lunsford and Lunsford (2008) summarized the long tradition of such studies and observed that the proportion of spelling errors were stable across different studies and that most of the participants were college freshmen writing in their first language (L1). The spelling skills of L2 learners have been mostly studied in relation to the spelling skills of L1 writers. A series of studies in the 1990s and the early 2000s consistently reported that the developmental patterns of spelling skills in bilingual English learners were comparable to the patterns observed with native speakers (D'Angiulli et al., 2001; Da Fontoura & Siegel, 1995; Limbos & Geva, 2001; Tompkins et al., 1999; Wade-Woolley & Siegel, 1997). Lesaux et al. (2006) conducted a systematic review of the L2 English literacy development literature and argue that, as a group, L2 English learners could achieve mastery of spelling skills. However, based on a review of 27 empirical studies on the influence of L1 on the development of L2 spelling skills, Figueredo (2006) observed the influence of L1 in the development and mastery of L2 spelling skills; the closer the orthographic distance between L1 and L2 was, the more likely for a learner to master L2 spelling. Considered together, the systematic reviews of relevant literature by Lesaux et al. and Figueredo suggest that L2 English learners whose L1 is orthographically close to the alphabet system of English could achieve native-like spelling mastery. This conjecture is supported by Zhao et al.'s (2016) meta-analysis.

Recently, spelling errors of L2 writers have been studied with the goal of developing spell checkers customized for L2 learners (e.g., Hovermale, 2008; Lawley, 2015; Mitton & Okada, 2007; Rimrott & Heift, 2005, 2008). The findings from these studies consistently showed that the auto-correction performance of generic spell checkers was inferior with spelling errors made by L2 writers than with L1 writer errors (however, see Flor & Futagi, 2012, for contrasting results). This difference in spell checker performance may be attributable to the different nature of spelling errors made by L1 and L2 writers (Leacock et al., 2015). Bestgen and Granger (2011) analyzed the spelling errors in 223 essays included in the International Corpus of Learner English and reported that L2 English learners tended to make more spelling errors overall, as well as more errors due to their lack of English mastery. Bestgen and Granger concluded that the spelling errors of L2 learners could be a meaningful predictor of the overall quality of the corresponding texts. However, Flor et al. (2015) showed that the distinction between L1 and L2 English writers was confounded with the overall writing proficiency of a given writer. Based on their review of a corpus of TOEFL® iBT and GRE® essays, Flor et al. observed a relationship between spelling errors and the overall quality of the texts regardless of whether English was the L1 of a writer or not. They also noted that the severity of spelling errors depended largely on the overall quality of a given essay rather than the writer's L1. In Flor et al.'s study, once the overall essay quality was controlled, difference between L1 and L2 writers only manifested in the frequency of intended correct forms.

Spell Checkers for Language Learners

Spell checkers for developing writers are often included in automated writing evaluation (AWE) systems, such as *Criterion®*, *My Access!®*, and *WriteToLearn™*. Classroom uses of such AWE systems have yielded research reports showing how learners respond to feedback from spell checkers. Warschauer and Grimes (2008) observed middle- and high-school classrooms using *Criterion* and *My Access* and found that AWE feedback led to revisions involving spelling, word choice, and grammar errors. Kim (2010) reports a similar finding in her survey of college-level Korean learners of English; her students perceived *Criterion*'s feedback on mechanical aspects such as spelling as the system's strongest feature. As Backer (2014) observed with college students in an engineering course, learners also favored spell checkers as the strength of AWE systems when writing about specialized academic content. However, this does not necessarily mean that learners would always act on spelling-related AWE feedback. Lavolette et al. (2015) noted that undergraduate students in their study did not always correct spelling errors detected by AWE systems. In sum, these findings suggest that, while learners prefer to have feedback on their spelling, how they would utilize such feedback is not always clear.

The lack of research on L2 spelling in general has been frequently noted in the literature (e.g., Bestgen & Granger, 2011; Figueredo, 2006; Lesaux et al., 2006). Therefore, it is not surprising that little has been studied about the use of a spell checker by L2 learners in an assessment context. Although Flor et al.'s (2015) study analyzed L2 English writer essays collected from two standardized writing assessments, neither of the two assessments provided test takers with a spell checker.

Research Questions

The goal of this study was to gain a preliminary understanding of how a spell checker would be used in an L2 writing assessment context and how the spell checker use would impact test taker responses. Because not every test taker would choose to use the spell checker, the question of how it would be used implied a preceding question of how often test takers would choose to use it. If spell checker use would bring about changes in how test takers respond to writing tasks, the changes would manifest in what they submit (i.e., the product of writing) and how they write their responses (i.e., the process of writing). The impact of spell checker use was then investigated through comparisons between spell checker users and non-users, in terms of the product and process of their writing. In sum, this study was guided by the following two research questions:

1. How often did test takers of an L2 writing assessment choose to use a spell checker and how did they use it?
2. What were the impacts of the spell checker use on the product and process of test takers' writing?

Methods

Participants

This study was part of a larger project to develop and evaluate innovative writing assessment tasks for adult English learners. We contacted TOEFL iBT test takers via email to introduce the project and ask for voluntary participation. A total of 61 adult English learners who took the TOEFL iBT test volunteered to participate. They took the operational TOEFL iBT test first and then responded to two experimental writing tasks (details provided in the following subsection) at the same testing center in which they took the operational test. Most of them (57 out of 61) agreed to provide their linguistic background and gender. These test takers spoke 10 different languages including Chinese (21), Spanish (12), Thai (7), Korean (4), Arabic (3), German (3), Vietnamese (3), Greek (2), Farsi (1), and Nepali (1). None of them identified themselves as an English bilingual. The

proportions of male and female participants were 42 and 58 percent, respectively. To investigate the impact of the spell checker use, the test takers were randomly divided into two groups. The first group (hereafter Group 1), consisting of 31 test takers, responded to two experimental writing tasks (which we describe in detail in the following subsection) without a spell checker. The second group (hereafter Group 2), consisting of the remaining 30 test takers, were given a spell checker while they were writing for the same two tasks.

Instruments

The written responses of the test takers were elicited through two experimental writing tasks. Both tasks were computer-delivered and provided source materials to facilitate writing. The first task (hereafter Task 1) presented a reading passage about a science-related topic, followed by an audio interview that provided counter-arguments against the points made in the passage. The test takers were then asked to summarize the topic based on the passage and the interview. The second task (hereafter Task 2) was situated in a hypothetical online forum, in which two commentators wrote posts advancing opposing views on a given topic. The test takers were asked to provide their own opinion on the topic, with reference to either or both of the two previous posts. The directions and source materials for the tasks were evaluated and revised through three previous prototype trials, each involving approximately 10 adult English learners. The test takers were given 20 and 25 minutes to complete Task 1 and Task 2, respectively. These time limits were also determined based on the results of the prototype trials to ensure that the test takers would have enough time to develop their content.

We developed, for this study, a spell checker that only detects spelling errors without providing suggestions. The 30 test takers in Group 2 had access to the detection-only spell checker. The spell checker could be turned on and off by clicking on a button. The default was set at the off position, because we wanted to know when test takers turned it on. The Group 2 test takers, who had access to the spell checker, were allowed to turn it on and off anytime while they were writing. When the spell checker was on, it underlined words that contained spelling errors, as can be seen in Figure 1.

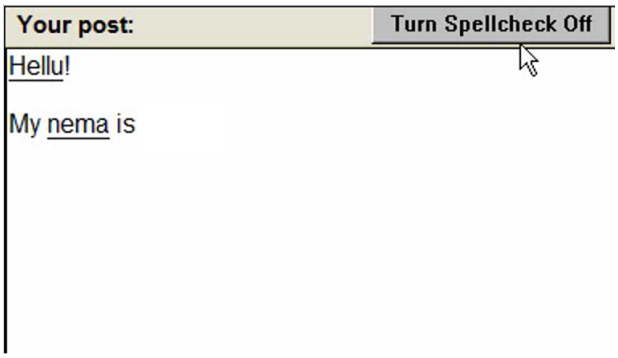

Figure 1. Spell checker on/off button and flagged (underlined) errors.

Data

The written responses to the two experimental writing tasks comprised the main source of data for this study. All test takers had taken the TOEFL iBT test immediately before responding to the experimental tasks, and their iBT scores (ranging from 0 to 120; see https://www.ets.org/toefl/ibt/scores/understand/ for the description of TOEFL iBT scores) were also included as a covariate. Another important source of data was keystroke logs, which captured the entire writing process of the test takers by recording which key action took place at which cursor position, as well as the timestamp for each key action. The use of the spell checker was also captured in the logs. The multiple sources of data allowed us to examine the writing product and process of the test takers.

Analysis

In order to address the first research question, we examined the keystroke logs to find out how many test takers in Group 2 chose to use the spell checker and whether there were any differences between those who chose to use the spell checker and those who did not. We focused on whether there was strong evidence indicating a systematic relationship between the spell checker use and general English proficiency (measured by the TOEFL iBT test). We also investigated the patterns of spell checker use by examining at which point of writing each user turned on and off the spell checker. The use patterns were examined separately for each task to identify meaningful commonalities and/or differences between the two tasks.

We addressed the second research question from two angles, which were the product and process of writing. The impact of spell checker use on writing product was investigated in terms of the number of words and spelling errors in submitted responses (i.e., final products). The choice of the two outcome variables was based on our expectation that, while spell checker use would reduce the amount of spelling errors, it would also decrease the length of responses due to time spent correcting spelling errors.

We regressed the two outcome variables on the spell checker use that was dummy coded into two indicators: (1) test takers who were not given the spell checker (i.e., Group 1) versus the reference group, and, (2) those who were given the spell checker but chose not use it versus the reference group, with those who used it designated as the reference group. The impact of spell checker use on writing process was examined by reconstructing the entire process of each test taker based on the keystroke logs. The reconstructed processes were then inspected visually, using the time series of word counts at the group level (using the above three-category classification). We also reviewed editing behaviors made while the spell checker was used, with the goal of identifying noteworthy spell checker uses and/or misuses.

Results

Research question 1: Spell checker use patterns

We scanned the keystroke logs of Group 2 test takers to identify those who used the spell checker and for how long. The results of the scan are given in Figure 2. Figure 2 shows that 11 and 14 test takers (out of the 30 test takers who were given the spell checker) turned on the spell checker at least once while responding to Task 1 and Task 2, respectively. A few of the uses, however, were too short to be considered meaningful. For example, when responding to Task 1, test takers T19 and T59 turned off the spell checker immediately after they turned it on and did not turn it on again. Their total time of use was less than two seconds. Such short, one-off uses would not have given the test takers anything other than a quick glance of how the spell checker looked. We examined the reconstructed writing processes of test takers who used spell checkers for less than a minute and found that the shortest spell checker use during which any keystroke event occurred lasted for 33 seconds (T29, Task 1). Therefore, we decided to consider only the test takers with at least 30 seconds of use as meaningful users of the spell checker, resulting in 7 and 11 meaningful users for Task 1 and Task 2, respectively. In all subsequent analyses of the spell checker use impact, only these 7 and 11 test takers were considered to be spell checker users. On the other hand, the other Group 2 test takers who chose not to (meaningfully) use the spell checker and Group 1 test takers will be collectively referred to as the non-users (54 non-users for Task 1 and 50 non-users for Task 2).

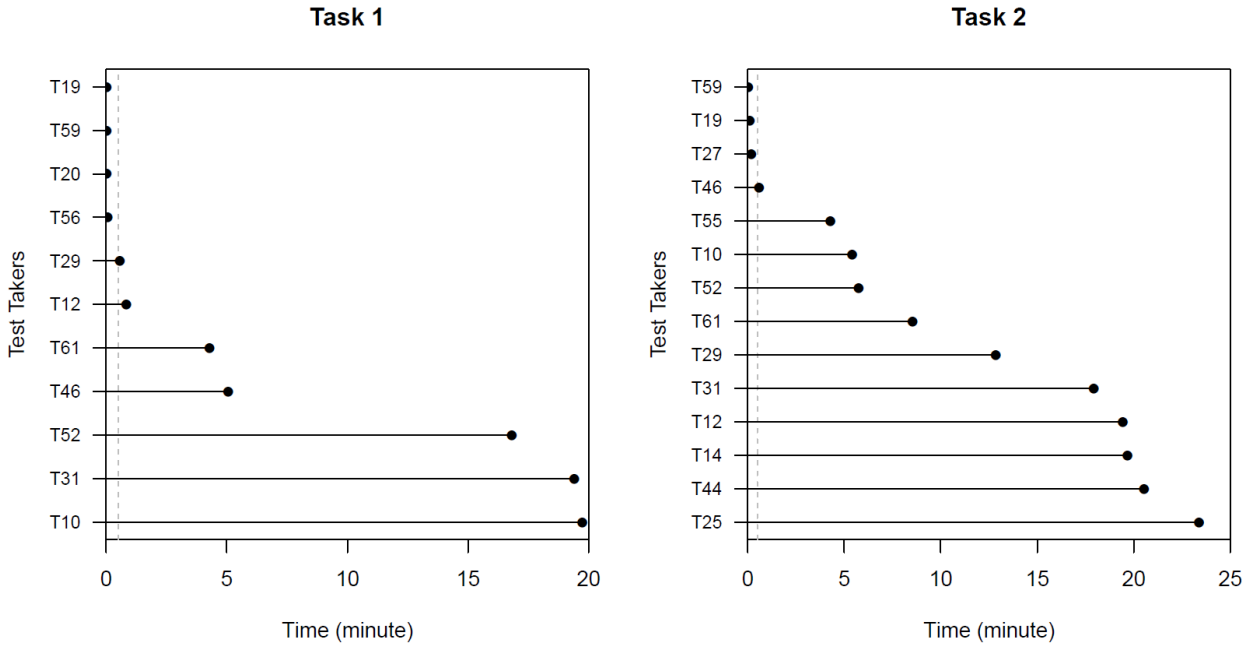

Figure 2. Total duration of spell checker use. The dotted lines represent the 30-second points that were used to exclude extremely short uses.

We compared the spell checker users with the non-users in terms of the TOEFL iBT scores, as can be seen in Figure 3. Despite the differences in group size, the TOEFL iBT score distributions of the three groups appeared similar. Each of the three score distributions showed similar spreads, with most of the test takers scoring in the 60 to 100 range, which corresponds to the CEFR levels from B1 to C1 or above (Papageorgiou et al., 2015). The means and inter-quartile ranges of the three distributions were also comparable, which supported the visual evaluation of Figure 3. We interpreted this result as indicating no systematic difference between the spell checker users and non-users in their English proficiency.

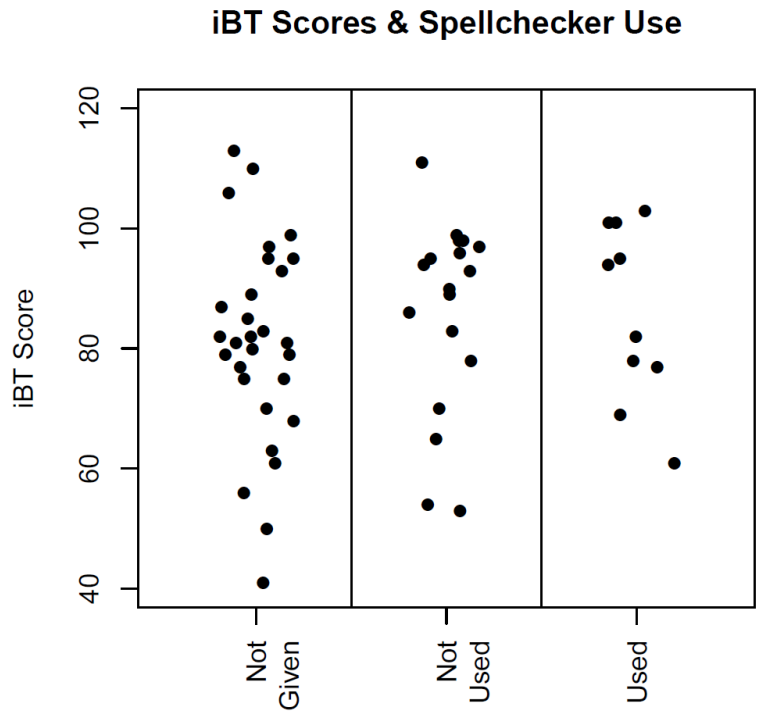

Figure 3. TOEFL iBT scores of Group 1 ("Not Given"; n=29) test takers, Group 2 test takers who chose not to use the spell checker ("Not Used"; n=18), and Group 2 spell checker users ("Used"; n=10). Out of a total of 61 participants, 57 agreed to disclose their iBT scores (n=57 for this plot).

We then examined at which point of writing the spell checker was used. Figure 4 provides a visual summary of the results and shows that every spell checker user submitted the response with the spell checker turned on. Figure 4 also suggests three broad groupings among the users. The first group of users turned on the spell checker at the beginning and used it throughout the writing. This group includes T31 and T10 for Task 1, and T44 and T25 for Task 2. The second group wrote most of their responses without the spell checker, turned on the spell checker right at the end and used it for a short while, and submitted their responses. The test takers who belonged to this group are T29 and T12 for Task 1, and T46 for Task 2. The remaining users comprised the third group, who wrote for a while (typically for the first several minutes) without the spell checker and turned it on and used it until the end.

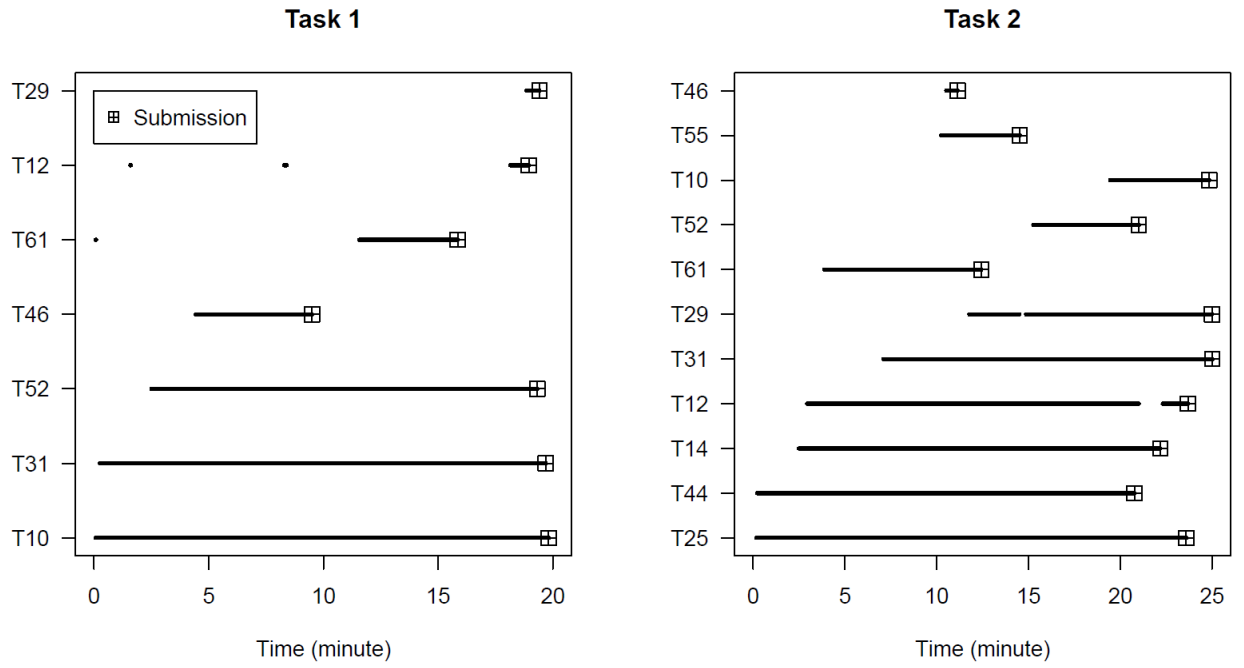

Figure 4. The spell checker use patterns and the submission points of the 7 and 11 users for Task 1 and Task 2, respectively. The solid lines represent the window of spellchecker use. The users are sorted on the y-axis in the ascending order of total use duration.

Regarding the spell checker use as a sign of being mindful of spelling errors, one explanation for these patterns can be the following. The first use pattern may represent test takers who prefer not to make spelling errors to begin with. They would check every word they write and correct errors as they go on. The second use pattern, on the other hand, may be considered as the opposite strategy, in which test takers first draft their response without much concern about spelling errors and then address all the errors at the end. The third pattern may be regarded as a middle point between the first and second patterns; the drafting stage similar to the second pattern is observed here, but spellchecking is not so delayed as in the second pattern.

Research question 2: Impact of spell checker on writing process and product

In order to understand the impact of spell checker use on the amount of writing, we regressed the final word counts of written responses on the spell checker use. We also used two covariates, TOEFL iBT scores and the total time of writing, to control for other factors that could affect the amount of writing. We fit a separate regression model for each of the two tasks, because the two tasks were different in their design characteristics and genre. Given the relatively small sample size of 57 (4 out of 61 did not provide their iBT scores), the estimates of model coefficients were expected to be associated with large standard errors. Accordingly, our goal for this analysis was to identify the sign and

approximate size of the spell checker use impact, rather than to estimate its precise magnitude or make statistical inferences. The regression model results are given in Table 1.

Table 1. Estimated coefficients and standard errors from the regression of total number of words

	Task 1		Task 2	
	Coefficients	S.E.†	Coefficients	S.E.
Intercept	149.96*	17.13	162.26*	14.94
Not given (vs. Used)	37.27*	18.81	55.20*	17.22
Did not use (vs. Used)	49.76*	19.29	51.74*	18.95
TOEFL iBT score	2.09*	0.34	1.58*	0.42
Total time of writing	8.37*	1.32	4.43*	1.13
R^2 (adjusted R^2)	0.59 (0.55)		0.35 (0.30)	

Notes: *: significant at the α=0.05 level; †: Standard Error.

The spell checker users wrote, on average and conditional on their iBT scores and writing time, approximately 40 to 50 fewer words than those who did not use the spell checker. This result aligns well with our prior expectation; all else being equal, as test takers spend more time on checking and correcting spelling errors, they would end up with less time for producing content and thus write fewer words. Those who were not given the spell checker and those who chose not to use it were similar in the average number of words conditional on the other two predictors. As expected, these estimates had large standard errors and thus should not be interpreted as the precise magnitude of the spell checker use impact.

The regression models accounted for approximately 58 and 35 percent of the observed variance in word counts for Task 1 and Task 2, respectively. When the spell checker use indicators were not included as predictors, the amount of variance accounted for decreased by 5 and 14 percent (both of which were significant at the 95 percent level), respectively for Task 1 and Task 2. The size of the estimated coefficients for the spell checker use indicators and the magnitude of the R^2 value changes suggest that the spell checker use impact was larger for Task 2 than Task 1. The differential impact of spell checker use may be attributed to the different nature of the tasks themselves. Task 1 requires a summary of two sources, and therefore, test takers may be relying more heavily on the words in the sources than they would for Task 2, which asks for test takers' opinions. The heavy reliance on the sources could then yield a smaller need for checking and correcting spelling errors, leading to an attenuated impact on the amount of words written. However, this interpretation of the differential impact is merely one of many possible explanations.

We also examined the impact of spell checker use on another product-related aspect: the number of spelling errors included in submitted responses. Unlike the number of total words that were symmetrically distributed, the number of errors were highly skewed with relatively small interquartile ranges as well as several outliers. This distributional characteristic of the spelling errors presented a challenge for fitting a linear regression model. Therefore, to address this challenge, we first transformed the number of errors using the natural log (which made the distribution more symmetrical) and then regressed the log-transformed error counts on the spell checker use, with TOEFL iBT scores and total word counts as covariates. Similarly to the analysis of word counts, we fit a separate model for each of the two tasks. The results of the regression models are given in Table 2.

Table 2. Estimated coefficients and standard errors from the regression of the number of spelling errors (log-transformed)

	Task 1		Task 2	
	Coefficients	S.E. †	Coefficients	S.E.
Intercept	-1.74*	0.70	-0.28	0.51
Not given (vs. Used)	3.02*	0.78	1.48*	0.61
Did not use (vs. Used)	3.14*	0.80	1.80*	0.64
TOEFL iBT score	-0.01	0.02	-0.01	0.01
Total number of words	0.01*	0.00	0.01*	0.00
R^2 (adjusted R^2)	0.37 (0.33)		0.26 (0.22)	

Notes: *: significant at the α=0.05 level; †: Standard Error.

The positive coefficients for those who did not use the spell checker indicated that, on average and conditional on the covariates, the responses of the spell checker users contained significantly fewer spelling errors than those of the test takers who did not use the spell checker. The models were able to account for approximately 37 and 26 percent of the variance in the log-transformed error counts for Task 1 and Task 2, respectively. The impact of TOEFL iBT scores and word counts on the number of spelling errors appeared negligible; the estimated coefficients for the two covariates were very small. When the spell checker use indicators were removed from the models, the R^2 values were reduced to 16 and 15 percent for Task 1 and Task 2, respectively.

The impact of spell checker use on writing process was examined at the group level by visually inspecting the time series of word count increase. The time series were constructed by counting the number of words at every 10 seconds for each test taker, and the average time series of the three groups were compared. In order to focus on test takers who spent comparable amounts of time for writing, we excluded from the averaging process test takers who spent less than 15 minutes writing. The averages represented by the lines are expected to fluctuate more after the 15-minute point because this portion of the time-series was based on progressively fewer individuals as test takers submitted

their responses. The resulting group-average time series of word count increase, for Task 1 and Task 2, are given in Figure 5.

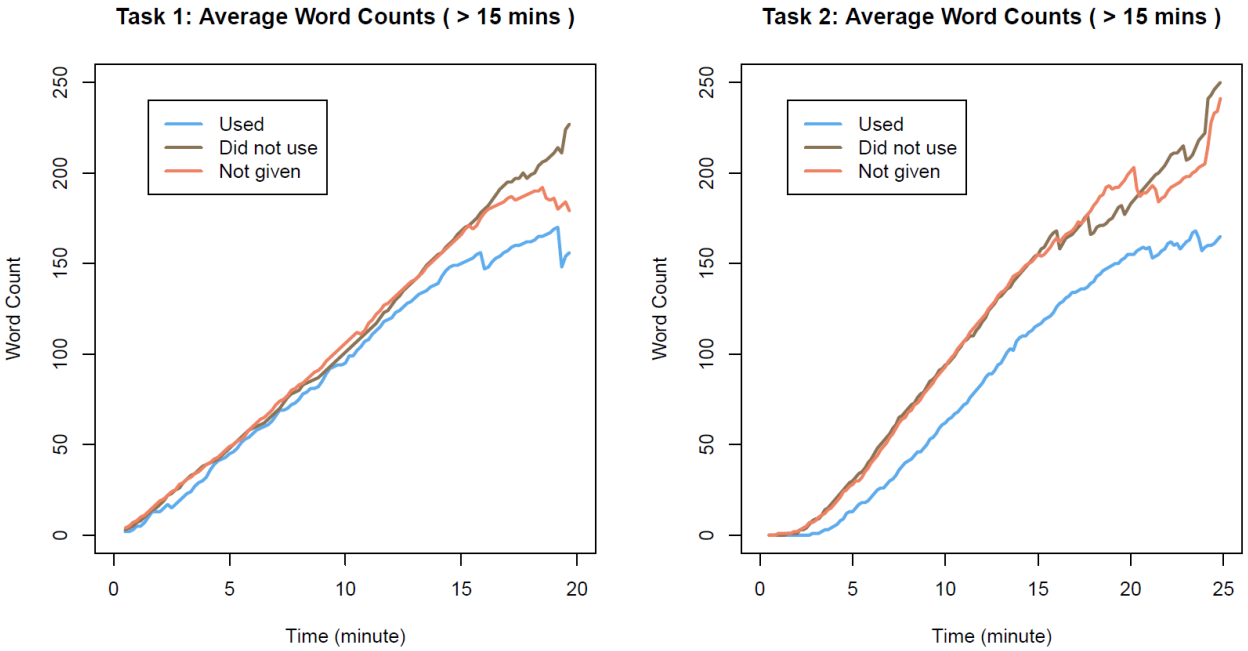

Figure 5. Averaged time series of word counts for test takers who spent more than 15 minutes on the tasks; the spell checker users ("Used"), the non-users in Group 2 ("Did not use"), and Group 1 test takers ("Not given").

Figure 5 shows that the spell checker use appeared to have impacted the writing processes for Task 1 and Task 2 differently. For Task 1, all three groups showed comparable average productivity in words produced up to the 12-minute point. On the other hand, the spell checker users for Task 2, on average, appeared to have written their responses at a slower pace than the non-users. The difference appeared early on and became more pronounced as the writing progressed. The differential impact on Task 1 and Task 2 writing processes is in line with the findings about the writing products in that, overall, the impact of spell checker use was larger for Task 2 than for Task 1.

Lastly, we reviewed how the spell checker users corrected spelling errors. A round of thorough reviews yielded the following qualitative observations. The majority of corrections involved adding and/or subtracting characters, changing character orders, and/or replacing a character with alternatives, until the indication of a given spelling error (i.e., an underline) disappeared. When several correction attempts turned out to be unsuccessful, the spell checker users often moved on from the given error. Although many spell checker users dealt with one error at a time, it was not uncommon to find "multi-tasking" cases in which test takers took stabs at multiple errors while alternating back and forth between the errors.

However, there were a few error correction attempts that lasted for a relatively long time. This could be potentially damaging to a test taker if he or she was under time pressure. A few of such extensive correction attempts even led the test taker to inappropriate word choices, such as the example in Figure 6.

```
<Key 1135>: ... as long as you have ascess to both of them ...
<Key 1136>: ... as long as you have asces to both of them y...
<Key 1137>: ... as long as you have asce to both of them yo...
<Key 1138>: ... as long as you have asc to both of them you...
<Key 1139>: ... as long as you have as to both of them you ...
<Key 1140>: ... as long as you have a to both of them you c...
<Key 1141>: ... as long as you have  to both of them you ca...
<Key 1142>: ... as long as you have a to both of them you c...
<Key 1143>: ... as long as you have as to both of them you ...
<Key 1144>: ... as long as you have asc to both of them you...
<Key 1145>: ... as long as you have asce to both of them yo...
<Key 1146>: ... as long as you have asces to both of them y...
<Key 1147>: ... as long as you have ascess to both of them ...
<Key 1148>: ... as long as you have asces to both of them y...
<Key 1149>: ... as long as you have asce to both of them yo...
<Key 1150>: ... as long as you have asc to both of them you...
<Key 1151>: ... as long as you have as to both of them you ...
<Key 1152>: ... as long as you have a to both of them you c...
<Key 1153>: ... as long as you have ac to both of them you ...
<Key 1154>: ... as long as you have ace to both of them you...
<Key 1155>: ... as long as you have aces to both of them yo...
<Key 1156>: ... as long as you have acess to both of them y...
<Key 1157>: ... as long as you have aces to both of them yo...
<Key 1158>: ... as long as you have ace to both of them you...
<Key 1159>: ... as long as you have ac to both of them you ...
<Key 1160>: ... as long as you have a to both of them you c...
<Key 1161>: ... as long as you have  to both of them you ca...
<Key 1162>: ... as long as you have e to both of them you c...
<Key 1163>: ... as long as you have ex to both of them you ...
<Key 1164>: ... as long as you have exc to both of them you...
<Key 1165>: ... as long as you have exce to both of them yo...
<Key 1166>: ... as long as you have exces to both of them y...
<Key 1167>: ... as long as you have excess to both of them...
```

Figure 6. An excerpt from a spell checker user showing the process of spelling correction leading an inappropriate word choice.

The sequence in Figure 6 focuses on the test taker's attempt to correct a misspelled word, "ascess". It is clear from the context that the correct form for the misspelled word was "access". The test taker made a few correction attempts, including “ascess” (Keystrokes

1141 through 1147) and “acess” (Keystrokes 1152 through 1156). As these attempts were still flagged as errors, the test taker made another attempt and satisfied the spell checker by opting for a different word, "excess" (Keystrokes 1162 through 1167). We believe that this "solution" could be more damaging than the original spelling error to the communicative effectiveness of the surrounding sentence. In Keystroke 1135, the sentence contained a simple spelling error for which a rater would not have much difficulty inferring the intended correct form. A total of 33 keystrokes later, the sentence became more uncertain with an inappropriate word choice. The test taker spent approximately two percent of the allotted writing time to make the response arguably worse, and this might not have happened without the spell checker.

However, we note that the sequence in Figure 6 represents an extreme, arguably the most damaging spell checker use we observed in our data. Even when several consecutive attempts failed to correct a given error, most spell checker users moved on. Persistent error correction attempts, let alone those that ended up with an inappropriate word choice, were very rare.

Discussion and Conclusions

Through multiple analyses we addressed the two research questions of this study: (1) spell checker use patterns, and (2) the impact of spell checker use. About a third of the test takers who were given the option to use the spell checker chose to use it. The spell checker users did not appear to differ from the non-users in terms of general English proficiency. Overall, the spell checker users either turned it on at the beginning and left it on throughout the writing process or turned it on after having completed a certain amount of writing. The impact of spell checker use was observed in both the product and process of writing. The spell checker users, on average, wrote fewer words and made fewer spelling errors than the non-users. For Task 2, they also tended to write at a slower pace than the non-users. The impact of spell checker use appeared larger for Task 2 than for Task 1. Overall, excessive and/or damaging uses of the spell checker that might raise measurement concerns were rarely observed.

The quality of writing has to do with a much larger set of factors than word counts and spelling errors. Therefore, we do not claim that our findings about word and error counts can be translated into the quality of writing in a straightforward manner. Our goal was to gain a preliminary understanding of the spell checker use impact in an L2 writing assessment context, and our overall interpretation of the findings is that the spell checker use did not bring about a substantial change in what the tasks were designed to measure and how test takers responded to them. To our knowledge, this is the first empirical

evidence on the impact of using a spell checker during an L2 assessment for adult learners. The patterns of spell checker use and their impact we observed in this study carry an important implication for assessment designers: our findings suggest few potential fairness concerns when providing a detection-only spell checker in the context of an adult L2 writing assessment. There appeared to be no differential advantage to using the spell checker, and the choice of whether to use the spell checker did not appear closely related to general English proficiency. We believe these findings suggest that the fairness concerns should not preclude a decision to implement a simple spell checker in an assessment setting and its potential advantages, including enhanced authenticity and reduced test anxiety.

This study also has a methodological implication for empirical studies on spell checker use in general. Previous studies have focused on the type of spelling errors learners make and on the relationship between spelling errors and learners' linguistic background. As a result, *when* and *how* learners use a spell checker have received less attention. We found analyzing keystroke logs to be a useful way of understanding the impact of spell checkers on writing process. By examining the keystroke logs of the participants, we were able to reconstruct and analyze their entire writing process including when they used the spell checker and how they interacted with it. This led to important observations that would not have been available had we only examined submitted responses. Capturing keystroke logs has become more feasible as writing in a digital environment becomes more common and many types of writing take place on a computer. The findings from our writing process examination demonstrated that the collection and analysis of keystroke logs can offer substantive insights in understanding the impact of spell checker use on writing process.

Individual cases can be valuable if they suggest a potentially systematic issue, and, in this regard, peculiar cases of spell checker use can be noteworthy. We have presented an extreme use case in which the test taker ended up with an inappropriate word by trying to correct a simple spelling error. Such a case does not represent a beneficial use of a spell checker. Given the often insignificant role of spelling in writing assessments for adult learners, spending much time on spelling errors is not efficient, and inappropriate word choices can arguably be more detrimental than simple spelling errors. Although such an extreme use was rare, it would be prudent to be proactive in preventing potential side effects to ensure no unintended consequences of allowing a spell checker during an L2 writing assessment. Therefore, we recommend that, when an L2 writing assessment allows a spell checker, test takers be reminded about potentially detrimental impacts of spell checker use on the product and process of writing, especially when the amount of time for writing is relatively short.

The findings and implications of this study should be considered along with its limitations. Our data were from a relatively small number of participants in an experimental study with a custom-developed spell checker. Therefore, the nature of the findings may differ from what we would observe in an operational assessment context with a much larger sample. As pointed out by an anonymous reviewer, none of our participants would qualify as a low-proficiency or basic learner (e.g., CEFR level A1 or A2), who may demonstrate different patterns of spell checker use. We did not examine the impact of spell checker use on scores, for the tasks and the corresponding scoring rubrics were experimental and the measurement errors due to human scoring were unknown. Consequently, we cannot claim whether the net impact of spell checker use on final scores would be positive or negative. Due to the small sample size, it was also not feasible to examine whether the different use patterns we observed were related to important background variables, such as test takers' English proficiency, gender, and/or first language.

The findings and the limitations of this study point to multiple lines of research that can benefit from a more fine-grained examination based on larger data. The side effects of spell checker use were not frequently observed among the test takers in this study, but it is premature to rule out concerns about misuses. With a larger sample size, different types of misuse cases may emerge, and the proportion of such misuse cases can be estimated. Moreover, replicating the experiment in this study with low-proficiency learners who have not yet acquired a proper control of English spelling will help understand whether and how the impact of spell checker use interacts with proficiency. Another background variable of interest would be the alphabetical system of learners' L1. As noted in the literature (Figueredo, 2006; Lesaux et al., 2006; Zhao et al., 2016), orthographical similarity between L1 and L2 is an important factor for the mastery of L2 spelling skill. Therefore, it would be particularly interesting to recruit learners whose L1s differ in terms of alphabetical systems and compare their spell checker use patterns. More broadly, the different patterns of spell checker use comprise a promising line of future research. If a set of specific use patterns lead to better (or worse) quality responses, test takers would benefit from being advised to adopt (or avoid) such patterns, and assessment developers should consider them in developing task prompts and scoring rubrics. The potential relationship between spell checker use patterns and response quality as well as test taker background variables, therefore, will carry significant implications to major stakeholders.

Acknowledgements

Any opinions expressed in this paper are those of the authors and not necessarily of Educational Testing Service. We would like to thank Heather Buzick, Larry Davis,

Danielle Guzman-Orth, and Guangming Ling for their comments on an earlier version of this paper. We would also like to thank the editors of the journal and anonymous reviewers for their constructive feedback. Any remaining flaws are our own.

References

Backer, P. R. (2014). Effectiveness of an online writing system in improving students' writing skills in engineering. *Computers in Education Journal, 5*, 14-27.

Bestgen, Y., & S. Granger. (2011). Categorising spelling errors to assess L2 writing. *International Journal of Continued Engineering Education and Life-Long Learning, 21*, 235-252. https://doi.org/10.1504/IJCEELL.2011.040201

Cullen, P., French, A., & Jakeman, V. (2014). *The official Cambridge guide to IELTS*. Cambridge: Cambridge University Press.

Da Fontoura, H. A., & Siegel, L. S. (1995). Reading, syntactic, and working memory skills of bilingual Portuguese–English Canadian children. *Reading and Writing, 7*, 139–153. https://doi.org/10.1007/BF01026951

D'Angiulli, A., Siegel, L. S., & Serra, E. (2001). The development of reading in English and Italian in bilingual children. *Applied Psycholinguistics, 22*, 479–507. http://dx.doi.org/10.1017/S0142716401004015

ETS. (2012). *The official guide to the TOEFL® test* (4th ed.). New York, NY: McGraw-Hill,

Figueredo, L. (2006). Using the known to chart the unknown: A review of first-language influence on the development of English-as-a-second-language spelling skill. *Reading and Writing, 19*, 873-905. https://doi.org/10.1007/s11145-006-9014-1

Flor, M., & Futagi, Y. (2012). On using context for automatic correction of non-word misspellings in student essays. In *Proceedings of The 7th Workshop on Innovative Use of NLP for Building Educational Applications*, 105-115, at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 3-8, 2012, Montréal, Canada.

Flor, M., Futagi, Y., Lopez, M., & Mulholland, M. (2015). Patterns of misspellings in L2 and L1 English: A view from the ETS Spelling Corpus. *Bergen Language and Linguistics Studies, 6*, 107–132. http://dx.doi.org/10.15845/bells.v6i0.811

Green, A., & Youngs, B. E. (2001). Using the web in elementary French and German courses: Quantitative and qualitative study results. *CALICO Journal, 19*, 89-123. http://dx.doi.org/10.1558/cj.v19i1.89-123

Hovermale, D. J. (2008). SCALE: Spelling correction adapted for learners of English. *Workshop presented at CALICO 2008 ICALL SIG*, March 18-19, 2008, San Francisco, USA.

Kim, T-E. (2010). Reflection on using the Criterion online writing evaluation service. *Multimedia-Assisted Language Learning, 13*(3), 59-83.

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology, 19*(2), 50–68. http://dx.doi.org/10125/44417

Lawley, J. (2015) Spelling: computerised feedback for self-correction. *Computer Assisted Language Learning, 29*, 868-880. https://doi.org/10.1080/09588221.2015.1069746

Leacock, C., Chodorow, M., & Tetreault, J. (2015). Automatic grammar- and spell-checking for language learners. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (Cambridge Handbooks in Language and Linguistics, pp. 567-586). Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139649414.025

Lesaux, N., Koda, K., Siegel, L., & Shanahan, T. (2006). Development of literacy. In D. August, & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the national literacy panel on language-minority children & youth* (pp. 75-122). Mahwah. NJ: Erlbaum.

Limbos, M., & Geva, E. (2001). Accuracy of teacher assessments of second-language students at risk for reading disability. *Journal of Learning Disabilities, 34*, 136–151. https://doi.org/10.1177/002221940103400204

Lunsford, A. A., & Lunsford, K. J. (2008). "Mistakes are a fact of life": A national comparative study. *College Composition and Communication, 59*, 781–806.

MacArthur, C. A. (1999). Overcoming barriers to writing: Computer support for basic writing skills. *Reading & Writing Quarterly, 15*, 169–192. https://doi.org/10.1080/105735699278251

Mitton, R. & Okada, T. (2007). *The adaptation of an English spellchecker for Japanese writers.* London: Birkbeck ePrints.

Ndiaye, M., & Vandeventer Faltin, A. (2003). A spell checker tailored to language learners. *Computer Assisted Language Learning, 16*, 213-232. http://dx.doi.org/10.1076/call.16.2.213.15881

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM-15-06). Princeton, NJ: Educational Testing Service.

Pearson, H. (2012). Issues in the assessment of spelling. *Literacy Learning: The Middle Years, 20*, 29–33.

Rimrott, A. & Heift, T. (2005). Language learners and generic spell checkers in CALL. *CALICO Journal, 23*, 17-48.

Rimrott, A. & Heift, T. (2008). Evaluating automatic detection of misspellings in German. *Language Learning & Technology, 12*, 73-92.

Tompkins, G. E., Abramson, S., & Pritchard, R. H. (1999). A multilingual perspective on spelling development in third and fourth grades. *Multicultural Education, 6*, 12–18.

Wade-Woolley, L., & Siegel, L. S. (1997). The spelling performance of ESL and native speakers of English as a function of reading skill. *Reading and Writing, 9*, 387–406. https://doi.org/10.1023/A:1007994928645

Warschauer, M. (2010). New tools for teaching writing. *Language Learning & Technology, 14*, 3-8. http://dx.doi.org/10125/44196

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. Pedagogies: An International Journal, 3, 22-36. https://doi.org/10.1080/15544800701771580

Zhao, J., Quiroz, B., Dixon, L. Q., & Joshi, R. M. (2016). Comparing bilingual to monolingual learners on English spelling: A meta-analytic review. *Dyslexia, 22*, 193-213. https://doi.org/10.1002/dys.1530