Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

Language Education & Assessment, 1 (2), 45-58 (2018)
https://dx.doi.org/10.29140/lea.v1n2.61

The Impact of Spelling Errors on Trained Raters' Scoring Decisions

IKKYU CHOI [a]

YEONSUK CHO [b]

[a] *Educational Testing Service, USA*
Email: ichoi001@ets.org

[b] *Educational Testing Service, USA*
Email: ycho@ets.org

Abstract

Second language (L2) writing assessments seldom allow a spellchecker and often have a time limit. Naturally, test takers often submit responses with spelling errors. However, little is known about whether and how spelling errors in test taker responses affect trained raters' scoring decisions. In this study, we investigated the impact of spelling errors on trained raters' holistic evaluation of response quality. We selected 148 responses to four L2 writing tasks and created error-free versions of the responses by correcting spelling errors in the original responses. Both the original and corrected responses were randomly assigned to trained raters, who scored the responses according to the same holistic scoring rubrics. We compared the resulting scores of original and corrected responses to gauge the impact of spelling errors. We also examined whether the impact of spelling errors varied across different task types and spelling error characteristics. The results showed that the spelling error correction led to an average of more than a half score point increase. The score gains from the error correction varied across task types and the quantity of corrected errors. These findings have multiple implications, including suggestions for rater training and assessment development.

Keywords: error correction; human rater scoring; L2 writing assessment; spelling errors

Introduction

Spelling errors had often been viewed as an indicator of having poor writing skills. However, this view of spelling errors is undergoing a change as more and more writing takes place in digital environments, in which writers have easy access to spellcheckers. Many readers now believe that the writers themselves have the responsibility of correcting spelling errors using spell checkers (Figueredo & Varnhagen, 2005). However, not every digital writing environment provides a spellchecker. A prominent example is standardized second language (L2) writing assessments in which test takers

Data Availability Statement: *All relevant data are within this paper.*

respond to a set of assessment tasks in a controlled setting. Such assessments seldom allow a spellchecker and often have a time limit. Consequently, test takers do not always have the means (other than their knowledge of spelling) or time to correct spelling errors before submitting their responses. The resulting spelling errors will be transparent to raters who score the responses.

Most L2 writing assessments rely on human raters for scoring test taker responses. Raters often give a single score representing the overall quality of a given response (called a holistic score), which is operationalized in scoring rubrics. Spelling errors are typically considered as a minor concern for writing in higher-education contexts (e.g., Anson & Anson, 2017; Johnson, Wilson, & Roscoe, 2017), and thus often not explicitly mentioned in scoring rubrics. Under such assessment conditions, raters should adhere to rubrics by disregarding spelling errors in assigning scores. Although raters are trained to score responses according to scoring rubrics, the nature of holistic scoring makes it difficult to examine whether and how a specific aspect, such as spelling errors, would affect scoring decisions. Research in writing assessment contexts has rarely focused on spelling errors, and relevant findings mostly involved the anxiety of test takers (Pearson, 2012) and the consistency and severity of raters in evaluating surface level errors during rater training (Elder, Barkhuizen, Knoch, & von Randow, 2007). Little is known how trained raters process spelling errors when making scoring decisions, especially when they are not supposed to penalize spelling errors according to the rubrics.

In this study, we investigate the impact of spelling errors on trained raters' holistic evaluation of written responses by L2 test takers. Our data consisted of 148 adult L2 learners' written responses to experimental writing tasks. We created error-free versions of the responses by correcting spelling errors in the original responses. Both the original and corrected responses were then randomly assigned to trained raters, who scored the responses according to holistic scoring rubrics. We then compared the scores of original and corrected responses to gauge the impact of spelling errors. We also examined whether the impact of spelling errors varied across different task types and error characteristics. We believe that the findings of this study have substantive implications for understanding the impact of spelling errors in a controlled assessment setting, as well as practical implications for rater training and L2 writing assessment development.

Literature Review

The impact of spelling errors on the perceived quality of texts has been examined with several different groups of readers[1]. Varnhagen (2000) examined how second-, fourth-, and sixth-graders in the U.S. felt when they were reading stories with spelling errors. Her participants viewed the stories and their authors negatively when spelling errors were present. Figueredo and Varnhagen (2005) showed that the negative impact of spelling errors observed by Varnhagen (2000) could be generalized to adult readers. They examined 270 undergraduate students' evaluation of texts with spelling errors and reported that, similarly to the young students in Varnhagen's (2000) study, the undergraduates gave harsher evaluations of writing ability when texts contained spelling errors. Similarly, Kreiner, Schnakenberg, Green, Costello, and McClin (2002) conducted several experiments with 82 college students, and observed that the participants were critical of the authors of texts with spelling errors. Martin and Ranson (1990) provided an observation that might explain the negative impact of spelling errors for readers of varying ages. In particular, they pointed out that spelling errors prevented readers from focusing on the message of a given text because readers were distracted from the message while mentally restoring intended correct forms from words with spelling errors. Considered together, these

[1] In order to distinguish study participants who read and evaluated texts with spelling errors from professional raters who were trained to assign scores to texts for an assessment, we call the former "readers" and the latter "raters" throughout this paper.

findings suggest that spelling errors had a negative impact on untrained readers' perceptions of text quality and the authors' writing ability.

Researchers have also examined whether the magnitude of the negative impact is affected by the characteristics of spelling errors. One of the experimental conditions of Kreiner *et al.* (2002) involved the number of spelling errors included in texts. They found a positive relationship between the quantity of spelling errors and their negative impact; texts with only a few spelling errors did not receive significantly lower scores from the readers, whereas a large number of spelling errors evidently led the readers to give poor scores. The contribution of spelling error types has shown rather mixed results. Figueredo and Varnhagen (2005) reported that their undergraduate readers gave more critical ratings for non-homophone errors than for homophone errors. However, Boland and Queen (2016) found both simple typos and homophonous grammatical errors similarly damaging. These mixed results may have to do with the diversity of spelling errors found in texts. Wilcox, Yegelski, and Yu (2014) examined high school students' writing and noted that the eight most frequently observed spelling error types accounted for only about a half of the entire errors in their sample, indicating a considerable level of diversity. With such a diverse set of spelling errors, consistent classification of errors is very difficult.

As Lunsford and Lunsford (2008) pointed out, most spelling error studies have focused on L1 writing of college students. L2 learners' spelling errors have often been studied with young learners in relation to the spelling errors of young L1 writers. Lesaux, Koda, Siegel, and Shanahan (2006) conducted a systematic review of studies about L2 English learners' spelling skill mastery, and conclude that, in many empirical comparisons, young L1 and L2 writers demonstrated comparable spelling performance. Lesaux *et al.* also noted studies indicating that L2 learners' spelling skill could be predicted by the same set of predictors used to predict that of L1 writers (e.g., Da Fontoura & Siegel 1995; Chiappe & Siegel, 1999; Arab-Moghaddam & Sénéchal, 2001; Abu-Rabia & Siegel, 2002). Zhao, Quiroz, Dixon, and Joshi (2016) offered a similar conclusion based on their meta-analysis of research comparing the spelling skill of young English monolinguals and bilinguals. These reviews suggest that acquisition of spelling skill is similar between young L1 and L2 writers. However, the findings about young L2 learners' spelling skill mastery do not necessarily generalize to spelling performance of adult L2 learners. A recent study by Doolan (2017) suggests that adult L2 learners have more difficulty in learning correct spelling than young learners.

Spelling errors of adult L2 learners have been examined in several recent studies in the context of spellchecker development. Although spellcheckers are one of the most widely used tools for writing in digital environments (MacArthur, 1999), spellcheckers designed for L1 writers may not perform as well for L2 writers (Rimrott & Heift, 2005). Researchers arguing for spellcheckers for L2 learners have carefully observed spelling errors made by L2 learners and compared them with spelling errors of L1 writers (e.g., Mitton & Okada, 2007; Hovermale, 2008; Rimrott & Heift, 2008). Bestgen and Granger (2011) examined 223 essays written by adult English L2 learners as responses to assessment tasks and reported that the L2 writers in their sample tended to make more spelling errors than L1 writers, and that the types of L2 learners' spelling errors were different from those of L1 writers. They also note that L2 writers' spelling errors were a significant predictor of raters' evaluation of essay quality. Flor, Futagi, Lopez, and Mullholland's (2015) survey of a corpus consisting of GRE and TOEFL iBT responses also indicates the inverse relationship between the proportion of spelling errors in a given response and the score assigned to the response. However, Flor *et al*. also report that several characteristics of spelling errors varied as a function of scores rather than following the L1 vs. L2 distinction. In particular, they note that the severity of spelling errors was more strongly affected by the writing proficiency of a test taker rather than whether the test taker was an L1 or L2 writer.

Despite the interest in spelling errors of L2 learners, relatively little is studied about the impact of

spelling errors on raters' scoring decisions in an adult L2 writing assessment context. This is partly due to the general notion that spelling errors are not necessarily indicative of adult L2 writers' overall writing ability, and therefore, should not be the focus of an adult L2 writing assessment. This view is reflected in the fact that spelling errors are rarely mentioned or mentioned only when they interfere with meaning in many scoring rubrics for large-scale writing assessments for university admission purposes, such as the IELTS Academic Test (Cullen, French, & Jakeman, 2014) and the TOEFL iBT® Test (ETS, 2012). However, research has indicated that trained raters do not arrive at a final score strictly following a given set of evaluation criteria. Lumley (2002), for example, demonstrated that even trained and experienced raters struggle between their personal impression and the scoring criteria when evaluating written texts, and that their judgments could be affected by surface features, such as spelling errors. Moreover, it has been acknowledged that different raters appear to adopt different processes to make their scoring decision (e.g., Cumming, 1990; Weigle, 1999; Barkaoui, 2010). The findings from the study conducted by Flor *et al*.(2015) suggest that even trained rater's judgment of response quality may be affected by the presence of spelling errors. However, to our knowledge, there has been no direct investigation into whether trained raters would assign different scores to the same set of responses with and without spelling errors in an adult L2 writing assessment context.

Research Questions

The goal of this study was to investigate whether spelling errors would have an impact on trained raters' scoring decisions. If a meaningful impact exists, it was of our interest to estimate its size. We also examined whether the potential impact of spelling errors would remain consistent across different assessment tasks and varying degrees of error quantity and quality. In sum, this study was guided by the following research questions:

1. Do spelling errors have an impact on trained raters' scoring decisions for responses to L2 writing assessment tasks?
2. Does the impact of spelling errors on trained raters' scoring decisions vary across assessment tasks and error characteristics?

Method

Instruments

This study was part of a larger project investigating experimental writing tasks for adult L2 learners with two types of writing tasks. The first type, which we call "source synthesis," presented a reading passage about a topic, followed by an audio-recorded interview conversation about the same topic, and asked test takers to write a summary synthesizing the two sources for a naive reader who is unfamiliar with the topic. The second type, on the other hand, was contextualized in an online discussion forum for a hypothetical class, in which a discussion about a topic was taking place. Test takers were given two previous forum posts representing opposite views, and were instructed to post their own view about the topic while drawing from either or both of the two previous posts. In the remainder of this paper, we call this second type "the online forum task." Each task type was represented by two tasks designed to be parallel with each other, with the only difference being the topic (i.e., each task had its own topic). The four experimental tasks were assembled into two computer-delivered forms (Form 1 and Form 2). Each form consisted of a source synthesis task and an online forum task, in that order. Because the two tasks within a task type were designed to be parallel, the two forms were also designed to parallel. The forms were administered in multiple testing centers under a standardized setting, and up to 45 minutes were allowed for completion: 20 minutes for the source synthesis task and 25 minutes

for the online forum task.

Responses

A total of 788 adult L2 learners participated as test takers in the larger project. Each test taker was randomly assigned to one of the two forms, and responded to the tasks in the assigned form, resulting in a total of 1,576 responses. Given the focus of this study on spelling errors, we excluded responses without any spelling errors and responses with only obvious typographical errors (i.e., spelling error caused by the proximity of the location of key and a careless typing, such as *ans* for *and, freedon* for *freedom*). This process left 148 responses, which we analyzed for this study. The length of the selected responses (i.e., the number of total words in a response) ranged from 51 to 311, with mean 188 and standard deviation 62. The number of spelling errors in the 148 responses ranged from 1 to 30, with mean 12.7 and standard deviation 5.5. The corresponding proportion of spelling errors ranged from 1 percent to 20 percent, with mean 7 percent and standard deviation 3 percent. A breakdown of these 148 responses by form and task type in Table 1 shows that the selected sample contained more Form 1 responses than Form 2, and more source synthesis responses than online forum, as can be seen in Table 1. However, Table 1 also shows that even the smallest cell had more than 20 responses. The 148 responses were written by test takers from 22 different countries, with individuals from Mexico, Colombia, and China being the most common in the sample. The test takers also spoke a total of 15 different L1s, with Spanish (35 percent), Portuguese (11 percent), and Chinese (11 percent) comprising the majority of the sample and were roughly balanced in gender (52 percent female and 48 percent male).

Table 1. *Breakdown of analyzed responses into form and task type*

	Source Synthesis	Online Forum
Form 1	54	29
Form 2	41	24

Error Correction

We evaluated the impact of spelling errors by comparing human rater scores on two different versions of each response: the original version, which included the original spelling errors made by test takers, and the corrected version, in which the spelling errors were corrected. The corrected versions were obtained in the following manner. We first reviewed the original responses using ConSpel (Flor & Futagi, 2013), a computer program for automatic spelling error detection and correction. The resulting flagged errors were manually examined to address both false-positives (e.g., uncommon proper names flagged as an error) and false-negatives (e.g., writing "dan" for "and", but not flagged because "dan" is a common first name). The combination of the automatic detection and the manual examination yielded a total of 1,809 errors across the 148 responses. We then classified the errors into two groups. The first group (Group 1 errors), which comprised the majority of the errors (92 percent), consisted of clear typos and/or spelling errors (e.g., *biirds* for *birds*, *sso* for *so*). The second group (Group 2 errors) contained cases for which the correct form could be figured out from the context, but the errors were attributed to lexical, morphological, or grammatical errors or combination of such (e.g., preparate for *prepared*, *beautifulest* for *beautiful*). Therefore, Group 1 errors asked for little cognitive load for correction (or might not even register as errors in some cases), Group 2 errors required complex inferences regarding the intended correct forms. We note that this classification was far from a comprehensive taxonomy of L2 English writers' spelling errors; instead, our goal was to have a binary classification that represents the difficulty of error correction (error quality as described in the next subsection) in a reliable manner. More example errors in each group are given in the Appendix.

Although Group 2 errors accounted for a small portion of the total errors (8 percent), about a half of the responses (53 percent) included at least one such error.

Error Characteristics – Quantity & Quality

We operationalized the quantity of errors for each response in two ways: the total number of corrected errors in the response, and the number of corrected errors divided by the total number of words in the response (called "the error density" in the remainder of this paper). We also examined multiple aspects of error quality. The first aspect addressed the impact of different types of spelling errors in the original version, and was operationalized using the two groups of errors we introduced earlier. In particular, we distinguished responses with only Group 1 errors and the other responses with at least one non-Group 1 error. The second aspect of error quality addressed the severity of an error, which we quantified with edit distance (Levenshtein, 1966). Edit distance represents the minimum number of edits needed to correct a word with spelling error(s) to obtain the corresponding correct form. A spelling error with edit distance of one indicates that the deletion, addition, or substitution of one character was required to correct that spelling error. The larger the edit distance is, the larger the discrepancy between the incorrect and correct forms. In other words, large edit distance represents a severe error. We calculated edit distance of each corrected error and created two measures capturing the error severity of a given response based on individual edit distance: the proportion of errors with edit distance greater than one (hereafter "the severe error proportion"), and the largest edit distance among the corrected errors of a given response (hereafter "the max distance"). The severe error proportion was to capture overall error severity by measuring the share of errors that required multiple editing for correction, while the max distance was designed to monitor the role of the most severe error that was corrected in a given response.

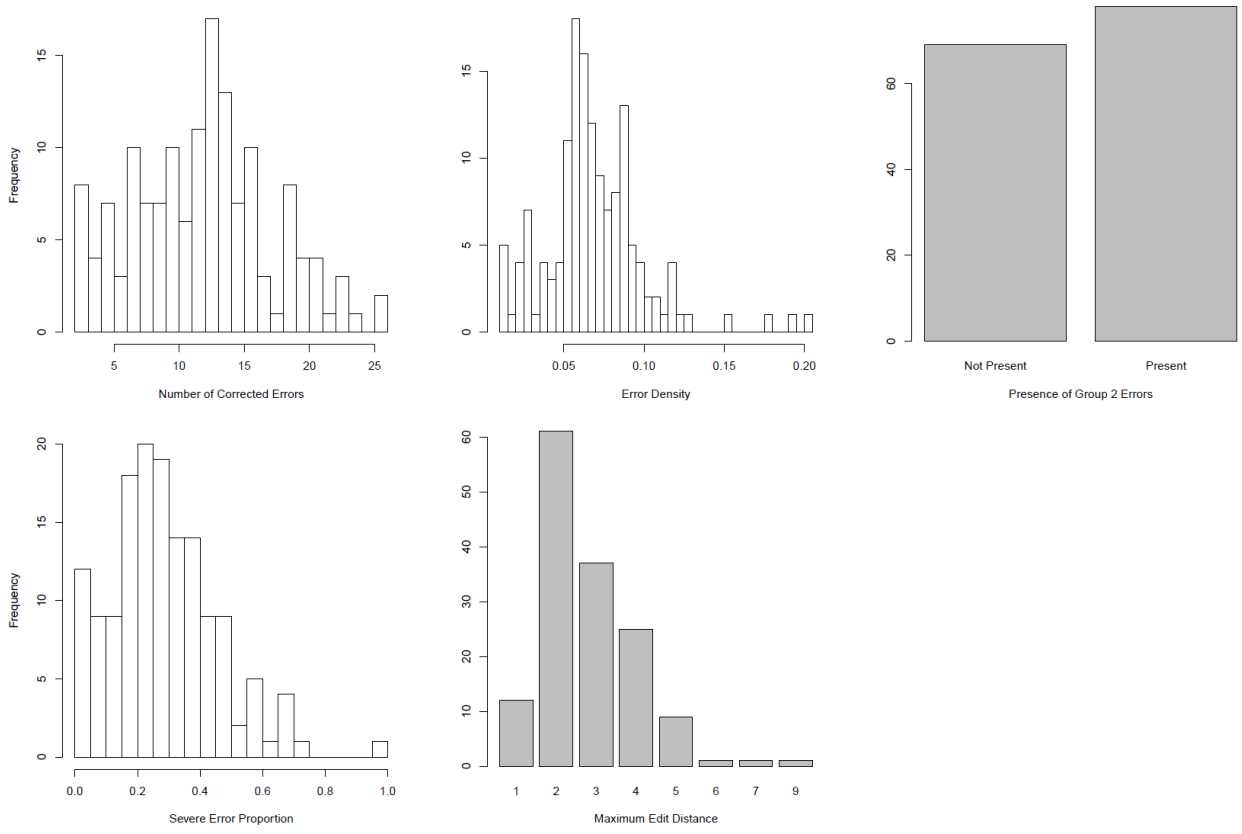

Figure 1. Distribution of error quantity and quality predictors

We then examined how the error quantity and quality variables were distributed, as can be seen in

Figure 1. Figure 1 shows that the distributions of the total number of corrected errors and the error density were both roughly symmetrical. The responses were evenly divided between the two groups of errors, with 47 percent having only Group 1 errors and 53 percent with at least one Group 2 error. The severe error proportion was slightly skewed with a right tail; most of the responses (95 percent) had less than 60 percent of errors with edit distance greater than one. The max distance distribution was also slightly skewed with a right tail, with the majority ranging from two to four. Overall, the univariate distributions of the error quantity and quality variables did not signal any irregularities that would discourage their use as predictors in the subsequent modeling procedure.

Scoring

Both the original and corrected versions were scored by a group of raters. All raters had a Bachelor's degree or higher and had experience in professionally scoring adult L2 learner essays for English language proficiency assessments. Before scoring the responses, the raters went through online training in which they practiced scoring responses to the experimental writing tasks with benchmark samples and scoring rubrics. Each response was scored independently by two raters based on scoring rubrics developed for the experimental tasks. The scoring rubric for the source synthesis tasks included three key aspects: Language Control, Completeness/Accuracy of source information, and Organization/Coherence. On the other hand, the online forum scoring rubric focused on two aspects, which were Language/Expression and Contribution to the Discussion. Both rubrics were designed to evaluate a given response in a holistic manner on a zero to five scale. The raters went through a training session with several practice scoring examples and were given benchmark samples for each score category. Because spelling errors were not part of the evaluation criteria, no attention was drawn to spelling errors during training and the directions to raters did not make any reference to how spelling errors should be considered.

After the training, each version of each response was scored by two randomly assigned raters, and no rater scored both versions of the response from the same test taker. The participating raters were aware that they were scoring for a research project but did not know the purpose of the research. Thus, the raters who were scoring the corrected responses did not know that spelling errors had been corrected. The inter-rater agreement between the two raters was measured with quadratically weighted Kappa (Cohen, 1968): 0.67 for the original responses and 0.66 for the corrected responses. We used the average of two rater scores as the holistic score of each response. Therefore, the holistic scores were on the scale of zero to five, with an increment of 0.5. For convenience, we call the holistic scores on the original and corrected versions "the original response score" and "the corrected response score," respectively, in the remainder of this paper.

Analysis

We addressed the two research questions by comparing the original and corrected response scores. In order to answer the first research question, we examined whether the mean original response score differed significantly from the mean corrected response score using a paired-samples t test. For the second research question, we used a linear regression model to examine whether the score differences between the two response versions could be explained by task types and error characteristics. In particular, we modeled the score difference between the two versions of each response, which we obtained by subtracting the original response score from the corrected response score (hereafter "the gain score"), with predictors representing task types and characterizing the quantity and quality of spelling errors. Although the gain scores were ordinal variables with a boundary set by the scale points (i.e., cannot be larger than 5 or smaller than -5), we decided to rely on a linear model because of the relatively large number of score categories (in our data, ranging from -0.5 to 3 with 0.5 increase).

Results

Comparison between Original and Corrected Response Scores

We first examined how the original and corrected response scores were distributed. The scatter plot in Figure 2 below shows their bivariate distribution, with each dot representing a response.

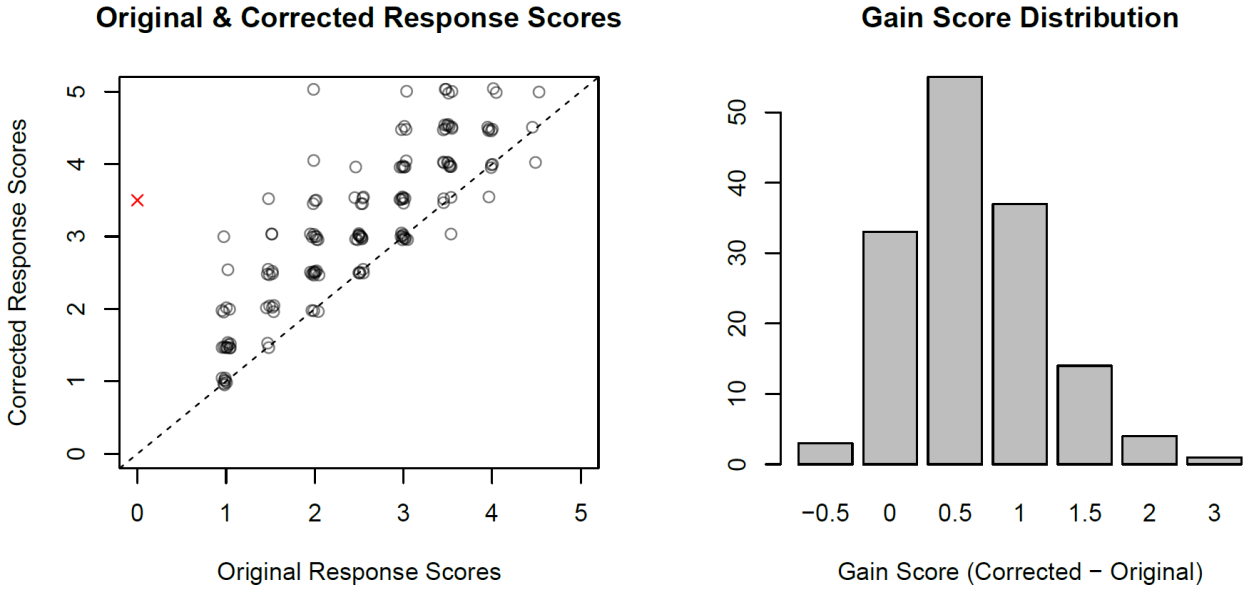

Figure 2. Bivariate distribution of original and corrected response scores (with random jitters to show density; left) and gain score distribution (right).

The bivariate distribution shows one outlier (denoted by "x"), which was excluded in the gain score distribution in the right panel. The scatter plot shows one outlier with its original response score equal to zero (red colored "x" in the plot). Upon reading this response, it was unclear why the original version of this response received the extreme score, which is reserved for an empty (i.e., blank submission) or irrelevant response. The response was not empty (> 200 words), and it partially addressed the given topic, as supported by the non-zero score given to its corrected version. Because of the relatively small size of our sample, this outlier with a somewhat ambiguous reason for the extreme score could have had shifted analyses results. Therefore, we excluded this outlier from all subsequent analyses. After removing the outlier, we calculated the gain score of each response, and examined how the resulting gain scores were distributed. The distribution of the gain scores is also given in Figure 2.

Both the scatter plot and the bar plot in Figure 2 indicate a strong pattern that, in general, the corrected responses received higher holistic scores than the corresponding original responses. Most of the dots in the scatter plot lay above the 45 degree line, and the bar plot shows that almost all responses had a non-negative gain score. The mean of the gain scores was 0.65, with the standard deviation of 0.56. Given the double scoring of each version, one of the two raters giving one score point higher leads to a 0.5 increase of the corresponding holistic score. The average gain of 0.65 indicates that the score gain from the error correction, on average, was larger than the minimum score increment of 0.5. The paired t test showed that the mean difference between the original and corrected scores was statistically significant ($t = 13.87$ with 146 degrees of freedom; $p < .01$). The corresponding effect size (Cohen's D, corrected for dependence following Morris and DeShon (2002)) was 1.16, indicating a large effect.

Relationship of Score Gains with Task Types and Error Characteristics

The strong pattern of positive and significant gain scores and the large effect size indicated that the

raters generally evaluated the corrected responses more favorably than the original responses. With this finding, we proceeded to address the second research question by modeling the gain scores with our knowledge of the tasks and the corrected errors. In particular, we examined how the gain scores varied across different tasks, and whether the magnitude of a gain score could be accounted for by the quantity and quality of spelling errors. As we mentioned earlier, the four tasks were classified by type ("source synthesis" and "online forum") and form ("Form 1" and "Form 2"). We used dummy coding for both the task type (0 for "source synthesis" and 1 for "online forum") and form (0 for "Form 1" and 1 for "Form 2") and included the resulting task and form indicators as predictors in modeling the gain scores.

We then predicted the gain scores using the task type and form and five error predictors (i.e., the total number of corrected errors, the error density, the presence/absence of Group 2 errors (0 for absence and 1 for presence), the severe error proportion, and the max distance). We also included the original response scores as an additional predictor to control for the quality of original responses. Because the intercept of a linear regression model represents the average outcome variable when all predictor variables are 0, we centered the error predictors such that the resulting predictors have zero means, and shifted the original version score downward by one (resulting a scale of 0 to 4) to facilitate the interpretation of the model intercept. The regression model results are given in Table 2 below.

Table 2: *Regression of gain scores on task and error variables*

	Coefficient*	Standard Error
Intercept	**0.98**	**0.12**
Online Forum (vs. Source Synthesis)	**-0.23**	**0.09**
Form 2 (vs. Form 1)	0.03	0.09
Number of Corrected Errors	**0.04**	**0.01**
Error Density	**-5.28**	**2.25**
Max Distance	**0.09**	**0.04**
Severe Error Proportion	-0.40	0.29
Group 2 Error Present (vs. Absent)	-0.03	0.09
Original Version Score	**-0.16**	**0.06**
R^2	0.15	

Notes: *: bolded coefficients were significant at the α=0.05 level

Table 2 shows that the fitted regression model accounted for approximately 15 percent of the variance of the gain scores. Given the small number of predictors and the inherent uncertainty in human rater scores, the large remaining variance in the gain scores is not surprising. Because of the centering and shifting of the predictors, the estimated intercept of 0.98 in Table 2 represents the average gain score of the responses to the source synthesis task (task code 0) in Form 1 (form code 0), with the average number of corrected errors (which was 12.2), the average error density (which was 0.07), the average severe error proportion (which was 0.29), the average maximum edit distance (which was 2.8), and the average original response score (which was 2.5), without any Group 2 errors. The average gain score of online forum responses was lower than that of the source synthesis task by 0.23 (coefficient for Online Forum in Table 2), holding all other predictors constant. This difference was statistically significant. The two error quantity predictors (i.e., the number of corrected errors and error density) were both significant, although the directions were different. On average, responses with more

corrected errors tended to have higher gain scores. The model estimated that each additional corrected error was associated with 0.04 unit increase in the gain score (coefficient for Number of Corrected Errors in Table 2). On the other hand, responses with higher error density, on average and holding all other predictors constant, had lower gain scores. The coefficient for the error density was large in absolute value (-5.28 in Table 2), but this was an artifact of the error density scale; the estimated coefficient represented the average decrease in gain score by 0.0528 associated with the one percent increase in the error density by one percent. The max distance was the only significant predictor among the three measures of error quality. The sign of the slope coefficient for the max distance was in the expected direction in that corrected responses with large max distance values tended to receive larger gains. Similarly, the negative and significant impact of the original response score was reasonable considering the ceiling effect; when the original score was already high (e.g., 4 or 4.5), there is not much room to get a higher score. This trend was also shown in the left panel of Figure 2.

Summary and Discussion

We observed that the mean score of the corrected responses was significantly higher than the mean score of the original responses. The size of this mean difference (0.65) was larger than one standard deviation of the gain scores (0.56), indicating a large effect size of our spelling correction. This finding answers the first research question with empirical evidence that spelling errors can affect trained raters' scoring decisions. We then fit the regression model in Table 2 to the gain scores to explore the relationships among the impact of spelling error correction and the quality and quantity of corrected errors. The model accounted for approximately 15 percent of the variance in the gain scores. We observed that the gain scores significantly differed across the two task types, the number of corrected spelling errors, the error density, the max distance, and the original response scores. On the other hand, the size of the score gain due to spelling error correction remained stable across test forms, and was not significantly predicted by the presence/absence of the Group 2 errors or the severe error proportion. These results from the regression model suggest that the second research question can be answered with a conditional statement. In our data, the impact of spelling errors on trained raters' scoring decisions varied across task types, the quantity of corrected errors, the severity of the biggest error corrected (i.e., the max distance), and original response scores.

It was noteworthy that the impact of spelling error correction differed across the two task types. Although we do not have data to identify the reasons for the differential impact, our conjecture is that it may have to do with the different genres of responses elicited by the two task types, as well as the different scoring rubrics. The source synthesis tasks asked for a summary of two source materials, and therefore, accuracy of content was explicitly mentioned as a key aspect in the scoring rubrics. On the other hand, the online forum tasks elicited an argumentative essay, and the corresponding scoring rubrics did not explicitly mention accuracy. Moreover, raters might have spotted spelling errors for the source synthesis tasks more easily than for the online forum tasks, because the two sources would have allowed them to form clear expectations for the content of responses.

Another significant predictor of score gains was the quantity of corrected errors. We believe that this finding can be interpreted in a straightforward manner; the more spelling errors corrected, the larger the magnitude of score gain due to the correction. The difference between the largest (26) and the smallest (2) number of corrected errors in our data was 24 (note that these numbers differ from the minimum and maximum number of all errors, which were 1 and 30, because not every error was corrected). Given the estimated coefficient for the number of corrected errors (i.e., 0.04), our model predicts that this extreme difference would lead to a one-point score gain on a five point scale from spelling error correction. The amount of spelling errors has been identified as a factor contributing to untrained readers' perceptions of L1 writing quality (Kreiner *et al*., 2002). Although we cannot directly

compare our estimate with the estimates from the previous studies due to the differences in scoring rubrics, our finding was consistent with findings from previous studies despite the contextual differences. However, we also observed that the error density had a negative impact on the gain score, which was not expected. This might reflect the possibility that responses with a large number of spelling errors relative to its word count might have other issues (e.g., grammar, word choice, or organization) than spelling errors, although we do not have data to examine this conjecture.

It was also noteworthy that the severe error proportion or the presence of the Group 2 errors was not a significant predictor of score gains. We do not have data to identify reasons for these results. The raters might not have disproportionately penalized severe errors in the original responses to begin with, which could explain the lack of a significant effect of these severity measures. This may also indicate the difference between untrained readers and trained raters. While the untrained readers in a previous study evaluated essays with certain types of spelling errors more harshly (Figueredo & Varnhagen, 2005), the trained raters in our study did not appear to be affected by the quality of spelling errors. The scoring rubrics used by the raters in this study did not explicitly mention spelling errors at all, and the training could have helped the raters treat typographical errors of different degrees of severity more or less the same, which can be an encouraging sign from a rater training perspective.

Implications

In this study, we have directly compared rater scores on adult EFL learners' writing samples with and without spelling errors. To our knowledge, an empirical comparison between trained raters' judgment on assessment responses with and without spelling errors has not been reported in the literature. The findings of this study, therefore, have several substantive and practical implications. From a substantive perspective, our findings indicate that some findings from untrained readers' perception on naturalistic writing with spelling errors can be generalized to a more formal assessment setting with trained raters. We observed that, overall, the trained raters in this study evaluated corrected responses without spelling errors more favorably than the corresponding original responses with spelling errors, as untrained readers had done in previous studies. Moreover, the magnitude of score gains from spelling error correction was significantly impacted by the quantity of corrected errors, which was also in line with the findings of the previous studies with untrained readers. However, the trained raters in this study differed from untrained readers in a previous study (Figueredo & Varnhagen, 2005) in that their scoring decisions did not appear to have been influenced by the types or proportion of spelling errors, which could be potentially attributed to their training and scoring rubrics. In sum, we believe that our findings have shown the importance of empirical investigations for generalizing previous findings about the impact of spelling errors on text quality judgment to a new context, especially when the new context differs considerably from the previously studied contexts.

The significant difference between the holistic scores of original and corrected responses has an important bearing on the training of raters and the meaning of the score. We believe that depending on the design principles of a writing assessment, spelling errors can be regarded differently, ranging from a major contributing factor for the broader concept of accuracy to an isolated mechanical aspect irrelevant to the intended interpretation of an assessment score. Our findings suggest that trained raters' scoring decisions may be affected by the presence of spelling errors, even when the scoring rubrics do not explicitly mention spelling errors. Assessment developers should decide whether such a tendency is in line with the design principles of their assessment. If such a tendency is well aligned to the assessment's design principles, then it would be desirable to specifically instruct raters to consider spelling in their scoring decision, and to inform test takers of the potential impact of spelling errors. On the other hand, if the impact of spelling errors on scoring decisions represents construct irrelevant variance, rater training should be designed to discourage raters from using spelling errors as one of the

criteria for evaluating responses.

We believe that the spelling errors we corrected in this study are likely to be corrected by test takers if they are given a simple spellchecker that only flags typographical errors (without flagging grammatical errors). Therefore, the findings of this study can also be interpreted as a potential impact of allowing the use of spellcheckers on rater scores. We observed that, even after controlling for the error quantity and quality predictors, the magnitude of the average gain score was smaller for responses with high original scores. In other words, the impact of error correction could differ across different levels of original scores. This is not surprising given the ceiling effect and the magnitude of the effect was not large. However, we interpreted this as an indication that allowing a simple spellchecker might introduce differential impact for test takers at different levels of writing proficiency. We also note that allowing a spellchecker can introduce other complicating factors, such as whether it should provide suggestions, who chooses to use a spellchecker, and how well test takers use it, which should be studied carefully.

The magnitude of the overall score gain from spelling error correction is also pertinent to the introduction of a spellchecker in an existing writing assessment. The spelling correction led to an average of more than a half point score gain. This positive impact of spelling error correction can be an issue if a spellchecker is introduced in the middle of the life cycle of a writing assessment. Assuming most test takers would use a spellchecker if available, our findings suggest that test takers who use a spellchecker to correct errors may receive a higher score than test takers who possess comparable proficiency levels but took the test before the spellchecker was allowed, just because these new test takers would have fewer spelling errors. This can undermine the comparability of scores across different test administrations, which is particularly problematic for standardized assessments. Based on our findings, we recommend that assessment developers carefully consider the impact of spelling error correction on rater scores in their own contexts to gauge the magnitude of potential score gains and use the empirical results to establish procedures to ensure the comparability of scores from different administrations.

The size of the error correction impact in this study is specific to the scoring rubrics we used. We acknowledge that the size of the impact cannot translate neatly into the scales of other rubrics. Given the lack of a universal scale for evaluating L2 learners' responses to assessment tasks, our view is that the impact of spelling error correction should be studied and interpreted within a specific scale used for an empirical investigation. Therefore, we believe that assessment developers who are interested in training raters to ignore spelling errors or providing a spellchecker for test takers should conduct an empirical investigation in their context. This study can be a template for such studies in terms of data collection and analysis methodologies, as well as the interpretation of findings. We also believe that an empirical investigation into raters' perception of spelling errors would make a promising line of research. More information about whether and how much raters vary in their recognition and processing of spelling errors will be of great practical importance for rater training and scoring rubric development.

References

Abu-Rabia, S., & Siegel, L. S. (2002). Reading, syntactic, orthographic, and working memory skills of bilingual Arabic-English speaking Canadian children. *Journal of Psycholinguistic Research, 31*, 661-678. https://doi.org/10.1023/A:1021221206119

Anson, I. G., & Anson, C. M. (2017). Assessing peer and instructor response to writing: A corpus analysis from an expert survey. *Assessing Writing*, *33*, 12-24. https://doi.org/10.1016/j.asw.2017.03.001

Arab-Moghaddam, N., & Sénéchal, M. (2001). Orthographic and phonological processing skills in reading and spelling in Persian/English bilinguals. *International Journal of Behavioral Development, 25*, 140-147. https://doi.org/10.1080/01650250042000320

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience, *Language Assessment Quarterly, 7*, 54-74. https://doi.org/10.1080/15434300903464418

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

Boland, J. E., & Queen, R. (2016). If you're house is still available, send me an email: Personality influences reactions to written errors in email messages. *PLoS ONE, 11* (3), e0149885. https://doi.org/10.1371/journal.pone.0149885

Chiappe, P., & Siegel, L. S. (1999). Phonological awareness and reading acquisition in English- and Punjabi-speaking Canadian children. *Journal of Educational Psychology, 91*, 20-28. http://dx.doi.org/10.1037/0022-0663.91.1.20

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70* (4), 213-220. http://dx.doi.org/10.1037/h0026256

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

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31-51. https://doi.org/10.1177/026553229000700104

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

Doolan, S. M. (2017). Comparing patterns of error in generation 1.5, L1, and L2 first-year composition writing. *Journal of Second Language Writing, 35*, 1-17. https://doi.org/10.1016/j.jslw.2016.11.002

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*, 37-64. https://doi.org/10.1177/0265532207071511

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

Figueredo, L., & Varnhagen, C. K. (2005). Didn't you run the spell checker? Effects of type of spelling error and use of a spell checker on perceptions of the author. *Reading Psychology, 26*, 441-458. https://doi.org/10.1080/02702710500400495

Flor M., & Futagi Y. (2013). Producing an annotated corpus with automatic spelling correction. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use – Proceedings 1* (pp. 139-154). Louvain-la-Neuve: Presses universitaires de Louvain.

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

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.

Johnson, A. C., Wilson, J., & Roscoe, R. D. (2017). College student perceptions of writing errors, text quality, and author characteristics. *Assessing Writing*, *34*, 72-87. https://doi.org/10.1016/j.asw.2017.10.002

Kreiner, D. S., Schnakenberg, S. D., Green, A. G., Costello, M. J., & McClin, A. F. (2002). Effects of spelling errors on the perception of writers. *Journal of General Psychology, 129*, 5-17. https://doi.org/10.1080/00221300209602029

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.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*, 707-710.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*, 246-276. https://doi.org/10.1191/0265532202lt230oa

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

Martin, C. L., & Ranson, D. E. (1990). Spelling skills of business students: An empirical investigation. *The Journal of Business Communication, 27*, 377-400.

Mitton, R., & Okada, T. (2007). *The adaptation of an English spellchecker for Japanese writers.* London: Birkbeck ePrints. http://eprints.bbk.ac.uk/archive/00000592

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105-125. http://dx.doi.org/10.1037/1082-989X.7.1.105

Pearson, H. (2012). Issues in the assessment of spelling. *Literacy Learning: The Middle Years., 20*, 29-33. https://search.informit.com.au/documentSummary;dn=419212829571994;res=IELHSS

Rimrott, A., & Heift, T. (2005). Language learners and generic spell checkers in CALL. *CALICO Journal, 23*, 17-48. http://www.jstor.org/stable/24156231

Rimrott, A., & Heift, T. (2008). Evaluating automatic detection of misspellings in German. *Language Learning & Technology, 12*, 73-92. http://dx.doi.org/10125/44156

Varnhagen, C. K. (2000). Shoot the messenger and disregard the message? Children's attitudes toward spelling. *Reading Psychology, 21*, 115-128. https://doi.org/10.1080/02702710050084446

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*, 145-178. https://doi.org/10.1016/S1075-2935(00)00010-6

Wilcox, K. C., Yagelski, R., & Yu, F. (2014). The nature of error in adolescent student writing. *Reading and Writing, 27*, 1073-1094. https://doi.org/10.1007/s11145-013-9492-x

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

Author Biodata

Ikkyu Choi is a research scientist at Educational Testing Service. His research interests include second language development profiles, test taking processes, and scoring of constructed responses.

Yeonsuk Choi is a research scientist at Educational Testing Service, Princeton, NJ, USA. Her current work focuses on the development and validation of language tests for adults and young learners.

Acknowledgements

We thank Larry Davis, Michael Flor, Jakub Novak, and Don Powers for their advice on an earlier draft. Any remaining flaws are entirely our own.