The Impact of Misspelled Words on Automated Computer Scoring: A Case Study of Scientific Explanations

Minsu Ha & Ross H. Nehm

Journal of Science Education and Technology

ISSN 1059-0145

J Sci Educ Technol
DOI 10.1007/s10956-015-9598-9

Springer

J Sci Educ Technol
DOI 10.1007/s10956-015-9598-9

The Impact of Misspelled Words on Automated Computer Scoring: A Case Study of Scientific Explanations

Minsu Ha[1] · **Ross H. Nehm**[2]

Abstract Automated computerized scoring systems (ACSSs) are being increasingly used to analyze text in many educational settings. Nevertheless, the impact of misspelled words (MSW) on scoring accuracy remains to be investigated in many domains, particularly jargon-rich disciplines such as the life sciences. Empirical studies confirm that MSW are a pervasive feature of human-generated text and that despite improvements, spell-check and auto-replace programs continue to be characterized by significant errors. Our study explored four research questions relating to MSW and text-based computer assessments: (1) Do English language learners (ELLs) produce equivalent magnitudes and types of spelling errors as non-ELLs? (2) To what degree do MSW impact concept-specific computer scoring rules? (3) What impact do MSW have on computer scoring accuracy? and (4) Are MSW more likely to impact false-positive or false-negative feedback to students? We found that although ELLs produced twice as many MSW as non-ELLs, MSW were relatively uncommon in our corpora. The MSW in the corpora were found to be important features of the computer scoring models. Although MSW did not significantly or meaningfully impact computer scoring efficacy across nine different computer scoring models, MSW had a greater impact on the scoring algorithms for naïve ideas than key concepts. Linguistic and concept redundancy in student responses explains the weak connection between MSW and scoring accuracy. Lastly, we found that MSW tend to have a greater impact on false-positive feedback. We discuss the implications of these findings for the development of next-generation science assessments.

Keywords Computer scoring · Open-ended assessment · Misspelled words · Machine learning · Misclassification · Computers · Assessment

Introduction

The new *Framework for Science Education* (NRC 2012) emphasizes three major competencies necessary for student participation in twenty-first-century science: (1) engaging in scientific practices such as explanation and argumentation, (2) utilizing cross-cutting concepts such as cause/effect and structure/function, and (3) grounding disciplinary reasoning in core ideas (e.g., natural selection and plate tectonics). Scientific practices such as explanation and argumentation are foundational scientific activities because they can help to foster deep engagement with the processes of science through debate, discussion, analysis, and critique (NRC 2012). These practices require oral or written language, and because of this, careful analysis of language by teachers and researchers is increasingly central to science assessment (Federer et al. 2014). To this end, science teachers, education researchers, and assessment developers are increasingly relying upon short answer and essay tasks (Beggrow et al. 2014; Bridgeman et al. 2012; Haudek et al. 2012; Linn et al. 2014). However, the time, expertise, and costs associated with assessing short answer and essay responses are formidable. Inconsistent human scoring and

✉ Minsu Ha
msha@kangwon.ac.kr

[1] Division of Science Education, College of Education, Kangwon National University, Hyoja-dong, Chuncheon-si, Gangwon-do 200-701, South Korea

[2] Center for Science and Math Education, Department of Ecology and Evolution, Stony Brook University (SUNY), 092 Life Sciences Building, Stony Brook, NY 11794-5233, USA

Published online: 30 January 2016

delayed feedback to students pose additional challenges to analyzing open-ended text (Nehm et al. 2012). To addresses these long-standing problems, automated computer scoring systems (ACSSs) are being developed for both formative (classroom-based) and summative (high-stakes) assessments of arguments and explanations (see Moharreri et al. 2014 for a review). In addition, research groups are forming in order to develop and use ACSS in STEM education (see Haudek et al. 2011; Linn et al. 2014).

Regardless of the assessment context (formative or summative), ACSS faces the long-standing and seemingly trivial problem of misspelled words (MSW). Although humans can effortlessly and automatically identify the equivalence of the word "explanation" and the MSW "explination," computers cannot. "Spell-check" and "autocorrect" functions are standard components of many word processing programs, but they are not standard parts of most ACSS. There are many reasons for this situation. One reason is that educators would like students to learn how to spell words correctly independent of technological scaffolds. A second is that spell-check and autocorrect systems are not perfect; they can introduce errors in both spelling detection and spelling alternatives (as many smartphone users are keenly aware). A third is that such a system could be used to suggest terms or words during a high-stakes test, thereby introducing construct-irrelevant variance into knowledge measurement (cf. AERA et al. 2014).

Even if we brush aside the problem of MSW as a problem that will inevitably be solved by technology, the fact remains that the spell-check systems currently available are *not* integrated into most educational technology systems, raising the question of the degree to which current ACSS is impacted by this simple but important issue (Ha et al. 2011). No study to our knowledge has explored the degree to which MSW impact the automated analysis of biological explanations, which will be a growing target of text analysis given the widespread adoption of the Next Generation Science Standards (NGSS; NRC 2013). The overarching aim of this study is to empirically explore the degree to which MSW impact ACSS accuracy and discuss the implications of these findings for automated analysis of scientific practices in educational settings.

Misspelled Words: Frequency and Identification in Written Text

Misspelled words are common in student writing (Connors and Lunsford 1988), and because of this there has been much prior work on this topic in English writing assessments and natural language processing research. Misspelled words have been identified as one of the most frequent errors in students' essays. For example, Connors and Lunsford (1988) reported that approximately 25 % of all errors found in 300 students' essays were misspelled words. More surprisingly, Lunsford and Lunsford (2008) found that spelling errors comprised more than 6 % of all errors detected in a national sample of 3000 college composition essays (even though students were allowed to use spell-check systems). Clearly, MSW are a common feature of human-generated text.

Because MSW are ubiquitous features of written text, natural language processing (NLP) research has developed automatic detection and correction systems to address this issue (Flor and Futagi 2012). One approach for automatic detection of MSW is achieved by comparing human-generated text to the words in a dictionary. Computational algorithms suggest candidate words for MSW using edit distance (e.g., Damerau 1964; Levenshtein 1966) and phonetic similarity (e.g., Flor and Futagi 2012). Flor and Futagi's (2012) empirical study indicated that state-of-the-art automatic MSW *detection* programs worked well, exceeding 99 % recall measures (i.e., the percentage of MSW detected by programs divided by the total number of MSW) and nearly reaching 99 % precision measures (i.e., the percentage of cases correctly labeled as MSW among MSW labeled by the program).

Although new versions of automatic MSW *correction* programs (i.e., ConSpel-B) are characterized by significant improvements over previous programs (e.g., Aspell, and MS words, see Flor and Futagi 2012), they continue to display significant limitations (precision and recall <80 %). That is, the automatic program did *not* correct more than 20 % of the MSW in the text (i.e., <80 % recall) and more than 20 % of the suggestions by the correction program were wrong (i.e., <80 % precision). Moreover, the dictionaries that spell-check programs are often based upon commonly lack domain-specific words. For example, this manuscript is being prepared in Microsoft Word for Mac 2011, and common biology words such as "macromutation," "megastrobili," and "paedomorphosis" are all identified as MSW (even though they are not). Biology terms commonly used for more than 50 years, such as "mammalogy" are also shown to be MSW. Clearly, MSW detection and correction programs have errors of their own and currently are not able to eliminate the problem of MSW in human-generated text, particularly in terms of biological language.

Automated Scoring of Text Using Machine Learning Methods

Machine learning is a major area of computer science research that explores how computers can utilize patterns in human behaviors and judgments to build predictive models that can be used to categorize future cases (for a review, see Abu-Mostafa 2012). Machine learning is

becoming an increasingly central part of humans' everyday lives: movie recommendations, online product suggestions, and targeted advertising are common applications. In order to describe how machine learning methods work, a simple example may be helpful. If a restaurant owner had a large data set listing menu orders along with the time of day that they were ordered (afternoon, early evening, late evening), she could use machine learning to detect patterns in the data and build an algorithm that accounts for these patterns. Then, the model could be used to predict which menu items would be most likely to be ordered in future days depending on the time of day.

Similarly, machine learning methods may be used to build computer scoring models of students' written responses. *Scoring* a written response is a human judgment based on rules. For example, if a human grader classified students' written responses as being emblematic of a "positive attitude" or a "negative attitude" using a scoring rubric, the machine could build a model that accounts for the text associated with each category (positive or negative). Then, using a new set of responses, the machine could make an appropriate prediction about which responses are likely to have that attribute.

Figure 1 summarizes how machine learning can develop a scoring algorithm using human-scored text. First, the software extracts attributes or features from a text corpus (e.g., words, combinations of words) that might discriminate positive and negative cases (e.g., the presence or absence of the idea of biological variation in a text response). Then, the machine learning software determines which features differentiate positive and negative cases. For example, "mutation," "variation," "genetic difference" could be text features that identify the presence of *biological variation* in a text response. Software is then used to build a statistical model using these distinguishing features. This model is the scoring algorithm that is applied to future cases. It is typically necessary to refine the model as more and more data are applied to the model. Iterative cycles of feature extraction, model building, and testing are used until a robust model is produced (see Moharreri et al. 2014 for a more detailed description).

Machine learning methods are able to use linguistic corpora (e.g., databases of written text). MSW, spacing errors, and other issues can limit the accurate detection of a concept or idea in written text (Flor and Futagi 2012). For example, if a response includes "motation [sic]" or "geneticmutation [sic]" instead of "mutation," human graders will most likely be able to recognize that these cases are MSW and classify them appropriately (e.g., as a science concept being present in an response). However, computers cannot infer the true meaning of these MSW or spacing errors and could erroneously classify a case as lacking an idea (thus, mislabeling the data). Misclassified data hinder appropriate algorithm extractions in the machine learning process (Flor and Futagi 2012; Nagata et al. 2011; Muhlenbach et al. 2004).

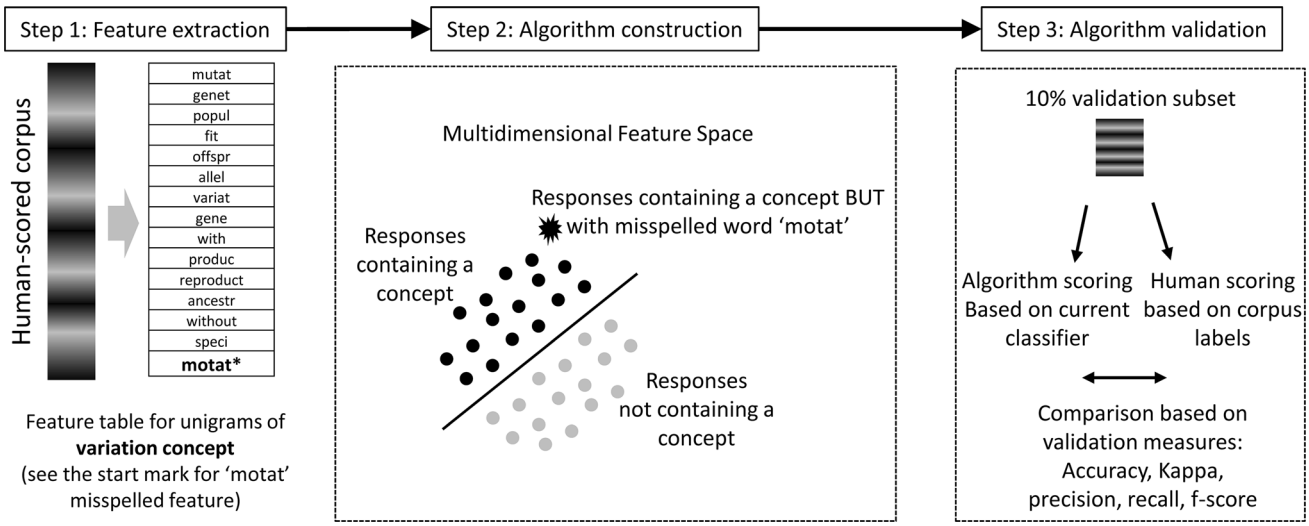


Fig. 1 The process of machine learning to build automated computer scoring models showing the potential problem of misspelled words. Step 1: The available features (words, word stems, etc.) are extracted from the corpus. Very uncommon features shown <5 times and stop words (the most common words in English, e.g., "the") are not included in the feature table; Step 2: Machine learning software builds an algorithm using statistical analyses that detect differences of features between responses containing a certain concept (labeled by a human grader) and responses not containing a certain concept (labeled by human grader); Step 3: The algorithm generated is tested using a fresh sample (typically a 10 % set aside data set). Machine learning software compares its algorithm performance with human scoring

Impact of MSW on Computer Scoring Systems

Four issues motivated our study of the impact of MSW on ACSS. The first is the need to improve the scoring accuracy of ACSS (cf. Nagata et al. 2011). ACSS relies upon consistent language features. If the training corpus (i.e., human-scored student responses) contains MSW, but the testing corpus (the new responses to be scored) does not, then the algorithm will misclassify responses. Similarly, if the testing corpus (i.e., students' written responses to be scored) contains MSW, but the training corpus does not, then the algorithm cannot appropriately classify the student responses with MSW. If cases of MSW are manually removed, then the machine loses data that might be relevant for building a robust algorithm for new sets of responses. Consequently, a less accurate scoring algorithm could be generated. In short, MSW can impact the efficacy of the scoring algorithms.

A second issue relates to the usability of ACSS in classrooms employing electronic response systems. Recent literature on the use of "clickers" in large classes (e.g., introductory biology, with >500 students) has revealed that rapid and immediate feedback by an instructor plays a vital role in improving students' learning and metacognition (Chen et al. 2010; Brady et al. 2013). Although using clickers with open-ended text is not common, it will likely increase given the growing use of smart phones and other devices in classrooms (Kucirkova et al. 2014). If an instructor needs to correct students' MSW prior to analysis in an ACSS, then such rapid feedback will not be possible. If responses are collected for review in the next class, spell correction will still take time. For example, in some biology student response corpora ($n = 2000$), about 550 MSW are typically found (Authors, unpublished data). Assuming it takes 5 s to correct a MSW using an automated spell-check function in a program (e.g., MS Excel), it would take an instructor about 45 min make corrections prior to analysis.

The third issue is that English language learners (ELLs) comprise an increasingly large proportion of US science classrooms (Flynn and Hill 2005), and they are more likely to make spelling errors than native speakers (Bebout 1985; Flor and Futagi 2012; Haggan 1991). Learners of English as a second language (ESL) tend to generate more MSW than native English speakers for many reasons (Haggan 1991), and so many scholars have noted that language proficiency could bias the assessment of ELLs using ACSS (see Abedi 2004; Abedi et al. 2004). Given that ELLs tend to make more spelling errors, and spelling errors may impact the efficacy of ACSS, the impact of spelling on scoring becomes an important assessment topic in need of empirical investigation.

Lastly, as the automated scoring of text moves into more interactive contexts, such as personalized cognitive tutors, the impact of MSW on feedback becomes particularly important. False-positive and false-negative feedback could be differentially impacted by MSW. In false-positive feedback, ACSS informs the test-takers that they are successful even though they were not. In false-negative feedback, ACSS informs test-takers that they are *not* successful even though they were. ACSS appears to produce equal numbers of false-positive and false-negative feedback (see Ha and Nehm 2012), although empirical studies of the impact of spelling on such feedback are limited. The type of feedback that is provided to students is known to impact their self-efficacy. Karl et al. (1993), for example, argued that positive feedback tends to promote self-efficacy more than negative feedback. This idea has been supported by some empirical tests (see, for example, Holroyd et al. 1984). Although Holroyd et al.'s (1984) study did not examine the effect of false-negative feedback, Fitzsimmons et al. (1991) explored this issue. Their results showed that athletes that were provided with false-negative feedback were able to lift less weight and showed poorer bench press performance relative to those who were provided with false-positive feedback. Although these findings are unlikely to directly translate into educational settings, they do suggest that false-negative feedback *could* differentially impact student self-efficacy. In sum, MSW are not a trivial problem in ACSS and deserve more empirical attention as assessments become increasingly text-based.

Research Questions

Our study explored four research questions relating to misspelled words and text-based computer assessments of biological explanations:

- RQ1 Do ELLs produce equivalent magnitudes and types of spelling errors as non-ELLs?
- RQ2 What percentage of MSW relate to concept-specific scoring rules?
- RQ3 What impact do MSW have on computer scoring accuracy?
- RQ4 Are MSW more likely to impact false-positive or false-negative feedback?

Instrument Used to Generate the Text Responses

A universally recognized core idea in the sciences is natural selection (e.g., AAAS 2011; NRC 2012), and explaining evolution by natural selection is a core disciplinary practice (NRC 2012; Opfer et al. 2012). Evolution, like most fields of biology, is jargon-rich and provides a

fruitful context for studying the impacts of spelling errors on computer scoring. EvoGrader (Moharreri et al. 2014) is a new ACSS designed to measure the degree to which students use core scientific ideas (or "key concepts") to build normative evolutionary explanations. EvoGrader automatically scores text answer to the ACORNS (Assessing Contextual Reasoning about Natural Selection) instrument (Nehm et al. 2012). The scoring models in the current version of EvoGrader were built using a human-scored corpus that lacked MSW.

The ACORNS is an open-response formative assessment instrument that prompts students to explain how changes in populations or species occur through time. The item prompts are isomorphic: "A [species/population] of *X* [lacks/has] *Y*. How would biologists explain how a [species/population] of *X* [with or without] *Y* evolved from an ancestral *X* [species/population] [with or without] *Y*?" (Note that $X =$ taxa and $Y =$ traits; many different options are available in the ACORNS assessment). The data used in the current study are from students' written responses to 21 different ACORNS items (e.g., evolution of poisonous snails, penguins inability to fly, and roses with and without thorns). See www.evograder.org for additional details.

Human and Computer Scoring of Text Responses

Student responses to the ACORNS prompts were scored in two ways. First, the presence or absence of nine core concepts was noted in each student's written response (see Table 1 for details). Six normative scientific concepts relating to natural selection (i.e., variation, heritability, competition, limited resources, differential survival/reproduction, and non-adaptive ideas) and three "misconceptions" or non-normative naïve ideas (i.e., needs/goals, use/disuse, and adapt/acclimation) were scored. These concepts have been shown to be important for measuring students' reasoning about evolution (e.g., Bishop and Anderson 1990; Nehm and Reilly 2007; Nehm and Schonfeld 2007; Nehm et al. 2010) (See Table 1). A variety of forms of validity evidence have been gathered for the ACORNS instrument (see Beggrow et al. 2014; Moharreri et al. 2014 for recent reviews).

In addition to noting the presence or absence of the nine concepts in each written response, answers were coded into one of four possible reasoning patterns or models (i.e., a "scientific model" [including only normative scientific ideas], a "mixed model" [including both scientific and naïve ideas], a "naïve model" [including only non-normative naive ideas], or "no model" [rephrasing the question, not answering with relevant information]. Two human raters (a Ph.D. student in biology education and an evolutionary biologist) used the scoring rubrics of Nehm et al. (2010) to code all responses. All human-scored data exhibited acceptable inter-rater agreements (>0.8 κ for all normative scientific ideas and non-normative naive ideas). Consensus scores were established in cases of disagreement based upon discussions between the two raters.

EvoGrader (the ACSS used in our study) currently has nine scoring models built using machine learning methods that align with the nine concepts that were scored by humans (see above and Table 1). EvoGrader was built using machine learning tools (for details, see Moharreri et al. 2014). The software package LightSIDE (see Mayfield and Rosé, in press) forms the core of the EvoGrader system. LightSIDE is a free, open-source machine learning software package available from LightSide Labs (see http://lightsidelabs.com). EvoGrader has been shown to be able to score students' responses as accurately as trained human raters (see Beggrow et al. 2014 and Moharreri et al. 2014 for details).

Table 1 Concept types, names, and descriptions of scoring models

Concept type	Concept name	Concept description
Normative scientific idea	Variation	The presence and causes (mutation/recombination/sex) of variation
	Heritability	The heritability of variation (The degree to which a trait is transmitted from parents to offspring)
	Competition	A situation in which two or more individuals struggle to get resources that are not available to everyone
	Limited resources	Limited resources related to survival/reproduction, such as food and predators, and reproduction (such as pollinators)
	Differential survival	The differential reproduction and/or survival of individuals
	Non-adaptive idea	Genetic drift and related non-adaptive factors contributing to evolutionary change
Non-normative naïve idea	Adapt/acclimation	Adjustment or acclimation to circumstances (which may subsequently be inherited)
	Need/goal	Goal-directed change; needs as a direct cause of evolutionary change
	Use/disuse	The use (or lack of use) of traits directly causes their evolutionary increase or decrease

The nine concepts that we studied are characterized by different words and linguistic structures. For this reason, the machine learning parameters that are used to extract features from the corpus (and build the scoring algorithms) need to be different for each of the nine concepts. These parameters include N-Gram selection (i.e., the number of contiguous feature sequences), stemming (i.e., grouping words based on common stems), and removing stop words (e.g., very common words such as "a," "the," "and"). Prior research has established the most effective machine learning parameters for these nine concepts (see Nehm et al. 2012). The scoring models (derived from the extraction parameters) were built using tenfold validation using Sequential Minimal Optimization (SMO; see Platt 1999 for details). Given that each scoring model for each concept is built using *different* parameters, it is likely that MSW could *differentially* impact scoring models (and resulting scoring efficacy). For this reason, our analyses focus on the impact of MSW collectively as well as for each concept.

Scales of Analysis: Measuring the Impact of Misspelled Words

The scale of analysis will impact measures of the impact of MSW on scoring efficacy. We can consider four different analysis scales when interpreting the impact of MSW on scoring efficacy: *class level, student level, individual response level,* and *individual concept level* (see Table 2). For example, an instructor who teaches an undergraduate biology class with 300 students might administer four ACORNS items as a diagnostic pretest. This will result in 1200 written responses. The ACSS will then predict the presence/absence of nine concepts and produce 10,800 predictions. The accuracy of these predictions will depend on how the instructor chooses to use the results; that is, accuracy measures will depend on whether the results are used to characterize the class, a student (all items), an individual response from one item, or an individual concept. Different statistical tests are needed for different levels of analysis (Table 2).

There are four levels that can be used to compare the impact of MSW on ACSS performance (that is, scores produced by the computer using text response either *with* student-generated MSW or *without* [i.e., human-corrected] MSW): *class, student, response,* and *concept levels* (see Table 2). The largest grain size of analysis is the *classroom level*. In this case, total class scores are compared in the two conditions. The second level of analysis is the *student level*; here, the ranks or percentiles of students are compared in the two conditions. The third level is the *response level.* Here total scores are compared across different items in the two conditions. (It is important to note that *student level* and *response level* will be the same in cases in which only one item is used). The fourth level is the *concept level*. This level focuses on differences between specific concept scores in the two conditions (e.g., "variation" scores) (see Table 2). As is apparent, there are many ways of measuring the impact of MSW on computer scoring performance.

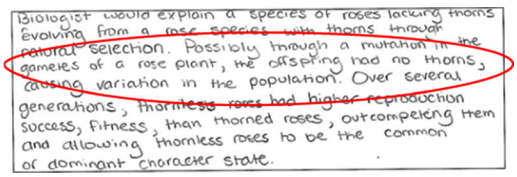
Different comparison levels require different methods of analysis (Table 2, right column). The most widely used inter-rater reliability measure is Cohen's kappa, which can be used to examine the degree to which the two conditions align for each *concept* in each *item* (i.e., the *individual concept* level of analysis). Thus, Cohen's kappa is the finest grain size measure of scoring efficacy. Moving up to the *individual response* level (that is, scores for *a single answer* from an individual *student*), Spearman correlations of the *total* key concept scores (normative scientific concepts) or *total* naive idea scores (non-normative "misconceptions") of each item between the human and computer scores or the Cohen's kappa for the four reasoning model types can be calculated. Although the range of key concept (0–6) and naïve idea (0–3) scores in each *response* is limited, at the *student* level (among items for one student) the range of key concepts is 0–24 and the range of naïve ideas is 0–12 (given four items) and well suited to Pearson correlation methods. (It should be noted that Pearson correlation methods for *student-level* comparisons were not performed in Study 3 because only one item was used) Finally, at the *class level*, calculations of percentage differences were made. For example, if the computer scoring model generated 230 wrong predictions out of 10,800 cases, the percentage of incorrectly predicted cases would be 2.12 %. We used these different levels and different methods to quantify the impact of MSW on computer scoring efficacy. Recall that *all* responses in our study were scored by the computer *and* by trained human raters. We conducted three studies as outlined below.

Methods

Study 1

The sample for Study 1 included information on whether student participants were ELL or native speakers of English. The goal of Study 1 was to determine whether the magnitudes and types of MSW differed between ELL and non-ELL students. The sample used for Study 1 included 1988 college students (1529 native English speakers and 459 ELLs). This group of participants generated written responses to two ACORNS items. As in all of our studies, student responses were scored by both human raters (see above) and the ACSS. Percentages and types of MSW were characterized for both groups. In addition, each MSW was examined to determine whether it was related to a

Table 2 Four different analysis levels used in this study to compare computer scoring performance with and without misspelled words

Level	Statistical method to check correspondence between human grader and ACSS
Class 	Total number of errors caused by the ACSS divided by the total ACSS predictions for one class
Student 	Pearson correlation of student total score Cohen's kappa to compare students' group (e.g., percentile group, different performance group) labeled by ACSS and human grader
Response 	Spearman rho correlation of item total score Cohen's kappa to compare reasoning model type of each response labeled by ACSS and human grader
Concept 	Cohen's kappa Agreement percentage Precision Recall F_1 score

computer scoring rule or not. Summary statistics and percentages were calculated and compared in both groups.

Study 2

The sample for Study 2 did not include information on whether students were ELL or non-ELL. The goals of Study 2 were to quantify how the presence of MSW in the training corpus (i.e., the corpus to be used for building the scoring algorithm) and the testing corpus (i.e., the corpus to be used to examine the robustness of the scoring algorithm) impacted the measurement of student understanding.

Study 2 used 10,270 written responses to 21 ACORNS items in order to test the impact of MSW on computer scoring efficacy. To assemble the testing corpus, we randomly selected 2000 responses (500 students $\times$ 4 items) from our total corpus (10,270 written student responses). The remaining set of responses ($n = 8270$ responses) was used as the training corpus. Recall that all responses in this corpus were human-scored (see above) and computer scored.

Statistical Methods

We used Cohen's kappa, raw percentage agreement, precision, recall, and F_1 scores to measure correspondence magnitudes between human and computer scores for the nine concepts at the individual *response level* (see Table 2). Specifically, we calculated correspondence measures in four different conditions: (1) non-spelling checked (NSC) training corpus and non-spelling checked testing corpus, (2) NSC training corpus and spelling checked (SC) testing corpus, (3) SC training corpus and NSC testing corpus, and (4) SC training corpus and NSC testing corpus.

Repeated-measures ANOVAs were used to examine the magnitudes of impact of MSW in both the training and the testing corpora on these correspondence measures (e.g., kappa, agreement percentage) across the nine concepts. We also used Cohen's kappa and raw percentage agreement to measure the correspondence of human and computer scoring for the four reasoning model types in order to

explore the impact of MSW on the accuracy of the ACSS at the individual response level (scientific, mixed, naïve, none; see above). Pearson correlations were also used to measure correspondence magnitudes between human and computer scores for total key concept and naïve idea scores at the student level. Finally, in order to evaluate the overall impact of MSW on ACSS efficacy we calculated false-positive and false-negative prediction percentages among the total predictions associated with MSW frequency.

In this study, five correspondence measures were used: (1) Cohen's kappa, (2) raw agreement percentage, (3) precision, (4) recall, and (5) F_1 score (Bejar, 1991). Cohen's kappa is a widely used measure for inter-rater reliability because it accounts for chance agreements. Different cutoff scores for Cohen's kappa have been introduced in the literature. We used the benchmarks suggested by Landis and Koch (1977): "moderate" agreement (Kappa between 0.41 and 0.60), substantial agreement (Kappa between 0.61 and 0.80) and "almost perfect" agreement (Kappa between 0.81 and 1.00; Landis and Koch 1977). Agreement percentage is a simple measure; it is the percentage of total agreements between human and ACSS scores relative to the total number of cases. Precision, recall, and F_1 scores are widely used measures in information retrieval studies (Su, 1994). Precision indicates the percentage of correct predictions among total positive predictions, whereas recall indicates the percentage of correct predictions among actual positive cases. For example, if the ACSS predicts concept A occurs in students' explanations 100 times, and 80 cases were correctly identified, then precision is 80 %. In this situation (precision is low), the ACSS *overestimates* students' performance. If students' explanations contain 100 instances of concept A but the ACSS identifies only 80 of these cases, then recall is 80 %. In this situation (when recall is low), the ACSS *underestimates* students' performance. The F_1 score is a weighted average of precision and recall.

Pearson correlation coefficients were used to examine score correspondences for two different "holistic" variables (i.e., total key concept scores and naïve idea scores). Pearson correlation coefficients have been used for evaluating correspondence magnitudes in many different fields (see Sato et al. 2005; Zhu et al. 2002). Many studies consider Pearson correlation coefficients >0.9 to be "nearly identical" (e.g., Sato et al. 2005; Zhu et al. 2002); this is the cutoff that we adopted.

Study 3

The sample for Study 3 did not include information on whether students were ELL or non-ELL. The goal of Study 3 was to quantify the impact of MSW on scoring accuracy using a different corpus than that used in Study 2, thereby testing the generalizability of findings. In Study 3, the testing corpus included students' responses to novel ACORNS-like questions (similar in form but differing in exact language from standard items). The ACSS in this case was trained by students' responses to the typical ACORNS instrument.

The sample for Study 3 included 285 undergraduate students' written responses to two novel ACORNS-like items (evolution of white beetles and rose thorns—Corpus A) from one institution and 1112 undergraduate students' written responses to one ACORNS item (evolution of orchids—Corpus B), but it was collected at a greater diversity of academic institutions ($n = 59$). The scoring models used in Study 3 were trained using all of the 10,270 human-scored written responses used in Study 2. It should be noted that the training set for Study 3 is smaller than Study 2 (e.g., 8270 vs. 10,270).

Statistical Methods

The same correspondence measures were used as in Study 2. Repeated-measures ANOVAs were used to examine the impact of MSW on scoring model efficacy. Given relatively small sample sizes, prior to performing the ANOVAs we performed tests of normality. Kolmogorov–Smirnov tests and Shapiro–Wilk tests, along with descriptive statistics of skewness and kurtosis, revealed that the vast majority of cases met appropriate assumptions of the ANOVA. However, we found two cases that did not meet such assumptions. We decided to use repeated-measure ANOVAs for all analyses but supplement the two non-normal cases with results from the nonparametric Wilcoxon signed ranks test. As we report below, the results produced by the two methods did not differ.

We did not test the effect of MSW in the training corpus in Study 3 because this research question was already examined in Study 2. However, as in Study 2, we examined the impact of MSW on the testing corpus. Another difference between the two studies is that in Study 3 students were only asked to write a response to one item (therefore, student-level and response-level results were the same). We used Spearman correlation coefficients to calculate human–computer correspondence for total key concepts and naive ideas in each response. Like Study 2, we also calculated the percentage of wrong predictions among total predictions associated with MSW frequency in order to evaluate the overall impact of MSW on the efficacy of the ACSS.

Results

Study 1

For RQ1, we compared the number of and types of MSW between ELL and non-ELL students. Capitalization errors

(e.g., "THis"), apostrophe errors (e.g., "werent," "didnt"), stop words (e.g., "a," "the," "and"), or species or trait names (e.g., pluegone, labiatae) were not characterized as MSW. As shown in Table 3, our analysis revealed that ELLs generated approximately two times as many MSW compared to native English speakers (0.824 % for ELL and 0.420 % for non-ELLs). In addition, the percentage of MSW that were scoring-rule-related was also two times higher for ELLs than non-ELLs (45.5 % for ELLs and 24.2 % for non-ELLs). Overall, MSW were quite uncommon in both groups (Table 3).

Study 2

Overall, MSW were also relatively uncommon in the corpus used for Study 2. As in Study 1, we did not consider capitalization errors (e.g., "THis"), apostrophe errors (e.g., "werent," "didnt"), stop words (e.g., "a," "the," "and"), or species or trait names (e.g., pluegone, labiatae) as MSW. Using this approach, 2675 words (1.9 %) out of 143,018 words in the training corpus ($n = 8270$ responses) contained MSW. Out of 143,018 words, the number of *scoring-rule-related* words (that is, words central to identifying the concept in the rubrics, see Table 4) was 35,787 (25.02 %). Among the MSW in the training corpus, 772 (28.9 % of the total MSW) were *scoring-rule-related* words. The 772 *scoring-rule-related* MSW included 396 different MSW; that is, on average, each MSW appeared twice (1.95 times). Similar to what we found in the training corpus, the testing corpus ($n = 2000$) contained 542 MSW (1.7 %) out of 32,380 total words. Out of 32,380 words, the number of *scoring-rule-related* words was 5217 (16.11 %). Among the MSW, 135 (24.91 %) were *scoring-rule-related* (Table 4). The 135 *scoring-rule-related* MSW included 93 different types of MSW; that is, on average each MSW appeared 1.45 times. The most common MSW in the training corpus was "environment." This word is an important feature for the scoring algorithms, particularly for the concepts of "limited resources" and "adapt/acclimation" (see Table 1). Students also made a substantial number of MSW for "competition," "heritable," "advantageous," "necessary," "in order" (e.g., "inorder"—spacing errors) that are core terms related to key concepts and naïve ideas.

At the *individual concept level* (see Table 2), repeated-measure ANOVAs indicated that there were no significant effects of MSW in the training and testing corpora, and no significant interaction effects on kappa, agreement percentage, precision, and F_1 score (see Table 5; Kappa: training data: $F[1,8] = 4.372$, $p = 0.070$, $\eta_p^2 = 0.353$, testing data: $F[1,8] = 0.224$, $p = 0.649$, $\eta_p^2 = 0.027$, interaction: $F[1,8] = 0.590$, $p = 0.464$, $\eta_p^2 = 0.069$; Agreement percentage: training data: $F[1,8] = 2.51$, $p = 0.152$, $\eta_p^2 = 0.239$, testing data: $F[1,8] = 1.14$, $p = 0.318$, $\eta_p^2 = 0.124$, interaction: $F[1,8] = 0.80$, $p = 0.397$, $\eta_p^2 = 0.091$; Precision: training data: $F[1,8] = 3.05$, $p = 0.119$, $\eta_p^2 = 0.276$, testing data: $F[1,8] = 1.05$, $p = 0.337$, $\eta_p^2 = 0.116$, interaction: $F[1,8] = 0.94$, $p = 0.360$, $\eta_p^2 = 0.105$; F_1 score: training data: $F[1,8] = 4.45$, $p = 0.068$, $\eta_p^2 = 0.358$, testing data: $F[1,8] = 0.14$, $p = 0.722$, $\eta_p^2 = 0.017$, interaction: $F[1,8] = 0.59$, $p = 0.465$, $\eta_p^2 = 0.068$).

At the individual concept level, we found significant but weak effects of MSW in the testing corpus, no significant effects of MSW in the training corpus, and no interaction effects (Recall: training data: F [1,8] $= 3.05$, $p = 0.119$, $\eta_p^2 = 0.276$, testing data: $F[1,8] = 6.05$, $p = 0.039$,

Table 3 The comparison of misspelled words in student explanations between English native and ELL students

	N	Total words (excluding stop words)	# of words per answer	# of total MSW	Percentage of MSW (%)	Scoring-rule-related words in MSW	% of Scoring-rule-related words in MSW (%)
English native	1529	35,502	23.2	149	0.420	36	24.2
ELL	459	9346	20.4	77	0.824	35	45.5

Table 4 The numbers of words, MSW, and MSW related to concepts in three evolutionary responses data sets in Study 2

Data sets	Total words (A)	Total words except stop words (B) (B/A)	Scoring-rule-related words in total words (B)	MSW (C) (C/B)	Scoring-rule-related words in MSW (D) (D/C)	Total # of kind of MSW
Training corpus (n of essay = 8270)	266,640	143,018 (53.64 %)	35,787 (25.02 %)	2675 (1.87 %)	772 (28.86 %)	396
Testing corpus ($n = 2000$)	60,602	32,380 (53.43 %)	5217 (16.11 %)	542 (1.67 %)	135 (24.91 %)	93

Table 5 The differences of efficacy (e.g., kappa and agreement percentage) of automated computer scoring models for non-spelling checked (NSC) and spelling checked (SC) training and testing corpora (TR: Training corpus and TE: Testing corpus) in Study 2

Concept	NSC–TR/NSC–TE	NSC–TR/SC–TE	SC–TR/NSC–TE	SC–TR/SC–TE
Kappa				
Variation	0.872	0.875	0.876	0.880
Heritability	0.838	0.838	0.851	0.848
Competition	0.962	0.962	0.962	0.962
Limited resources	0.948	0.949	0.956	0.957
Differential survival	0.846	0.845	0.844	0.843
Non-adaptive idea	0.978	0.968	0.978	0.968
Need/goal	0.841	0.848	0.838	0.845
Use/disuse	0.752	0.760	0.768	0.776
Adapt/acclimation	0.726	0.721	0.726	0.732
Mean	0.863	0.863	0.867	0.868
SD	0.089	0.087	0.087	0.083
Repeated-measure ANOVA statistics	Training data: $F[1,8] = 4.372$, $p = 0.070$, $\eta_p^2 = 0.353$			
	Testing data: $F[1,8] = 0.224$, $p = 0.649$, $\eta_p^2 = 0.027$			
	Interaction: $F[1,8] = 0.590$, $p = 0.464$, $\eta_p^2 = 0.069$			
Agreement percentage				
Variation	94.7	94.9	94.9	95.1
Heritability	97.2	97.2	97.4	97.4
Competition	99.9	99.9	99.9	99.9
Limited resources	98.3	98.3	98.5	98.6
Differential survival	92.5	92.4	92.4	92.3
Non-adaptive idea	99.9	99.9	99.9	99.9
Need/goal	94.3	94.6	94.2	94.5
Use/disuse	98.2	98.2	98.3	98.3
Adapt/acclimation	95.6	95.5	95.6	95.7
Mean	96.7	96.7	96.8	96.8
SD	2.6	2.6	2.7	2.6
Repeated-measure ANOVA statistics	Training data: $F[1,8] = 2.51$, $p = 0.152$, $\eta_p^2 = 0.239$			
	Testing data: $F[1,8] = 1.14$, $p = 0.318$, $\eta_p^2 = 0.124$			
	Interaction: $F[1,8] = 0.80$, $p = 0.397$, $\eta_p^2 = 0.091$			
Precision				
Variation	91.2	91.4	91.7	91.9
Heritability	96.4	96.4	96.5	96.5
Competition	100.0	100.0	100.0	100.0
Limited resources	97.2	96.8	98.1	97.9
Differential survival	93.7	93.3	93.7	93.4
Non-adaptive idea	100.0	97.9	100.0	97.9
Need/goal	92.5	93.2	92.3	92.8
Use/disuse	72.8	73.2	73.5	73.8
Adapt/acclimation	76.3	75.4	76.3	75.8
Mean	91.1	90.8	91.4	91.1
SD	9.9	9.8	9.8	9.6
Repeated-measure ANOVA statistics	Training data: $F[1,8] = 3.05$, $p = 0.119$, $\eta_p^2 = 0.276$			
	Testing data: $F[1,8] = 1.05$, $p = 0.337$, $\eta_p^2 = 0.116$			
	Interaction: $F[1,8] = 0.94$, $p = 0.360$, $\eta_p^2 = 0.105$			
Recall				
Variation	90.6	90.9	90.8	91.1
Heritability	76.5	76.5	78.4	77.9

Table 5 continued

Concept	NSC–TR/NSC–TE	NSC–TR/SC–TE	SC–TR/NSC–TE	SC–TR/SC–TE
Competition	92.9	92.9	92.9	92.9
Limited resources	94.7	95.2	95.0	95.4
Differential survival	88.7	89.1	88.5	88.7
Non-adaptive idea	95.8	95.8	95.8	95.8
Need/goal	83.5	83.9	83.3	83.9
Use/disuse	79.7	81.1	82.4	83.8
Adapt/acclimation	73.7	73.7	73.7	75.4
Mean	86.2	86.6	86.7	87.2
SD	8.2	8.1	7.7	7.4
Repeated-measure ANOVA statistics	Training data: $F[1,8] = 3.05$, $p = 0.119$, $\eta_p^2 = 0.276$			
	Testing data: $F[1,8] = 6.05$, $p = 0.039$, $\eta_p^2 = 0.430$			
	Interaction: $F\ [1,8] = 0.52$, $p = 0.493$, $\eta_p^2 = 0.061$			
F_1 score				
Variation	90.9	91.2	91.2	91.5
Heritability	85.3	85.3	86.5	86.2
Competition	96.3	96.3	96.3	96.3
Limited resources	96.0	96.0	96.5	96.6
Differential survival	91.2	91.1	91.0	91.0
Non-adaptive idea	97.9	96.8	97.9	96.8
Need/goal	87.8	88.3	87.6	88.1
Use/disuse	76.1	76.9	77.7	78.5
Adapt/acclimation	75.0	74.6	75.0	75.6
Mean	88.5	88.5	88.9	89.0
SD	8.4	8.2	8.1	7.8
Repeated-measure ANOVA statistics	Training data: $F[1,8] = 4.45$, $p = 0.068$, $\eta_p^2 = 0.358$			
	Testing data: $F[1,8] = 0.14$, $p = 0.722$, $\eta_p^2 = 0.017$			
	Interaction: $F[1,8] = 0.59$, $p = 0.465$, $\eta_p^2 = 0.068$			

$\eta_p^2 = 0.430$, interaction: $F\ [1,8] = 0.52$, $p = 0.493$, $\eta_p^2 = 0.061$). The statistical power for the significant effects that we found was <0.6 (0.58), and so this finding should be interpreted with caution.

Recall is a measure of the percentage of correct predictions among total positive cases. For the nine concepts that we studied, the scoring algorithm for "use/disuse" exhibited the lowest recall (1.35 %) followed by the scoring algorithms for "adapt/acclimation" (0.84 %), "needs/goals" (0.51 %), and "limited resources" (0.46 %). However, the frequency of the "use/disuse" concept in the testing corpus was very low ($n = 74$); thus, these results (1.35 % of the 74 cases) are explained by only one instance. Similarly, only 1.5 (0.84 %) of the 179 "adapt/acclimation" cases were false negatives, and 2.5 (0.51 %) of the 490 "needs/goals" cases were false negatives. In sum, 1492 MSW in the training corpus, and 135 MSW in the testing corpus were scoring-rule-related, but these MSW contributed to about three error cases (out of 2000).

At the *individual response* level (see Table 2), we also examined kappa values and raw agreement percentages for the four *reasoning models* (e.g., scientific model, mixed model, naïve model, and no model). We found that the impact of MSW in the training and testing corpora was miniscule [non-spelling checked (NSC) training/NSC testing: 87.9 %, and 0.819 κ; NSC training/spell-checked (SC) testing: 88.1 % and 0.821 κ; SC training/NSC testing: 87.8 % and 0.818 κ; SC training/SC testing: 88.0 % and 0.820 κ].

At the *student response level* (see Table 2), the results of the Pearson correlations indicated that MSW did not meaningfully impact the efficacy of the scoring algorithm (Table 6). Compared to human-scored key concept scores, the NSC training corpus and the NSC testing corpus had correlations of 0.957 ($p \ll 0.001$, $n = 500$), whereas NSC training corpus and SC testing corpus had correlations of 0.958 ($p \ll 0.001$, $n = 500$). Likewise, the SC training corpus and the NSC testing corpus had correlations of 0.957 ($p \ll 0.001$, $n = 500$), whereas SC training corpus and SC testing corpus had correlations of 0.958 ($p \ll 0.001$, $n = 500$). The differences between these correlation coefficients was approximately 0.001.

Table 6 Correlation results for conditions in Study 2

Data set		Pearson correlation coefficients	
Training data	Testing data	Key concept score	Naïve idea score
Non-spelling check	Non-spelling check	0.956**	0.892**
Non-spelling check	Spelling check	0.957**	0.895**
Spelling check	Non-spelling check	0.957**	0.895**
Spelling check	Spelling check	0.958**	0.898**

** $p < 0.01$

At the *class level* (see Table 1), the number of total prediction errors by the ACSS using *spell-checked* training *and* testing corpora was 582 [3.23 % out of 18,000 predictions (2000 responses × 9 concepts)]. The number of total prediction errors by the ACSS with *non-spell-checked* training and *spell-checked* testing corpora was 586 (3.26 %). The number of total prediction errors by the ACSS using *spell-checked* training and *non-spell-checked* testing corpora was 591 (3.28 %). Lastly, the number of prediction errors by the ACSS with *non-spell-checked* training and *non-spell-checked* testing corpora was also 591 (3.28 %). Thus, the prediction errors caused by MSW totaled nine cases out of 18,000 predictions (0.05 %).

Study 3

In corpus A, we found 214 MSW (2.25 %) out of 9511 total words (excluding stop words). Of the 9511 words in the corpus, 613 (6.45 %) were *scoring-rule-related.* Of the MSW, 30 (14.0 % MSW) were classified into *scoring-rule-related* words (Table 7). The 30 *scoring-rule-related* MSW included 25 different MSW; that is, on average each MSW appeared 1.20 times. Corpus B contained 651 MSW (2.7 %) out of 24,354 total words (excluding stop words). Among 24,354 words, 4994 (20.51 %) were *scoring-rule-related.* Of the MSW in corpus B, 246 (37.79 %) were *scoring-rule-related.* The 246 *scoring-rule-related* MSW included 153 different types of MSW; that is, on average, each MSW appeared 1.61 times. The most common MSW in the training corpus was "environment." In addition, many students misspelled "competition," "beneficial," "advantageous," and "necessary" (see Table 7).

The correspondence measures between human- and computer-scored responses for the two testing corpora are given in Table 8. At the *individual concept level*, repeated-measures ANOVAs indicated that there were significant but very weak effects of MSW on kappa, agreement percentage, and F_1 scores. Between-group analyses (i.e., the two corpora) did not reveal any significant differences, which indicates similar patterns occurred across the two corpora (Kappa: $F[1,16] = 6.31$, $p = 0.023$, $\eta_p^2 = 0.283$, Between group: $F[1,16] = 0.07$, $p = 0.791$, $\eta_p^2 = 0.005$, Agreement percentage: $F[1,16] = 5.73$, $p = 0.029$, $\eta_p^2 = 0.264$, Between group: $F[1,16] = 0.07$, $p = 0.795$, $\eta_p^2 = 0.004$, F_1 score: $F[1,16] = 5.86$, $p = 0.028$, $\eta_p^2 = 0.268$, Between group: $F[1,16] = 0.19$, $p = 0.667$, $\eta_p^2 = 0.012$).

However, it must be noted that the significant (but weak) effects shown in Kappa, agreement percentage, and F_1 scores exhibited very weak power (<0.7). We found significant effects of MSW on the recall measure with acceptable power ($F[1,16] = 9.01$, $p = 0.008$, $\eta_p^2 = 0.360$, power = 0.805, Between group: $F[1,16] = 0.63$, $p = 0.441$, $\eta_p^2 = 0.038$). Lastly, using ANOVAs, we did not find significant effects of MSW on precision measures ($F[1,16] = 0.22$, $p = 0.647$, $\eta_p^2 = 0.013$, Between group: $F[1,16] = 1.63$, $p = 0.220$, $\eta_p^2 = 0.092$). Due to non-normality of the precision measures, we also performed non-parametric tests (Wilcoxon signed ranks test) that confirmed nonsignificant effects revealed using ANOVA ($z = 0.000$, $p = 1.000$).

The largest difference in recall measures between the non-spell-checked (NSC) and spell-checked (SC) corpora was found in the scoring algorithm for the "needs/goals" concept in corpus A (4.5 %, 3 cases among 67). In

Table 7 The numbers of words, MSW, and MSW related to concepts in three corpora in Study 3

Data sets	Total words (A)	Total words except stop words (B) (B/A)	Scoring-rule-related words in total words (B)	MSW (C) (C/B)	Scoring-rule-related words in MSW (D) (D/C)	Total # of kind of MSW
Corpus A ($n = 285$)	17,675	9511 (53.81 %)	613 (6.45 %)	214 (2.25 %)	30 (14.02 %)	25
Corpus B ($n = 1112$)	45,925	24,354 (53.03 %)	4994 (20.51 %)	651 (2.67 %)	246 (37.79 %)	153

contrast, for corpus B, the largest difference in recall measures between NSC and SC corpora was for the "adapt/acclimation" idea (4.5 %, 5 cases among 110).

We also examined the impact of MSW on kappa and raw agreement percentages for the four reasoning models (e.g., scientific, mixed, naïve, and no model) at the *individual response level*. It should be noted that the *individual response level* and the *student level* are the same in Study 3 because students only responded to one item (in each corpus). The results show that MSW in both corpora are associated with scoring accuracy differences of approximately 5 %, and differences of kappa values of 0.03 (corpus A: NSC: κ 0.816 (89.8 %), SC: κ 0.847 (91.6 %); corpus B: NSC: κ 0.770 (85.0 %), SC: κ 0.787 (86.2 %). At the *individual response* level, the Spearman correlations indicated that MSW did not meaningfully impact scoring efficacy for key concept scores (NSC: $r = 0.936$, $p \ll .001$; SC: $r = 0.939$, $p \ll .001$ for corpus A; NSC: $r = 0.892$, $p \ll .001$; SC: $r = 0.898$, $p \ll .001$ corpus B). In contrast, MSW impacted the efficacy of naïve idea scores (NSC: $r = 0.839$, $p \ll .001$, SC: $r = 0.873$, $p \ll .001$ for corpus A; NSC: $r = 0.840$, $p \ll .001$, SC: $r = 0.852$, $p \ll .001$ for corpus B).

Table 8 The differences of efficacy (e.g., kappa, agreement percentage) of automated computer scoring models for non-spelling checked and spelling checked Corpus A ($n = 285$) and Corpus B ($n = 1112$)

	Kappa		Agreement		Precision		Recall		F_1 score	
	NSC	SC	NSC	SC	NSC	SC	NSC	SC	NSC	SC
Corpus A (n = 285)										
Variation	0.937	0.944	96.8	97.2	100.0	100.0	93.9	94.6	96.9	97.2
Heritability	0.855	0.867	95.4	95.8	98.0	98.0	80.3	82.0	88.3	89.3
Competition	1.000	1.000	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Limited resources	0.944	0.958	97.2	97.9	100.0	100.0	94.5	95.9	97.2	97.9
Differential survival	0.888	0.881	94.4	94.0	97.2	97.2	92.2	91.5	94.6	94.3
Non-adaptive idea	0.707	0.707	98.6	98.6	83.3	83.3	62.5	62.5	71.4	71.4
Need/goal	0.845	0.894	94.4	96.1	87.0	90.0	89.6	94.0	88.2	92.0
Use/disuse	0.796	0.796	99.3	99.3	80.0	80.0	80.0	80.0	80.0	80.0
Adapt/acclimation	0.829	0.829	95.8	95.8	89.7	89.7	81.4	81.4	85.4	85.4
M	0.867	0.875	96.9	97.2	92.8	93.1	86.0	86.9	89.1	89.7
SD	0.088	0.090	2.1	1.9	7.9	7.7	11.3	11.6	9.3	9.4
Corpus B (n = 1112)										
Variation	0.839	0.832	94.8	94.5	92.5	90.8	82.4	82.8	87.2	86.7
Heritability	0.847	0.859	95.1	95.4	97.5	97.6	79.8	81.5	87.8	88.8
Competition	0.932	0.932	99.7	99.7	100.0	100.0	87.5	87.5	93.3	93.3
Limited resources	0.815	0.843	94.2	95.0	98.9	99.0	74.6	78.2	85.1	87.4
Differential survival	0.852	0.861	92.6	93.1	96.7	96.4	89.8	90.9	93.1	93.6
Non-adaptive idea	1.000	1.000	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Need/goal	0.864	0.876	94.6	95.1	95.8	95.9	85.1	86.6	90.1	91.0
Use/disuse	0.693	0.693	97.4	97.4	68.6	68.6	72.9	72.9	70.7	70.7
Adapt/acclimation	0.662	0.701	95.1	95.6	95.2	95.5	53.6	58.2	68.6	72.3
M	0.834	0.844	95.9	96.2	93.9	93.8	80.6	82.1	86.2	87.1
SD	0.105	0.098	2.5	2.4	9.8	9.8	13.0	11.8	10.4	9.7
Repeated-measure ANOVA statistics	$F[1,16] = 6.31$, $p = 0.023$, $\eta_p^2 = 0.283$, between group: $F[1,16] = 0.07$, $p = 0.791$, $\eta_p^2 = 0.005$		$F[1,16] = 5.73$, $p = 0.029$, $\eta_p^2 = 0.264$, between group: $F[1,16] = 0.07$, $p = 0.795$, $\eta_p^2 = 0.004$		$F[1,16] = 0.22$, $p = 0.647$, $\eta_p^2 = 0.013$, between group: $F[1,16] = 1.63$, $p = 0.220$, $\eta_p^2 = 0.092$ Wilcoxon signed ranks test: $z = 0.000$, $p = 1.000$		$F[1,16] = 9.01$, $p = 0.008$, $\eta_p^2 = 0.360$, power = 0.805, between group: $F[1,16] = 0.63$, $p = 0.441$, $\eta_p^2 = 0.038$		$F[1,16] = 5.86$, $p = 0.028$, $\eta_p^2 = 0.268$, between group: $F[1,16] = 0.19$, $p = 0.667$, $\eta_p^2 = 0.012$	

Finally, at the *class level*, 72 total prediction errors occurred in spell-checked corpus A [2.81 % out of 2565 predictions (285 responses × 9 concepts)], whereas the number of total prediction errors with the non-spelling checked corpus was 80 (3.12 %). Thus, the prediction errors caused by MSW in corpus A comprised only eight cases out of 2565 (0.31 %). The number of total prediction errors with SC corpus B comprised 381 [3.8 % out of 10,008 predictions (1112 responses × 9 concepts)], whereas the number of total prediction errors with the NSC corpus was 406 (4.06 %). Thus, the prediction errors caused by MSW in corpus B included 25 cases out of 10,008 (0.25 %).

Discussion

RQ1: Do ELLs Produce Equivalent Magnitudes and Types of Spelling Errors as Non-ELLs?

A recent report emphasized that English language learners (ELLs) comprise an increasingly large proportion of US science classrooms (Flynn and Hill 2005), and they are more likely to make spelling errors than native speakers of English (Bebout 1985; Flor and Futagi 2012; Haggan 1991). Although a convincing logical argument may be advanced that greater numbers of MSW will occur in ELLs' text responses, it is important to empirically test this claim across educational levels and subject areas (e.g., college biology). ELL students in our sample produced twice as many MSW as native speakers. In addition, these errors were scoring-rule-related (see Table 3). Our answer to RQ1 is unsurprising but nevertheless important and aligns with previous work in other areas (e.g., Bebout 1985). Given that the MSW generated by ELLs were shown to be related to scoring rules, it is important to examine the relationship between MSW and computer scoring accuracy.

RQ2: What Percentage of MSW Relate to Concept-Specific Scoring Rules?

This study empirically explored the impact of misspelled words (MSW) on computer scoring accuracy using large samples (>10,000) of biological explanations written by undergraduate students. In the corpora that we studied, MSW were relatively uncommon. Specifically, students misspelled approximately two percent of words (excluding stop words) in their written responses. However, 29.0 % of these MSW were scoring-rule-related; that is, words central to the machine learning algorithms that automatically scored the text. Notably, all four of our data sets displayed the same patterns (see "Results" section). The most common MSW in the corpora were "environment," "competition," "beneficial," "advantageous," and "necessary," which are central terms for detecting both scientific and naïve ideas of natural selection (e.g., differential survival, need/goal idea, see scoring rubrics of Nehm et al. 2010 and Federer et al. 2014). In sum, MSW were relatively uncommon in our corpora, although those words that were misspelled were central to scoring biological explanations relating to evolution.

RQ3: What Impact Do MSW Have on Computer Scoring Accuracy?

Our findings from Study 2 and Study 3 demonstrate that MSW do *not* meaningfully impact computer scoring efficacy across nine different computer scoring algorithms for scientific explanations relating to evolution. However, it should be noted that reduced accuracy caused by MSW was shown in several cases (e.g., for concepts such as "use/disuse," "need/goal," "adapt/acclimation," and "limited resources"). However, the more important question is whether these decreases are large enough to matter in a practical sense. The findings from Study 2 and Study 3 are not in complete alignment, and for this reason we must be careful about attempting to generalize our findings.

In the sample of 2000 student explanations, the total number of concepts detected (i.e., six key concepts and three naïve ideas) was 2931. Of these, the total number of concepts that the computer system detected in the spell-checked corpus—but failed to detect in the non-spell-checked corpus—was 17. Thus, error introduced by MSW could be estimated to be 0.6 %. If one considers that 135 MSW were related to scoring rules (see Table 4) and that the number of concepts that the computer system failed to capture because of MSW was 17, then it is clear that there is not a direct relationship between the number of MSW and scoring accuracy. Indeed, approximately 90 % of the MSW related to scoring rules did *not* impact computer scoring efficacy. Two student responses relating to the "variation" key concept and the "needs/goals" naïve idea help to illustrate why this is so.

> If evolved by natural selectionmutation [sic] arises introducing poisonous trait, *variation in original population* poisonous and non-poisonous, differential survival in an environment favoring poisonous trait-increases [sic] in frequency, non-poisonous *trait* eventually is fixed. If by *genetic* drift, see next question." [Variation example]:

> They could use the evolution theory, and show how hte [sic] plants needed to adapt innorder [sic] to survive their environment." [Needs/goals example].

One reason why MSW do not always impact scoring efficacy is because single words are not typically used to detect a concept; a wide range of words can help the scoring model detect the same concept. In the first quote shown above, the student made two *spacing errors* (for *mutation* and *trait*). These two words are very important for scoring the "variation" concept (see the rubrics of Nehm et al. 2010). However, the scoring model detected three additional features of the response (shown in italics above: variation, trait, genetic) and correctly scored the response for this concept (despite the presence of MSW). The second example (shown above) also illustrates this point. In this instance, the MSW "inorder" is used, which is an important feature of the "needs/goals" concept. Nevertheless, the scoring model detected *need* and *to survive* features, which also suggest that the student has this particular naïve idea. These two examples help to illustrate the point that MSW relating to a scoring rule do not guarantee that a scoring error will occur.

Similar results were found with the other corpora. Using the corpus of 1397 student explanations (i.e., Corpora A and B used in Study 3), total number of concepts (six key concepts and three naïve ideas) was 2495. The total number of concepts that the computer correctly detected in the spell-checked (SC) corpora but failed to capture in NSC corpora was 47. Thus, the error rate is less than two percent (1.88 %). In addition, the total number of MSW related to computer scoring rules was 276. Therefore, the majority of the MSW related to the scoring rules did *not* impact scoring accuracy. Although our different corpora produced slightly different statistical results, both clearly demonstrate that MSW do not have as large of an impact (<2 % error) as one might expect.

RQ4: Are MSW More Likely to Impact False-Positive or False-Negative Feedback?

The findings from Study 2 and Study 3 indicate that MSW have the largest impact on recall measures, which means that they are more likely to generate false-negative feedback (i.e., students are informed that they did *not* possess particular concepts even though those concepts are in fact included in their responses). For example, MSW resulted in about 4 % of the "use/disuse" concepts to *not* be detected by the computer. Such false-negative feedback, despite being rare, may nevertheless generate negative downstream consequences (e.g., decreases in motivation, self-efficacy, and performance). It is important to point out that the scoring algorithms for naïve ideas were impacted by MSW to a greater degree than were normative scientific concepts. For example, the first, second, and third highest ranked concepts (in terms of recall) were "use/disuse," "adapt/acclimation," and "needs/goals" in Study 2—all of which are naïve ideas (see Table 1). Likewise, in Study 3, the scoring models most impacted by MSW were "needs/goals" (corpus A) and "adapt/acclimation" (corpus B). Unusually, the false-negative feedback relating to naïve ideas (that is, the computer is more likely to detect that a student does *not* have a naïve idea) is—somewhat ironically—much like false-positive feedback. Given that MSW impact the scoring algorithm for naïve ideas to a greater extent than key concepts, the effects of false-negative feedback caused by MSW is not as problematic as it might first appear.

Solutions to the Problem of Misspelled Words

Although all of our empirical studies demonstrated that MSW: (1) are uncommon in large corpora of written biological explanations of evolutionary change by undergraduate students and (2) do not meaningfully impact computer scoring of text, it is by no means clear whether these findings generalize to high school students, samples with greater percentages of ELLs, or other scientific domains. Spell-check and word replacement systems are one obvious solution of the challenge of MSW. Unfortunately, existing computer technology is not sufficiently well developed to solve the challenge, particularly in jargon-rich domains like biology.

There may be other approaches to address the challenge of MSW. We may be able to manually design ACSS that include common MSW. For example, in our study, we could empirically identify common MSW in students' written responses and integrate them into the training corpora of ACSS. For example, we found "enviroment [sic]" 38 times and "enviornment [sic]" 25 times in our corpora. The training corpus could be modified to incorporate common MSW that could be included in the feature libraries of machine learning scoring models.

Finally, we could build upon recent work on computerized educational "early warning systems" (Lee et al. 2015). Although the corpus used in Study 2 was collected using a system that included a spell-check tool, most students clearly did not use it. Computer systems could be designed (much like Microsoft Word) that notify writers that their text might contain MSW. However, as our study found, the availability of spell-check systems does not mean that students will use them. However, if students are made aware that their instructors will be using ACSS to grade their responses, and grading accuracy is impacted by spelling, students may be more likely to correct MSW. It is clear that new technology will demand changes in student and instructor behavior.

Study Limitations

Our study displays limitations that should be considered when interpreting the results. First, we focused on corpora

written by college students in the domain of biology (specifically, evolutionary explanations). It is not known whether these findings will generalize to other domains at different grade levels (e.g., high school chemistry). For example, misspellings may be higher or lower in other domains, and consequently, scoring efficacy could be impacted in different ways than we have documented. However, given that no studies to our knowledge have been conducted on the impact of MSW on computer scoring in college biology, this study provides important insights into an issue that will become more important as technology advances (e.g., voice recognition software, cell phones used as classroom clickers). Second, the large corpora that we studied tended to have low levels of MSW. This is partly a consequence of the approach we used to define MSW (e.g., we excluded apostrophe errors, spacing errors, and stopword errors). Samples containing much higher percentages of MSW could alter our conclusions. More empirical work using more diverse samples of learners would help to determine how broadly our findings hold. Finally, our study has conceptualized MSW as a construct-irrelevant feature of written explanations. However, MSW may not be randomly distributed across student responses and could in principle provide information relevant to the measurement of student abilities. Future work should explore how this question could be empirically examined.

Acknowledgments We thank the reviewers for helpful and thought-provoking comments on an earlier version of the manuscript. Financial support was provided by a National Science Foundation TUES grant (1322872). Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF.

References

Abedi J (2004) The no child left behind act and English language learners: assessment and accountability issues. Educ Res 33(1):4–14

Abedi J, Hofstetter CH, Lord C (2004) Assessment accommodations for English language learners: implications for policy-based empirical research. Rev Educ Res 74(1):1–28

Abu-Mostafa YS (2012) Machines that think for themselves. Sci Am 307(1):78–81

Agarwal S, Godbole S, Punjani D, Roy S (2007) How much noise is too much: a study in automatic text classification. In: Seventh IEEE international conference on Data mining, 2007. ICDM 2007, pp 3–12. IEEE

American Association for the Advancement of Science (AAAS) (2011) Vision and change in undergraduate biology education. AAAS, Washington, DC

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) (2014) The standards for educational and psychological testing. AERA Publications, Washington, DC

Bebout L (1985) An error analysis of misspellings made by learners of English as a first and as a second language. J Psycholinguist Res 14(6):569–593

Beggrow EP, Ha M, Nehm RH, Pearl D, Boone WJ (2014) Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance? J Sci Educ Technol 23(1):160–182

Bejar II (1991) A methodology for scoring open-ended architectural design problems. J Appl Psyc 76(4):522–532

Bishop BA, Anderson CW (1990) Student conceptions of natural selection and its role in evolution. J Res Sci Teach 27(5):415–427

Brady M, Seli H, Rosenthal J (2013) "Clickers" and metacognition: a quasi-experimental comparative study about metacognitive self-regulation and use of electronic feedback devices. Comp Educ 65:56–63

Bridgeman B, Trapani C, Attali Y (2012) Comparison of human and machine scoring of essays: differences by gender, ethnicity, and country. Appl Measur Educ 25:27–40

Chen JC, Whittinghill DC, Kadlowec JA (2010) Classes that click: fast, rich feedback to enhance student learning and satisfaction. J Eng Educ 99(2):159–168

Connors RJ, Lunsford AA (1988) Frequency of formal errors in current college writing, or Ma and Pa Kettle do research. Coll Compos Commun 39(4):395–409

Damerau FJ (1964) A technique for computer detection and correction of spelling errors. Commun ACM 7(3):171–176

Federer MR, Nehm RH, Opfer JE, Pearl D (2014) Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations. Res Sci Educ 45(4):527–553

Fitzsimmons PA, Landers DM, Thomas JR, van der Mars H (1991) Does self-efficacy predict performance in experienced weightlifters? Res Quart Exerc Sport 62(4):424–431

Flor M, Futagi Y (2012) On using context for automatic correction of non-word misspellings in student essays. In: Proceedings of the seventh workshop on building educational applications Using NLP, pp 105–115. Association for Computational Linguistics

Flynn K, Hill J (2005) English language learners: a growing population. Policy brief mid-continent research for education and learning, pp. 1–12

Ha M, Nehm RH (2012) Using machine-learning methods to detect key concepts and misconceptions of evolution in students' written explanations. Paper to be presented at the National Association for Research in Science Teaching, Indianapolis, IN

Ha M, Nehm RH, Urban-Lurain M, Merrill JE (2011) Applying computerized scoring models of written biological explanations across courses and colleges: prospects and limitations. CBE Life Sci Educ 10:379–393

Haggan M (1991) Spelling errors in native Arabic-speaking English majors: a comparison between remedial students and fourth year students. System 19(1):45–61

Haudek KC, Kaplan JJ, Knight J, Long T, Merrill J, Munn A, Nehm RH, Smith M, Urban-Lurain M (2011) Harnessing technology to improve formative assessment of student conceptions in STEM: forging a national network. CBE Life Sci Educ 10(2):149–155

Haudek KC, Prevost LB, Moscarella RA, Merrill J, Urban-Lurain M (2012) What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. CBE Life Sci Educ 11(3):283–293

Holroyd KA, Penzien DB, Hursey KG, Tobin DL, Rogers L, Holm JE, Marcille PJ, Hall JR, Chila AG (1984) Change mechanisms in EMG biofeedback training: cognitive changes underlying improvements in tension headache. J Consult Clin Psychol 52(6):1039–1053

Karl KA, O'Leary-Kelly AM, Martocchio JJ (1993) The impact of feedback and self-efficacy on performance in training. J Organ Behav 14(4):379–394

Kucirkova N, Messer D, Sheehy K, Panadero CF (2014) Children's engagement with educational iPad apps: insights from a Spanish classroom. Comp Educ 71:175–184

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:1159–1174

Lee UJ, Sbeglia GC, Ha M, Finch SJ, Nehm RH (2015) Clicker score trajectories and concept inventory scores as predictors for early warning systems for large STEM Classes. J Sci Ed Tech 24(6):848–860

Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. Soviet Phys Doklady 10(8):707–710

Linn MC, Gerard L, Ryoo K, McElhaney K, Liu OL, Rafferty AN (2014) Computer-guided inquiry to improve science learning. Science 344(6180):155–156

Lunsford AA, Lunsford KJ (2008) "Mistakes are a fact of life": a national comparative study. Coll Compos Commun 59(4):781–806

Moharreri K, Ha M, Nehm RH (2014) EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. Evolut Educ Outreach 7(1):1–14

Muhlenbach F, Lallich S, Zighed DA (2004) Identifying and handling mislabelled instances. J Intell Inf Syst 22(1):89–109

Nagata R, Whittaker E, Sheinman V (2011) Creating a manually error-tagged and shallow-parsed learner corpus. Proceedings of the 49th annual meeting of the association for computational linguistics. ACL, Stroudsburg, pp 1210–1219

National Research Council (2012) A framework for K-12 science education: practices, crosscutting concepts, and core ideas. The National Academies Press, Washington, DC

National Research Council (2013) Next generation science standards: for states, by states. The National Academies Press, Washington, DC

Nehm RH, Reilly L (2007) Biology majors' knowledge and misconceptions of natural selection. Bioscience 57(3):263–272

Nehm RH, Schonfeld IS (2007) Does increasing biology teacher knowledge of evolution and the nature of science lead to greater preference for the teaching of evolution in schools? J Sci Teach Educ 18(5):699–723

Nehm RH, Ha M, Rector M, Opfer JE, Perrin L, Ridgway J, Mollohan K (2010) Scoring guide for the open response instrument (ORI) and evolutionary gain and loss test (ACORNS). Technical report of National Science Foundation REESE project 0909999

Nehm RH, Ha M, Mayfield E (2012) Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. J Sci Educ Technol 21(1):183–196

Opfer JE, Nehm RH, Ha M (2012) Cognitive foundations for science assessment design: knowing what students know about evolution. J Res Sci Teach 49(6):744–777

Platt J (1999) Fast training of support vector machines using sequential minimal optimization. In: Schölkopf B, Burges CJC, Smola AJ (eds) Advances in Kernel methods—support vector learning. MIT Press, Cambridge, pp 185–208

Sato T, Yamanishi Y, Kanehisa M, Toh H (2005) The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21(17):3482–3489

Su LT (1994) The relevance of recall and precision in user evaluation. J Am Soc Inf Sci 45(3):207–217

Zhu Z, Pilpel Y, Church GM (2002) Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. J Mol Biol 318(1):71–81