The Importance of Grammar and Mechanics in Writing Assessment and Instruction: Evidence from Data Mining

Scott Crossley
Georgia State University
34 Peachtree Ave, Ste 1200
Atlanta, GA 30303
01+404-413-5179
scrossley@gsu.edu

Kris Kyle
Georgia State University
34 Peachtree Ave, Ste 1200
Atlanta, GA 30303
01+404-413-5200
kkyle3@gsu.edu

Laura Allen
Arizona State University
PO Box 872111
Tempe, AZ 85287
01+404-413-5200
LauraKAllen@asu.edu

Danielle S. McNamara
Arizona State University
PO Box 872111
Tempe, AZ 85287
01+404-413-5200
dsmcnamara1@gmail.com

ABSTRACT
The current study examined relationships between expert human judgments of text quality and grammar and mechanical errors in student writing. A corpus of essays (N = 100) written by high school students in the W-Pal system was collected, coded for grammar and mechanical errors, and scored by expert human raters. Results revealed weak relations between grammar errors and holistic essay scores and stronger relations between mechanics and holistic essay scores. Implications for essay scoring algorithms and providing feedback to writers are discussed.

Keywords
Intelligent tutoring systems, grammar and mechanics, automated feedback, automatic essay scoring

1. INTRODUCTION
The Writing Pal (W-Pal; [7, 9, 10]) is an intelligent tutoring system (ITS) that provides students with instruction and game-based practice on how to use writing strategies. The system also gives students opportunities to write essays, receive automated feedback on these essays, and revise the essays. The purpose of this study is to examine the importance of errors in grammar and mechanics (e.g., punctuation and spelling) for predicting holistic scores of essay quality and how the relationship between grammar and mechanics and essay quality can be used to help develop instructional modules and feedback algorithms within W-Pal. Our particular interest in the consequences of considering grammar and spelling in instructional modules and in providing automated feedback to students stems primarily from concerns expressed by writing instructors who have used W-Pal in their classes. Currently, W-Pal focuses on providing students with feedback that centers on using strategies to more effectively compose essays, including strategies to plan essays, write more effective introductions, essay bodies, and conclusions, and to revise their essays. These strategies have proven successful; however, some teachers remain concerned that students primarily need feedback on lower level aspects of writing such as grammar, punctuation, and spelling.

Although research supports the teaching of mechanics to students [5, 8], meta-analyses of effective writing instruction have demonstrated that grammar instruction is among the least effective types of student interventions [6]. On the other hand, teachers report that correct grammar and mechanics are important elements of writing instruction and writing quality. For example, in a study by Cutler and Graham [3], over 75% of surveyed teachers indicated that they taught grammar skills at least several times a week at the expense of teaching essay writing, planning, and revising. Additional evidence for the perceived importance of grammar skills in the classroom can also be found in writing textbooks, which dedicate large sections to grammar instruction [8].

Our main design and pedagogical questions in the context of W-Pal are whether to include a module that explicitly teaches grammar and mechanics, whether to provide grammar and mechanics feedback to the students who use W-Pal, and whether to incorporate grammar and mechanic indices in our automatic scoring algorithms. Fully answering these questions will likely require behavioral or intervention studies. However, an initial step in assessing the importance of grammar and mechanics is to use data mining techniques to assess relationships between grammar and mechanical accuracy and essay quality. Thus, in this study, we examine a corpus of essays written by students who were provided instruction in W-Pal. The essays were scored by expert raters for overall essay quality as well as grammatical and mechanical accuracy. The essays were also coded for grammatical and mechanical accuracy by a separate set of expert raters. We specifically seek to address the following three research questions:

1. To what extent are expert analytic scores of grammar and mechanics related to holistic scores of essay quality?
2. To what extent are expert analytic scores of grammar and mechanics associated with the number and type of errors observed in an essay?
3. To what extent are expert scores of holistic essay quality associated with the number and type of errors observed in an essay?

Our underlying presumption is that the answers to these questions will enhance our understanding of essay writing and expert judgments of essay quality. In turn, these answers will aid in the design and development of W-Pal by providing information about the importance of grammar and mechanical errors in assessing writing quality. If grammar and mechanical errors are important indicators of writing quality, then there may be value in providing instructional modules that help students avoid making grammatical and spelling errors, in providing feedback to learners about the number and types of errors that occur in their writing, and in including automated measures of grammar and mechanics in the scoring algorithms used by W-Pal. The results of the study will also strengthen our understanding of the linguistic features that underlie writing quality.

2. METHODS

To address the research questions for this study, a corpus of essays was hand coded to identify grammar and mechanical errors and these errors were then regressed onto the expert ratings of grammar and mechanics and the expert judgments of essay quality. In addition, correlations were conducted between the expert ratings of grammar and mechanics and the holistic judgments of essay quality.

2.1 Corpus

We selected 100 essays from an on-line writing study conducted in the W-Pal ITS. The essays were written by public high school students in the metro Phoenix area. The students ranged in age from 14 to 19 and the majority of the students in the study (62%) were female; 56% of the students identified themselves as native speakers of English, with the remaining participants identifying themselves as non-native speakers of English. Participants attended 10 sessions (1 session/day) over a 2-4 week period. Participants wrote a pretest essay during the first session and a posttest essay during the last session. The essays were written on two prompts (on the value of competition and on the role of images/appearances). The prompts were counterbalanced across the pretest and posttest essays. The essays used in this study were selected from the pretest essays only.

Two expert raters with at least 4 years of experience teaching freshman composition courses at a large university rated the quality of the essays using a standardized SAT rubric and an analytic rubric that contained four subsections: introduction, body, conclusion, and correctness (see [2] for more details on the rubric). The correctness subsection consisted of one rating that asked reviewers to judge an essay's grammar and mechanical accuracy. Both the SAT and the analytic rubric generated a rating with a minimum score of 1 and a maximum of 6. Raters were informed that the distance between each score was equal. The raters were first trained to use the rubric with 20 similar essays taken from another corpus. The final interrater reliability for all essays in the corpus was $r > .70$. The mean score between the raters was used as the final value for the quality of each essay. The essays selected for this study had a scoring range between 1 and 4.5. The mean score for the essays was 2.9 and the median score was 3.0. The scores were normally distributed.

2.1 Hand-Coding of Errors

An error tag-coding scheme was developed to investigate the grammar, mechanics, word use, and spelling in the 100 selected essays. The coding scheme was based on an error-tagging manual reported in Dagneaux, Dennes, Granger, and Meunier [4]. The manual consists of subsections related to form (spelling and morphology), grammar (nouns, adjectives, and verbs), lexico-grammar (complementation, dependent prepositions), lexical choices (single, phrases, connectors, and conjunctions), and word problems (redundant and missing words). Two expert raters were trained on this manual. After reviewing a training set of essays and the manual, new codes were incorporated that related to punctuation, spelling, sentence fragments, and ambiguous referents. These codes were not available in the original coding scheme but errors in the essays necessitated them. After training was completed, the raters coded each essay independently and codes between raters were compared. Differences in coding were adjudicated between the two raters until agreement was reached. Final raw scores were provided for each essay for each code. In addition, a score based on text length was computed (a normalized score). Component scores were calculated for all form errors (spelling and morphology), all grammar errors, all lexico-grammar errors, all lexical choice errors, all word problem errors, and all punctuation errors. Lastly, a total count of all errors in the essay was computed.

2.2 Statistical Analyses

Statistical analyses using SPSS were conducted to investigate the role that grammar and mechanics play in explaining human scores of essay quality. A correlation was calculated between holistic essay scores and expert scores of grammar and mechanics to examine links between holistic and analytic scores. A regression model was then used to assess the accuracy of the expert scores for grammar and mechanics by investigating associations between the hand-coded error counts and the expert judgments. Finally, a regression model was used to examine the associations between the hand-coded errors and the expert scores for holistic essay quality. For both regression models, a training and test approach was used. SPSS syntax does not select an exact percentage for training and test sets and thus training sets in SPSS may range from 63-71% of the corpus.

3. Results

3.1 Expert Scores

A correlation was calculated between the expert ratings for the holistic score and the expert ratings for grammar and mechanics (the analytic score). The resulting correlation, $r(100) = .388$, $p < .001$, reflects a positive, medium effect between the holistic and analytic scores.

3.2 Grammar Scores

Correlations were calculated between the hand-coded errors and the expert scores for grammar and mechanics to examine the strength of the relationship between these two variables. Prior to this analysis, the hand-coded error scores were also checked for multi-collinearity. The analyses demonstrated that there were 26 hand-coded errors that demonstrated at least a small effect size ($r > .10$) with the expert ratings and did not demonstrate strong multi-collinearity with each other (defined as $r > .90$). The majority of these variables were related to overall errors and mechanics, but not to grammar.

A stepwise linear regression analysis was conducted including the 26 hand-coded errors in which these variables were regressed onto the raters' evaluations for the 71 essays randomly selected by SPSS for the training set. The linear regression using the 23 variables yielded a significant model, $F(2, 69) = 20.980$, $p < .001$, $r = .615$, $r^2 = .378$. The test set yielded $r = .653$, $r^2 = .426$. Two variables were significant predictors in the regression: total

number of errors (raw) and punctuation errors (raw). The regression model for the training set is presented in Table 1.

Table 1: Regression analysis predicting expert grammar and mechanics scores

Entry	Variable added	r	R^2
Entry 1	Total errors raw	0.572	0.327
Entry 2	Punctuation errors raw	0.615	0.378

3.3 Holistic Scores

Correlations were calculated between the hand-coded errors and the expert holistic scores to assess the strength of the relationship between errors and the holistic rating of essay quality and to check for multi-collinearity between the hand-coded errors. These analyses showed that there were 22 hand-coded errors that demonstrated at least a small effect size with the expert ratings of essay quality and did not demonstrate strong multi-collinearity with each other. The majority of the errors that demonstrated medium or close to medium effect sizes were related to spelling, punctuation, and lexical errors.

A stepwise linear regression analysis was conducted with the 22 variables in which the variables were regressed onto the raters' evaluations for the 71 essays randomly selected by SPSS for the training set. The regression model for the training set is presented in Table 2. The linear regression using the 22 variables yielded a significant model, $F(2, 69) = 8.043$, $p < .010$, $r = .435$, $r^2 = .189$. The test set yielded $r = .456$, $r^2 = .208$. Two variables were significant predictors in the regression: total number of errors normalized and logical connector errors normalized.

Table 2: Regression analysis predicting expert holistic scores

Entry	Variable added	r	R^2
Entry 1	Total errors normalized	0.350	0.122
Entry 2	Logical connector errors normalized	0.435	0.189

3.4 Post-Hoc Analysis

We conducted a post-hoc analysis in which we removed the total errors variable. We conducted this analysis to examine if, in the absence of a total error count, errors related to grammar or mechanics (or both) were predictive of essay quality. As in the previous analyses, a stepwise linear regression analysis was conducted with the remaining 21 variables from the holistic score analysis. These 21 variables were regressed onto the raters' evaluations for the 64 essays randomly selected by SPSS for the training set. The linear regression using the 21 variables yielded a significant model, $F(1, 63) = 9.601$, $p < .010$, $r = .364$, $r^2 = .132$. The test set yielded $r = .293$, $r^2 = .086$. One variable was a significant predictor in the regression: form errors normalized (i.e., errors related to spelling and morphology errors normalized for text length). The remaining 20 variables, including all of the grammar variables, did not significantly add to the model and were left out. The regression model for the training set is presented in Table 3.

Table 3: Regression analysis predicting expert holistic scores without total errors

Entry	Variable added	r	R^2
Entry 1	Form errors normalized	0.364	0.132

4. Discussion

We have taken a corpus-based data mining approach to investigating the importance of grammatical and mechanical features in predicting the quality of students' essays. The results of this study indicate that expert ratings of grammar and mechanical accuracy are positively correlated to essay score and that the total number of errors and the number of punctuation errors in an essay are predictive of human judgments of grammar and mechanical accuracy. The findings also indicate that if grammatical errors in essays have any effect on expert judgments of essay quality, they are small. In contrast, errors related to spelling, punctuation, and lexical choices showed relatively strong correlations. These findings call into question the need to design instructional modules to teach grammar in W-Pal, as well as in other tutoring systems which focus on helping students to improve their writing quality.

In reference to relations between expert judgments of essay quality and expert judgments of grammar and mechanics, the findings report a moderate correlation that explains 15% of the variance in overall essay quality. Previous studies have shown similar results for the strength of grammar and mechanic judgments to predict essay quality [1, 2]. Importantly, these studies have indicated that human ratings of grammar and mechanics are the least predictive analytic ratings of essay quality (behind analytic judgments related to text organization, perspective, unity, conviction, and other elements). The regression analysis between coded errors in essays and human judgments of grammar and mechanic errors demonstrated that total errors and punctuation errors explained 43% of the variance in the human judgments for the test set. Such a finding indicates that expert ratings of grammar and mechanics are not solely based on overt errors in essays (i.e., over 50% of the variance in these judgments are not explained by grammar, spelling, and punctuation errors in the essay).

In reference to relations between grammatical errors and overall essay quality, the strongest correlation reported for a grammatical variable (article errors) demonstrated only a small effect size with holistic scores (and one that was not significant). In total, only four grammatical errors demonstrated at least small effect sizes with holistic scores of essay quality (i.e., article errors, verb morphology errors, noun errors, and verb errors). In no instances were grammatical error variables included in regression models that predicted essay quality. Thus, the findings point toward a weakness of grammatical errors in explaining writing quality and provide little evidence to support the inclusion of a grammar instruction module in the W-Pal system or include grammar indices in the automatic scoring algorithms contained in W-Pal. Additionally, since grammar errors in the essays are not strongly linked to overall scores of essay quality, there appears to be no strong evidence to provide feedback to W-Pal users concerning grammatical errors.

Correlations between holistic scores and the hand coded errors yielded the strongest associations for spelling errors. However, only a few spelling error variables showed medium effects sizes with essay quality and only one index of combined spelling and morpheme errors (form errors) was included in a regression model

that explained essay quality (this index explained 13% of the variance in essay quality). The majority of mechanical errors demonstrated only small effects with human judgments of essay quality and most of these errors did not reach significance.

Thus, while the evidence for mechanical instruction is a bit stronger, the findings do not strongly support the need to design instructional modules to teach mechanics in W-Pal. From a practical standpoint, designing a module that covers all potential spelling and punctuation errors in English is also too ambitious for a single ITS. In addition, research has demonstrated that students learn to spell best when they correct their own mispellings under the guidance of a teacher. This is especially true for students who have developed spelling skills (such as the adolescent writers targeted by W-Pal); these students should be able to predict spelling difficulties and apply previous knowledge to correct present spelling errors [11]. Therefore, the results of this study combined with design limitations and previous studies suggest that explicit spelling instruction may not be beneficial or practical.

In contrast to grammar errors, however, relations between spelling errors and holistic essay scores do appear strong enough to justify changes to the W-Pal automatic scoring algorithms and to the automatic feedback system. Automatically counting the number and types of spelling errors in an essay may improve the accuracy of the current scoring algorithm. In addition, compiling incidence scores for the number and types of punctuation normalized by the number of clauses or sentences may also increase the accuracy of the scoring algorithm. From a feedback perspective, highlighting spelling errors for W-Pal users may allow them to correct mispelled words more naturally. If, after highlighting spelling errors, users cannot still correctly spell the word, a drop-down menu with suggested spellings could be provided. In this way, spelling feedback that resembles best practices could be provided to W-Pal users. Of course such feedback mechanisms need to be assessed experimentally to better understand the relationship between spelling feedback and essay quality.

5. Conclusion

The results from this study, in combination with previous research, indicate that the explicit instruction of grammar in an ITS like W-Pal is likely unnecessary. In addition, providing feedback to users in reference to grammatical errors in their writing appears unwarranted (mostly because grammatical errors do not demonstrate strong relationships with essay quality). The same cannot be said for spelling and punctuation, which yield stronger relationships with judgments of writing quality. Thus, future versions of the W-Pal system will likely need to be sensitive to students' spelling and punctuation errors. However, we realize that the expectations of the scoring rubric used in this study may differ from the expectations found in an actual classroom and that the rubric itself may help in determining the importance of grammar and mechanics for the raters. The findings also indicate that human ratings of grammar and mechanics go beyond overt grammar, punctuation, and spelling errors as found in the text. A better understanding of what textual elements humans attend to when assessing grammar and mechanics would assist in more accurately identifying errors, which would be helpful in developing instructional techniques more strongly grounded in teacher cognition. Overall, the findings from this study provide important implications for system development and design that are based on real learning in practice. The findings also promote a number of future research areas.

6. ACKNOWLEDGMENTS

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES. We also thank Rod Roscoe, Tanner Jackson, Erica Snow, and Jianmin Dai for their help with the data collection and analysis and developing the ideas found in this paper. We lastly thank the readers that found our misspellings of misspell and will perpetually wonder if they judged the quality of this paper differently as a result.

7. REFERENCES

[1] Crossley, S. A. & McNamara, D. S. 2010. Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 984-989). Austin, TX: Cognitive Science Society.

[2] Crossley, S. A., & McNamara, D. S. 2011. Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* (pp. 1236-1241). Austin, TX: Cognitive Science Society.

[3] Cutler, L., & Graham, S. 2008. Primary grade writing instruction: A national survey. Journal of Educational Psychology, 100, 907 – 919.

[4] Dagneaux, E., S. Denness, S. Granger & Meunier, F. 1996. *Error Tagging Manual*, Version 1.1. Center for English Corpus Linguistics. Louvain-la-Neuve: Université Catholique de Louvain.

[5] Graham, S. 1983. Effective spelling instruction. *Elementary School Journal, 83* (5), 560-567.

[6] Graham, S., & Perin, D. 2007. A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99,* 445-476.

[7] McNamara, D., Raine, R., Roscoe, R., Crossley, S., Jackson, G., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Carney, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P., and Graesser, A. 2012. The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 298-311). Hershey, P.A.: IGI Global.

[8] Morris, D., Blanton, L., Blanton, W., & Perney, J. (1995). Spelling instruction and achievement in six classrooms. *Elementary School Journal, 96,* 145–162.

[9] Roscoe, R. & McNamara, D. in press. Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*.

[10] Roscoe, R. D., Varner, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. in press. The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition.*

[11] Schoephoerster, H. 1962. Research into variations of the test-study plan of teaching spelling. *Elementary English, 39*, 460-462.