Writing Quality, Knowledge, and Comprehension Correlates of Human and Automated Essay Scoring

[1]Rod D. Roscoe, [2]Scott A. Crossley, [3]Erica L. Snow, [3]Laura K. Varner, and [3]Danielle S. McNamara

[1]Arizona State University, Human and Environmental Systems, 7271 E. Sonoran Arroya Mall, SANCA 150D, Mesa, AZ 85212
[2]Georgia State University, Applied Linguistics/ESL, 34 Peachtree St. Suite 1200, Atlanta, GA 30303
[3]Arizona State University, Psychology, Learning Sciences Institute, PO Box 8721111, Tempe, AZ 85287
rod.roscoe@asu.edu, sacrossley@gmail.com, erica.l.snow@asu.edu, laura.varner@asu.edu, dsmcnamara@asu.edu

Abstract

Automated essay scoring tools are often criticized on the basis of construct validity. Specifically, it has been argued that computational scoring algorithms may be unaligned to higher-level indicators of quality writing, such as writers' demonstrated knowledge and understanding of the essay topics. In this paper, we consider how and whether the scoring algorithms within an intelligent writing tutor correlate with measures of writing proficiency and students' general knowledge, reading comprehension, and vocabulary skill. Results indicate that the computational algorithms, although less attuned to knowledge and comprehension factors than human raters, were marginally related to such variables. Implications for improving automated scoring and intelligent tutoring of writing are briefly discussed.

Automated Essay Scoring

Automated writing evaluation (AWE) uses computational tools to grade and give feedback on student writing (Shermis & Burstein, 2003). Studies have reported scoring reliability and accuracy in terms of significant, positive correlations between human and automated ratings, high perfect agreement (i.e., exact match between human and automated scores), and adjacent agreement (e.g., human and automated scores within one point on a 6-point scale). AWE has been touted as a means to efficiently score large numbers of essays and enable classroom teachers to offer more writing assignments (Shermis & Burstein, 2003). Several systems now offer scoring, feedback, and class management tools, such as *Criterion* (Attali & Burstein, 2006), *WriteToLearn* (Landauer, Lochbaum, & Dooley, 2009), and *MyAccess* (Grimes & Warschauer, 2010).

AWE tools gather large amounts of data about texts, such as structure, word use, syntax, cohesion, and semantic similarity (e.g., Landauer, McNamara, Dennis, & Kintsch, 2007; McNamara, Graesser, McCarthy, & Cai, in press; Shermis & Burstein, 2003). Importantly, students' ability to produce text that is technically correct (e.g., spelling), structured (e.g., paragraphs), and lexically proficient (e.g., use of rare words) is related to writing quality (Deane, 2013). Thus, scoring tools can and do leverage the links between linguistic features and overall writing quality to generate scores that are reliably similar to human ratings.

Despite such computational power, AWE necessarily excludes aspects of writing that are difficult to detect automatically, which tend to correspond to higher-level issues (e.g., comprehension). As a result, the proliferation of AWE has met with justifiable criticism, with perhaps the strongest objections pertaining to construct validity (Anson et al., 2013; Condon, 2013; Deane, 2013). A common concern is that AWE tools are not able to assess the most meaningful aspects of good and poor writing, such as writers' demonstrated knowledge and understanding of the essay topic, the persuasiveness of arguments, or an engaging style. In short, many argue that automated scoring fails "to measure meaningfulness of content, argumentation quality, or rhetorical effectiveness" because AWE systems "do not measure the full writing construct, but rather, a restricted construct" (Deane, 2013, p. 16). These deficits may be further exhibited in the feedback that AWE systems can give to developing writers. If a system cannot detect higher-level features of writing, then that system also cannot provide intelligent, personalized assistance on those same features to struggling students.

To begin exploring the issue of construct validity in automated scoring, this paper uses available data to examine how and whether human and automated scores for prompt-based essays are correlated to students' knowledge and related literacy skills of reading comprehension and vocabulary. Although this is not an exhaustive list of the qualities of good writers, knowledge and literacy skills offer a meaningful point of departure. Proficient writers not only adhere to rules of grammar and spelling, they also display skillful use of their knowledge and understanding of the topic (Chesky & Hiebert, 1987; McCutchen, 2000). Similarly, students' ability to write comprehensibly is related to their skills in comprehending text and vocabulary

(Fitzgerald & Shanahan, 2000; Shanahan & Lomax, 1986). Along these lines, the SAT scoring rubric (Camara, 2003) assigns a score of "6" (highest score) to an essay "that demonstrates outstanding critical thinking, using clearly appropriate examples, reasons and other evidence to support its position" and "exhibits skillful use of language, using a varied, accurate, and apt vocabulary." In contrast, an essay scored as a "1" (lowest) exhibits "fundamental errors in vocabulary… no viable point of view on the issue, or provides little or no evidence to support its position."

We hypothesize that scores assigned by trained, expert raters will correlate with measures of general knowledge, reading skill, and vocabulary. That is, human raters may be able to assess whether student writers possess relevant knowledge (e.g., appropriate historical references) and can incorporate that knowledge within comprehensible, well-worded text (e.g., showing that they have understood the prompt). The central question of the current study is whether these nuances are beyond the scope of what is captured by automated essay scores. Specifically, we consider whether algorithms that predict essay quality are correlated with measures of knowledge and reading skill, and how these relations compare to human ratings.

Method

The Writing Pal

Automated scoring in this study was powered by Writing Pal (W-Pal; Roscoe, Brandon, Snow, & McNamara, 2013; Roscoe & McNamara, 2013; Roscoe, Varner, Weston, Crossley, & McNamara, in press). W-Pal provides strategy instruction across multiple writing phases (i.e., prewriting, drafting, and revising) via short, animated lesson videos and educational practice games. W-Pal also allows students to practice writing argumentative essays and receive feedback. These essays are scored via algorithms that generate an overall holistic score and drive automated, formative feedback. This formative feedback provides actionable recommendations and strategies for students to revise their work. Evaluations of W-Pal have indicated that it facilitates gains in students' writing proficiency, essay revising, strategy acquisition, and self-perceptions of their writing ability (Crossley, Roscoe, & McNamara, 2013; Roscoe & McNamara, 2013; Roscoe et al., 2013).

Participants

In a prior study, high school students (*n* = 87) from the southwest United States enrolled in a 10-session program with W-Pal. Ethnically, 5.7% of students identified as African-American, 12.5% Asian, 19.3% Caucasian, and 54.5% Hispanic. Average age was 15.6 years with 62.1% female. Average grade level was 10.4 with 40.2% of students reporting a GPA of ≤ 3.0. Most students identified as native English speakers (*n* = 49), although many identified as English Language Learners (ELL, *n* = 38).

Participants began by writing a prompt-based pretest essay on the topic of either *Competition* or *Impressions* and completing demographic and individual differences measures (e.g., reading comprehension, vocabulary, and writing attitudes). Eight training sessions allowed students to learn and practice writing strategies with automated feedback. On the last day, students wrote a posttest essay on *Competition* or *Impressions* (counter-balanced with pretest) and completed attitude and perception surveys.

The original aim of this evaluation was to contrast learning with the complete W-Pal (i.e., lessons, games, and essays) versus a writing-intensive version (i.e., more writing practice but without lessons and games). Data on students' knowledge, reading, and vocabulary were collected as part of that research, and the current study uses those available data to begin assessing the construct validity of W-Pal's automated scoring. However, future work on construct validity will require that we incorporate a more comprehensive set of writing construct measures.

Corpus and Scoring

We examine the essays that students wrote *prior* to training. This corpus allows us to address our questions separately from the effects of instruction or feedback.

Expert raters were trained to assign a single holistic score to each essay on a scale of 1 (lowest) to 6 (highest). Each essay was scored by two raters (IRR ≥ .70) and the final holistic score was an average of the two expert ratings. Raters also used a scoring rubric comprising ten subscales on different aspects of writing (Table 1). Lead, Purpose, and Plan subscales pertain to the essay introduction and how writers establish key claims and concepts. Topic Sentences, Transitions, Organization, and Unity subscales relate primarily to the body of the essay and how writers elaborate, support, and connect their ideas to support the thesis. Perspective and Conviction subscales capture how writers summarize their main points in the conclusion and relate these ideas to the reader or to broader issues. Importantly, the above subscales depend upon writers' demonstration of their understanding of the prompt and upon development of their ideas and arguments. Finally, the Correctness subscale refers to the technical and mechanical quality of the writing.

Automated scoring was driven by algorithms designed to assess high school student essays in W-Pal. The algorithm includes diverse variables spanning lexical, syntactic, and semantic text features. The full algorithm is not presented here, but examples are provided to communicate the scope of the algorithm. For instance, lexical measures consider the use of varied wording (e.g., lexical diversity) and specificity (e.g., hypernymy). Syntactically, the algorithm assesses sentence construction and the use of various parts of speech. Semantically, the algorithm evaluates relatedness between the essay and the writing prompt (e.g., LSA comparisons) and the use of thematic words (e.g., words related to emotion). The algorithm does not explicitly detect properties such as the appropriateness of examples and evidence, accuracy of ideas, or logical presentation of arguments. To our knowledge, no published algorithms currently enable subjective judgments of this kind.

Table 1. Human essay scoring rubric subscales.

Subscale	Description
Effective Lead	Introduction begins with a surprising statistic, a quotation, a vivid description, an engaging fragment of dialog, or device to *grab the reader's attention* and point toward the thesis.
Clear Purpose	The introduction includes one or two sentences that provide key *background and establish the significance* of the discussion.
Clear Plan	The introduction ends with a *thesis statement* that provides a claim about the topic and a preview of the support and organizational principle to be presented in the essay body.
Topic Sentences	*Each paragraph includes a sentence (often at the start) that connects with the thesis* and makes a comment on one of the points outlined in the introduction.
Paragraph Transitions	Each topic sentence is preceded by a phrase, clause, or sentence that *links the current and prior paragraphs*, stressing the relationship between the two.
Organization	The body paragraphs *follow the plan set up in the introduction*, underscoring the organizational principle.
Unity	The details presented throughout the body *support the thesis* and do not stray from the main idea.
Perspective	The writer *summarizes the key points* that sustain the thesis and stress its significance.
Conviction	The author *re-establishes the significance* of the discussion as it pertains to the thesis.
Correctness	The writer *employs correct Standard American English*, avoiding errors in grammar, syntax, and mechanics.

The algorithm was developed with a corpus of 556 essays written by high school students and scored (on 1-6 scale) by expert raters. Accuracy was tested by examining how well automated ratings predicted human ratings. The algorithm accounted for 63% of the variance in human ratings for a training set of essays and 55% of the variance in human ratings for a test set. We observed a perfect agreement of 57% and an adjacent agreement of 98%.

Knowledge and Comprehension Measures

General Knowledge Test. To estimate students' general knowledge, multiple-choice questions regarding science, history, and literature were generated in several phases. Questions were taken from prior work to select predictive items with moderate difficulty (i.e., 30-60% of students could answer correctly). Each question was correlated with individual difference measures (e.g., reading skill) along with performance on comprehension tests. Questions with low correlations were eliminated. Second, additional items of moderate difficulty were obtained from test preparation item banks. Further questions were then generated by sampling topics in high school textbooks. In this process, 55 multiple-choice questions (i.e., 18 science, 18 history, and 19 literature) were piloted with 15 undergraduates to test item performance. Thirty questions (10 per domain) were selected such that no items selected exhibited either a ceiling (> .90) or floor effect (< .25, chance level). Scores on the 30 knowledge test items are summed to provide a single knowledge score. Examples are given in Table 2.

Table 2. Examples of knowledge question and answers.

Domain	Question and Answer Choices
Science	The poisons produced by some bacteria are called… a) antibiotics, b) toxins, c) pathogens, d) oncogenes.
History	A painter who was also knowledgeable about mathematics, geology, music, and engineering was… a) Michelangelo, b) Cellini, c) Titian, d) da Vinci.
Literature	Which of the following is the setting used in "The Great Gatsby"… a) New York, b) Boston, c) New Orleans, d) Paris

Gates-MacGinitie Reading. Reading comprehension skill was tested with the Gates-MacGinitie (4th ed.) reading skill test (form S) level 10/12 (MacGinitie & MacGinitie, 1989). The test consisted of 48 multiple-choice questions assessing students' comprehension of 11 short passages. Each passage was associated with two to six questions, which measured shallow comprehension as well as deeper comprehension that required the reader to make inferences about the text. The participants were administered the standard instructions with 20 minutes to complete the test.

Gates-MacGinitie Vocabulary. The vocabulary section of the Gates-MacGinitie (4th ed.) test (form S) level 10/12 (MacGinitie & MacGinitie, 1989) was used to assess vocabulary skill. The test comprised 45 sentences or phrases, each with an underlined vocabulary word. For each underlined word, participants were asked to select the most closely related word from a list of five choices. The items were designed to provide no contextual information about meaning. Participants were administered the standard instructions with 10 minutes to complete the test.

Results

Essay Scores

Human and automated holistic scores were positively correlated, $r = .63$, $p < .001$. Essays that were judged to be higher quality by human raters were also rated as higher quality by the scoring algorithm.

We then conducted a correlation and regression analysis to examine the relations among human holistic scores and subscale ratings (Table 3). If expert raters are assessing writers' knowledgeable and comprehensible expression and defense of their arguments and ideas, then many of the subscale ratings should be related to holistic scores.

Table 3. Essay score and subscale correlations.

	Holistic Ratings	
Essay Ratings	Human	Automated
Automated Holistic	.63[a]	--
Human Holistic	--	.63[a]
Human Subscales		
Lead	.43[a]	.37[a]
Purpose	.63[a]	.44[a]
Plan	.77[a]	.51[a]
Topic Sentences	.72[a]	.62[b]
Transitions	.56[a]	.41[a]
Organization	.79[a]	.67[a]
Unity	.80[a]	.52[a]
Perspective	.51[a]	.51[a]
Conviction	.45[a]	.46[a]
Correctness	.53[a]	.31[b]

Note. [a]$p < .001$. [b]$p < .01$. [c]$p < .05$.

Correlations ranged from $r = .43$ (Lead) to $r = .80$ (Unity); all *p*-values were below .001. A stepwise linear regression assessed how much of the variance in human ratings could be explained by subscales. The model was significant, $F(10, 86) = 72.23$, $p < .001$, $R^2 = .84$, accounting for about 84% of the variance. Significant predictors included the Unity ($\beta = .374$, $p < .001$), Plan ($\beta = .356$, $p < .001$), Perspective ($\beta = .171$, $p = .002$), Lead ($\beta = .125$, $p = .015$, Correctness ($\beta = .118$, $p = .028$), and Transitions $\beta = .115$, $p = .038$) subscales. Human ratings thus seemed to relate to whether students could engage the reader with details, establish a clear thesis, communicate main ideas, and support their arguments.

We next conducted correlation and regression analyses to examine relations between automated scores and the human subscales. The extent to which automated scores correlate with human subscale ratings provides an indicator of whether automated scores are capturing students' skillful expression of ideas and arguments.

Correlations ranged from $r = .31$ (Correctness) to $r = .67$ (Organization), which were smaller in magnitude than correlations between human holistic scores and subscales. A stepwise linear regression assessed the amount of the variance in automated ratings explained by subscales. The model was significant, $F(3, 86) = 30.94$, $p < .001$, $R^2 = .53$, accounting for about 53% of the variance. Significant predictors included the Organization ($\beta = .475$, $p < .001$), Perspective ($\beta = .286$, $p = .001$), and Lead ($\beta = .173$, $p = .039$) subscales. These results imply that human raters may have been more sensitive to higher-level aspects of writing but automated scores were also linked to these factors.

Altogether, these results are perhaps to be expected; correlations should be stronger between ratings made at the same time from the same source (i.e., *human* holistic and *human* subscale ratings) than from different sources (i.e., *automated* holistic ratings versus *human* subscale ratings). However, these findings suggest that humans were better attuned to how students conveyed and argued their ideas. In subsequent sections, we turn to external measures of general knowledge and reading skill. Students who are more skilled at demonstrating their knowledge, thoughtful reading and understanding of the prompt, and appropriate vocabulary should be able to write more proficiently. In other words, human-assigned scores should be positively correlated to measures of knowledge and reading skill. The extent to which automated scores are, or are not, correlated with these factors sheds further light on how or whether automated scores capture a more limited writing construct.

General Knowledge

Correlations were conducted to examine relations between general knowledge and essay ratings (Table 4). Human ratings were significantly correlated with knowledge scores ($r = .30$, $p = .005$). Knowledge was also related to Unity, Plan, Organization, Transitions, and Correctness subscales. As expected, students' knowledge was related to human judgments of essay quality and specific aspects of the essays. Students who knew more about science, literature, and history, and who might reference such details in their text, were judged to be more proficient writers.

Automated scores were not significantly related to measures of general knowledge ($r = .20$, $p = .07$), although they were positively correlated. The algorithm may have been somewhat less attuned to how or whether students incorporated their world knowledge into their writing.

Table 4. Essay ratings and knowledge score correlations.

Essay Ratings	Knowledge
Automated Holistic	.20
Human Holistic	.30^{b}
Human Subscales	
Lead	.13
Purpose	.20
Plan	.29^{b}
Topic Sentences	.16
Transitions	.24^{c}
Organization	.28^{b}
Unity	.30^{b}
Perspective	.15
Conviction	.14
Correctness	.22^{c}

Note. $^{a}p < .001$. $^{b}p < .01$. $^{c}p < .05$.

Reading Comprehension Skill

Correlations were conducted to examine relations between students' reading comprehension skill and essay ratings (Table 5). Reading comprehension was correlated with human holistic ratings (r = .46, p < .001) and with the Lead, Purpose, Plan, Transitions, Organization, Unity, and Correctness subscales. Thus, students' ability to comprehend text was related to many aspects of proficient writing (see also Fitzgerald & Shanahan, 2000). Skilled comprehenders were more likely to write essays that engaged the reader, stated and supported clear arguments, and presented ideas in organized and coherent manner.

Reading comprehension scores were also correlated with automated holistic ratings (r = .34, p = .001), although to a lesser magnitude than human holistic scores. Human raters may have been more attuned to students' apparent understanding of the ideas discussed in their essays.

Vocabulary Skill

Measures of vocabulary skill were correlated with human and automated holistic ratings (Table 5). Vocabulary was correlated with human ratings (r = .42, p < .001) and with Lead, Purpose, Plan, Topic Sentences, Transitions, Organization, Unity, and Correctness subscales. Similarly, vocabulary was correlated with automated scores (r = .29, p = .006). Thus, both human and automated essay ratings were positively correlated with vocabulary skill, although correlations with human ratings were again higher.

In practice, automated tools may be more applicable to detecting students' sophisticated word use. For instance, WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) databases may allow systems to estimate whether students' word choices tend toward uncommon versus common wording, or vague versus specific wording. One might expect automated ratings to correlate well with vocabulary measures. However, incorporating "advanced words" is no guarantee that students understand or use the terms correctly. Human holistic scores seemed to be more able to capture subtle failures of students' vocabulary skill than were the automated holistic scores (Table 5).

Table 5. Essay ratings and reading measure correlations.

	Reading Skill Measures	
Essay Ratings	Comprehension	Vocabulary
Automated Holistic	.34^{b}	.29^{b}
Human Holistic	.46^{a}	.42^{a}
Human Subscales		
Lead	.42^{a}	.41^{a}
Purpose	.37^{a}	.38^{a}
Plan	.48^{a}	.46^{a}
Topic Sentences	.26^{c}	.27^{c}
Transitions	.28^{c}	.28^{b}
Organization	.40^{a}	.40^{a}
Unity	.38^{a}	.32^{b}
Perspective	.18	.15
Conviction	.20	.17
Correctness	.43^{a}	.38^{a}

Note. $^{a}p < .001$. $^{b}p < .01$. $^{c}p < .05$.

Discussion

Despite the rapid proliferation of automated essay scoring programs, and despite evidence of reliability or accuracy, automated scoring will continue to meet resistance unless better construct validity can be obtained (Deane, 2013). Critics have argued that automated tools are "unable to recognize or judge those elements that we most associate with good writing" such as "logic, clarity, accuracy, ideas relevant to a specific topic, innovative style, effective appeals to audience, different forms of organization, types of persuasion, quality of evidence, humor or irony, and effective uses of repetition" (Anson et al., 2013).

In this study, we considered this issue in the context of available data and the scoring algorithms used by the W-Pal intelligent tutor. Specifically, we sought to examine how expert human ratings and automated ratings of students' essays were related to each other and to measures of general knowledge and reading skill. As expected, and as argued by experts in writing assessment, human raters seemed to take a variety of higher-level factors into account when judging the quality of an essay. Human raters seemed to consider how and whether arguments and ideas were presented in an engaging, organized, and compelling manner, and how these ideas were supported by meaningful evidence and linked to broader issues. Additionally, human raters seemed to be influenced by how students' essays demonstrated students' knowledge, reading skill, and vocabulary skill.

The more fundamental question of the current study was how and whether automated essay scores correlated with writing quality, knowledge, and comprehension. A key finding is that *automated scoring in W-Pal was related to higher-level aspects of writing quality and argumentation*. Automated holistic ratings (Table 3) were correlated with all subscales, and were specifically predicted by students' engagement of the reader in the introduction (e.g., Lead subscale), support and organization of ideas in the body of the essay (e.g., Organization subscale), and summarizing key ideas in the conclusion (e.g., Perspective subscale). Automated holistic ratings were also correlated to students' reading comprehension and vocabulary skills. These outcomes suggest that automated scoring algorithms were at least partially able to capture elements of student writing tied to their knowledge or understanding of the topics.

Importantly, although the data show that the algorithms were related to several higher-level features, results also reveal ways in which automated scores were limited compared to humans. Correlations between automated holistic ratings and measures of writing quality (Table 3), knowledge (Table 4), and reading skill (Table 5) tended to be lower than correlations between these same measures and human holistic ratings. Such findings add credence to concerns about the validity of automated scores. Automated scoring in W-Pal may overlook aspects of writing quality related to how students use their world knowledge to establish or defend their ideas.

To improve the efficacy of W-Pal or other intelligent systems for writing instruction, a necessary innovation will be expanded algorithms that explicitly address higher-level features of writing quality. Although available scoring methods demonstrate accuracy, improved construct validity will enable more meaningful and interpretable scores, thus overcoming some students' and teachers' resistance to AWE (e.g., Grimes & Warschauer, 2010). Expanded algorithms also unlock novel possibilities for delivering formative feedback related to crucial writing processes, such as actionable strategies for crafting compelling and well-supported arguments. As new algorithms become more conceptually integrated with core writing constructs, the next generations of AWE tools will become more valuable and accepted by teachers who strive to nurture sophisticated and skillful student writers.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080589 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Anson, C., Filkins, S., Hicks, T., O'Neill, P., Pierce, K. M., & Winn, Maisha. (2013). *NCTE position statement on machine scoring: Machine scoring fails the test*. National Council of Teachers of English. Retrieved from http://www.ncte.org.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning and Assessment,4*.

Camara, W. J. (2003) *Scoring the essay on the SAT writing section.* College Entrance Examination Board, New York.

Chiesky, J., & Hiebert, E. H. (1987). The effects of prior knowledge and audience on high school students' writing. *Journal of Educational Research*, *80*, 304-313.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*, 100-108.

Crossley, S. A., Roscoe, R. D., & McNamara, D. S. (2013). Using natural language processing algorithms to detect changes in student writing in an intelligent tutoring system. *Proceedings of the 26th International Florida Intelligence Research Society Conference* (pp. 208-213). Menlo Park, CA: AAAI Press.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7-24.

Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist*, *35*, 39-50.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning and Assessment, 8*.

Landauer, T. K., Lochbaum, K. E., & Dooley, S. (2009). A new formative assessment technology for reading and writing. *Theory into Practice, 48*, 44-52.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds). (2007). *Handbook of latent semantic analysis*. New York: Routledge.

MacGinitie, W. H., & MacGinitie, R. K. (1989). *Gates-MacGinitie reading tests*. Chicago: Riverside Publishing.

McCutchen, D. (2000). Knowledge, processing, and working memory: Implications for a theory of writing. *Educational Psychologist, 35*, 13-23.

McNamara, D. S., Graesser, A. C., McCarthy, P., Cai, Z. (in press). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge: Cambridge University Press.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An online lexical database. *International Journal of Lexicography, 3*, 235-244.

Roscoe, R. D., Brandon, R. D., Snow, E. L., & McNamara, D. S. (2013). Game-based writing strategy practice with the Writing Pal. In K. Pytash & R. Ferdig (Eds.), *Exploring technology for writing and writing instruction* (pp. 1-20). Hershey, PA: IGI Global.

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*.

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*.

Shanahan. T., & Lomax, R. G. (1986). An analysis and comparison of theoretical models of the reading-writing relationship. *Journal of Educational Psychology, 78*, 116-123.

Shermis, M. D., & Burstein, J. C. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Psychology Press.