The Journal of Writing Assessment
Volume 13, Issue 1: 2020

The BABEL Generator and E-Rater: 21st Century Writing Constructs and Automated Essay Scoring (AES)

Les Perelman, Massachusettes Institute of Technology

Automated essay scoring (AES) machines use numerical proxies to approximate writing constructs. The BABEL Generator was developed to demonstrate that students could insert appropriate proxies into any paper, no matter how incoherent the prose, and receive a high score from any one of several AES engines. Cahill, Chodorow, and Flor (2018), researchers at Educational Testing Service (ETS), reported on an Advisory for the e-rater AES machine that can identify and flag essays generated by the BABEL Generator. This effort, however, solves a problem that does not exist. Since the BABEL Generator was developed as a research tool, no student could use the BABEL Generator to create an essay in a testing situation. However, testing preparation companies are aware of e-rater's flaws and incorporate the strategies designed to help students exploit these flaws. This test prep does not necessarily make the students stronger writers just better test takers. The new version of e-rater still appears to reward lexically complex, but nonsensical essays demonstrating that current implementations of AES technology continue to be unsuitable for scoring summative, high stakes writing examinations.

Keywords: Automated essay scoring (AES), BABEL generator, writing constructs, writing assessments, fairness

---

Automated essay scoring (AES) was first developed by Ellis Batten Page in the 1960s. Page (1966) coined two terms: *trin* and *prox*.

> A *trin* is the intrinsic variable of real interest to us. For example, we may be interested in a student's "aptness of word choice," or "diction." A *prox*, on the other hand, is some variable which it is hoped will approximate the variable of true interest. For example, the student with better diction will probably be the student who uses a less common vocabulary. At present, the computer cannot *measure directly the semantic aptness of expression in context, or "diction." But* it can discover the proportion of words not on a common word list, and this proportion may be a prox for the trin of diction. (p. 240)

However, rather than use Page's obscure terminology, I will use the more common terms of *construct* in place of *trin* and *proxy* in place of *prox*. Despite the insertion of a complex psychometric vocabulary, AES has not really progressed beyond Page's (1966) formulation. Testing companies freely use the term *artificial Intelligence*, but most of the systems appear to produce a holistic score largely through summing weighted proxies. Fifty years after Page's article, ETS's e-rater 2.0, for example, followed Page's formulation by calculating the construct of *lexical complexity* through two quantitative proxies, the frequency of infrequently used words and the average number of characters in a word: the more rarely used and long words, the higher the numerical value of the construct (Attali & Burstein, 2006). Meaning is irrelevant.

These values that comprise the writing construct are not arbitrarily assigned, but derive from analyzing very large corpora of student essays scored by human readers and formulating a weighted string of proxies, the sum of which will most closely approximate the set of scores given by human readers. The problem is that, although these proxies may be associated with a construct, they are independent of it. If one enters a house and goes into large library with shelves of books lining each of the three walls, there is a high probability that the individual who lives in the house is well read. However, someone can also buy the books to give the impression they are well read when they never read at all. Indeed, interior decorators can buy books by the unit of shelf-foot. Similarly, students can memorize long rarely used words to increase their scores on essays graded by a machine.

To test the hypothesis that AES systems would award high scores to essays that contained the right proxies in sufficient number but contained no meaning, in 2014, I and three then undergraduates, Louis Sobel and Damien Jiang from MIT and Milo Beckman from Harvard, developed the Basic Automatic BS Essay Language (BABEL) Generator (Kolowich, 2014; Sobel, Beckman, Jiang, & Perelman, 2014). Our purpose was to demonstrate through an extreme example that a student could achieve high scores on machine graded essays simply by memorizing and providing the right proxies (Anson & Perelman, 2017; Kolowich, 2014; Perelman, 2016a). We based our design largely on the proxies described for e-rater V.2 (Attali & Burstein, 2006).

The machines were much less sophisticated than we expected. Meeting once a week, in only four weeks, we developed an application that was able, using the input of one to three words, to generate essays of complete gibberish that received high scores from the four scoring engines we could access, *e-rater*, *Lightside*, *My Access*, a classroom version of *Intellimetric* by Vantage Technologies, and *EASE*, developed at MIT for EdX. We were unable to access Pearson Technologies' Intelligent Essay Assessor (IEA). In 2012, Mike Winerip, a reporter for *The New York Times,* asked Peter Foltz, Vice President of Cognitive Computing in Pearson's AI and Products Solutions, to give me access to the IEA. Dr. Foltz turned down Winerip's request because I wanted "to show why it doesn't work" (Winerip, 2012). We were able to gain access to e-rater by paying to take web-based practice versions of the two essays contained in the Graduate Record Examination (GRE) at the *Score It Now* website.

The remainder of this article focuses specifically on e-rater, utilizing data generated by submissions of BABEL-generated essays to Score It Now. One reason for focusing on e-rater's scoring of the GRE is that the same year as the development of the BABEL Generator, ETS researchers published a study that showed only a modest gain in scores by substituting longer less frequently used words in place of shorter ones (Bejar, Flor, Futagi, & Ramineni, 2014). However, in the ETS study, only 5% of the words in an essay, for example 10 words in a 200-word essay, were replaced with longer, less frequently used words.

Because some of the other AES engines had a 500-word limit, BABEL was programmed to generate essays of approximately 500 words. However, since e-rater employs essay length in words as one of its major proxies (Attali & Burstein, 2006; Perelman, 2012, 2013, 2014) and would accept essays of up to 1,000 words, longer submissions of up to 999 words were created by generating two

essays from the same set of keywords and then randomly interleaving paragraphs from the two texts. I generated 19 pairs of essays (one argumentative essay and one essay analyzing an argument) which received scores ranging from 4-4 to 6-6 on the standard 1-6 scale.

Score It Now replicates the GRE Writing Test. Each test consists of a set of two essays. The first essay, which ETS defines as the *Issue Essay*, asks the test-taker to write an argumentative essay responding to a specific assertion. The second essay, which ETS defines as the *Argument Essay*, requires a written analysis of a short argument.

The two following excerpts from BABEL-generated responses to the *Issue Essay* and *Argument Essay*, each of which received the highest score of 6, illustrate several glaring deficiencies in e-rater:

> ***Issue Essay***
>
> Educatee on an assassination will always be a part of mankind. Society will always authenticate curriculum; some for assassinations and others to a concession. The insinuation at pupil lies in the area of theory of knowledge and the field of semantics. Despite the fact that utterances will tantalize many of the reports, student is both inquisitive and tranquil. Portent, usually with admiration, will be consistent but not perilous to student. Because of embarking, the domain that solicits thermostats of educatee can be more considerately countenanced. Additionally, programme by a denouncement has not, and in all likelihood never will be haphazard in the extent to which we incense amicably interpretable expositions. In my philosophy class, some of the dicta on our personal oration for the advance we augment allure fetish by adherents. The reprimand should, in any case, be rapacious, abhorrent, and enthusiastic.
>
> ***Argument Essay***
>
> Theatre on proclamations will always be an experience of human life. Humankind will always encompass money; some for probes and others with the taunt. Money which seethes lies in the realm of philosophy along with the study of semiotics. Instead of yielding, theatre constitutes both a generous atelier and a scrofulous contradiction. As I have learned in my reality class, theatre is the most fundamental analysis of human life. Gravity catalyzes brains to transmit pendulums to remuneration. Although the same gamma ray may receive two different pendulums at the study of semiotics, a plasma processes interference. Simulation is not the only thing an orbital implodes; it also inverts on theater. If some of the reprimands which whine comment those involved, epitome at recommendation can be more zealously enlightenmented. The more solicitation placates a affirmation, the more an injunction contends and may be assemblage but utters explanations. As I have learned in my theory of knowledge class, theater is the most fundamental amygdala of humanity. The neuron by pedant catalyzes neutrinoes to transmit information. Though the same gamma ray may emit two different pendulums, gravity with the report for circumscriptions produces the brain. ("BABEL Generated GRE Essays {excerpts}," 2016)

First, argument and, indeed, meaning are irrelevant to e-rater. Second, e-rater's scoring algorithms for the two rhetorically very different writing tasks appear to be identical. Moreover, e-rater does not notice some of the most egregious grammatical mistakes, especially article and preposition errors. Computers are notorious for failing to identify grammar errors correctly. At best, computers can identify a little over half the grammatical errors in a text without generating a substantial number of false positives (Gamon, Chodorow, Leacock, & Tetreault, 2013; Leacock, Chodorow, Gamon, & Tetreault, 2014; Perelman, 2016b). In addition, AES machines do not treat all errors equally. Computers can much more reliably identify some errors, such as verb formations, than it can correctly and consistently identify article and prepositions errors (Dikli & Bleyle, 2014). In identifying preposition use errors, *Criterion,* the classroom implementation of e-rater, only identifies about 25% of the errors present in texts, and about 20% of its tags on preposition use are false positives (Chodorow, Gamon, & Tetreault, 2010; Leacock et al., 2014; Tetreault & Chodorow, 2008). In detecting article errors, Criterion correctly identifies only about 40% of the errors while 10% of its reported errors are false positives (Han, Chodorow, & Leacock, 2006). E-rater's inability to detect half of all article errors may contribute to its giving significantly higher scores to native Mandarin speaking students on the GRE essays than the scores of human raters. Conversely, the GRE essays of African Americans, whose forms of English often differ from Standard English by offering more precise verb tenses, are given lower scores by e-rater, which is quite proficient in identifying non-standard verb formations, than the scores they receive from human graders (Bridgeman, Trapani, & Attali, 2012; Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012). Similarly, in one study, Criterion, the classroom implementation of e-rater, was an accurate predictor of course grades for students in all groups except African Americans (Elliot & Klobucar, 2013; Klobucar, Deess, Rudniy, & Joshi, 2013).

Another major characteristic of AES algorithms is that length matters. In the BABEL experiments on the practice GRE website, 16 of the 38 essays received the highest score of 6. Of those 16 essays, all but two had a length between 950-999 words. Length also matters with other AES machines. Mark D. Shermis, a strong advocate of AES, reported that he has run the Gettysburg Address, what Gary Wills (1992) has called "the words that remade America," through several early AES machines, and the 271 word document received only 2s and 3s (Bloom, 2012).

The focus on length is a central element of e-rater's algorithm for calculating a holistic score. Development and organization are calculated simply by counting the number of *discourse elements* (the ETS term for paragraphs) in an essay and their average length (Attali & Burstein, 2006; Attali & Powers, 2008; Quinlan, Higgins, & Wolff, 2009). Criterion, the classroom adaptation of e-rater, flags any paragraph with fewer than four sentences.

Perhaps the most troubling revelation from the BABEL experiments is that the writing construct privileged by e-rater is antithetical to over 100 years of widely-held standards of what constitutes clear and effective English prose. Instead of following Strunk and White's (1979) advice to "omit needless words" (p. 6), the passage above contains sentences such as "Despite the fact that utterances will tantalize many of the reports, student is both inquisitive and tranquil." Instead of adhering to Sir Ernest Gowers'

(1973) dictum that "the adjective is the enemy of the noun" (p. 37) and Orwell's (1954) rule, "Never use a long word where a short one will do" (p. 155), e-rater gave a top score of 6 to this essay, which contains the following two sentences:

> Still yet, armed with the knowledge that the report with infusion can petulantly be the injudicious stipulation, none of the lamentations by my circumstance compel inconsistency but agree. In my experience, many of the quips at our personal admonishment on the allocation we countenance collapse or disrupt risibly unsophisticated precincts.

Finally, the entirety of any of these BABEL-generated essays stands in sharp contrast to Stephen Pinker's (2014) advice to use infrequently used words, but only employ them judiciously.

Because e-rater's scoring of BABEL-generated essays provides graphic and easily understood examples of both e-rater's obliviousness to meaning and the kind of prose it favors, the BABEL Generator provides strong and easily comprehended evidence of AES's primary deficiency: that it fails to measure any reasonable formulation of the writing construct.

ETS has employed Advisories in conjunction with e-rater for over a decade (Bejar et al., 2014; Higgins, Burstein, & Attali, 2006; Higgins & Heilman, 2014; Zhang, Chen, & Ruan, 2016). The main purpose of these Advisories has been to prevent test takers gaming the machine by identifying essays that possess what the psychometricians dub *construct irrelevant features* students could produce that would elicit a high score from the machine but not from human readers (Zhang et al., 2016). ETS Advisories identify such elements as excessive repetition of words, phrases, or sentences, being off-topic, restatement of prompt text, the essay being too short or too long, and atypical organizational structures that cannot be parsed by the machine (Zhang et al., 2016). The importance of developing such Advisories is the expectation by testing organizations that the expansion of national testing at all levels along with cost and time required for human readers will soon necessitate machines such as e-rater becoming the primary evaluator of student writing in high stakes testing (Zhang et al., 2016).

In a recent article in the *Journal of Writing Analytics*, Cahill, Chodorow, and Flor (2018) reported on the development of an Advisory to flag essays that may have been developed using the BABEL Generator. The BABEL Advisory differs from the other Advisories in that it offers a solution in search of a problem. Students will never be able to use the BABEL Generator in real test situations. The BABEL Generator was not created to aid students in tests but to expose the core deficiencies of the reductive approach employed by AES machines that is antithetical to commonly accepted constructs of writing. The only problem that the study does try to solve is one of public relations by intimating that the Advisory resolves all the issues raised by the BABEL Generator's success (Seabrook, 2019).

While e-rater may now be able to currently detect BABEL-generated essays, it almost certainly will not be able to detect essays written by students employing the strategies used by the BABEL generator, particularly inflating essay length and memorizing infrequently used long words. Students can be *undoubtedly* and *indubitably* taught to insert such elements as long adverbial and adjectival pairs that are both *repetitious* and *redundant* (such as the two pairs in this sentence). What is tested becomes what is taught (Fanetti, Bushrow, & DeWeese, 2010; Higgins, Miller, & Wegmann, 2006; Jennings & Bearak, 2014; Posner, 2004). If machines score high stakes writing tests, teachers will be forced to teach to the test by having students memorize word lists and learn how to inflate prose. Rather than teaching students to communicate effectively to readers, teachers will be forced to emphasize the writing of verbose and pretentious prose within the cognitive straightjacket of the five-paragraph essay (Fanetti et al., 2010; White, 2008). Indeed, almost 40 years ago, Bruffee (1983) noted that, in a set of CUNY student placement essays, human readers immediately saw evidence for test preparation, evidence AES programs would not see as such and would, in fact, reward: "During one testing period at Brooklyn college, for example, readers found a whole batch of papers in which coherence had been attempted by repeating the same set of adverbial connectors in the same order, paragraph by paragraph: 'however,' 'accordingly,' 'therefore'" (p. 3).

The effects of being able to understand the limited construct that AES machines can assess has consequences for test takers. Indeed, ETS researchers themselves acknowledge the susceptibility of e-rater to both coaching and gaming when discussing e-rater's scoring mainland Chinese on average over a half a point higher than human raters ($d = 0.60$) on the GRE issue essay:

> Another possible explanation for the greater discrepancy between human and machine scores for essays from mainland China may be the dominance of coaching schools in mainland China that emphasize memorizing large chunks of text that can be recalled verbatim on the test. Human raters may assign a relatively low score if they recognize this memorized text as being somewhat off topic, though not so far off topic as to generate a score of 0. On the other hand, this grammatical and well-structured memorized text would receive a high score from e-rater. (Bridgeman et al., 2012, pp. 36–37)

The difference in scores for mainland Chinese given by humans and e-rater as well as e-rater's scoring African Americans, particularly African American males, lower than they are scored by human raters are issues that ETS researchers have noted (Bridgeman et al., 2012; Ramineni et al., 2012) but that, apparently, ETS has not addressed.

In addition, the data reported in Cahill et al.'s (2018) study appear to confirm that meaning is irrelevant to e-rater and offer strong evidence for questioning the use of AES for high-stakes tests. Although now flagged, the BABEL-generated essays still receive moderate to high scores from the new version of e-rater. Cahill et al. employed five large data sets. The first two were, or were similar to, the two essays included in the Graduate Record Examination: The first involves making an argument on a topic; the second asks for the analysis of an argument. The third set may have been the argumentative essay of the PRAXIS examination for teachers or a very similar exercise. The fourth set is almost certainly the Test of English as a Foreign Language (TOEFL) Independent Essay, asking students to make an argument, and the fifth set is probably the TOEFL Integrated Essay, which asks to students to compare two different sources on the same subject.

All data sets were scored on the standard 1-6 scale. For the first and second data sets, the new version of e-rater split the scores in the middle, giving a majority of 4s with the rest 3s for the issue essay and a majority 3s with the rest 4s for the argument essay. The scores would almost certainly have been higher if the outputs from the BABEL Generator were doubled as they were in the BABEL experiments to expand the length of the essays closer to the 1,000 word limit. In the third data set, argument essays from the teachers proficiency test, there were fewer 5s and more 4s, although 5s still comprised about 75% of the 20,000 essays. Most frightening, however, are the scores on the two English-as-a-foreign language essays. The scores on the argumentative essay dropped from 5s to still-upper-half 4s. However, the scores on comparative essay remained the same, almost entirely 5s, a very high score, especially for incoherent gibberish. These facts alone should be evidence enough to convince ETS and other testing companies that AES does not work.

Another crucial question is how does the new 2016 version of e-rater perform overall? Does it match the 2012 version in agreement with human readers as measured by the Pearson *r* and the Quadratic Weighted Kappa? In addition, does the new version still give higher scores than do human readers for mainland Chinese Mandarin speakers and lower scores than do human readers to African Americans as did the 2012 version (Bridgeman et al., 2012; Ramineni et al., 2012)?

The research article also demonstrates the severe limitations of big-data analytics, especially in writing research. The machine will only notice those features it is programmed to notice. In addition, missing from the article's literature review are any references to the many critiques of machine scoring, many of which directly address the issue key to the BABEL experiments: that machines do not understand meaning. This issue is a recurring theme in at least one collection of essays on machine scoring (Freitag Ericsson & Haswell, 2006) as well other articles (Condon, 2013; Herrington & Moran, 2001, 2012; McCurry, 2010; Perelman, 2012) in addition to other ones not cited here.

Deane (2013) articulated a validity argument that outlines how e-rater assesses vocabulary, accuracy, and fluency. These are necessary skills for a writer to master thereby to free cognitive capacity for successfully engaging higher order cognitive and rhetorical elements of the writing construct such as argumentation and audience awareness. E-rater's construct for vocabulary, also includes *lexical complexity*, which consists of two features: 1) the relative frequency of words in the essay based on the Lexile corpus (Quinlan et al., 2009)—the less frequently used the word, the higher its contribution to the feature score; and 2) "the mean average [sic] number of characters within words" (Quinlan et al., 2009, p. 35).

Deane's (2013) notion of accuracy refers to correctness of spelling, grammar, usage, and mechanics. The previous examples from the BABEL experiments as well as the literature review have already exposed e-rater's ineptness in identifying errors. On the other hand, a previous study (Perelman, 2016b) illustrates e-rater's proclivity towards false-positives.

The ubiquity of computers as the primary instrument college students employ to transcribe writing leads us to consider another of Deane's (2013) components of the writing construct assessed by e-rater, fluency. By demonstrating they can write lengthy paragraphs that form long essays, he argues, students are demonstrating a fluency in text production that frees up cognition to address higher order skills. There are, however, two major problems with this argument.

First, much of this claim is based on the work of McCuthen and her colleagues written between 1994 and 2000 (McCutchen, 1996, 2000; McCutchen, Covill, Hoyne, & Mildes, 1994). These studies employed samples of very short timed writing, 12 or 15 minutes for each essay, on populations of elementary and middle school students in middle-class suburban neighborhoods (McCutchen et al., 1994). Moreover, most of the subsequent studies of cognitive demands on short term memory in writing focused primarily on the cognitive demands of spelling and handwriting (McCutchen, 1996; Swanson & Berninger, 1996).

Second, the ecology of text production has changed radically. Entering the third decade of the 21st century, handwriting, for some students, may be only a transitional skill. Computer spell checkers are used primarily as aids in text production not, as its implementation in e-rater, as a measure of autonomous orthographic knowledge. Moreover, spell checkers are useful in correcting unintentional keyboarding errors, such as omitting a space between two words and typing an extra period, errors that e-rater does not correct but penalizes. Given that students have 30 minutes to write each of the two GRE essays, they have too little time to proofread and also produce the substantial text necessary for a high score. Keyboarding with e-rater becomes as much a typing test as it is a writing exercise. Basic research on how students and adults at various ages and from different populations now compose and inscribe is needed before any argument, such as the one offered by Deane (2013), can be considered.

As fairness becomes a more widely considered part of test taking, all applicants should be informed of e-rater's scoring criteria. Right now, the ETS web page states:

> For the Analytical Writing section, each essay receives a score from at least one trained rater, using a six-point holistic scale. In holistic scoring, raters are trained to assign scores on the basis of the overall quality of an essay in response to the assigned task. The essay is then scored by *e-rater*®, a computerized program developed by ETS that is capable of identifying essay features related to writing proficiency. If the human and the *e-rater* scores closely agree, the average of the two scores is used as the final score. If they disagree, a second human score is obtained, and the final score is the average of the two human scores. (Educational Testing Service, n.d.-a, n.d.-b)

The phrase "essay features related to writing proficiency" provides applicants with no useful information. The test preparation industry, however, certainly knows what features e-rater is privileging, further advantaging wealthier applicants who can afford their services.

It is the following paragraph, however, that contains the most egregious instance of misinformation. "The primary emphasis in scoring the Analytical Writing section is on your critical thinking and analytical writing skills rather than on grammar and mechanics." (Educational Testing Service, n.d.-a, n.d.-b). E-rater provides half the final score. Yet, e-rater does not emphasize

"critical thinking and analytic writing skills." Indeed, it is completely oblivious to them. Its closest approximation is its highly reductive feature of *development*, which is calculated by the number of sentences in each paragraph and the number of paragraphs in the essay. Furthermore, grammar and mechanics compose a significant portion of the features included in e-rater's calculations. Low-income students will believe these statements and focus on critical thinking and analytic skills. Affluent students who have taken test preparation classes, on the other hand, will be coached to provide e-rater with the proxies that will inflate their scores.

The BABEL experiment along with the other arguments presented here provide evidence that e-rater and similar AES machines are questionable for scoring summative writing assessments. E-rater is unable to consistently assess correct grammar and usage and posits a construct for diction that is antithetical to simplicity and clarity in language. Furthermore, the construct of fluency as described by Deane (2013) and others is anachronistic and may have little relevance to 21st-century writing practices. Rather than challenge these conclusions, the results of Cahill et al.'s (2018) study reinforce them. That even the new version of e-rater still gives high scores to BABEL-generated gibberish for what are almost certainly TOEFL and PRAXIS essays is clear and convincing proof that the use of e-rater in scoring these tests should cease immediately. E-rater's deficiencies in assessing grammaticality alone make it unsuitable to assess the English language skills of English language learners in the TOEFL Examination and, in conjunction with e-rater's inability to comprehend meaning, even more unsuited to assess the writing abilities of novice teachers in the PRAXIS Test.

These arguments should not be construed as blanket objections to the use of computers in the writing classroom. Computers have immense potential in writing instruction as aids to revision, platforms for peer review, and even for formative assessments. However, because the technology cannot pass through "the barrier of meaning" (Mitchell, 2018, 2019; Rota, 1985), computers, given their current limitations, should never be employed in summative writing assessments. Writing is, in essence, a mechanism for the transfer of meaning from one mind to another. Moreover, the evaluation of all writing assessments, both formative and summative, requires substantial research into how students, especially young adult students in the case of the GRE , compose and inscribe in the 21st century.

Acknowledgements

I wish to thank the editors of *The Journal of Writing Assessment* for their patience and guidance in the development of this article. I also wish to thank the three anonymous reviewers, whose comments helped to improve and refine my arguments. Finally, I wish to thank Professor Richard Haswell for his review and input into the article and, especially his contribution of the Bruffee reference and quotation.

Author Note

Les Perelman (perelman@mit.edu) retired after thirty years of directing writing programs at Tulane University and the Massachusetts Institute of Technology, where he also served as an Associate Dean of Undergraduate Education. He served on the Executive Committee of the Conference on College Composition and Communication and co-chaired that organization's Committee on Assessment. He has been a consultant to over twenty colleges and universities on the assessment of writing, program evaluation, and writing-across-the-curriculum. Recently, the New South Wales Teacher Federation honored him with their Champion of Public Education award for his efforts in helping to revise the Australian national primary and secondary writing assessments and, in particular, his role in persuading the Education Council of Australia to postpone indefinitely the use of automated essay evaluation (AES) in scoring national writing tests.

References

Anson, C., & Perelman, L. (2017). Machines can evaluate writing well. In C. E. Ball & D. M. Loewe (Eds.), *Bad ideas about writing* (pp. 286–278). Morgantown, WV: Digital Publishing Institute: West Virginia University Libraries. Retrieved from https://textbooks.lib.wvu.edu/badideas/badideasaboutwriting-book.pdf#page=289

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Attali, Y., & Powers, D. E. (2008). *A developmental writing scale* (Report No. ETS RR-08-19). Princeton, NJ: ETS.

BABEL Generated GRE Essays {excerpts}. (2016). Retrieved March 4, 2020, from http://lesperelman.com/wp-content/uploads/2019/12/R.pdf

Bejar, I. I., Flor, M., Futagi, Y., & Ramineni, C. (2014). On the vulnerability of automated scoring to construct-irrelevant response strategies (CIRS): An illustration. *Assessing Writing*, *22*, 48–59. https://doi.org/10.1016/j.asw.2014.06.001

Bloom, M. (2012). Computers grade essays fast ... but not always well. *National Public Radio*. Retrieved from https://www.npr.org/2012/06/07/154452475/computers-grade-essays-fast-but-not-always-well

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27–40.

Bruffee, K. A. (1983). Some curricular implications of the CUNY Writing Assessment Test. *Notes from the National Testing Network in Writing*, *2*, 4–5.

Cahill, A., Chodorow, M., & Flor, M. (2018). Developing an e-rater Advisory to detect Babel-generated essays. *Journal of Writing Analytics, 2*. Retrieved from https://wac.colostate.edu/docs/jwa/vol2/cahill.pdf

Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing*, *27*(3), 419–436. https://doi.org/10.1177/0265532210364391

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*(1), 100–108. https://doi.org/10.1016/j.asw.2012.11.001

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, *22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006

Educational Testing Service. (n.d.-a). How the GRE General Test is scored (for institutions). Retrieved March 2, 2020, from https://www.ets.org/gre/institutions/about/general/scoring/

Educational Testing Service. (n.d.-b). How the GRE General Test is scored (for test takers). Retrieved March 1, 2020, from https://www.ets.org/gre/revised_general/scores/how/

Elliot, N., & Klobucar, A. (2013). Automated essay evaluation and the teaching of writing. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 16–35). London: Routledge.

Fanetti, S., Bushrow, K. M., & DeWeese, D. L. (2010). Closing the gap between high school writing instruction and college writing expectations. *The English Journal*, *99*(4), 77–83. Retrieved from http://www.jstor.org/stable/27807171

Freitag Ericsson, P., & Haswell, R. H. (Eds.). (2006). *Machine scoring of student essays: Truth or consequences*. Logan UT.

Gamon, M., Chodorow, M., Leacock, C., & Tetreault, J. (2013). Grammatical error detection in automatic essay scoring and feedback. *Handbook of automated essay evaluation: Current applications and new directions* (pp. 251–266). London: Routledge.

Gowers, E., & Fraser, B. (1973). *The complete plain words* (Rev. ed.). London: HM Stationary Office.

Han, N. R., Chodorow, M., & Leacock, C. (2006). Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, *12*(2), 115–129. https://doi.org/10.1017/S1351324906004190

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English*, *63*(4), 480-499. Retrieved from http://www.jstor.org/stable/378891

Herrington, A., & Moran, C. (2012). Writing to a machine is not writing at all. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 219–232). New York: Hampton Press.

Higgins, B., Miller, M., & Wegmann, S. (2006). Teaching to the test…not! Balancing best practice and testing requirements in writing. *The Reading Teacher*, *60*(4), 310–319. https://doi.org/10.1598/RT.60.4.1

Higgins, D., Burstein, J. C., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, *12*(2), 145–159. https://doi.org/10.1017/S1351324906004189

Higgins, D., & Heilman, M. (2014). Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice*, *33*(3), 36–46. https://doi.org/10.1111/emip.12036

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, *43*(8), 381–389. https://doi.org/10.3102/0013189X14554449

Klobucar, A., Deess, P., Rudniy, O., & Joshi, K. (2013). Automated scoring in context: Rapid assessment for placed students. *Assessing Writing*, *18*(1), 62–84. https://doi.org/10.1016/J.ASW.2012.10.001

Kolowich, S. (2014, April 28). Writing instructor, skeptical of automated grading, pits machine vs. machine. *The Chronicle of Higher Education*.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). *Automated grammatical error detection for language learners* (2nd ed.). San Rafael CA: Morgan and Claypool. https://doi.org/10.2200/S00562ED1V01Y201401HLT025

McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing*, *15*(2), 118–129. https://doi.org/10.1016/j.asw.2010.04.002

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychological Review*, *8*(3), 299–325.

McCutchen, D. (2000). Knowledge, processing, and working memory: Implications for a theory of writing. *Educational Psychologist*, *35*(1), 13–23. https://doi.org/10.1207/S15326985EP3501_3

McCutchen, D., Covill, A., Hoyne, S. H., & Mildes, K. (1994). Individual differences in writing: Implications of translating fluency. *Journal of Educational Psychology*, *86*(2), 256–266. https://doi.org/10.1037/0022-0663.86.2.256

Mitchell, M. (2018, November 5). Artificial intelligence hits the barrier of meaning. *New York Times*.

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. New York: Farrar, Straus and Giroux.

Orwell, G. (1954). *A collection of essays*. New York: Harcourt Brace Jovanovich.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238-243. https://doi.org/10.2307/20371545

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In A. Bazerman, C. Dean, C. Early, J. Lunsford, K. Null, S. Rogers, & P. Stansell (Eds.), *International advances in writing research* (pp. 121–131). Fort Collins, Colorado: The WAC Clearinghouse and Parlor Press. Retrieved from https://wac.colostate.edu/books/wrab2011/chapter6.pdf

Perelman, L. (2013). Critique of Mark D. Shermis & Ben Hammer, "Contrasting state-of-the-art automated scoring of essays: Analysis." *The Journal of Writing Assessment*, *6*(1). Retrieved from http://journalofwritingassessment.org/article.php?article=69

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing*, *21*. https://doi.org/10.1016/j.asw.2014.05.001

Perelman, L. (2016a). The BABEL Generator. Retrieved from http://lesperelman.com/writing-assessment-robo-grading/babel-generator/

Perelman, L. (2016b). Grammar checkers do not work. *WLN: A Journal of Writing Center Scholarship*, *40*(7–8), 11–20. Retrieved from http://lesperelman.com/wp-content/uploads/2016/05/Perelman-Grammar-Checkers-Do-Not-Work.pdf

Pinker, S. (2014). *The sense of style: The thinking person's guide to writing in the 21st century!* New York: Viking.

Posner, D. (2004). What's wrong with teaching to the test? *Phi Delta Kappan*, *85*(10), 749–751. https://doi.org/10.1177/003172170408501009

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct coverage of the e-rater scoring engine* (Report No. ETS RR-09-01). Princeton, NJ: ETS. Retrieved from https://files.eric.ed.gov/fulltext/ED505571.pdf

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater ® scoring engine for the GRE ® issue and argument prompts* (Report No. ETS RR-12-02). Retrieved from https://www.ets.org/Media/Research/pdf/RR-12-02.pdf

Rota, G. C. (1985). The barrier of meaning. *Letters in Mathematical Physics*, *10*(2–3), 97–99. https://doi.org/10.1007/BF00398144

Seabrook. (2019, October). The next word: Could a computer write this article? *The New Yorker*, 52–63.

Sobel, L., Beckman, M., Jiang, D., & Perelman, L. (2014). BABEL Generator. Retrieved from https://babel-generator.herokuapp.com/

Strunk, W., & White, E. B. (1979). *The elements of style* (4th ed.). New York: Pearson. https://doi.org/10.2307/355984

Swanson, H. L., & Berninger, V. E. (1996). Individual differences in children's working memory and writing skill. *Journal of Experimental Child Psychology*, *63*(2), 358–385.

Tetreault, J., & Chodorow, M. (2008). The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics* (pp. 865–872). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from https://dl.acm.org/doi/10.5555/1599081.1599190

White, E. (2008). My five-paragraph-theme theme. *College Composition and Communication*, *59*(3), 524–525. Retrieved from http://www.jstor.org/stable/20457018

Wills, G. (1992). *Lincoln at Gettysburg: The words that remade America*. New York: Simon and Shuster.

Winerip, M. (2012, April 23). Facing a robo-grader? No worries. Just keep obfuscating mellifluously. *New York Times*. Retrieved from http://www.nytimes.com/2012/04/23/education/robo-readers-used-to-grade-test-essays.html

Zhang, M., Chen, J., & Ruan, C. (2016). Evaluating the Advisory flags and machine scoring difficulty in the e-rater® automated scoring engine. *ETS Research Report Series*, *2016*(2), 1-14. https:// doi.org.10.1002/ets2.12116