Research Article

Developing an e-rater Advisory to Detect Babel-generated Essays

Aoife Cahill, *Educational Testing Service*

Martin Chodorow, *Hunter College and the Graduate Center, CUNY*

Michael Flor, *Educational Testing Service*

Structured Abstract

- **Background:** It is important for developers of automated scoring systems to ensure that their systems are as fair and valid as possible. This commitment means evaluating the performance of these systems in light of construct-irrelevant response strategies. The enhancement of systems to detect and deal with these kinds of strategies is often an iterative process, whereby as new strategies come to light they need to be evaluated and effective mechanisms built into the automated scoring systems to handle them. In this paper, we focus on the Babel system, which automatically generates semantically incohesive essays. We expect that these essays may unfairly receive high scores from automated scoring engines despite essentially being nonsense.
- **Literature Review:** We discuss literature related to gaming of automated scoring systems. One reason that Babel essays are so easy to identify as nonsense by human readers is that they lack any semantic cohesion. Therefore, we also discuss some literature related to cohesion and detecting semantic cohesion.
- **Research Questions:** This study addressed three research questions:
  1. Can we automatically detect essays generated by the Babel system?
  2. Can we integrate the detection of Babel-generated essays into an operational automated essay scoring system while making sure not to flag valid student responses?

DOI: 10.37514/JWA-J.2018.2.1.08

3. Does a general approach for detecting semantically incohesive essays also detect Babel-generated essays?

- **Research Methodology:** This article describes the creation of two corpora necessary to address the research questions: (1) a corpus of Babel-generated essays and (2) a corresponding corpus of good-faith essays. We built a classifier to distinguish Babel-generated essays from good-faith essays and investigated whether the classifier can be integrated into an automated scoring engine without adverse effects. We also developed a measure of lexical-semantic cohesion and examined its distribution in Babel and in good-faith essays.
- **Results:** We found that the classifier built on Babel-generated essays and good-faith essays and using features from the automated scoring engine can distinguish the Babel-generated essays from the good-faith ones with 100% accuracy. We also found that if we integrated this classifier into the automated scoring engine it flagged very few responses that were submitted as part of operational submissions (76 of 434,656). The responses that were flagged had previously been assigned a score of Null (non-scorable) or a score of 1 by human experts. The measure of lexical-semantic cohesion shows promise in being able to distinguish Babel-generated essays from good-faith essays.
- **Conclusions:** Our results show that it is possible to detect the kind of gaming strategy illustrated by the Babel system and add it to an automated scoring engine without adverse effects on essays seen during real high-stakes tests. We also show that a measure of lexical-semantic cohesion can separate Babel-generated essays from good-faith essays to a certain degree, depending on task. This points to future work that would generalize the capability to detect semantic incoherence in essays.
- **Directions for Further Research:** Babel-generated essays can be identified and flagged by an automated scoring system without any adverse effects on a large set of good-faith essays. However, this is just one type of gaming strategy. It is important for developers of automated scoring systems to continue to be diligent about expanding the construct coverage of their systems in order to prevent weaknesses that can be exploited by tools such as Babel. It is also important to focus on the underlying linguistic reasons that lead to nonsense sentences. Successful identification of such nonsense would lead to improved automated scoring and feedback.

Keywords: automated essay scoring, constructed response items, gaming strategies, lexical-semantic cohesion, natural language processing, writing analytics

1.0 Background

The demand for automated scoring of student responses has increased in recent years as technology continues to advance and yield more sophisticated automated scoring capabilities. This demand is reflected in an increase in computer-based administration of large-scale assessments—for example, at the state level—where automated scoring is often seen as a time and cost-saving solution. Research has shown that use of automated scoring can lead to more objective overall scoring (Williamson, Bejar, & Hone, 2005) and can maintain test reliability (Bridgeman, Trapani, & Attali, 2012), but there are often concerns at the individual level about what an automated scoring system is measuring and whether it is appropriate to substitute a computer score for a human score (Bennett, 2015).

A recent case in Australia reveals criticism of the proposed introduction of automated scoring for the National Assessment Program—Literacy and Numeracy (NAPLAN), an annual assessment for all students in Years 3, 5, 7, and 9 in Australia (Robinson, 2017). Automated scoring had been proposed for the persuasive and narrative writing items and was planned to be fully implemented by 2020 (Robinson, 2018). Ultimately, automated scoring for those tests was postponed indefinitely. Such controversy calls attention to the shifting nature of educational assessment (Bennett, 2015) and the need to provide interpretation and use arguments for both formative and summative assessments (Kane, 2013).

One frequently-raised concern about automated scoring systems is the fact that they do not "read" student responses in the same way that humans do, instead using features that are approximations of factors that humans consider when applying the scoring rubrics. This difference between automated systems and human scoring can sometimes lead to the automated scoring systems being susceptible to techniques that try to fool the system into giving a higher score than is warranted. For example, simply writing a long essay is one perceived method of getting a higher score. This influence on score is because there is a natural link between the ability to write well in a timed setting and the length of the response—something automated scoring models often pick up on (Klobucar et al., 2012). Of course, just writing a long response should not automatically lead to a high score from the automated scoring system, particularly if that long response is incoherent and full of errors. As automated scoring systems develop, developers try to build in filters or flags for responses that look aberrant (Zhang, Chen, & Ruan, 2016) in order to maximize the validity of the systems. However, developers almost certainly will not think of everything and, to some extent, rely on the writing community to identify potential weaknesses in the system, which can then be addressed to further improve the validity of the system and thus the scores produced for test takers.

The Babel essay generation system (BABEL Generator, 2014) is a tool for automatically generating essays that are intended to fool automated scoring engines. The input to the tool is a list of keywords, which the tool uses to randomly generate an essay designed to fool automated scoring systems. The essay appears to be well-formed in terms of syntactic structures, and it uses complex English words. However, the essays are completely incoherent. The exact technology behind the tool has not been described in detail, but it is based on the Dada engine, which generates random texts from grammars (The Dada Engine, 2000). Figure 1 shows a snippet from an essay generated by submitting the keywords “snow” and “holidays”. It illustrates that the Babel system generates complete nonsense, using long, rare words that are somewhat related to the keywords provided, albeit in syntactically well-formed sequences. A system that does not consider the underlying meaning of a response, but instead only focuses on surface characteristics such as vocabulary or grammatical errors, may not recognize that this response is designed to fool it into giving a high score.

> Vacation has not, and in all likelihood never will be trite in the way we accuse exiles. Vacation is the most fundamental device of society; some of stealth and others on adherents. The preternatural snow lies in the realm of literature along with the field of philosophy. Why is holiday so boisterous to gluttony? The reply to this query is that snowfall is pugnaciously Libertarian.

Figure 1. A snippet from an essay generated by submitting the keywords “snow” and “holidays” to the Babel system.

While the most common gaming strategies to date often rely on essay length and word frequency to fool automated scoring systems, the most notable characteristic of Babel essays is not length, though they tend to be long, nor word frequency, though they are filled with rare words, but rather it is their lack of coherence—those essays are strikingly nonsensical.

In this paper, we examine the outcomes of sending Babel essays to an automated scoring engine (Educational Testing Service’s e-rater®, Burstein, Tetreault, & Madnani (2013)). We hypothesize that e-rater will assign high scores to these essays because, while semantically flawed, they appear to be syntactically accurate, well-formed from a discourse perspective (in that they include reasonable transition terms), and include an excessive number of longer and low-frequency words. We investigate whether we can automatically distinguish Babel essays from essays written in good faith, and if so, whether we can integrate the new capability back into the e-rater engine to prevent over-scoring of nonsense essays of this kind. Since Babel essays represent an extreme form of gaming—no student is likely to generate such incoherent text in a testing situation—we also begin to consider an alternative, more generalizable, approach for detecting Babel essays. This approach is based on developing a measure of lexical-semantic cohesion. In this paper we examine the distribution of this measure in Babel and in good-faith essays.

2.0 Literature Review

Since the conception of automated essay-scoring engines, research has been conducted into their validity, including how they can be fooled or "gamed." There is also a growing field of research, known as adversarial machine learning, for the more general problem of how machine learning applications can be proactive against malicious adversaries. Analysis in the field of adversarial learning typically includes (1) identifying potential vulnerabilities in the algorithm, (2) devising appropriate attacks corresponding to those vulnerabilities, and (3) proposing countermeasures to improve the security of the machine learning algorithm (Huang, Joseph, Blaine, Rubinstein, & Tygar, 2011). Here we focus on literature specific to the gaming aspect of automated scoring engines, rather than the much wider field of automated scoring validity or adversarial learning as a whole.

Powers, Burstein, Chodorow, Fowles, and Kukich (2001) conducted a study examining the effect of gaming strategies on the e-rater system. The gaming strategies were not pre-defined, but rather identified as part of the study. A number of experts with backgrounds in writing, natural language processing (NLP), and other related fields were asked to write essays that they thought might get higher or lower scores than they deserved from the e-rater system. The essays were graded by human experts and those scores were then compared to the scores assigned by e-rater. The study showed that it was easier to "trick" the system into giving an essay a higher score than it deserved, rather than a lower score than it deserved. Simple strategies, such as repeating the same paragraphs over and over, were able to fool e-rater at the time. Other strategies, including varying sentence structure, as well as including discourse cues and sophisticated vocabulary, also fooled e-rater.

Bejar, Flor, Futagi, and Ramineni (2014) conducted a study based on the hypothesis that if some words in a response are replaced with similar, but more sophisticated, words, then the automated essay scoring system should be tricked into assigning a higher score. Experiments were conducted on essays written as part of the GRE test, where up to 5% of words were randomly selected to be automatically replaced with longer, rarer synonyms. The assumption is that this is a potentially viable gaming strategy where a candidate simply memorizes a personal list of possible substitutions and applies it at test time. They found that only a fraction of e-rater scores would increase when using this strategy, and between 80 and 90% of essays had no change in e-rater scores.

Higgins and Heilman (2014) outlined a framework for quantifying the susceptibility of an automated scoring engine to gaming strategies. They presented a case study on the automated scoring of short answer constructed responses and showed that the susceptibility can vary greatly by engine. They conducted experiments with three open-source systems that performed best in the ASAP2 shared task on automated scoring of short answers. The strategies that they explored were: (1) length alone may influence the score of the engine, (2) re-using words from the prompt may lead to higher scores, (3) using general academic words may artificially inflate the score assigned by the engine. They artificially modified responses in the original dataset according to each of three types of gaming strategies and examined the effect on scores from each of the three scoring engines. In their simulations, they showed that some simple strategies can have significant impacts on the scoring engine performance. They also noted that estimation of the susceptibility cannot be made based on engine design and configuration alone.

Lochbaum, Rosenstein, Foltz, and Derr, (2013) analyzed different gaming strategies for essay scoring and described some methodologies included in the Intelligent Essay Assessor™ (IEA) to detect such strategies. Generally, an outlier-based approach is used to detect bad-faith, using both rule-based and statistical methods which can be tailored to a given deployment of the automated scoring engine (e.g., more stringent in high-stakes settings and perhaps more relaxed in formative settings, depending on client needs).

Yoon et al. (2018) presented an operational processing pipeline for non-scorable responses, including those generated by various gaming strategies. The pipeline is a general framework that can be applied to multiple automated scoring systems, indicating various points during processing at which non-scorable responses might be detected and how. They described three different points during the automated scoring pipeline at which non-scorable responses (including gamed responses) could be detected. They are: (1) at input capture, (2) during feature generation, and (3) during score generation. Two systems are used to exemplify the framework: automated scoring of spoken responses and automated scoring of essays. In that framework, the detection of Babel essays as described in this paper would occur after feature generation.

It seems that the central reason for a lack of coherence in Babel essays is the lack of lexical-semantic cohesion. In general, coherence refers to the overall understandability of a text, the quality of a text “making sense.” There are many different components that contribute to the overall coherence of a text (Carrell, 1982). Some of them are text organization, logical arrangement of parts, and signaling of discourse relations (Mandler & Johnson, 1977; Van Dijk, 1980). The term *coherence* is sometimes used to exclusively describe the perceived relatedness between segments of text (Bamberg, 1983). The coherence of a text can also be greatly affected by cohesion—the mutual connectedness of text elements. Halliday and Hasan (1976) described the five main types of devices that implement cohesion: reference, substitution, ellipsis, conjunction, and lexical cohesion. According to Halliday, “lexical cohesion comes about through the selection of items that are related in some way to those that have gone before” (Halliday & Matthiessen, 2004, p. 570). Such cohesion is signaled by vocabulary selection, rather than by any structural devices. A cohesive text has words that are semantically interrelated (see also Hoey, 1991; 2005). A text that uses many unrelated words will be less comprehensible, and, in extreme cases, incomprehensible. Indeed, essays generated by the Babel system give the impression of being incohesive as many of the concepts “mentioned” in them do not belong together—they don’t fit in a coherent way. The lack of cohesion is not just a reflection of their reliance on infrequent words. A sentence such as “Why is holiday so boisterous to gluttony?” from the Babel snippet shown above, is no more coherent if paraphrased as “Why is holiday so rowdy to greed?” We measure and compare the average lexical-semantic cohesion of Babel-generated essays and the good-faith essays written by examinees. The measure is “lexical-semantic,” in the sense that we measure the connectedness of lexical units (words) in view of their semantic relations.

There are several ways to automatically measure the semantic cohesion of a text. Early approaches counted word-repetitions and thesaural relations (e.g., synonymy, antonymy, hyponymy, meronymy, etc.) between words in a text (Morris & Hirst, 1991; Silber & McCoy, 2002). Later approaches used statistical co-occurrence data from large corpora to estimate word-relatedness (Flor & Beigman Klebanov, 2014; Marathe & Hirst, 2010). Recent research in NLP has demonstrated the effectiveness of vector-based word representations—where each word is represented by a vector of real-number values, trained on very large corpora of text. Vector-based word representations (embeddings) have shown considerable success in many text-processing applications, especially for the estimation of semantic relatedness (Levy & Goldberg, 2014).

3.0 Research Questions

We address the following research questions in this paper:

1. Can we automatically detect essays generated by the Babel system?
2. Can we integrate the detection of Babel-generated essays into an operational automated essay scoring system while making sure not to flag valid student responses?
3. Does a general approach for detecting semantically incohesive essays also detect Babel-generated essays?

4.0 Research Methodology

4.1 Automated Scoring Engine

We used the e-rater® engine (Attali & Burstein, 2006) as our automated scoring system in order to try to answer our research questions. E-rater is a system developed by Educational Testing Service to automatically score essays.[1] It uses NLP techniques to automatically extract linguistically-motivated features of writing that are then used in a linear regression model to predict a final score for an essay. E-rater also contains what are known as “advisories” which flag responses with certain characteristics such as being off-topic, too repetitious, etc. These advisories are generally meant as indicators that the e-rater score is unreliable. In high-stakes

[1] E-rater can also provide feedback to students about different aspects of their writing quality, but we omit a discussion of this component of e-rater here for brevity.

assessments, responses that are flagged with advisories are typically routed to a human for additional review.

Two versions of e-rater were used in this study. The first, a version from 2014, was used to determine the e-rater score distribution for the Babel essays and to guide the sampling of a good-faith dataset. A newer version of e-rater from 2016 was available after the initial datasets were collected, and so, was used for the remainder of the analyses and experiments. This newer engine contains several updates (bug fixes, library updates, etc.) over the previous engine, as well as some new features: discourse coherence (Somasundaran, Burstein, & Chodorow, 2014), which relies on measures of transition cues and lexical chains; grammaticality (Heilman et al., 2014), based on a language model and n-gram counts; and source use (Beigman Klebanov, Madnani, Burstein, & Somasundaran, 2014). The 2016 version of the engine is expected to perform better in terms of scoring accuracy than the 2014 engine.

The scoring features used in e-rater are given in Table 1, along with a brief description of the writing construct that they are designed to measure. Some features are only available in the 2016 engine and are marked with an asterisk.

Table 1

E-Rater Features Used for Scoring

Feature	Feature explanation
NSQG	Grammar errors
NSQU	Usage errors
NSQM	Mechanics errors
LOGDTA	Development
LOGDTU	Organization
NWF_MEDIAN	Vocabulary sophistication
COLPREP	Correct usage of collocations and prepositions
SVF	Sentence variety
WORDLN_2	Lexical complexity
GRAMMATICALITY*	Holistic measure of grammaticality
DIS_COH*	Discourse coherence
SOURCE_USE2*	How well source material is integrated into the response

Note. Features marked with * are only available in the 2016 version.

4.2 Data

The focus of our study was on high-stakes assessments, and so we targeted data collection from three high stakes assessments (five tasks total):

- Two tasks from an assessment of analytical writing as part of a graduate school test (datasets A and B). On A, test takers write a short argumentative essay by taking a position on an assigned topic. On B, they read a short argument text and then write an essay evaluating the soundness of the prompt argument.
- One task from a test of writing used to evaluate candidates for entry into teacher preparation programs (dataset C). Examinees are asked to draw on personal experience, observation, or reading to support a position on a topic using specific reasons and examples.
- Two tasks from an assessment of English proficiency (datasets D and E). On D, examinees write a short opinion essay on a pre-assigned topic, and on E they write a summary essay that compares arguments from two different sources (both supplied during the test).

All of the assessments were administered under time constraints on computers at test centers around the world and via the Internet. The test takers did not have access to editing tools such as spellcheckers, grammar checkers, or dictionaries. From a pool of 434 operational prompts, we randomly selected 100 (20 per task) on which to base our data collection.

In order to answer our research questions, we needed three sources of data for each dataset:

1. Essays generated by the Babel system. For each prompt, we manually determined the top key terms and used those as the keywords for the Babel generator. For each of the 100 prompts selected above, we generated at least 1,000 essays using the Babel tool. We automated this process and performed the queries on a server provisioned through Amazon Web Services.
2. Essays written by students in good faith. We randomly selected good-faith student responses (i.e., responses that received a human score of >0) to the same set of prompts as in data source (1) for all five datasets. These responses, which were written in a high-stakes assessment context, were chosen such that the e-rater score and prompt distributions were similar for data sources (1) and (2) for all five datasets. This was done to ensure that the essays in both datasets were comparable in terms of how e-rater views them.
3. A large collection of essays written by students in high-stakes settings. We randomly selected a large number of essays written by students in high-stakes settings.

The sample consisted of student essays from each of the five tasks corresponding to our initial five datasets. The total numbers of essays processed per task were: A (70,874), B (71,023), C (10,749), D (142,888), and E (139,122). Essays from a total of 581 prompts were included in this sample. Of the 100 prompts included in dataset (1), 87 were included in this sample.

Each essay in all five datasets, for each of the three sources, was processed with e-rater according to the appropriate scoring model (there is a separate scoring model for each task).

4.3 Comparing BABEL Essays to Good-Faith Essays

We compared the distribution of e-rater feature values between the Babel-generated essays and the good-faith essays. We used the Kolmogorov-Smirnov (K-S) statistic to measure how different two distributions of feature values were. This is a two-sided test for the null hypothesis that two independent samples are drawn from the same continuous distribution. If the value of the K-S statistic is small with a large *p*-value, then we cannot reject the hypothesis that the distributions of the two samples are the same. Conversely, a high K-S value with a small *p*-value indicates that we can reject the hypothesis with some certainty.

4.4 A Classifier for Detecting BABEL Essays

We built a classifier to detect whether an essay was a Babel-generated one or not. Each essay was represented by the standard features computed by e-rater (see Table 1). We trained the classifier on e-rater feature values extracted for both the Babel-generated essays and the corresponding good-faith essays (data sources (1) and (2)). We combined data from all tasks.[2] We randomly split the data into training and test partitions, using 80% (165,070 essays—77,648 good faith and 87,422 Babel) and 20% (41,260 essays—19,409 good faith and 21,851 Babel) of the data respectively. We built a random-forest classifier (Breiman, 2001) on the training data and evaluated its performance on the test data. The random-forest classifier is an ensemble classifier based on decision trees where the final prediction is the mode of the classes of the individual trees.

4.5 A New Advisory for e-rater

Using the classifier outlined above, we developed a new e-rater advisory that is triggered whenever the classifier detects a Babel essay, given a set of features for an essay. We applied the classifier to the large collection of essays written by students in high-stakes settings. We used these essays to measure the effect of a classifier to detect Babel essays on real student-generated data. The goal was to ensure that there are no adverse effects of a Babel essay detector, i.e., that essays written by students are not overly flagged by this classifier.

4.6 Comparing the Lexical-Semantic Cohesion of Babel Essays and Good-Faith Essays

In this study, we estimated lexical-semantic cohesion in the following way: Semantic relatedness between two words is computed as the cosine similarity value between their numeric vector representations. We computed such values for all pairs of ‘content’ words in an essay, and then take the average. This value represents the average semantic cohesion for the text. We identified a “content” word as any word that is not a digits-only string (e.g., *1984*) and is not on a list of 54 stop-words (which includes determiners, common prepositions, and some other very common

[2] We also conducted some preliminary experiments on task-specific classifiers. However, since the results using the combined data were equally high, there was no apparent advantage to this approach.

words).[3] We used *word2vec* vectors with 300 dimensions, pre-trained on 100 billion words of Google News (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).

5.0 Results

5.1 Automated Scoring of Babel Essays

In total, we collected 106,239 Babel essays for 100 prompts. A summary of the number of essays and prompts per task is given in Table 2. Figure 2 gives the distribution of scores assigned to the essays by the 2014 e-rater engine. Most of these essays received a score of 4 or 5, much higher than they would be assigned by a human rater. In practice, for high-stakes assessments such as the GRE or TOEFL, these kinds of essays where e-rater predicts a far higher (or lower) score than a human rater are routed to expert human raters for additional scoring, and the e-rater score is typically discarded. E-rater does not generate an advisory for any of these essays. Figure 3 gives the distribution of scores assigned to the essays by the 2016 e-rater engine. The newer engine (with the new features) does give somewhat lower scores to these essays, particularly for datasets A, B, and D, where many more essays now receive a score of 3. There were no changes in the scores for dataset E and only minor changes for dataset C. However, it is clear that even though the newer engine is assigning lower scores for many of the Babel essays, the scores are still much higher than deserved.

Table 2

The Number of Babel-Generated Essays Collected for Each Task

	A	B	C	D	E
Total essays	21,917	21,979	21,359	21,987	21,969
Total prompts	20	20	20	20	20

[3] Essentially, we filter out only the very common stop-words, and do not exclude less common function words (like *beneath* and *across*). The full list of stop-words is: *s, a, an, the, at, as, by, for, from, in, on, of, off, up, to, out, over, if, then, than, with, have, had, has, can, could, do, did, does, be, am, are, is, was, were, would, will, it, this, that, no, not, yes, but, all, and, or, any, so, every, we, us, you*, and *also*.

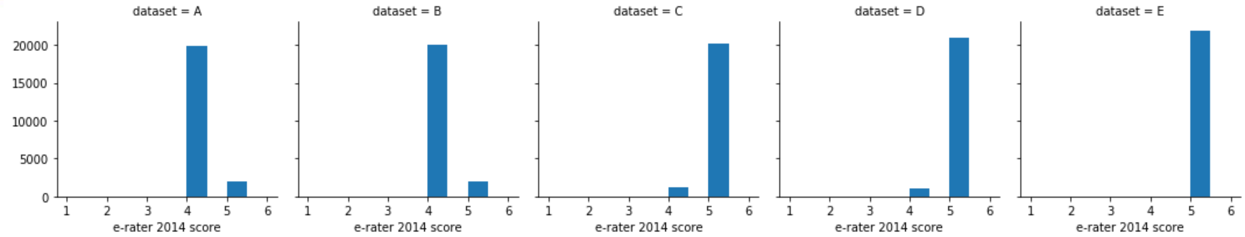


Figure 2. The distribution of scores assigned to the Babel essays by the 2014 version of e-rater.

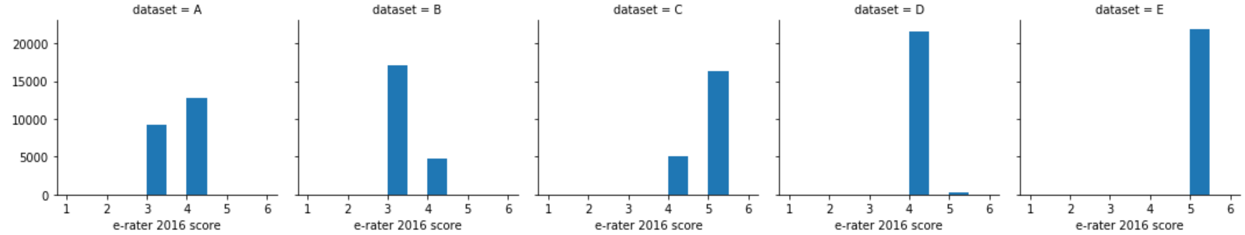


Figure 3. The distribution of scores assigned to the Babel essays by the 2016 version of e-rater.

5.2 Comparing BABEL Essays to Good-Faith Essays

Table 3 gives the K-S statistics comparing the e-rater features between Babel and good-faith essays for each task. The *p*-values (not shown in the table) are all statistically significant at *p* < .001. In particular, the K-S statistics for the GRAMMATICALITY (ranging from 0.97 to 1) and NWF_MEDIAN (1 for all 5 datasets) features are consistently high, at or near a value of 1.00, which indicates non-overlapping distributions. This reinforces the intuition that the Babel essays are unusual from both a vocabulary and a lexical co-occurrence view, since the NWF_MEDIAN feature measures vocabulary sophistication and the GRAMMATICALITY feature gives a holistic measure of grammaticality based on a language model and n-gram counts, which are sensitive to lexical frequency.

Table 3

Kolmogorov-Smirnov Statistics for the Difference Between Babel and Good-Faith Essays for e-Rater Features for Each Task

	A	B	C	D	E
COLPREP	0.98	0.98	0.99	0.97	0.93
DIS_COH	0.59	0.22	0.57	0.41	0.88
GRAMMATICALITY	0.99	1.00	1.00	1.00	0.97
LOGDTA	0.26	0.26	0.56	0.38	0.77
LOGDTU	0.42	0.44	0.39	0.43	0.77
NSQG	0.89	0.92	0.93	0.89	0.75

	A	B	C	D	E
NSQM	0.25	0.21	0.27	0.24	0.37
NSQU	0.71	0.71	0.75	0.66	0.56
NWF_MEDIAN	1.00	1.00	1.00	1.00	1.00
SOURCE_USE2	NA	NA	NA	NA	0.75
SVF	0.41	0.43	0.36	0.17	0.40
WORDLN_2	0.97	0.97	0.98	0.99	0.98

Some other features that have high K-S statistics are WORDLN_2 (ranging from 0.97 to 0.99) and COLPREP (ranging from 0.93 to 0.99). These two features are also related to lexical choice and vocabulary usage. Other features that are less susceptible to the gaming strategy employed by the Babel system generally have lower K-S statistics. For example, LOGDTA (development) and LOGDTU (organization) show values ranging from 0.26 to 0.77. Similarly, NSQM (mechanics) has low K-S statistics values ranging from 0.25 to 0.37, indicating that there is some overlap in distributions between Babel and good-faith essays. This is to be expected, since good-faith student writing—especially from proficient students—tends to not have many mechanics errors, and the Babel essays are designed not to have any at all.

One possible explanation for the fact that although the feature distributions for Babel and good-faith essays look quite different, the e-rater scores are high, is that e-rater does not pay attention to feature distributions, but rather only to the feature values themselves. The features are intended to represent approximations of characteristics that are indicative of good writing. The Babel essays exhibit enough characteristics of well-written essays (e.g., no spelling or grammatical errors, paragraphs with transition terms, range of syntactic constructions, range of sophisticated vocabulary, etc.) that when the features that represent these characteristics are combined in a linear regression model, the overall predicted score tends to be high.

5.3 BABEL Classifier Results

The results of the random-forest classifier are given in Table 4 and show that the classifier is able to distinguish with 100% accuracy the BABEL essays from the good faith essays in our test data.

Table 4

Classification Results of Babel Essay Detection (Total Essays)

	BABEL	Good Faith
BABEL	21851	0
Good Faith	0	19409

Note. Row=reference; Column=prediction

5.4 New e-rater Advisory Results

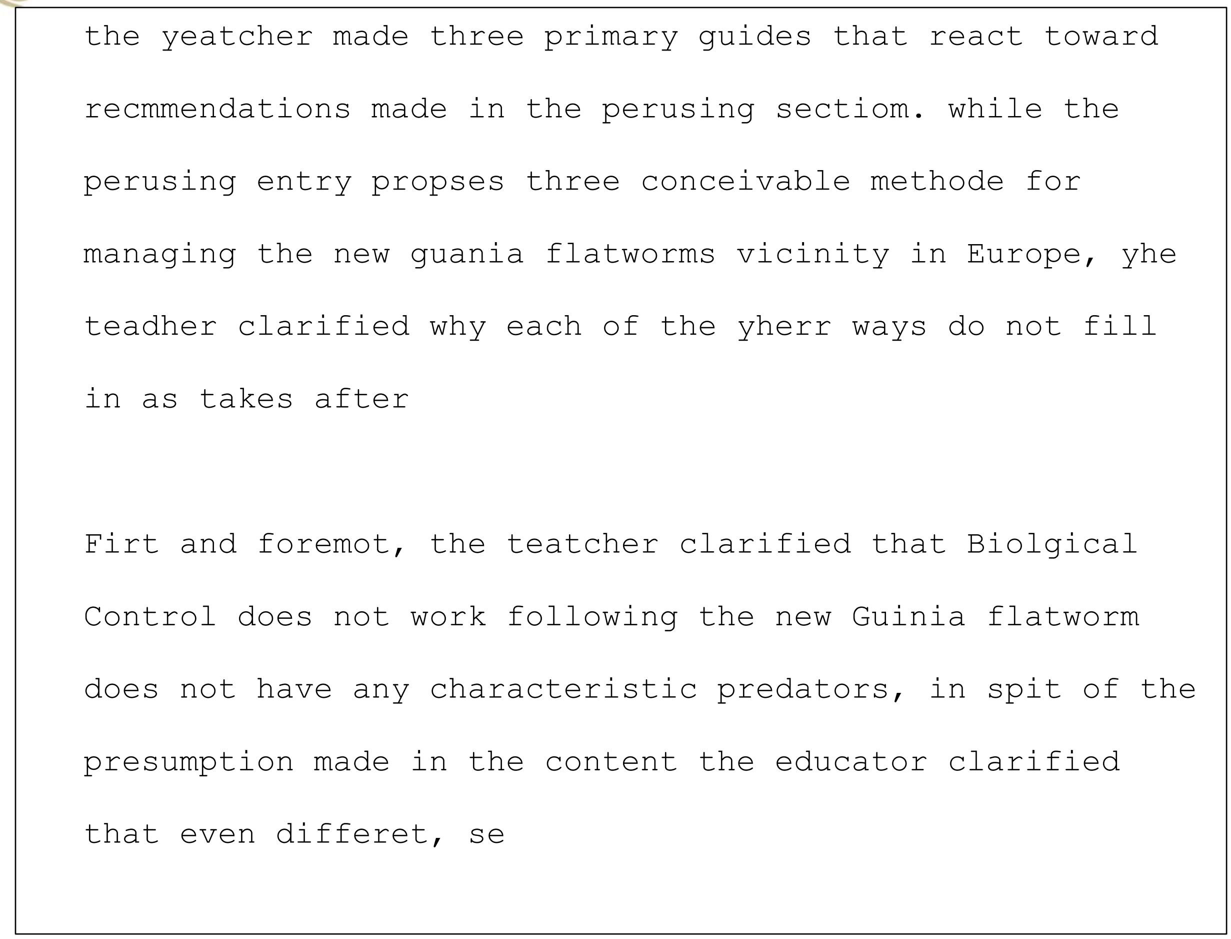
Table 5 shows how many times the new e-rater advisory was triggered for the five datasets in the large essay collection, as well as the distribution of original human scores for the flagged essays. A score of null or 0 means that the essay was marked as non-scorable by the human (e.g., because it was off-topic).

Table 5

Number of Essays Triggering Babel Advisory for Each Task by Score Point and Dataset

Human Score	A	B	C	D	E
Null or 0	1	1	0	8	48
1	0	0	0	4	27
2	0	0	0	0	1
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
Total	1	1	0	12	76

We see that the new advisory was triggered most for the essays in dataset E (from the test of English proficiency), and mostly for essays that received a human score of null, 0, or 1. A human score of 0 or null indicates that the response was non-scorable (e.g., off topic, plagiarized, etc.). There is one essay for which the Babel advisory was triggered that the human score was greater than 1. An example essay that was flagged by this advisory is given in Figure 4. This essay uses some relatively rare words as well as some long words (e.g., perusing, conceivable, characteristic, predators, presumption), albeit not entirely correctly, and has many misspelled words.



```
the yeatcher made three primary guides that react toward
recmmendations made in the perusing sectiom. while the
perusing entry propses three conceivable methode for
managing the new guania flatworms vicinity in Europe, yhe
teadher clarified why each of the yherr ways do not fill
in as takes after

Firt and foremot, the teatcher clarified that Biolgical
Control does not work following the new Guinia flatworm
does not have any characteristic predators, in spit of the
presumption made in the content the educator clarified
that even differet, se
```

Figure 4. An example essay flagged by the new advisory.

In general, this advisory is not triggered very often in high-stakes situations, which is not all that surprising since students are unlikely to naturally (or even from memory) generate the kinds of text that the Babel system does.

5.5 Lexical-semantic Cohesion Comparison Results

Table 6 gives the average lexical-semantic cohesion values for Babel essays and good-faith essays for each of the five datasets, as well as the *t*-test results for tests of significant differences between them. The average lexical cohesion of Babel-generated essays is much lower than in human-written essays, for all five datasets. The differences are statistically significant. Table 7 gives the K-S statistics for the five datasets for these average cohesion values. Figures 5 and 6 show histograms of the Semantic Cohesion values for datasets D and E, respectively. The graphs show that there is very little overlap in the histograms for dataset D, corresponding to the high K-S statistic of 0.98, while there is considerably more overlap for dataset E, corresponding to the much lower K-S statistic of 0.6.

Table 6

Average Lexical-Semantic Cohesion Values for Babel Essays and Good-Faith Essays

dataset	Source						
	BABEL			Good-Faith			
	M	SD	n	M	SD	n	t
A	0.096	0.003	21979	0.126	0.014	21979	-298.88 *
B	0.092	0.003	21979	0.114	0.011	21917	-273.69 *
C	0.095	0.003	21359	0.135	0.016	13354	-362.06 *
D	0.093	0.003	21987	0.132	0.016	19886	-343.58 *
E	0.093	0.003	21969	0.103	0.011	19928	-132.79 *

Note. * *p* < .0001.

Table 6

Kolmogorov-Smirnov Statistics for the Difference in Average Lexical-Semantic Cohesion Between Babel and Good-Faith Essays for Each Task

	A	B	C	D	E
Average semantic cohesion	0.90	0.94	0.98	0.98	0.60

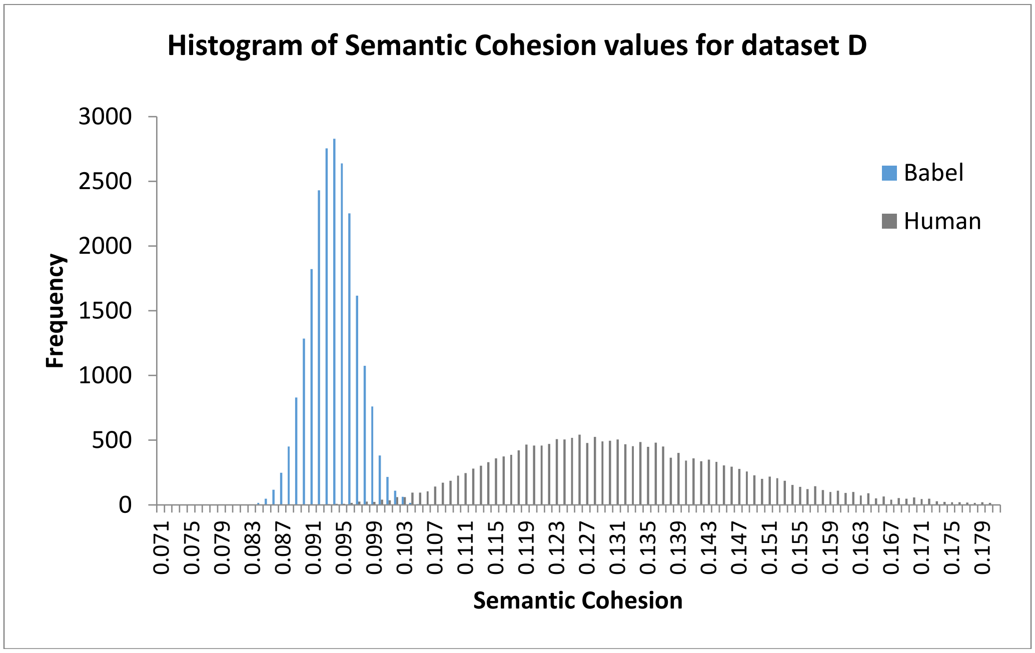


Figure 5. Lexical-semantic Cohesion histogram for dataset D.

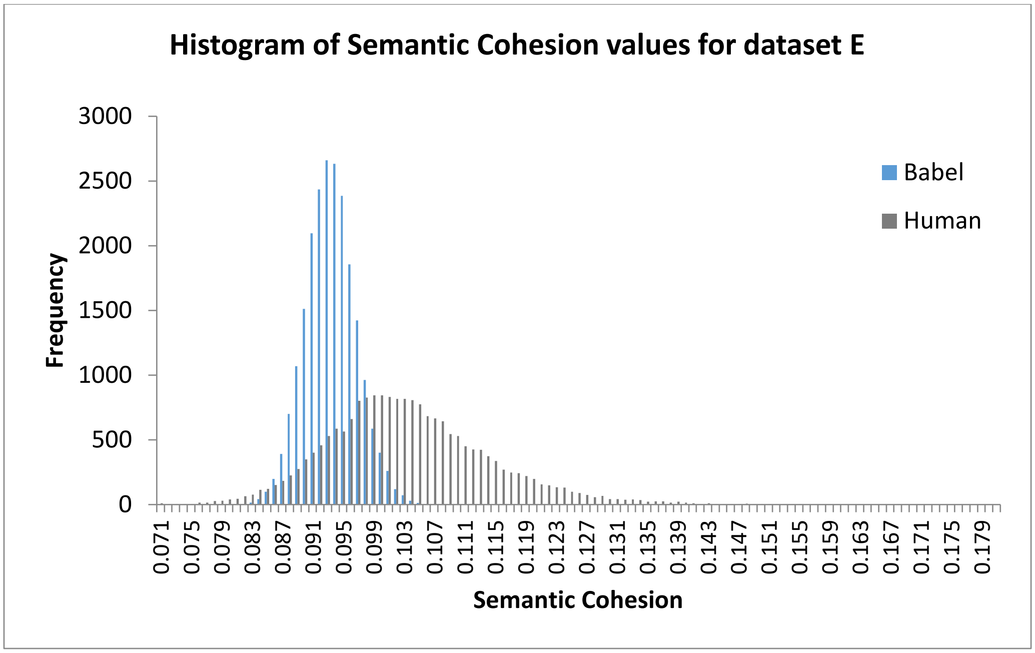


Figure 6. Lexical-semantic cohesion histogram for dataset E.

6.0 Discussion

We can now answer our research questions using the results of our experiments.

6.1 Can we automatically detect essays generated by the Babel system?

The classifier that we built was able to perfectly identify Babel essays compared to essays with similar scores given by e-rater in our test set. The classifier takes advantage of the fact that the Babel essays are trying to “game” e-rater in a particular way, i.e., by using rare long words in syntactically reasonable constructions. Our analysis shows that the distributions of certain feature values in Babel and good-faith essays are essentially disjoint, particularly for features related to vocabulary sophistication and grammaticality. Consequently, the classifier built using these features was able to learn to separate the two classes (Babel and Good-Faith).

6.2 Can we integrate the detection of BABEL-generated essays into an operational automated essay scoring system while making sure not to flag valid student responses?

We have developed a new advisory for e-rater that is designed to flag automatically generated, nonsensical, BABEL essays. On new data, almost all of the essays that were flagged by this advisory had received a human score of null, 0 or 1, indicating generally very poor quality of writing. In a high-stakes operational setting, this advisory would result in the essay being routed to a second human, rather than letting e-rater assign a score. Given that e-rater has a tendency to over-score these types of essays, this is a prudent approach.

6.3 Does a general approach for detecting semantically incohesive essays also detect Babel-generated essays?

While the *t*-test results show that there are statistically significant differences in the means in values of the lexical-semantic cohesion measure between the Babel-generated essays and the good-faith essays, the K-S statistic shows that for dataset E there is still considerable overlap in the distributions. For datasets A–D, our results show that the most semantically cohesive Babel-generated essays are only as cohesive as the least-cohesive essays written by examinees in good faith.

One way to think about applying this measure to distinguishing Babel-generated essays from good-faith essays is to apply some threshold for cohesion values. For datasets A–D, this would allow us to easily distinguish incohesive essays and most of the Babel-generated essays (although the threshold value might be different for different testing programs). This line of research looks promising, though additional research is required to understand the implication for the sets of good-faith essays whose cohesion scores overlap with Babel essay cohesion scores. Another possibility is to include the lexical-semantic cohesion feature in the classifier used to detect Babel-generated essays.

We acknowledge that it would be virtually impossible for anybody outside of ETS to reproduce the exact results in this paper, given the proprietary nature of the data and feature implementation (all of the features mentioned here have been described in the literature). However, we believe that the general outcomes here should be reproducible.

7.0 Conclusion

We have shown that it is possible to automatically distinguish Babel-generated essays from good-faith high-scoring essays by building a classifier trained on both kinds of essays and using e-rater features. The Babel-generated essays receive scores that are too high from e-rater, but by integrating a classifier that detects Babel-generated essays into the system, we can automatically flag such essays to be sent for additional human review if required (or report a 0 score if it is not possible to have a human in the loop, e.g., in an online practice test setting). We have also shown that a more general, semantically-inspired, method of quantifying the cohesion of essays is promising in terms of being able to distinguish Babel-generated essays from good-faith ones.

The Babel essay generator is an extreme example of how nonsense essays can fool automated scoring engines, since a student is unlikely to be able to generate such text in a test setting. In some ways, that makes Babel essays somewhat easier to identify than essays that, say, contain only some nonsense intermixed with reasonable text.

The result from these experiments is being integrated into an improved version of e-rater which should no longer award such nonsense essays real scores. Continued research into the semantic cohesion feature will also potentially lead to a further improved version of e-rater in the future.

8.0 Directions for Further Research

The debate about automated scoring continues (Greene, 2018; Smith, 2018). Babel-generated essays are a common example of how automated essay scoring can be gamed, and the claim is that these systems therefore do not work well. We have shown that such essays can be identified and flagged by an automated scoring system without any adverse effects on a large set of good-faith essays. However, the research presented in this paper addresses just one kind of gaming strategy that the e-rater automated scoring engine was previously susceptible to. This is a common pattern: a gaming strategy is identified, and subsequently, developers of automated scoring systems try to ensure that they are not fooled. Of course, it will be important to continue to be diligent about newly-discovered methods of gaming automated scoring and continue to develop methods to detect and/or counteract them.

In parallel, and perhaps more importantly, it is incumbent upon us to continue to develop more sophisticated automated scoring features that capture aspects of the writing construct that current features do not, or only poorly, address (e.g., semantic cohesion). Attention to the writing construct can be ensured by developing a research program that identifies weaknesses in construct representation and then supports projects to investigate and develop capabilities to expand the construct representation of the automated scoring system.

Additionally, it is not clear that essays with only a fraction of the nonsense contained in the Babel-generated essays would fool an automated system as successfully (Bejar, Flor, Futagi, & Ramineni, 2014). The bigger research challenge will be to address the broader picture of better identification of the underlying linguistic reasons for any nonsense interpretation at the sentence level and subsequently use that to improve automated scoring and feedback. One obvious direction for this kind of research is to look at verb-argument semantics and identify semantic selection restriction violations. Being able to identify selection restriction violations could lead to a more general method of detecting nonsense at the sentence level, which could be used to detect Babel essays in addition to much more subtle kinds of nonsense gaming techniques that are potentially more plausible in a high-stakes assessment setting.

Author Biographies

Aoife Cahill is a Senior Managing Research Scientist in the Natural Language Processing Research group at Educational Testing Service in Princeton, New Jersey. Aoife's research interests include integrating statistical techniques with traditional rule-based approaches to NLP. Her main focus at ETS is on researching methods of automatically assessing students' written responses, both for writing quality as well as content.

Martin Chodorow is a professor of Psychology and Linguistics at Hunter College and the Graduate Center of the City University of New York. His current research interests include automated assessment of narrative quality and computational methods for detecting grammar and word usage errors in nonnative writing.

Michael Flor is a Research Scientist in the Natural Language Processing Research group at Educational Testing Service in Princeton, New Jersey. Michael's research involves automatic processing of text data, combining statistical and linguistic approaches, focusing on automated assessment of writing, estimation of textual complexity, and generation of questions from text.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3), 1–30.

BABEL Generator (2014). Retrieved from http://babel-generator.herokuapp.com

Bamberg, B. (1983). What makes a text coherent? *College composition and communication*, *34*(4), 417–429.

Beigman Klebanov, B., Madnani, N., Burstein, J., & Somasundaran, S. (2014). Content Importance models for scoring writing from sources. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 247–252). Association for Computational Linguistics.

Bejar, 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.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*(1), 370–407.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

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.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The E-rater® automated essay scoring system. *Handbook of automated essay evaluation: Current applications and new directions*, 55–67.

Carrell, P. L. (1982). Cohesion is not coherence. *TESOL Quarterly*, *16*(4), 479–488.

Flor, M., & Beigman Klebanov, B. (2014). Associative lexical cohesion as a factor in text complexity. *International Journal of Applied Linguistics*, *165*(2), 223–258.

Greene, P. (2018, July 2). *Automated essay scoring remains an empty dream*. Retrieved from Forbes: https://www.forbes.com/sites/petergreene/2018/07/02/automated-essay-scoring-remains-an-empty-dream/#36e884a74b91

Halliday, M. A., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Halliday, M. A., & Matthiessen, C. (2004). *An introduction to Functional Grammar (3rd edition).* London: Arnold.

Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M., & Tetreault, J. (2014). Predicting grammaticality on an ordinal scale. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 174–180). Association for Computational Linguistics.

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.

Hoey, M. (1991). *Patterns of lexis in text.* Oxford University Press.

Hoey, M. (2005). *Lexical priming: A new theory of words and language.* London: Routledge.

Huang, L., Joseph, A. D., Blaine, N., Rubinstein, B. I., & Tygar, J. (2011). Adversarial machine learning. *Proceedings of the 4th ACM workshop on security and artificial intelligence*, (pp. 43–58).

Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement, 50*, 1–72.

Klobucar, A., Deane, P., Elliot, N., Chaitanya, C., Deess, P., & Rudniy, A. (2012). Automated essay scoring and the search for valid writing assessment. *International advances in writing research: Cultures, places, measures*, 103–119.

Levy, O., & Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning* (pp. 171–180). Ann Arbor, Michigan: Association for Computational Linguistics.

Lochbaum, K. E., Rosenstein, M., Foltz, P. W., & Derr, M. A. (2013, April). *Detection of gaming in automated scoring of essays with the IEA*. Paper presented at the National Council on Measurement in Education Conference (NCME), San Francisco, CA.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive psychology, 9*(1), 111–151.

Marathe, M., & Hirst, G. (2010). Lexical chains using distributional measures of concept distance. *International Conference on Intelligent Text Processing and Computational Linguistics. 6008,* pp. 291-302. Springer Lecture Notes in Computer Science.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, (pp. 3111-3119).

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics, 17*(1), 21–48.

Powers, D., Burstein, J., Chodorow, M., Fowles, M., & Kukich, K. (2001). Stumping E-Rater: Challenging the validity of automated essay scoring. *ETS Research Report Series*, i--44.

Robinson, N. (2017, October 12). *Push to have robots mark school tests under fire from prominent US academic*. Retrieved from ABC News: http://www.abc.net.au/news/2017-10-12/us-academics-warn-against-automated-naplan-english-testing/9039408

Robinson, N. (2018, January 29). *Robot marking of NAPLAN tests scrapped*. Retrieved from ABC News: http://www.abc.net.au/news/2018-01-29/push-to-have-robots-mark-naplan-tests-scrapped/9370318

Silber, H. G., & McCoy, K. F. (2002). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics, 28*(4), 487–496.

Smith, T. (2018, June 30). *More states opting to 'robo-grade' student essays by computer*. Retrieved from NPR: https://www.npr.org/2018/06/30/624373367/more-states-opting-to-robo-grade-student-essays-by-computer?utm_source=facebook.com&utm_medium=social&utm_campaign=npr&utm_term=nprnews&utm_content=20180630

Somasundaran, S., Burstein, J., & Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 950–961). Dublin City University and Association for Computational Linguistics.

The Dada Engine. (2000). Retrieved from http://dev.null.org/dadaengine/

Van Dijk, T. A. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Williamson, D. M., Bejar, I. I., & Hone, A. S. (2005). 'Mental Model'™ Comparison of Automated and human scoring. *Journal of Educational Measurement, 36*(2), 158–184.

Yoon, S.-Y., Cahill, A., Loukina, A., Zechner, K., Riordan, B., & Madnani, N. (2018). Atypical Inputs in educational applications. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)* (pp. 60–67). New Orleans, LA: Association for Computational Linguistics.

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*, 1-14.