# Comparing the Robustness of Deep Learning and Classical Automated Scoring Approaches to Gaming Strategies

*Sue Lottridge, Ben Godek, Amir Jafari, and Milan Patel*

*Cambium Assessment, Inc.*

## Introduction

The state of the art in machine-learning scoring has evolved in recent years to achieve gains in accuracy in a number of predictive tasks. Older models used feature-based approaches whereby experts wrote algorithms to create features thought to be relevant to item scoring and predicted scores using these weights applied to these features. The state-of-the-art approach learns features alongside the predictive model using very large, multi-layered neural networks (often called *deep learning*).

Most automated essay scoring engines are built using the feature-based approach (Shermis, 2014). In these systems, feature extraction algorithms generate numeric vectors that are intended to represent indicators of good writing. Examples of feature values are the number and proportion of spelling errors, the number and proportion of grammar errors, sentence variety, the number of discourse elements, and essays represented in a dimensionality-reduced, term-document space. Feature functions are typically written by computational linguists or other domain experts. Aside from the more general criticisms that automated scoring engines do not understand writing, features in these systems typically suffer from two key weaknesses: 1) they use "bag of words" semantic analysis whereby word order over long sequences is ignored when building features; and 2) they over-rely on essay length in prediction (Perelman, 2014). Once feature vectors are generated, a subset of the feature values is used in a statistical model to estimate weights (during training) and to apply weights (during scoring) to predict the score a human would have provided.

In recent years, research-based engines have employed "feature-free" deep-learning methods to "learn" features during the calibration process rather than creating them prior to modeling. These engines use long short-term memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) or transformer networks (Vaswani et al., 2017) and have demonstrated improvements over expert feature-derived engines (Taghipur & Ng, 2016; Rodriguez, Jafari, & Ormerod, 2019) on the publicly available Kaggle essay dataset (Hewlett Foundation, 2012). These neural networks have demonstrated order-of-magnitude accuracy increases in other tasks such as language translation (Weiss et al., 2017), image recognition (Szegedy et al., 2016), speech recognition (Chiu et al., 2017), cancer detection (Esteva et al., 2017), and language "understanding" tasks (Vaswani et al, 2017; Devlin et al., 2018). Importantly, the LSTM networks and the transformer networks are designed to include sequence—in the case of text, word order—in the modelling process, and thereby are thought to better model language than bag-of-words methods. They do not explicitly model length because features are learned; however, engines do have limitations in the maximum sequence length they can model.

One of the frequent criticisms of automated essay scoring is that engines do not understand language and therefore can be "tricked" into giving higher scores than they should (Page, 2003; Woods, 2017). Engines have been found to be susceptible to these responses, but the impact of such responses varies by item (Zhang, Chen, & Ruan, 2016; Burkhardt & Lottridge, 2013) and feature set (Higgins & Heilman, 2014). Automated scoring of essays can be viewed negatively by the public in part because of how they identify and score unusual responses (Shermis & Lottridge, 2019). As a result, almost every operational automated essay scoring engine employs filters to identify aberrant responses, either flagging them as such or routing them for human review and scoring. One potential promise of deep learning models is that they are more robust to gaming behaviors because they consider word use in context and therefore may not require filters or may require fewer filters.

Given these developments, this study sought to examine the robustness of one deep learning method (using a transformer model—Bidirectional Encoder Representations from Transformers [BERT]; Devlin et al., 2018) to a traditional automated scoring method (using Cambium Assessment, Inc.'s [CAI's] automated scoring engine, Autoscore) on a set of gaming responses. The gaming responses considered were: shuffled text (referred to as Shuffled essays; to examine the assumption that word order is learned by neural network models); grammatically correct but nonsense essays (referred to as Babel essays; to examine the extent to which grammar is contributing to score); Off-Topic essays (to examine the extent to which essay meaning is contributing to score); and Duplicated essays (to examine the extent to which length is contributing to score, controlling for meaning).

## Background

This section is divided into three parts to discuss the component elements contributing to the study research questions and design. These components include: 1) a description of gaming responses and studies of them; 2) a description of feature-based engines; and 3) a description of feature-free engines.

### Gaming responses

In this study, we define "gaming" responses in a broad sense. Namely, we define any response as gaming if it is constructed in some way that it is not a valid attempt to answer an essay prompt and seeks to obtain a higher score from the engine than is warranted. The literature has described many examples of gaming behavior, including: duplication of text to increase length (Zhang et al., 2016; Higgins & Heilman, 2014; Lochbaum et al., 2013; Powers et al., 2001); extensive use of prompt text (Zhang et al., 2016; Lochbaum et al., 2013); the use of key topic-related words in an otherwise off-topic essay (Zhang et al., 2016; Higgins et al., 2014; Kolowich, 2014; Lochbaum et al., 2013); and, Off-Topic essays, including those written to other prompts, canned essays, or those that were original writing but not addressing the prompt (Zhang et al., 2016; Burkhardt & Lottridge, 2013; Higgins, Burstein, & Attali, 2006). Other types of unusual/aberrant responses might be classified less as gaming and more as a lack of willingness or ability to engage with the prompt. These include gibberish or nonsense essays in which the examinee randomly types on a keyboard for most or all of their response, essays written in another language even though the required language is English, and very short essays that reflect a personal view of the examinee about the item or test (e.g., "I don't know," "This is dumb"). This latter set of responses is typically captured using filters and would not be considered "gaming;" however, responses such as these that earn valid—and even high—scores can reduce the trust stakeholders have in automated scoring.

One of the earliest papers on gaming was by Powers et al. (2001), in which the authors used a novel approach to assess gaming of an early version of ETS' e-rater automated scoring engine by soliciting experts to write gaming essays of their own choosing. In the study, 63 essays were written by 27 participants to game the engine. These responses were then hand-scored. The hand-scorers agreed with one another at higher rates than the version of e-rater at the time (correlation of .82 between two readers, and correlations of .42 and .37 between e-rater and the readers). The essays with the highest discrepancy between the human and e-rater scores were highly duplicated, with or without minor revisions. Other essays receiving high scores from e-rater but low scores from human raters were long and on-topic but not substantive.

The inclusion of on-topic key words in otherwise Off-Topic essays has been researched by Higgins & Heilman (2014), Cahill et al. (2014), and Henderson & Andrade (2019). Techniques include adding several random words from the prompt or context materials (Higgins et al., 2014), adding random "academic" words (Higgins et al., 2014), or using automated response generating software such as the Babel generator[1]. Most studies investigated whether it was possible to build filters that detected these bad-faith essays, including the Babel-generated essays. Filters were built using the "gaming" essays and

[1] https://babel-generator.herokuapp.com/

"good-faith" essays; machine learning methods were used to create a model that predicted whether an incoming essay was a Babel essay or a "good-faith" essay. In investigating filters to detect Babel essays, Cahill et al presented example distributions of scores provided by two different ETS e-Rater systems (Burstein, Tetreault, & Madnani, 2013); these distributions showed that e-rater tended to give the Babel essays scores near the middle of the score distribution. The authors then inspected which features were responsible for assigning the higher scores by comparing their values to those produced for good-faith essays. The identified features measured grammaticality and vocabulary sophistication. These features were then used to predict whether a response was a Babel essay or a good-faith essay. The classification results showed perfect agreement with the expected score. Their results reinforce the idea that the types of features used in automated essay scoring engines can and do impact the extent to which those engines can be gamed. Henderson & Andrade (2019) also examined the success of building a Babel filter using Babel essays and actual responses to an international assessment, with the intention of using the filter during live scoring to route responses for human scoring. The filter processed over 100,000 responses to a 2019 administration and detected 31 unusual responses; a review of those results indicated that they did warrant human review.

Similarly, Bejar et al. (2014) examined the effect of replacing words with more advanced, longer synonyms on e-rater scoring to simulate the condition whereby examinees use memorized, complex words with the hope of improving their score. Specific words were targeted and 5% of the targeted words in an essay were substituted. The authors found that examinees had a higher chance of improving their score by one point than decreasing their score by one point, and most often received the same score.

The ability for engines to identify and route off-topic responses and engine's susceptibility to mis-scoring off-topic responses has also been studied. Off-topic responses could be written to another prompt (perhaps even accidentally), memorized from a canned generic prompt, or only vaguely (or not at all) written to the prompt in question. Examples of the latter include original stories provided by the examinee. Zhang, Chen, and Ruan (2016) examined the difference between scores for papers that were flagged as off-topic against scores for papers that were not flagged in two large-scale, high-stakes testing programs. The authors found that the flagged Off-Topic essays had higher mean scores than the unflagged papers, with the standardized mean differences of .15 and .25 on two items. The flagged papers had higher Quadratic Weighed Kappa (QWK) values than the unflagged papers. Higgins, Burstein, & Attali (2006) built filters intended to identify Off-Topic essays that were either written to other reference texts (called "unexpected topic" essays) or were written in bad faith (reflecting a "deliberate non-responsiveness" to the item).

To build filters, the authors used essays written to other prompts for the "unexpected topic" essays and used low-scoring essays (namely those earning a score of 0) for the bad-faith essays. The authors used different methods and training data for the two types of essays because the vocabulary of unexpected topic essays was expected to overlap with the reference items and the vocabulary of bad-faith essays were expected to have less vocabulary overlap and different writing characteristics. In their study, the bad-faith essays performed the worst, with varying false negative rates across items and filter designs (between 16.8% and 38%). Burkhardt & Lottridge (2013) examined the performance of a filter to detect Off-Topic essays written to known reference prompts and a filter to detect generic Off-Topic essays written on general topics. The authors found that the specific models (ones that used responses to known reference prompts) were more accurate than those using generic, off-topic text. The results varied by prompt, however. The generic models did not work well, underlying the need for a training set well-suited to the modelling task.

Duplicated responses have also been studied. Zhang, Chen, and Ruan (2016) examined the difference between scores for papers that were flagged for repetition against scores for papers that were not flagged in two large-scale, high-stakes testing programs. The authors found that the flagged duplicated essays had higher mean scores than the unflagged papers, with the standardized mean differences (SMDs) as .10, .12, and .15 on three items. The flagged papers had similar QWK values to the unflagged papers. Higgins & Heilman (2014) examined changes in engine performance for three engines used in the Automated Scoring Assessment Prize for Short Answers (Hewlett Foundation, 2013) when responses were duplicated two times. The authors found variation across engines ranging from .1 to .25 standard deviation units between predicted scores of original and Duplicated essays. This result again illustrates the impact of engine design, particularly around feature construction and use, on scoring gaming essays.

To our knowledge, no paper has addressed shuffling of words in the response. While this approach is not likely to be used in practice, we argue that the ability of an engine to detect and/or appropriately score such responses reflects a core validation principle. Automated scoring engines do not understand essays, but deep learning engines—such as the modern neural network engines that include word order as part of their modelling—are often described as "language models." Language models are conceived as a probability model over sequences of words; in other words, these models should assign low probabilities to unusual word combinations and these low probabilities should translate to lower scores in a properly trained model.

Thus, a few studies have examined the robustness of engines (particularly ETS e-rater) to some gaming conditions, but none so far have examined robustness across automated scoring engines, particularly those built using newer deep learning models.

## Feature-Based Automated Essay Scoring

Expert-derived feature algorithms form the core of engine performance in currently-deployed automated essay scoring engines. The feature algorithms transform the almost-infinite varieties of essays into vectors of values that are intended to accurately represent elements of writing and predict writing quality as measured by human raters. Automated scoring engines vary in their use of features and the level of feature design. In his description of vendor results from the Automated Scoring Assessment Prize essay scoring contest, Shermis (2014) presents the high-level feature spaces of the nine competing engines at the time of that contest, as outlined in Table 1. The article only generally describes the features, but the descriptions suggest features in spelling, usage, mechanics, semantics, development, vocabulary usage, and text complexity. It is our belief that most engines use bag-of-words methodologies whereby the occurrence or frequency of occurrence of words or word pairs are used in model prediction, and that these engines do not currently employ methods that explicitly model the sequence of words used in an essay.

Perelman (2014) argued that the achievement of the automated essay scoring engines in the ASAP essay scoring study was due primarily to essay length. In his paper, he computed the correlation of the number of words in an essay to seven vendor-produced scores and computed the correlation of the number of words with the human-assigned scores. Averaged across vendors, he found that the shared variance (i.e., correlation squared) for the vendor-produced scores ranged from 61.9% to 85.0% and the shared variance for the human-assigned scores ranged from 50.7% and 53.9%. These results suggest that automated scoring engines may be more susceptible to the artificial use of length than human raters.

In this study, we focus on CAI's automated essay scoring engine, Autoscore. As with the engines described earlier, Autoscore uses

*Table 1. Automated Essay Scoring Engines and their Descriptions (Adapted from Shermis, 2014)*

| Engine Name | Organization | Features |
|---|---|---|
| Autoscore | Cambium Assessment, Inc. | Latent Semantic Analysis, grammar/mechanics/spelling, textual features (sentence and paragraph variety). |
| LightSIDE | Carnegie Mellon | Term-document frequency matrices. |
| Bookette | DRC | 90 text-features classified as structural-, syntactic-, semantic-, and mechanics-based. |
| e-rater® | Educational Testing Service | Dozens of features grouped into conceptually similar sets, including: grammar, usage, mechanics, style, organization and development, lexical complexity, and content relevance. |
| Lexile Writing Analyzer® | MetaMetrics | Composition surface text features including degree of diverse use of vocabulary and greater vocabulary density. |
| Project Essay Grade (PEG) | Measurement, Inc. | Measures of structure, mechanics, organization, semantics and syntax. |
| Intelligent Essay Assessor (IEA) | Pearson | Latent Semantic Analysis, spelling, and grammar errors. |
| CRASE™ | ACT | Spelling, usage and mechanics errors, variation in sentence type, idea development, personal engagement, sentence beginnings, and overly-common words. |
| IntelliMetric | Vantage Learning | Multiple algorithms to determine specific writing features. |

computational linguistic and statistical methods to process essays and predict scores. Autoscore models are derived from the statistical analysis of a set of training papers that have been human scored. The core of Autoscore is the extraction of features that are aligned to broad categories observed in most writing rubrics: ideas, grammar, spelling, word choice, organization, and voice.

Autoscore's features include measures of syntax, grammatical/mechanical correctness, spelling correctness, paragraph quality, sentence variation and quality, and semantics. The measures of semantics rely on Latent Semantic Analysis (LSA) (Latent Semantic Analysis; Deerwester et al., 1990). LSA can be thought of as a principal components analysis on a matrix that is constructed from the set of training responses and the words in those responses. The components can be thought of as topics occurring in the sample. The contents of the matrix reflect the relative importance of the word in the collection of responses.[2] Note that responses are spell-corrected prior to LSA analysis but are not spell-corrected for the non-semantic features. The LSA algorithm, because it is based upon the tf-idf matrix, does implicitly model length. This can make LSA susceptible to gaming behaviors that involve deceptively increasing essay length (such as using duplication).

Approximately 20 non-semantic features and 40–100[3] LSA dimensions are used as predictors in any model. The non-semantic features do not explicitly measure the length

[2] Values in the matrix are calculated using the term frequency-document inverse frequency (tf-idf) statistic.

[3] Fewer LSA (0-20) dimensions are typically used in models built to score Conventions because semantic content is expected to have minimal impact on how raters—and thus—how the engine assigns scores.

of the response and are not affected by duplication of text. Autoscore features also do not measure word complexity, and thus do not reward or penalize simple or complex word choices. The use and weighting of feature values are customized for each rubric and score dimension, using responses to that rubric, to achieve an optimal statistical model of human scoring. Ordinal probit regression is used as the statistical model for all items and score dimensions. The score assigned to a response is the argmax of the probit probabilities, and the probability associated with the argmax is the probability associated with the score. As noted in the literature review, the combination and weighting of features, along with the words used in the training sample, can impact a model's susceptibility to some gaming behaviors, but it is difficult to predict the degree of the impact.

Autoscore uses a range of filters to flag unusual responses. Responses captured by these filters either receive a score of 0 or are routed to human review. Filters used to route responses to human review include: an "Out of vocab" filter, which captures responses that do not primarily overlap words with the training sample; a "Duplicate text" filter, which captures responses where at least 75% of the response is a duplicate of itself; and a "Non-specific" filter, which captures responses that are similar to off-topic, other language, or undecipherable responses flagged by human raters. Additionally, Autoscore computes a confidence measure in percentiles; typically, responses with the lowest confidence percentiles values (less than 15) are routed for human review but this can be configured to meet client preference.

## Deep Learning

The models used in deep learning have evolved quickly in the past five years. These models first map words in a piece of text to a word embedding, which is then modelled using neural network architectures, in particular recurrent neural networks and transformer networks. In this section, we describe these three elements: 1) word embeddings, 2) recurrent neural networks, and 3) Transformer networks.

### Word Embeddings

Word embeddings are vector representations of words and/or subwords (Milokov et al., 2013) and are a more efficient (i.e., reduced-dimensionality) method to represent words than a dummy-coded (i.e., 1 if word exists and 0 otherwise) representation of a vocabulary. Anecdotally, word embeddings are thought to represent semantic information, although this had not been systematically verified via empirical methods and is itself a difficult problem (Schnabel et al., 2015). Also note that words are associated with a single embedding, so that words that change in context of use (bank as in "steep bank" vs. bank as in "bank account") still have the same embedding. The idea underlying word embeddings is that words that appear in similar contexts should have high cosine-similarity (i.e., normalized dot product) values with one another, and those that do not appear in similar contexts should have low cosine-similarity.

Word embeddings are built using bodies of text (or *corpora*), typically based on publicly-available data such as Wikipedia entries, news articles, or books (Milokov et al., 2013; Pennington, 2014). Neural-network based word embeddings are built typically using one of two methods (Milokov et al, (2013): 1) divide text into context windows (e.g., 5 to 10 words), mask a word in the context, and predict that word from the other words in the context window (CBOW); or, 2) divide the text into context windows, mask a word in the context, and use the masked word to predict the probability that the words in the corpora are in the context window (Skip-gram). In either case, the neural network has an internal layer between the model inputs and outputs that is then used as the embedding. In essence, the prediction task is ignored. Embeddings can have any number of dimensions, but 50–300 dimensions is standard. This use of a single dimension to represent the vocabulary in a corpus offers benefits in modelling because the model can use a consistent size input (the word embedding dimension) rather than the number of words in a corpus. Once a word embedding is built, it can be used as input into a neural network. Note that word embeddings are built upon a specific corpus and it is possible that words appearing in an essay may not exist in the corpus; in this case, the words are either mapped to a word token representing an "unknown" embedding, or the words are divided into sub-words that are then mapped to known sub-word embeddings. Test items that use "unique" language may suffer when using word embeddings as the unique language may not be accurately captured in the word embedding vocabulary. Models often use text processing, such as spell-correction, prior to the word embeddings stage to aid in matching words to the embedding vocabulary; however, the nature of the processing varies by the type of word embedding used.

### Recurrent Neural Networks

Recurrent Neural Networks (RNNs; Williams, et al., 1986) represent an early approach in modelling language. RNNs model state-level information, where—in the context of essays—a state is the position of a word in an essay. In this sense, parameters are estimated to best model the sequence of words in a way that is predictive of score. LSTM networks are a refinement on RNNs that use gating mechanisms (i.e., weights applied to the state-level parameters) that allow for the model to "retain" and "forget" sequence data in order to optimize score prediction.
The addition to these parameters enabled the models to have more stable parameter estimates during engine calibration. Additionally, multiple copies (called "hidden units") of the LSTM are created in parallel during training and randomly initialized; the use of these copies enables the overall model to learn variations of sequences that can predict score and then combine those copies at the end of the modelling process to predict the final score. Layers of LSTM are also used to model sequences of prior modelled sequences. The inputs prior to score prediction might be thought of as dimensionality-reduced representations of the copies of the modelled essay. Note that throughout the process described above, all "features" in the model are learned during model calibration; there is no notion of an expert-derived feature. It should be noted that LSTMs and RNNs exhibit difficulty learning long sequences (such as essays) and so methods such as attention were used to focus the modelling on particular states or state sequences to optimize score prediction (Yoon et al., 2017). In recurrent networks, attention parameters are also learned; these parameters simply weight the state-level predictions to optimize the prediction.

### Transformer Networks

Transformer networks were devised to correct the known issues around long-sequence handling and calibration issues that have plagued the recurrent network approaches for some applications (Vaswani, 2017). Transformer networks have their roots in language translation and remove the reliance on the state-based recurrent networks to use only feed-forward neural networks. Feed-forward neural networks can be thought of as multi-variate linear regressions whose output is then transformed using an activation function to restrict the predicted output. Activation functions can be piece-wise continuous functions (i.e., 0 if the weighted linear sum is less than some value and 1 otherwise) or continuous functions (i.e., tanh). There are multiple types of transformer networks. In this section, we focus on BERT, because it was among the first transformer networks and is well-documented. In this paper, we describe BERT at a high level and from a functional standpoint. Interested readers are encouraged to read the original paper (Devlin et al., 2018) and blog posts describing BERT, particularly the Annotated Transformer (Rush, 2018) and the Illustrated Transformer (Alomar, 2018).

Users of BERT models are described as "fine-tuning" the originally calibrated model when they use the pre-trained model to predict scores on a novel task. There are a number of classes available for fine-tuning. In essay scoring, the BERTForSequenceClassification[4] is used, which allows for classification prediction. When we train the BERT engine on essays, we use the already-calibrated model and the parameter values associated with that model

[4] https://huggingface.co/transformers/v2.2.0/model_doc/bert.html#bertforsequenceclassification

serve as starting parameters. The training process consists of back-propagating changes to the parameter values in such a way as to optimize the prediction of the essay scores. The inputs to the model are BERT-based word embedding vectors and positional embedding vectors. Positional embedding vectors represent the word order of the response and are defined algorithmically.

The word embedding vocabulary in BERT has about 30,000 words and word pieces (sub-words), including single alphanumeric characters and punctuation. The embedding was built from Wikipedia entries and Books Corpus (Zhu et al., 2015) and has 768 dimensions. The rationale for the use of dimension size was not supplied in the original paper. The BERT vocabulary is relatively small compared to other word embedding vocabularies (such as Google's, which has about 3 million words). Note that embeddings are typically built upon much larger vocabularies and the public embeddings offer smaller vocabularies (presumably to provide more user-friendly embedding sizes for download and use). In BERT, users can elect to use cased (i.e., upper and lower case in words are retained) or uncased (i.e., case is ignored) versions.

Once BERT receives the word and position embedding vectors, the two vectors are summed to produce a model input vector. Additionally, 0/1 weights are applied so that for the essay, any word in the BERT vocabulary but not in the essay is set to have a weight of 0 (via multiplication by 0) and the words in the essay are multiplied by 1 to maintain their weight. This process is called *attention masking*. The essays then are processed through 12 layers, each with what is called *multi-headed self-attention* (12 heads), which are sums of dot-products of various weights whose purpose is to determine how words in the essay are related to other words in the essay (e.g., pronoun resolution). Each head is itself a "copy," or hidden unit of the self-attention, designed to capture distinct elements of the text (similar to the hidden units in the LSTM networks). The 12th layer is then used as input to a categorical soft-max prediction of scores.[5]

BERT models are quite large, with approximately 110 million parameters and 428 megabytes per model for $\text{BERT}_{\text{BASE}}$, which is used in this study.[6] Training of neural network models (including BERT) is typically conducted by segmenting the training sample into batches, and reserving a portion of the training sample (called the test *sample*) to evaluate the performance of the model by scoring the responses and examining agreement with the expected score at the end of a full run through the training sample (called an *epoch*). During calibration, predictions on each batch are made, averaged, and then errors are back-propagated through the network. When batch sizes exceed 1, this process is called *mini-batch gradient descent*; otherwise, it is called *stochastic full batch gradient descent*. Multiple epochs are used to train the engine the epoch that achieves the highest agreement on the test sample is stored and considered to be the best model. Hyper-parameters can be set to tune performance, including: the learning rate, the case of the embedding, and the number of layers to use in the model. The training of BERT models requires at least a graphics processing unit (GPU) with at least 16 GB memory.

BERT was originally trained on a word-masking task and a next-sentence prediction task. The goal of the word-masking task was to randomly mask or obscure a small portion of words and train the model to predict the masked word. The authors argue that their use of word-masking models the bidirectional use of language because the training task has to predict the masked word from the other words in the surrounding segment text. The goal of the next-sentence prediction task was to predict whether a sentence followed the earlier sentence. Each of these tasks is somewhat irrelevant to the fine-tuning process outlined earlier because—as with the word embedding modelling—the final layer of prediction is ignored

[5] BERT, like many machine learning algorithms, does not offer ordinal prediction functionality.

[6] $\text{BERT}_{\text{LARGE}}$ has 340 million parameters and has 24 layers, more hidden units (1024), and attention heads (16).

in the fine-tuning process and replaced with the prediction task at hand, in this case essay score modelling. Still, it is important to note that BERT modelled at the sentence level and our inputs into the BERT model during fine-tuning consider the entire essay to be a “sentence” in the BERT framework. The authors argue that one can use segments, rather than strictly sentences, with BERT; however, it is unknown whether the segment length or other characteristic matter theoretically or empirically, as there is no alternative model on which to test a hypothesis. This is because training (versus fine-tuning) a BERT model is prohibitive, requiring tensor processing units and long training times. Also, the BERT framework allows for essays (sentences) of no longer than 510 words, which means that adjustments need to be made to essays that are longer than 510 words, either by truncation or by averaging predictions across segments.

## Research Questions

As described in the Background section, Autoscore and BERT represent two very different frameworks for modelling essay scoring. Given this, we offer the following two research questions:

1. Will the BERT transformer model be more robust to gaming conditions than the classical essay scoring approach (using Autoscore), as measured by agreement with the expected score and by deviance from the expected score distribution?
2. Will the BERT transformer model be more robust, when controlling for length, to gaming conditions than the classical essay scoring approach (using Autoscore), as measured by agreement with the expected score and by deviance from the expected score distribution?

## Methods

### Items

Essay items were taken from an interim assessment program administered in grades 3–8 and used across multiple states. Four items were available per grade in the program. For each grade, two of the items assessed the Informative/Explanatory genre. One item in this genre was selected at random. Items in the paper are referred to by their grade.

The items were scored using a three-trait rubric that assessed Conventions (scored 0, 1, 2); Evidence & Elaboration (scored 1, 2, 3, 4); and Purpose, Focus, & Organization (scored 1, 2, 3, 4). The Conventions trait assesses the overall grammatical correctness (including spelling) of the response. The Purpose, Focus, & Organization assesses whether the essay states a claim, maintains focus around that claim and is organized appropriately to support that claim. The Evidence & Elaboration trait assesses the quality of the essay’s support of the claim. The rubric also allows for condition codes that are assigned to responses that do not meet the minimal rubric requirements. Three rubric-based condition codes are available and are described in Table 2. In the event of condition code U and N assignment, the essay would earn a 0 in each trait. In the event of O (off-topic) condition code assignment, the essay would earn a 0 in Evidence & Elaboration and in Purpose, Focus, & Organization, and a rubric-based score in Conventions. Essays with duplicated text or shuffled words are not described specifically in Table 2, although shuffled words may fall under code U. Similarly, the Babel essays could fall under the code U or O.

*Table 2. Condition Codes*

| Code | Description |
|---|---|
| U | Response is unintelligible or incomprehensible. *Examples:* random keystrokes, indecipherable text |
| N | Written in a language other than English |
| O | Off topic<br>Response will still be scored for Conventions/Editing.<br>*Example:* a student writes about spaceships when the prompt is about the weather |

## Data

Data were drawn from different sources for this study. This section describes the various types of data used in the study.

### Original

The Original data were taken from a field test used for item response theory (IRT) item calibration and scaling and were also used for engine training and validation. For the purposes of this study, the Original data were from the sample used to validate the engine. The responses were double scored by two independent human raters, with adjudication of non-exact scores. The final score out of the hand-scoring process was the adjudicated read, if it existed, or the rater reads (if identical).

The Original data were also examined for length (number of words) and the 30 closest essays to each five length percentiles were sampled at random to obtain a total of 150 essays. The five length percentiles were: 10th, 25th, 50th, 75th, and 90th. The average length of the responses at each percentile was then used as a basis for creating the other synthetic responses. The purpose of creating the length percentile-based samples was to examine engine performances across response length. Table 3 presents the average length of the essay in each length percentile bin by grade. Note that the essay length increases by grade at each of the percentile bins.

### Shuffled Essays

The Shuffled essays were generated from each of the 150 Original length percentile samples. Essays were tokenized into words (using nltk.tokenize[7]), and the words were randomly shuffled along with punctuation. The length of Shuffled essays was the same as the length of the Original essays. In practice, Shuffled essays would receive a condition code. However, if a rubric-based score is applied, we expect that the lowest possible rubric-based score would be used. Namely, this would be a 0 for Conventions, and 1 for Evidence & Elaboration and 1 for Purpose, Focus, & Organization. The Shuffled essays represent a theoretical challenge to automated essay scoring rather than a realistic challenge as they are typically not seen in operational scoring; however, the robustness of an engine to such essays may indicate improved language processing.

### Babel Essays

The Babel essays were obtained from Babel website[8] using an automated tool called Selenium.[9] The Babel site requires as input three key words and then generates grammatically correct but nonsense essays using complex language. Content experts at CAI provided the key words based upon the item passage and prompt. The same key words were used when generating each essay, and 150 essays per item were collected. The essays were

*Table 3. Length (Number of Words) of Essays in Each Percentile Bin*

| Grade | N. essays | 10th | 25th | 50th | 75th | 90th |
|---|---|---|---|---|---|---|
| 3 | 150 | 44.6 | 65.1 | 97.5 | 153.4 | 225.9 |
| 4 | 150 | 67.9 | 109.6 | 161.2 | 217.5 | 282.8 |
| 5 | 150 | 111.3 | 168.2 | 267.2 | 361.0 | 481.0 |
| 6 | 150 | 126.3 | 178.3 | 244.9 | 357.2 | 450.3 |
| 7 | 150 | 144.8 | 195.1 | 268.6 | 355.1 | 436.1 |
| 8 | 150 | 149.6 | 207.9 | 283.8 | 386.6 | 487.7 |

[7] https://www.nltk.org/_modules/nltk/tokenize.html

[8] https://babel-generator.herokuapp.com/

[9] https://selenium-python.readthedocs.io/

then truncated to match the average length of essays in each percentile grouping. Truncation is a worst-case scenario; a more “realistic” scenario would be to create shorter essays from across the full essay. In practice, Babel essays would receive a condition code; however, if a rubric-based score is applied, we might expect scores of 2 for Conventions (because they are grammatically correct), and scores of 1 for Evidence & Elaboration and 1 for Purpose, Focus, & Organization (because these essays are essentially non-sense essays). Table 4 presents the average lengths of the essays in each length percentile bin and the key words used in generating the essays, by grade.

### Off-Topic Essays

Off-Topic essays were chosen randomly from the pool of responses to the other three items at the same grade level and to match the percentile group lengths. Ten responses from each item at each length percentile bin were chosen. We would expect that the Off-Topic essays would receive a condition code in the Evidence & Elaboration and Purpose, Focus, & Organization traits and a rubric-based score in Conventions. Absent a condition code, the expected score of the Off-Topic essays is the original score received for Conventions (because Conventions is conceived as a general dimension beyond an individual item) and the lowest rubric score (1) for Evidence & Elaboration and Purpose, Focus, & Organization (because these essays do not address the item). Because only the essay’s training data are used when calibrating Autoscore models, we might expect Autoscore to be robust to Off-Topic essays, except where essay language may overlap across topics. It is unclear how transformers will behave on off-topic responses, in part because the transformer networks are built upon large corpora and then are fine-tuned to a particular prompt. Table 5 presents the average lengths of the essays in each length percentile, by grade.

### Duplicated Essays

Duplicated essays were created from the Original essays described earlier, and each essay was appended with a full copy of itself. Thus, the Duplicated essays were twice the length of the Original essays. Hand-scoring of the Duplicated essays is not specified specifically in the rubric or condition codes; however, we assume that the expected score of

*Table 4. Length (Number of Words) of Babel Essays in Each Percentile Bin and Key Words Used in Generating the Essays*

| Grade | Key words | N. essays | 10th | 25th | 50th | 75th | 90th |
|---|---|---|---|---|---|---|---|
| 3 | germs, wash, fight | 150 | 44.0 | 65.0 | 98.2 | 153.1 | 227.1 |
| 4 | frogs, skin, legs | 150 | 69.1 | 109.1 | 161.2 | 220.1 | 289.1 |
| 5 | camouflage, light, color | 150 | 110.0 | 171.2 | 267.1 | 363.2 | 474.3 |
| 6 | popcorn, water, heat | 150 | 124.1 | 179.2 | 245.1 | 359.1 | 451.1 |
| 7 | glass, heat, sand | 150 | 142.1 | 195.2 | 269.1 | 356.2 | 440.2 |
| 8 | memories, smell, imagery | 150 | 149.1 | 207.1 | 286.1 | 387.5 | 483.0 |

*Table 5. Length (Number of Words) of Off-Topic Essays in Each Percentile Bin*

| Grade | N. essays | 10th | 25th | 50th | 75th | 90th |
|---|---|---|---|---|---|---|
| 3 | 150 | 96.6 | 98.7 | 97.8 | 131.1 | 147.9 |
| 4 | 150 | 129.1 | 142.5 | 229.2 | 193.7 | 245.4 |
| 5 | 150 | 213.7 | 222.4 | 290.3 | 336.4 | 376.4 |
| 6 | 150 | 278.1 | 321.4 | 306.5 | 379.7 | 464.1 |
| 7 | 150 | 322.1 | 326.1 | 389.4 | 373.6 | 389.0 |
| 8 | 150 | 391.8 | 372.0 | 416.1 | 444.5 | 541.2 |

the Duplicated essay is the same as the original score for the essay. This is because the raters presumably would ignore the duplicated text and assign a score based upon the original text in the response. That said, we also might expect the raters to assign lower scores in Evidence & Elaboration and in Purpose, Focus, & Organization given the duplication; however, in this study, we assume the best-case scoring scenario whereby the raters simply ignore than extra text when assigning scores. Because Autoscore uses Latent Semantic Analysis and tf-idf methods, we expect Autoscore to be impacted by duplication; however, it is unclear to what extent the transformer networks will be affected.

### Engine Training and Validation

The same sample was used to train and validate the Autoscore and BERT engines, and 80% of the sample was used to train the engines, and 20% was used to validate the engines. The engine training and validation data only included "valid" student responses, or those responses that did not receive condition codes from human raters. The engine training and validation data did not, then, include any of the gaming type responses examined in this study.

In Autoscore, multiple models were built across LSA dimensions (0, 10, 20 for Conventions; 40–100 in increments of 10 for Evidence & Elaboration and Purpose, Focus, & Organization) and the best model was selected. Note that the set of features varied by trait and by item, as did the feature weights. That said, the LSA dimensions should offer more predictive power for Evidence & Elaboration and Purpose, Focus, & Organization relative to the grammatical/textual features, and less predictive power for Conventions. In BERT, 30 epochs were run and the model with the best quadratic weighted kappa measurement relative to the final human score was selected on an intermediate test set (chosen randomly from 20% of the training sample).

A validation sample was used to evaluate the performance of each engine. The key metrics used in the evaluation were: *Exact Agreement*, or the percentages of responses in which two scoring sources agreed; *Quadratic Weighted Kappa*, or the quadratically-weighted agreement above and beyond chance; and the absolute standardized mean difference (SMD) (using the pooled or averaged standard deviation of the two scoring sources). For threshold-based evaluations of performance, we allowed for the engine-final human score (AS-HS) exact agreement to be within 5.25% of human rater 1-human rater 2 (H1-H2) exact agreement, the AS-HS QWK within .1 of H1-H2 QWK, and the magnitude of the AS-HS SMD to be no greater than .15 (Williamson, Xi, and Breyer, 2012; McGraw-Hill Education CTB, 2014; Pearson and ETS, 2014).

### Evaluating Engine Results on Gaming Data

To evaluate the engine results on the gaming data, three metrics were used: exact agreement with the expected score; the Kullback-Leibler (KL) divergence metric of the engine score point distribution against the expected score distribution; and inspection of engine score point distributions relative to the expected score distribution.

Briefly, the Kullback-Leibler divergence measure provides an overall measure of how different one probability distribution (*P*) is from a reference probability distribution (*Q*). In this study, the expected score distribution was used as the reference probability distribution for both engine comparisons in the gaming condition under consideration. Values closer to zero indicate that the *P* distribution is similar to the *Q* distribution and greater values indicate the *P* distribution diverges from the *Q* distribution. The formula for computing the metric appears below and here *X* refers to the set of scores in the rubric and the probabilities are computed as percentages.[10]

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

[10] .0001 was added to all percentage values to ensure division by zero did not occur.

The agreement, KL metric, and the score distributions were computed for each engine and grade for each gaming condition overall and for each length percentile bin.

## Results

This section presents the results of the study, including the performance of the engines on the validation samples, the ability of Autoscore to detect the gaming responses using the full suite of filters, the performance of the two engines (without filters) on the gaming datasets overall and by percentile grouping.

### Performance of Engines on Validation Samples

Using the criteria outlined earlier, both Autoscore and BERT met the QWK and Exact Agreement criteria, generally exceeding human-human agreement for each trait (Table 6). Autoscore met the SMD criteria for 15 of the 18 item/trait combinations, and BERT met the SMD criteria for 14 of the 18 item/trait combinations. Autoscore and BERT failed the criteria on the

*Table 6. Performance of Humans, Autoscore, and BERT on Essay Item Validation Samples, by Trait*

| Conventions | | QWK | | | Exact Agreement (%) | | | SMD | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Grade | N | H1-H2 | AS-HS | BERT-HS | H1-H2 | AS-HS | BERT-HS | H1-H2 | AS-HS | BERT-HS |
| 3 | 366 | 0.59 | 0.68 | 0.73 | 63 | 71 | 73 | 0.15 | -0.04 | 0.11 |
| 4 | 362 | 0.58 | 0.61 | 0.74 | 62 | 67 | 77 | 0.00 | -0.05 | -0.05 |
| 5 | 382 | 0.60 | 0.62 | 0.64 | 74 | 80 | 81 | -0.14 | **0.23** | **0.18** |
| 6 | 381 | 0.60 | 0.63 | 0.66 | 65 | 72 | 73 | -0.03 | **0.18** | **0.17** |
| 7 | 360 | 0.69 | 0.71 | 0.74 | 82 | 82 | 84 | -0.05 | 0.08 | 0.15 |
| 8 | 373 | 0.59 | 0.66 | 0.72 | 79 | 83 | 85 | 0.06 | **0.15** | **0.16** |
| Averages | | 0.61 | 0.65 | 0.71 | 71 | 76 | 79 | 0.00 | 0.09 | 0.12 |
| Elaboration | | QWK | | | Exact Agreement (%) | | | SMD | | |
| Grade | N | H1-H2 | AS-HS | BERT-HS | H1-H2 | AS-HS | BERT-HS | H1-H2 | AS-HS | BERT-HS |
| 3 | 366 | 0.64 | 0.66 | 0.71 | 61 | 67 | 70 | 0.02 | -0.03 | -0.13 |
| 4 | 362 | 0.53 | 0.65 | 0.71 | 56 | 70 | 75 | -0.05 | -0.05 | 0.01 |
| 5 | 382 | 0.66 | 0.69 | 0.72 | 59 | 67 | 69 | -0.02 | 0.04 | -0.08 |
| 6 | 381 | 0.73 | 0.75 | 0.77 | 59 | 68 | 69 | 0.05 | 0.04 | 0.10 |
| 7 | 360 | 0.66 | 0.60 | 0.78 | 77 | 72 | 82 | -0.04 | -0.03 | 0.13 |
| 8 | 373 | 0.72 | 0.68 | 0.69 | 73 | 70 | 71 | 0.05 | -0.06 | -0.06 |
| Averages | | 0.66 | 0.67 | 0.73 | 64 | 69 | 73 | 0.00 | -0.01 | 0.00 |
| Organization | | QWK | | | Exact Agreement (%) | | | SMD | | |
| Grade | N | H1-H2 | AS-HS | BERT-HS | H1-H2 | AS-HS | BERT-HS | H1-H2 | AS-HS | BERT-HS |
| 3 | 366 | 0.65 | 0.62 | 0.67 | 62 | 66 | 66 | 0.03 | -0.09 | 0.13 |
| 4 | 362 | 0.56 | 0.67 | 0.70 | 59 | 69 | 69 | -0.02 | -0.11 | 0.01 |
| 5 | 382 | 0.58 | 0.55 | 0.65 | 59 | 65 | 69 | -0.06 | -0.08 | -0.05 |
| 6 | 381 | 0.73 | 0.72 | 0.72 | 61 | 66 | 65 | 0.09 | 0.00 | **0.20** |
| 7 | 360 | 0.66 | 0.70 | 0.70 | 67 | 70 | 69 | -0.05 | 0.00 | 0.09 |
| 8 | 373 | 0.73 | 0.72 | 0.75 | 68 | 70 | 73 | -0.03 | -0.04 | 0.02 |
| Averages | | 0.65 | 0.66 | 0.70 | 63 | 68 | 69 | -0.01 | -0.05 | 0.07 |

Note. H1=First human read. H2=Second human read. HS=Final human read. AS=Autoscore. BERT=BERT model. Values identified in bold red exceeded the criterion thresholds.

same items (grades 5, 6, and 8) and traits (Conventions). Across items, BERT generally outperformed Autoscore in the agreement measures (QWK and Exact Agreement) but had slightly larger SMDs than Autoscore.

Because essay length (number of words) may have an impact on engine performance in general and on the gaming responses, the correlation of the length with the final human, Autoscore, and BERT scores was examined (Table 7). The correlations were similar (within .1) across the three scoring sources for each of the three traits. The correlations were weak for Conventions (ranging between .24 and .37 for the human score), and moderate in Evidence & Elaboration and Purpose, Focus, & Organization (.50–.61 and .48–.65, respectively, for the human score). Given this, we should expect that the impact of length on Conventions scoring for either model should be small and impact of length on the other two traits should be similar.

*Table 7. Correlation of Essay Length (number of words) to Human, Autoscore, and BERT Scores on Essay Item Validation Sample, by Trait*

| Conventions | | | |
|---|---|---|---|
| **Grade** | **HS** | **Autoscore** | **BERT** |
| 3 | 0.28 | 0.36 | 0.30 |
| 4 | 0.27 | 0.36 | 0.35 |
| 5 | 0.28 | 0.29 | 0.33 |
| 6 | 0.24 | 0.33 | 0.28 |
| 7 | 0.27 | 0.28 | 0.29 |
| 8 | 0.37 | 0.31 | 0.29 |
| **Averages** | 0.29 | 0.32 | 0.31 |
| **Elaboration** | | | |
| **Grade** | **HS** | **Autoscore** | **BERT** |
| 3 | 0.58 | 0.66 | 0.63 |
| 4 | 0.56 | 0.49 | 0.52 |
| 5 | 0.59 | 0.63 | 0.60 |
| 6 | 0.61 | 0.69 | 0.71 |
| 7 | 0.50 | 0.62 | 0.48 |
| 8 | 0.62 | 0.65 | 0.57 |
| **Averages** | 0.58 | 0.62 | 0.58 |

*Table 7. Correlation of Essay Length (number of words) to Human, Autoscore, and BERT Scores on Essay Item Validation Sample, by Trait (continued)*

| Organization | | | |
|---|---|---|---|
| **Grade** | **HS** | **Autoscore** | **BERT** |
| 3 | 0.53 | 0.57 | 0.56 |
| 4 | 0.48 | 0.39 | 0.48 |
| 5 | 0.52 | 0.53 | 0.60 |
| 6 | 0.62 | 0.64 | 0.71 |
| 7 | 0.62 | 0.71 | 0.66 |
| 8 | 0.65 | 0.62 | 0.71 |
| **Averages** | 0.58 | 0.58 | 0.63 |

### Detection of “Gaming” Responses Using Filters

This section outlines the extent to which Autoscore, with the filters, was able to identify and appropriately route the responses. Autoscore flagged every Babel essay with an Out of Vocab condition code and flagged every Duplicated essay with a Duplicated condition code. The results for the two other conditions (Shuffled and Off-Topic), which were not identified by filters, are presented in this section.

### Shuffled Essays

Autoscore detected very few Shuffled essays using the condition code filters (Table 8, 1 in each trait). This is because that Autoscore has no filter in place to identify Shuffled essays. The essays that were assigned the Out of Vocab condition code was assigned that code because the word tokenization method in this study varied from that used in Autoscore. In Conventions, most of the Shuffled essays received a 0 from Autoscore (595 out of 900) across items, with the percentages of 0s decreasing as the grades increased. When Autoscore gave a score of 1 or 2, it was identified as low confidence for most responses. The grade 8 item showed the poorest performance in Conventions.

In Evidence & Elaboration, Autoscore assigned the lowest rubric score to just under half of the responses (442 of 900) across items. In grades 5, 6, and 7, when Autoscore assigned a score greater than 1, it was most often flagged as low confidence (identified as *Rtd* in the table). In the other grades (3, 4, and 8), this was not the case. In Purpose, Focus, & Organization, Autoscore assigned the lowest rubric score to just over half of the responses (549 of 900) across items. In grades 5, 6, and 7, when Autoscore assigned a score greater than 1, it was most often flagged as low confidence. In the other grades (3, 4, and 8), this was not the case.

Thus, the ability of Autoscore—with the confidence filters—showed generally good, although varied, performance across items. In grades where Autoscore assigned inappropriate scores with high confidence, it is likely that the LSA features were dominant in the score prediction.

*Table 8. Detection of Shuffled Essays by Autoscore, with Filters*

| | | | | | | Score and Routing Status | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Conventions | | | Condition Code Assigned | | | 0 | | 1 | | 2 | | | |
| Grade | N | Length | OOV | DUP | NSP | Rtd | AS | Rtd | AS | Rtd | AS | | |
| 3 | 150 | 117.3 | 0 | 0 | 0 | 6 | 135 | 5 | 4 | 0 | 0 | | |
| 4 | 150 | 167.8 | 1 | 0 | 0 | 4 | 145 | 0 | 0 | 0 | 0 | | |
| 5 | 150 | 277.7 | 0 | 0 | 0 | 59 | 59 | 27 | 5 | 0 | 0 | | |
| 6 | 150 | 271.4 | 0 | 0 | 0 | 50 | 44 | 49 | 6 | 1 | 0 | | |
| 7 | 150 | 279.9 | 0 | 0 | 0 | 35 | 39 | 68 | 0 | 8 | 0 | | |
| 8 | 150 | 303.1 | 0 | 0 | 0 | 1 | 18 | 22 | 42 | 23 | 44 | | |
| Sum | | | 1 | 0 | 0 | 155 | 440 | 171 | 57 | 32 | 44 | | |
| | | | | | | Score and Routing Status | | | | | | | |
| Elaboration | | | Condition Code Assigned | | | 1 | | 2 | | 3 | | 4 | |
| Grade | N | Length | OOV | DUP | NSP | Rtd | AS | Rtd | AS | Rtd | AS | Rtd | AS |
| 3 | 150 | 117.3 | 0 | 0 | 0 | 1 | 74 | 4 | 58 | 6 | 7 | 0 | 0 |
| 4 | 150 | 167.8 | 1 | 0 | 0 | 0 | 80 | 3 | 64 | 1 | 1 | 0 | 0 |
| 5 | 150 | 277.7 | 0 | 0 | 0 | 38 | 62 | 44 | 2 | 4 | 0 | 0 | 0 |
| 6 | 150 | 271.4 | 0 | 0 | 0 | 12 | 32 | 57 | 15 | 29 | 3 | 2 | 0 |
| 7 | 150 | 279.9 | 0 | 0 | 0 | 33 | 39 | 76 | 0 | 2 | 0 | 0 | 0 |
| 8 | 150 | 303.1 | 0 | 0 | 0 | 12 | 59 | 25 | 38 | 9 | 6 | 0 | 1 |
| Sum | | | 1 | 0 | 0 | 96 | 346 | 209 | 177 | 51 | 17 | 2 | 1 |
| | | | | | | Score and Routing Status | | | | | | | |
| Organization | | | Condition Code Assigned | | | 1 | | 2 | | 3 | | 4 | |
| Grade | N | Length | OOV | DUP | NSP | Rtd | AS | Rtd | AS | Rtd | AS | Rtd | AS |
| 3 | 150 | 117.3 | 0 | 0 | 0 | 0 | 93 | 11 | 46 | 0 | 0 | 0 | 0 |
| 4 | 150 | 167.8 | 1 | 0 | 0 | 1 | 138 | 3 | 7 | 0 | 0 | 0 | 0 |
| 5 | 150 | 277.7 | 0 | 0 | 0 | 28 | 62 | 58 | 2 | 0 | 0 | 0 | 0 |
| 6 | 150 | 271.4 | 0 | 0 | 0 | 28 | 38 | 63 | 12 | 9 | 0 | 0 | 0 |
| 7 | 150 | 279.9 | 0 | 0 | 0 | 34 | 39 | 69 | 0 | 8 | 0 | 0 | 0 |
| 8 | 150 | 303.1 | 0 | 0 | 0 | 26 | 62 | 19 | 40 | 1 | 2 | 0 | 0 |
| Sum | | | 1 | 0 | 0 | 117 | 432 | 223 | 107 | 18 | 2 | 0 | 0 |

Note. OOV=Out of Vocab Filter. DUP=Duplicate Text Filter. NSP=Non-Specific Filter. Rtd=Confidence percentile lower than 15 and would be routed for human review. AS=Autoscore.

### Off-Topic Essays

Autoscore detected very few Off-Topic essays using the condition code filters (Table 9, 9 in each trait). Note that Autoscore has no filter in place specifically to identify Off-Topic essays in these items because there were no Off-Topic codes assigned to the responses in the training and validation sample.[11] In Conventions, responses received a range of scores, with 26% (232/900) receiving confidence percentiles below 15 and routed for human review. Note that in a random sample, we expect about 15% of the responses to be routed; this higher percentage is likely due to the off-topic nature of

*Table 9. Detection of Off-Topic Essays by Autoscore, with Filters*

| | | | | | | Score and Routing Status | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Conventions | | | Condition Code Assigned | | | 0 | | 1 | | 2 | | | |
| Grade | N | Length | OOV | DUP | NSP | Rtd | AS | Rtd | AS | Rtd | AS | | |
| 3 | 150 | 114.4 | 4 | 0 | 0 | 0 | 12 | 5 | 91 | 3 | 35 | | |
| 4 | 150 | 188.0 | 2 | 1 | 1 | 0 | 36 | 15 | 94 | 1 | 0 | | |
| 5 | 150 | 287.8 | 1 | 0 | 0 | 0 | 3 | 6 | 64 | 20 | 56 | | |
| 6 | 150 | 349.9 | 0 | 0 | 0 | 1 | 8 | 21 | 27 | 64 | 29 | | |
| 7 | 150 | 360.0 | 0 | 0 | 0 | 2 | 2 | 13 | 16 | 68 | 49 | | |
| 8 | 150 | 433.1 | 0 | 0 | 0 | 0 | 5 | 0 | 10 | 13 | 122 | | |
| Sum | | | 7 | 1 | 1 | 3 | 66 | 60 | 302 | 169 | 291 | | |

| | | | | | | Score and Routing Status | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Elaboration | | | Condition Code Assigned | | | 1 | | 2 | | 3 | | 4 | |
| Grade | N | Length | OOV | DUP | NSP | Rtd | AS | Rtd | AS | Rtd | AS | Rtd | AS |
| 3 | 150 | 114.4 | 4 | 0 | 0 | 8 | 138 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 150 | 188.0 | 2 | 1 | 1 | 12 | 130 | 4 | 0 | 0 | 0 | 0 | 0 |
| 5 | 150 | 287.8 | 1 | 0 | 0 | 24 | 123 | 2 | 0 | 0 | 0 | 0 | 0 |
| 6 | 150 | 349.9 | 0 | 0 | 0 | 53 | 44 | 32 | 20 | 1 | 0 | 0 | 0 |
| 7 | 150 | 360.0 | 0 | 0 | 0 | 23 | 26 | 60 | 41 | 0 | 0 | 0 | 0 |
| 8 | 150 | 433.1 | 0 | 0 | 0 | 8 | 121 | 4 | 16 | 1 | 0 | 0 | 0 |
| Sum | | | 7 | 1 | 1 | 128 | 582 | 102 | 77 | 2 | 0 | 0 | 0 |

| | | | | | | Score and Routing Status | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Organization | | | Condition Code Assigned | | | 1 | | 2 | | 3 | | 4 | |
| Grade | N | Length | OOV | DUP | NSP | Rtd | AS | Rtd | AS | Rtd | AS | Rtd | AS |
| 3 | 150 | 114.4 | 4 | 0 | 0 | 6 | 135 | 2 | 3 | 0 | 0 | 0 | 0 |
| 4 | 150 | 188.0 | 2 | 1 | 1 | 9 | 128 | 7 | 2 | 0 | 0 | 0 | 0 |
| 5 | 150 | 287.8 | 1 | 0 | 0 | 21 | 119 | 5 | 4 | 0 | 0 | 0 | 0 |
| 6 | 150 | 349.9 | 0 | 0 | 0 | 38 | 44 | 47 | 20 | 1 | 0 | 0 | 0 |
| 7 | 150 | 360.0 | 0 | 0 | 0 | 57 | 41 | 25 | 26 | 1 | 0 | 0 | 0 |
| 8 | 150 | 433.1 | 0 | 0 | 0 | 10 | 136 | 3 | 1 | 0 | 0 | 0 | 0 |
| Sum | | | 7 | 1 | 1 | 141 | 603 | 89 | 56 | 2 | 0 | 0 | 0 |

Note. OOV=Out of Vocab Filter. DUP=Duplicate Text Filter. NSP=Non-Specific Filter. Rtd=Confidence percentile lower than 15 and would be routed for human review. AS=Autoscore.

[11] Presumably, this was due to an oversight in hand-scoring.

the responses, even in Conventions. Note that Autoscore models are trained for each prompt and there is no generic Conventions prompt; thus, we might expect Autoscore performance to be slightly worse when scoring responses from one item using the model from another item.

In Evidence & Elaboration, Autoscore assigned the lowest rubric score to most of the responses (708 of 900) across items. Across grades, when Autoscore assigned a score greater than 1, it was most often flagged as low confidence. Most of the scores above 1 occurred in grades 6 and 7. In Purpose, Focus, & Organization, Autoscore assigned the lowest rubric score to most of the responses (744 of 900) across items. Across grades, when Autoscore assigned a score greater than 1, it was most often flagged as "low confidence." As with the Evidence & Elaboration, most of the scores above 1 occurred in grades 6 and 7. The results in grades 6 and 7 may be due to the use of prompts that use more common language and so are less likely to be found as unusual by the flags.

### Scoring of Gaming Responses by Engine

Next, the performance of Autoscore and the BERT fine-tuned engine—without filters—were evaluated on each of the gaming essays. For each condition, the accuracy (exact agreement) of the engines with the expected score is provided, as are the score point distributions of each engine score. In addition, the Kullback-Leibler metric is provided to illustrate the discrepancy of each engine score distribution with the expected distribution. The results are presented by gaming condition, with the item-level results presented first (collapsing length percentiles), followed by the length percentile bins (collapsing items). Appendices A–D provide the results for each length percentile bin and item.

### Shuffled Essays

Recall that for the Shuffled essays, the expected score for the Conventions trait was 0, and the expected score for each of the Evidence & Elaboration and Purpose, Focus, & Organization trait was 1.

### Item-Level Results

For the shuffled essays, BERT and Autoscore showed fairly low agreement with the expected scores across items for each trait (Table 10). Autoscore agreement with the expected score exceeded 80% for only three items and traits; BERT showed better performance, with 9 items and traits exceeding 80%. Both Autoscore and BERT showed very low agreement (<30%) for a few items and traits, as well (2 for Autoscore, 4 for BERT). The averaged Kullback-Leibler values showed variance across items and traits without a discernable pattern.

Autoscore and BERT did not display similar patterns across items; namely, if one engine scored well, this was not necessarily associated to good scores by the other engine. This suggests that the scoring behavior was not a property of the item; rather, it is a property of the engine.

### Length Percentile Bin Results

When the same statistics are reviewed within length percentile bins across items, some trends around length emerge. Namely, as observed in Table 11, agreement with the expected score by both Autoscore and BERT is relatively high at the lowest length percentile bin and then decreases steadily as the length percentile bin increases. Autoscore shows a greater rate of decrease, exhibiting sharper drops in accuracy at the 75th and 90th length percentile bins. The Kullback-Leibler measure increases for both Autoscore and BERT across length percentile bins. Autoscore assigned more non-adjacent scores in Conventions at the 75th and 90th length percentile bins, and both engines tended to assign more non-adjacent scores in the other two traits at the 75th and 90th percentile bins.

This set of results indicate that Autoscore and BERT are not generally robust to shuffling, and that both engines use length to some

*Table 10. Autoscore and BERT Predictions on Shuffled Essays, by Item*

| Conventions | | Accuracy (%) | | Kullback-Leibler | | Expected Score | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade | N | AS | BERT | AS | BERT | | 0 | 1 | 2 | | 0 | 1 | 2 | |
| 3 | 150 | 94 | 98 | 0.33 | 0.09 | 0 | 94 | 6 | 0 | | 98 | 2 | 0 | |
| 4 | 150 | 100 | 100 | 0.00 | 0.00 | 0 | 100 | 0 | 0 | | 100 | 0 | 0 | |
| 5 | 150 | 79 | 7 | 1.45 | 8.35 | 0 | 79 | 21 | 0 | | 7 | 93 | 0 | |
| 6 | 150 | 63 | 100 | 2.74 | 0.00 | 0 | 63 | 37 | 1 | | 100 | 0 | 0 | |
| 7 | 150 | 49 | 15 | 3.80 | 7.37 | 0 | 49 | 45 | 5 | | 15 | 85 | 0 | |
| 8 | 150 | 13 | 60 | 7.06 | 2.95 | 0 | 13 | 43 | 45 | | 60 | 39 | 1 | |
| Average | | 66 | 63 | 2.56 | 3.13 | | 66 | 25 | 8 | | 63 | 36 | 0 | |
| Elaboration | | Accuracy (%) | | Kullback-Leibler | | Expected Score | Autoscore (%) | | | | BERT (%) | | | |
| Grade | N | AS | BERT | AS | BERT | | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 3 | 150 | 50 | 43 | 3.68 | 4.32 | 1 | 50 | 41 | 9 | 0 | 43 | 47 | 11 | 0 |
| 4 | 150 | 54 | 80 | 3.49 | 1.34 | 1 | 54 | 45 | 1 | 0 | 80 | 20 | 0 | 0 |
| 5 | 150 | 67 | 60 | 2.34 | 3.01 | 1 | 67 | 31 | 3 | 0 | 60 | 40 | 0 | 0 |
| 6 | 150 | 29 | 33 | 5.41 | 5.10 | 1 | 29 | 48 | 21 | 1 | 33 | 46 | 20 | 1 |
| 7 | 150 | 48 | 97 | 4.03 | 0.15 | 1 | 48 | 51 | 1 | 0 | 97 | 3 | 1 | 0 |
| 8 | 150 | 47 | 93 | 3.87 | 0.37 | 1 | 47 | 42 | 10 | 1 | 93 | 5 | 2 | 0 |
| Average | | 49 | 67 | 3.80 | 2.38 | | 49 | 43 | 8 | 0 | 67 | 27 | 6 | 0 |
| Organization | | Accuracy (%) | | Kullback-Leibler | | Expected Score | Autoscore (%) | | | | BERT (%) | | | |
| Grade | N | AS | BERT | AS | BERT | | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 3 | 150 | 62 | 99 | 2.84 | 0.02 | 1 | 62 | 38 | 0 | 0 | 99 | 1 | 0 | 0 |
| 4 | 150 | 93 | 99 | 0.37 | 0.02 | 1 | 93 | 7 | 0 | 0 | 99 | 1 | 0 | 0 |
| 5 | 150 | 60 | 15 | 3.01 | 7.01 | 1 | 60 | 40 | 0 | 0 | 15 | 72 | 13 | 0 |
| 6 | 150 | 44 | 49 | 4.28 | 4.03 | 1 | 44 | 50 | 6 | 0 | 49 | 51 | 0 | 0 |
| 7 | 150 | 49 | 93 | 3.86 | 0.38 | 1 | 49 | 46 | 5 | 0 | 93 | 6 | 1 | 0 |
| 8 | 150 | 59 | 21 | 3.05 | 6.20 | 1 | 59 | 39 | 2 | 0 | 21 | 48 | 31 | 0 |
| Average | | 61 | 63 | 2.90 | 2.94 | | 61 | 37 | 2 | 0 | 63 | 30 | 7 | 0 |

degree in their prediction of score. Notably, as the length increases, the engines typically perform worse on these items (although this is not always the case, as can be seen in Appendix A). These results suggest that the grammar features in Autoscore are not enough to mitigate the LSA bag of word features when responses are shuffled for most items. They also suggest that, although positional embeddings are used in the BERT model (recall that the BERT input embeddings are the sum of the word embeddings and the positional embeddings), they are not sufficient to overcome the value of the word embeddings in prediction for most items.

### Babel Essays

For the Babel essays, the expected score for the Conventions trait was 2 (because responses are grammatically correct), and the expected score for each of the Evidence & Elaboration and Purpose, Focus, and Organization trait was 1.

*Table 11. Autoscore and BERT Predictions on Shuffled Essays, by Length Percentile Bin*

| Conventions | | Accuracy (%) | | Kullback-Leibler | | Expected Score | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **% ile** | **N** | **AS** | **BERT** | **AS** | **BERT** | | **0** | **1** | **2** | | **0** | **1** | **2** | |
| 10th | 180 | 83 | 76 | 1.10 | 1.65 | 0 | 83 | 17 | 1 | | 76 | 24 | 0 | |
| 25th | 180 | 73 | 66 | 1.85 | 2.48 | 0 | 73 | 26 | 1 | | 66 | 34 | 0 | |
| 50th | 180 | 67 | 64 | 2.21 | 2.62 | 0 | 67 | 26 | 7 | | 64 | 36 | 0 | |
| 75th | 180 | 57 | 57 | 2.99 | 3.31 | 0 | 57 | 29 | 14 | | 57 | 43 | 0 | |
| 90th | 180 | 51 | 53 | 3.53 | 3.55 | 0 | 51 | 29 | 20 | | 53 | 46 | 1 | |
| **Average** | | 66 | 63 | 2.34 | 2.72 | | 66 | 25 | 8 | | 63 | 36 | 0 | |
| **Elaboration** | | **Accuracy (%)** | | **Kullback-Leibler** | | **Expected Score** | **Autoscore (%)** | | | | **BERT (%)** | | | |
| **% ile** | **N** | **AS** | **BERT** | **AS** | **BERT** | | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 10th | 180 | 93 | 95 | 0.41 | 0.26 | 1 | 93 | 7 | 0 | 0 | 95 | 05 | 0 | 0 |
| 25th | 180 | 71 | 81 | 2.06 | 1.30 | 1 | 71 | 29 | 0 | 0 | 81 | 19 | 0 | 0 |
| 50th | 180 | 46 | 61 | 4.22 | 2.88 | 1 | 46 | 52 | 3 | 0 | 61 | 38 | 1 | 0 |
| 75th | 180 | 23 | 55 | 6.21 | 3.27 | 1 | 23 | 66 | 11 | 0 | 55 | 38 | 07 | 0 |
| 90th | 180 | 13 | 46 | 7.00 | 3.92 | 1 | 13 | 61 | 24 | 2 | 46 | 33 | 21 | 1 |
| **Average** | | 49 | 67 | 3.98 | 2.33 | | 49 | 43 | 8 | 0 | 67 | 27 | 6 | 0 |
| **Organization** | | **Accuracy (%)** | | **Kullback-Leibler** | | **Expected Score** | **Autoscore (%)** | | | | **BERT (%)** | | | |
| **% ile** | **N** | **AS** | **BERT** | **AS** | **BERT** | | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 10th | 180 | 96 | 89 | 0.19 | 0.64 | 1 | 96 | 4 | 0 | 0 | 89 | 11 | 0 | 0 |
| 25th | 180 | 86 | 69 | 0.88 | 2.20 | 1 | 86 | 14 | 0 | 0 | 69 | 31 | 0 | 0 |
| 50th | 180 | 59 | 59 | 3.11 | 3.06 | 1 | 59 | 41 | 0 | 0 | 59 | 40 | 1 | 0 |
| 75th | 180 | 37 | 50 | 5.14 | 3.61 | 1 | 37 | 63 | 1 | 0 | 50 | 36 | 14 | 0 |
| 90th | 180 | 28 | 46 | 5.76 | 3.91 | 1 | 28 | 62 | 11 | 0 | 46 | 32 | 22 | 0 |
| **Average** | | 61 | 63 | 3.02 | 2.68 | | 61 | 37 | 2 | 0 | 63 | 30 | 7 | 0 |

### Item-Level Results

For the Babel essays, BERT and Autoscore showed fairly low agreement with the expected score across items for each trait (Table 12). Autoscore's agreement with the expected score exceeded 80% for seven items and traits; BERT showed slightly worse performance, with five items and traits exceeding 80%. Both Autoscore and BERT showed very low agreement (<30%) for a few items and traits, as well (2 for Autoscore, 7 for BERT).

BERT performed well at grades 5–8 in Conventions, and both engines performed poorly in Conventions for grades 3 and 4. This result may be due to the fact that the Babel sentences are relatively sophisticated and lie outside the typical writing observed for those grades, causing issues for both engines. Autoscore showed strong agreement for grades 3–5 in Evidence and Elaboration with worse performance at the higher grades. BERT showed poor agreement at the higher grades. Autoscore showed strong agreement for grades 3–5 and 8 in Purpose, Focus, & Organization. BERT showed poor agreement at all grades, except grade 4.

*Table 12. Autoscore and BERT Predictions on Babel Essays, by Item*

| Conventions | | Accuracy (%) | | Kullback-Leibler | | Expected Score | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade | N | AS | BERT | AS | BERT | | 0 | 1 | 2 | | 0 | 1 | 2 | |
| 3 | 150 | 4 | 9 | 8.67 | 8.11 | 2 | 0 | 96 | 4 | | 0 | 91 | 9 | |
| 4 | 150 | 0 | 0 | 9.17 | 8.98 | 2 | 1 | 99 | 0 | | 94 | 6 | 0 | |
| 5 | 150 | 75 | 96 | 1.71 | 0.20 | 2 | 0 | 25 | 75 | | 0 | 4 | 96 | |
| 6 | 150 | 66 | 81 | 2.49 | 1.29 | 2 | 0 | 34 | 66 | | 0 | 19 | 81 | |
| 7 | 150 | 97 | 98 | 0.16 | 0.09 | 2 | 0 | 3 | 97 | | 0 | 2 | 98 | |
| 8 | 150 | 100 | 100 | 0.00 | 0.00 | 2 | 0 | 0 | 100 | | 0 | 0 | 100 | |
| Average | | 57 | 64 | 3.70 | 3.11 | | 0 | 43 | 57 | | 16 | 20 | 64 | |
| **Elaboration** | | **Accuracy (%)** | | **Kullback-Leibler** | | **Expected Score** | **Autoscore (%)** | | | | **BERT (%)** | | | |
| **Grade** | **N** | **AS** | **BERT** | **AS** | **BERT** | | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 3 | 150 | 99 | 71 | 0.02 | 2.04 | 1 | 99 | 1 | 0 | 0 | 71 | 29 | 0 | 0 |
| 4 | 150 | 96 | 100 | 0.20 | 0.00 | 1 | 96 | 4 | 0 | 0 | 100 | 0 | 0 | 0 |
| 5 | 150 | 96 | 55 | 0.20 | 3.49 | 1 | 96 | 4 | 0 | 0 | 55 | 45 | 0 | 0 |
| 6 | 150 | 41 | 7 | 4.72 | 7.52 | 1 | 41 | 59 | 0 | 0 | 7 | 34 | 55 | 4 |
| 7 | 150 | 38 | 23 | 5.04 | 6.19 | 1 | 38 | 62 | 0 | 0 | 23 | 61 | 16 | 0 |
| 8 | 150 | 55 | 39 | 3.43 | 4.57 | 1 | 55 | 45 | 0 | 0 | 39 | 36 | 25 | 0 |
| Average | | 71 | 49 | 2.27 | 3.97 | | 71 | 29 | 0 | 0 | 49 | 34 | 16 | 1 |
| **Organization** | | **Accuracy (%)** | | **Kullback-Leibler** | | **Expected Score** | **Autoscore (%)** | | | | **BERT (%)** | | | |
| **Grade** | **N** | **AS** | **BERT** | **AS** | **BERT** | | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 3 | 150 | 74 | 40 | 1.82 | 4.52 | 1 | 74 | 26 | 0 | 0 | 40 | 50 | 4 | 6 |
| 4 | 150 | 100 | 63 | 0.00 | 2.78 | 1 | 100 | 0 | 0 | 0 | 63 | 37 | 0 | 0 |
| 5 | 150 | 79 | 33 | 1.39 | 5.57 | 1 | 79 | 21 | 0 | 0 | 33 | 67 | 0 | 0 |
| 6 | 150 | 55 | 0 | 3.49 | 8.41 | 1 | 55 | 45 | 0 | 0 | 0 | 35 | 61 | 4 |
| 7 | 150 | 61 | 10 | 2.89 | 7.37 | 1 | 61 | 39 | 0 | 0 | 10 | 33 | 57 | 0 |
| 8 | 150 | 100 | 7 | 0.00 | 7.90 | 1 | 100 | 0 | 0 | 0 | 7 | 76 | 17 | 0 |
| Average | | 78 | 25 | 1.60 | 6.09 | | 78 | 22 | 0 | 0 | 25 | 50 | 23 | 2 |

### Length Percentile Bin Results

When the same statistics are reviewed within length percentile bins and across items (Table 13), we observe increased agreement in Conventions for both engines as length percentile bins increase, and decreased agreement in the other two traits as the length percentile bins increase. Unlike the Shuffled essays, BERT tends to show a sharper decrease in performance as length percentile bins increase for Evidence & Elaboration and for Purpose, Focus, & Organization. The Kullback-Leibler measure increases for both Autoscore and BERT across the length percentile bins for Evidence & Elaboration and for Purpose, Focus, & Organization, but the opposite is true for conventions. Autoscore assigned no non-adjacent scores in any dimension and length percentile bin, and BERT assigned non-adjacent scores as length percentile bins increased for Evidence & Elaboration and for Purpose, Focus, & Organization. The magnitude of non-adjacent scores (around 15–17%) stayed constant across the length percentile bins for BERT.

*Table 13. Autoscore and BERT Predictions on Babel Essays, by Length Percentile Bins*

| Conventions | | Accuracy (%) | | Kullback-Leibler | | Expected Score | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| % ile | N | AS | BERT | AS | BERT | | 0 | 1 | 2 | | 0 | 1 | 2 | |
| 10th | 180 | 53 | 54 | 3.63 | 3.26 | 0 | 1 | 47 | 53 | | 16 | 30 | 54 | |
| 25th | 180 | 49 | 63 | 3.96 | 2.51 | 0 | 0 | 51 | 49 | | 17 | 21 | 63 | |
| 50th | 180 | 58 | 65 | 3.16 | 2.34 | 0 | 0 | 42 | 58 | | 16 | 19 | 65 | |
| 75th | 180 | 62 | 67 | 2.86 | 2.16 | 0 | 0 | 38 | 62 | | 15 | 18 | 67 | |
| 90th | 180 | 63 | 71 | 2.77 | 1.90 | 0 | 0 | 37 | 63 | | 14 | 15 | 71 | |
| Average | | 57 | 64 | 3.28 | 2.43 | | 0 | 43 | 57 | | 16 | 20 | 64 | |
| Elaboration | | Accuracy (%) | | Kullback-Leibler | | Expected Score | Autoscore (%) | | | | BERT (%) | | | |
| % ile | N | AS | BERT | AS | BERT | | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 10th | 180 | 90 | 81 | 0.60 | 1.30 | 1 | 90 | 10 | 0 | 0 | 81 | 19 | 0 | 0 |
| 25th | 180 | 91 | 61 | 0.56 | 2.88 | 1 | 91 | 09 | 0 | 0 | 61 | 38 | 01 | 0 |
| 50th | 180 | 76 | 48 | 1.70 | 3.77 | 1 | 76 | 24 | 0 | 0 | 48 | 32 | 21 | 0 |
| 75th | 180 | 52 | 32 | 3.71 | 5.07 | 1 | 52 | 48 | 0 | 0 | 32 | 36 | 29 | 2 |
| 90th | 180 | 47 | 24 | 4.22 | 5.90 | 1 | 47 | 53 | 0 | 0 | 24 | 46 | 29 | 1 |
| Average | | 71 | 49 | 2.16 | 3.78 | | 71 | 29 | 0 | 0 | 49 | 34 | 16 | 1 |
| Organization | | Accuracy (%) | | Kullback-Leibler | | Expected Score | Autoscore (%) | | | | BERT (%) | | | |
| % ile | N | AS | BERT | AS | BERT | | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 10th | 180 | 86 | 63 | 0.88 | 2.72 | 1 | 86 | 14 | 0 | 0 | 63 | 37 | 0 | 0 |
| 25th | 180 | 92 | 42 | 0.48 | 4.48 | 1 | 92 | 08 | 0 | 0 | 42 | 53 | 4 | 0 |
| 50th | 180 | 83 | 13 | 1.08 | 7.08 | 1 | 83 | 17 | 0 | 0 | 13 | 56 | 31 | 0 |
| 75th | 180 | 66 | 6 | 2.48 | 7.72 | 1 | 66 | 34 | 0 | 0 | 06 | 52 | 39 | 3 |
| 90th | 180 | 64 | 3 | 2.67 | 7.99 | 1 | 64 | 36 | 0 | 0 | 03 | 51 | 42 | 5 |
| Average | | 78 | 25 | 1.52 | 6.00 | | 78 | 22 | 0 | 0 | 25 | 50 | 23 | 2 |

These results suggest the grammar features and reliance on an item-specific training sample may be sufficient to make Autoscore robust to Babel essays for some items; however, Autoscore shows variability in performance and sensitivity to essay length. These results also suggest that, outside of Conventions on some items, the BERT model is not robust to Babel essays, particularly for longer essays. This may be due at least in part to the use of word embeddings built on item-agnostic corpora; Babel words may be mapped to points in the word embedding that are influencing score in ways that Autoscore (via the LSA) does not.

### Off-Topic Essays

For the Off-Topic essays, the expected score for the Conventions trait was the same as the score the raters assigned when scoring against the correct prompt and the expected score for each of the Evidence & Elaboration and Purpose, Focus, & Organization trait was 1.

### Item-Level Results

For the Off-Topic essays, we analyze the Conventions scores differently than the other two trait scores. For Purpose, Focus, & Organization, Autoscore and BERT showed different patterns of agreement and score distributions. Autoscore performance mimicked its performance for Evidence & Elaboration. BERT showed poor performance across all items in this trait, with lower levels of agreement with the expected score and more non-adjacent scores (Table 14).

For Conventions, we can compare the agreement results relative to those on the original sample. Both Autoscore and BERT showed lower overall agreement with the expected score relative to the original validation sample (61% vs. 71% for Autoscore and 68% vs. 76% for BERT). Autoscore and BERT performed similar to the original sample for most items, with BERT trending closer to the original agreements. Both Autoscore and BERT showed poor agreement for the grade 4 and grade 7 items. BERT showed less divergence with the expected score distribution than Autoscore.

For Evidence & Elaboration, BERT and Autoscore showed strong agreement with the expected score for grades 3 through 5 and

*Table 14. Autoscore and BERT Predictions on Off-Topic Essays, by Item*

| Conventions | | Accuracy (%) | | Kullback-Leibler | | Expected Score (%) | | | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Grade** | **N** | **AS** | **BERT** | **AS** | **BERT** | **0** | **1** | **2** | **0** | **1** | **2** | | **0** | **1** | **2** | |
| 3 | 150 | 52 | 72 | 0.26 | 0.04 | 5 | 35 | 61 | 10 | 65 | 25 | | 7 | 45 | 47 | |
| 4 | 150 | 33 | 30 | 0.60 | 0.55 | 9 | 45 | 46 | 26 | 73 | 1 | | 37 | 57 | 5 | |
| 5 | 150 | 66 | 83 | 0.11 | 0.01 | 2 | 25 | 73 | 2 | 47 | 51 | | 1 | 19 | 79 | |
| 6 | 150 | 67 | 72 | 0.02 | 0.03 | 7 | 41 | 53 | 6 | 32 | 62 | | 5 | 30 | 65 | |
| 7 | 150 | 63 | 69 | 0.12 | 0.14 | 9 | 36 | 55 | 3 | 19 | 78 | | 3 | 16 | 81 | |
| 8 | 150 | 87 | 83 | 0.07 | 0.03 | 3 | 20 | 77 | 3 | 7 | 90 | | 3 | 11 | 85 | |
| **Averages** | | 61 | 68 | 0.20 | 0.13 | 6 | 34 | 61 | 8 | 41 | 51 | | 10 | 30 | 61 | |
| **Elaboration** | | **Accuracy (%)** | | **Kullback-Leibler** | | **Expected Score (%)** | | | **Autoscore (%)** | | | | **BERT (%)** | | | |
| **Grade** | **N** | **AS** | **BERT** | **AS** | **BERT** | **1** | | | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 3 | 150 | 100 | 85 | 0.00 | 0.98 | 100 | | | 100 | 0 | 0 | 0 | 85 | 15 | 0 | 0 |
| 4 | 150 | 97 | 85 | 0.12 | 0.98 | 100 | | | 97 | 3 | 0 | 0 | 85 | 15 | 0 | 0 |
| 5 | 150 | 99 | 83 | 0.05 | 1.06 | 100 | | | 99 | 1 | 0 | 0 | 83 | 16 | 1 | 0 |
| 6 | 150 | 65 | 27 | 2.57 | 5.74 | 100 | | | 65 | 35 | 1 | 0 | 27 | 53 | 21 | 0 |
| 7 | 150 | 33 | 29 | 5.57 | 5.54 | 100 | | | 33 | 67 | 0 | 0 | 29 | 56 | 15 | 0 |
| 8 | 150 | 86 | 76 | 0.86 | 1.59 | 100 | | | 86 | 13 | 1 | 0 | 76 | 22 | 2 | 0 |
| **Averages** | | 80 | 64 | 1.53 | 2.65 | | | | 80 | 20 | 0 | 0 | 64 | 30 | 6 | 0 |
| **Organization** | | **Accuracy (%)** | | **Kullback-Leibler** | | **Expected Score (%)** | | | **Autoscore (%)** | | | | **BERT (%)** | | | |
| **Grade** | **N** | **AS** | **BERT** | **AS** | **BERT** | **1** | | | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 3 | 150 | 97 | 69 | 0.16 | 1.99 | 100 | | | 97 | 3 | 0 | 0 | 69 | 20 | 10 | 1 |
| 4 | 150 | 94 | 61 | 0.33 | 2.76 | 100 | | | 94 | 6 | 0 | 0 | 61 | 32 | 7 | 0 |
| 5 | 150 | 94 | 45 | 0.33 | 4.28 | 100 | | | 94 | 6 | 0 | 0 | 45 | 53 | 1 | 0 |
| 6 | 150 | 55 | 9 | 3.45 | 7.37 | 100 | | | 55 | 45 | 1 | 0 | 9 | 47 | 41 | 3 |
| 7 | 150 | 65 | 19 | 2.51 | 6.47 | 100 | | | 65 | 34 | 1 | 0 | 19 | 51 | 31 | 0 |
| 8 | 150 | 97 | 34 | 0.12 | 5.35 | 100 | | | 97 | 3 | 0 | 0 | 34 | 64 | 2 | 0 |
| **Averages** | | 84 | 39 | 1.15 | 4.70 | | | | 84 | 16 | 0 | 0 | 39 | 45 | 15 | 1 |

grade 8. Autoscore showed higher levels of agreement than BERT. BERT performed poorly for grades 6 and 7 and Autoscore performed poorly for grade 7. When Autoscore did not agree with the expected score, it almost always assigned an adjacent score. This was also true for BERT, except in grades 6 and 7.

For Purpose, Focus, & Organization, Autoscore and BERT showed different patterns of agreement and score distributions. Autoscore performance mimicked its performance for Evidence & Elaboration. BERT showed poor performance across all items in this trait, with lower levels of agreement with the expected score and more non-adjacent scores.

### Length Percentile Bin Results

When the same statistics are reviewed within length percentile bins across items (Table 15), BERT shows slightly higher agreement with the expected score compared to Autoscore and shows slightly better distributions across the length percentile bins in Conventions. Both Autoscore and BERT have decreasing percentages of non-adjacent scores in Conventions across the length percentile bins.

Autoscore shows good levels of agreement with the expected score across the length percentile bins for Evidence & Elaboration and for Purpose, Focus, & Organization. BERT shows lower levels of agreement compared to Autoscore and decreasing agreement as length percentiles increase. In addition, BERT shows poor agreement across the length percentile bins for Purpose, Focus, & Organization. Finally, Autoscore has very few non-adjacent scores relative to the expected score for Evidence & Elaboration and for Purpose, Focus, & Organization, while BERT increased from 4% to 8% in Evidence & Elaboration across the length percentile bins and varied between 12% to 18% in Purpose, Focus, & Organization.

These results suggest the grammar features for Conventions modelling and reliance on an item-specific training sample may be sufficient to make Autoscore robust to Off-Topic essays for some—but not all—items; this result suggests that the vocabulary used in the LSA has overlap with vocabularies elicited by other items. Also, Autoscore did not score the Off-Topic Conventions responses well relative to their original scores, suggesting that the prompt-specific approach used to build the Conventions models may not generalize well to other items. In general, outside of Conventions on some items, the BERT model is not robust to Off-Topic essays. As with the Babel essays, this result may be due to the use of word embeddings built on item-agnostic corpora. Finally, length did not seem to impact Autoscore behavior but did appear to affect BERT to a small degree.

### Duplicated Essays

For the Duplicated essays, the expected score for each trait was the same as the score the raters assigned when scoring using the unmodified essay.

### Item-Level Results

For the Duplicated essays, the results varied by engine and trait (Table 16). BERT and Autoscore showed good agreement with the expected scores for Conventions relative to the agreements in the Original essays (69% vs. 71% for Autoscore and 78% vs. 76% for BERT). BERT showed good similarity to the expected score point distribution as well, with KL values near 0; Autoscore showed more divergence with the expected score distribution, particularly for grades 6 through 8.

Both Autoscore and BERT showed relatively low agreement in the other two traits compared to the agreements observed in the Original essay sample. In Evidence & Elaboration, Autoscore average agreement on the Duplicated essays was 31% vs. 64% on the Original essays, and BERT average agreement on the Duplicated essays was 56% vs. 69% on the Original essays. In Purpose, Focus, & Organization, Autoscore average agreement on the Duplicated essays was 46% vs. 63% on the Original essays, and BERT average agreement on the Duplicated essays was 46% vs. 68% on the Original essays. The averaged Kullback-Leibler values showed slightly better distributions

*Table 15. Autoscore and BERT Predictions on Off-Topic Essays, by Length Percentile Bins*

| Conventions | | Accuracy (%) | | Kullback-Leibler | | Expected Score (%) | | | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **% ile** | **N** | **AS** | **BERT** | **AS** | **BERT** | **0** | **1** | **2** | **0** | **1** | **2** | | **0** | **1** | **2** | |
| 10th | 180 | 57 | 64 | 0.02 | 0.03 | 8 | 39 | 52 | 14 | 39 | 47 | | 15 | 31 | 54 | |
| 25th | 180 | 58 | 69 | 0.05 | 0.01 | 9 | 25 | 66 | 8 | 39 | 52 | | 11 | 31 | 59 | |
| 50th | 180 | 67 | 71 | 0.04 | 0.07 | 3 | 39 | 58 | 8 | 42 | 50 | | 10 | 29 | 61 | |
| 75th | 180 | 59 | 66 | 0.01 | 0.02 | 5 | 36 | 59 | 9 | 37 | 54 | | 8 | 29 | 63 | |
| 90th | 180 | 65 | 71 | 0.07 | 0.00 | 3 | 28 | 69 | 2 | 45 | 53 | | 4 | 31 | 65 | |
| **Averages** | | 61 | 68 | 0.04 | 0.03 | 6 | 34 | 61 | 8 | 41 | 51 | | 10 | 30 | 61 | |
| **Elaboration** | | **Accuracy (%)** | | **Kullback-Leibler** | | **Expected Score (%)** | | | **Autoscore (%)** | | | | **BERT (%)** | | | |
| **% ile** | **N** | **AS** | **BERT** | **AS** | **BERT** | **1** | | | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 10th | 180 | 82 | 71 | 1.15 | 1.97 | 100 | | | 82 | 17 | 1 | 0 | 71 | 24 | 5 | 0 |
| 25th | 180 | 83 | 69 | 1.06 | 2.12 | 100 | | | 83 | 16 | 1 | 0 | 69 | 27 | 4 | 0 |
| 50th | 180 | 81 | 63 | 1.26 | 2.55 | 100 | | | 81 | 19 | 0 | 0 | 63 | 30 | 7 | 0 |
| 75th | 180 | 77 | 62 | 1.61 | 2.68 | 100 | | | 77 | 23 | 0 | 0 | 62 | 31 | 7 | 0 |
| 90th | 180 | 76 | 56 | 1.65 | 3.14 | 100 | | | 76 | 24 | 0 | 0 | 56 | 36 | 8 | 0 |
| **Averages** | | 80 | 64 | 1.34 | 2.49 | | | | 80 | 20 | 0 | 0 | 64 | 30 | 6 | 0 |
| **Organization** | | **Accuracy (%)** | | **Kullback-Leibler** | | **Expected Score (%)** | | | **Autoscore (%)** | | | | **BERT (%)** | | | |
| **% ile** | **N** | **AS** | **BERT** | **AS** | **BERT** | **1** | | | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 10th | 180 | 83 | 53 | 1.13 | 3.32 | 100 | | | 83 | 17 | 0 | 0 | 53 | 33 | 13 | 2 |
| 25th | 180 | 88 | 44 | 0.72 | 4.05 | 100 | | | 88 | 11 | 1 | 0 | 44 | 39 | 16 | 1 |
| 50th | 180 | 85 | 39 | 0.96 | 4.60 | 100 | | | 85 | 15 | 0 | 0 | 39 | 48 | 12 | 0 |
| 75th | 180 | 81 | 31 | 1.26 | 5.32 | 100 | | | 81 | 19 | 0 | 0 | 31 | 51 | 18 | 0 |
| 90th | 180 | 82 | 29 | 1.21 | 5.46 | 100 | | | 82 | 18 | 0 | 0 | 29 | 52 | 18 | 1 |
| **Averages** | | 84 | 39 | 1.05 | 4.55 | | | | 84 | 16 | 0 | 0 | 39 | 45 | 15 | 1 |

for BERT in Conventions and Evidence & Elaboration compared to Autoscore, and similar distributions as Autoscore for Purpose, Focus, & Organization.

Reviewing the score point distributions for Autoscore and BERT relative to the expected score distributions, we see generally higher scores for Autoscore and BERT for each trait across the items with Autoscore assigning proportionally many more scores at the higher score range than BERT.

**Length Percentile Bin Results**

When the same statistics are reviewed within length percentile bins across items (Table 17), BERT shows better performance compared to Autoscore across all traits. With the exception of the 10th percentile bin, BERT showed good performance in Conventions in the length percentile bins with respect to agreement and divergence. Neither BERT nor Autoscore showed strong agreement in the other two traits across the length percentile bins. Interestingly, Autoscore showed worse performance as length

*Table 16. Autoscore and BERT Predictions on Duplicated Essays, by Item*

| Conventions | | Accuracy (%) | | Kullback-Leibler | | Expected Score (%) | | | Autoscore (%) | | | BERT (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Grade** | **N** | **AS** | **BERT** | **AS** | **BERT** | **0** | **1** | **2** | **0** | **1** | **2** | **0** | **1** | **2** |
| 3 | 150 | 65 | 72 | 0.04 | 0.03 | 20 | 37 | 43 | 11 | 37 | 53 | 11 | 44 | 45 |
| 4 | 150 | 59 | 78 | 0.05 | 0.00 | 19 | 55 | 26 | 20 | 41 | 39 | 16 | 57 | 27 |
| 5 | 150 | 76 | 79 | 0.07 | 0.02 | 2 | 29 | 69 | 2 | 13 | 85 | 1 | 21 | 77 |
| 6 | 150 | 63 | 71 | 0.23 | 0.03 | 4 | 43 | 53 | 3 | 12 | 85 | 3 | 31 | 66 |
| 7 | 150 | 80 | 85 | 0.14 | 0.03 | 5 | 19 | 77 | 3 | 2 | 95 | 3 | 11 | 87 |
| 8 | 150 | 71 | 81 | 0.23 | 0.05 | 3 | 27 | 69 | 2 | 3 | 95 | 3 | 14 | 83 |
| **Averages** | | 69 | 78 | 0.13 | 0.03 | 9 | 35 | 56 | 7 | 18 | 75 | 6 | 30 | 64 |

| Elaboration | | Accuracy (%) | | Kullback-Leibler | | Expected Score (%) | | | | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Grade** | **N** | **AS** | **BERT** | **AS** | **BERT** | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 3 | 150 | 47 | 58 | 0.33 | 0.15 | 31 | 47 | 19 | 2 | 15 | 45 | 20 | 20 | 16 | 44 | 40 | 0 |
| 4 | 150 | 38 | 60 | 0.54 | 0.14 | 33 | 48 | 17 | 1 | 13 | 37 | 27 | 21 | 13 | 57 | 30 | 0 |
| 5 | 150 | 44 | 53 | 0.18 | 0.20 | 27 | 48 | 19 | 6 | 12 | 39 | 31 | 19 | 14 | 45 | 41 | 0 |
| 6 | 150 | 15 | 38 | 0.97 | 0.41 | 12 | 43 | 40 | 5 | 1 | 14 | 31 | 53 | 1 | 15 | 62 | 22 |
| 7 | 150 | 21 | 67 | 2.35 | 0.34 | 24 | 64 | 12 | 0 | 5 | 32 | 35 | 29 | 11 | 65 | 19 | 5 |
| 8 | 150 | 24 | 58 | 1.30 | 0.17 | 30 | 55 | 14 | 1 | 8 | 35 | 19 | 37 | 25 | 38 | 37 | 0 |
| **Averages** | | 31 | 56 | 0.94 | 0.24 | 26 | 51 | 20 | 3 | 9 | 34 | 27 | 30 | 13 | 44 | 38 | 5 |

| Organization | | Accuracy (%) | | Kullback-Leibler | | Expected Score (%) | | | | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Grade** | **N** | **AS** | **BERT** | **AS** | **BERT** | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 3 | 150 | 52 | 42 | 0.38 | 0.75 | 28 | 53 | 17 | 1 | 19 | 45 | 16 | 20 | 10 | 45 | 14 | 31 |
| 4 | 150 | 66 | 58 | 0.06 | 0.08 | 38 | 45 | 15 | 2 | 33 | 39 | 20 | 8 | 24 | 45 | 27 | 5 |
| 5 | 150 | 66 | 45 | 0.03 | 0.28 | 15 | 57 | 25 | 3 | 9 | 54 | 30 | 7 | 3 | 36 | 59 | 3 |
| 6 | 150 | 33 | 45 | 0.48 | 0.29 | 11 | 45 | 36 | 8 | 3 | 19 | 37 | 40 | 0 | 22 | 61 | 17 |
| 7 | 150 | 25 | 41 | 0.68 | 0.57 | 16 | 51 | 29 | 3 | 3 | 25 | 36 | 35 | 3 | 17 | 80 | 0 |
| 8 | 150 | 35 | 45 | 0.59 | 0.30 | 25 | 46 | 27 | 1 | 8 | 39 | 27 | 25 | 5 | 35 | 60 | 0 |
| **Averages** | | 46 | 46 | 0.37 | 0.38 | 22 | 49 | 25 | 3 | 13 | 37 | 28 | 23 | 7 | 33 | 50 | 9 |

percentile bins increased, while BERT did not exhibit a trend or exhibited the opposite trend. With respect to score distributions, Autoscore showed more of a bias upward in score as the length percentile bins increased. BERT showed upward bias as well, but it is less severe.

In Autoscore, word length does not influence the grammar and textual features, but it does affect the LSA features. These features are clearly influenced by the duplicated text and this influence appears to be magnified with longer essays. While BERT is also affected to some degree, the impact of duplication appears less severe with BERT. Also, BERT performed well on the Conventions items. This suggests that the predictive weights may be more impacted by the presence of a word than by the frequency of the word.

*Table 17. Autoscore and BERT Predictions on Duplicated Essays, by Length Percentile Bin*

| Conventions | | Accuracy (%) | | Kullback-Leibler | | Expected Score (%) | | | | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| % ile | N | AS | BERT | AS | BERT | 0 | 1 | 2 | | 0 | 1 | 2 | | 0 | 1 | 2 | |
| 10th | 180 | 57 | 67 | 0.07 | 0.01 | 19 | 44 | 37 | | 16 | 29 | 54 | | 14 | 43 | 43 | |
| 25th | 180 | 66 | 76 | 0.07 | 0.03 | 12 | 35 | 53 | | 7 | 22 | 71 | | 6 | 31 | 63 | |
| 50th | 180 | 69 | 80 | 0.12 | 0.02 | 7 | 38 | 56 | | 4 | 17 | 79 | | 4 | 30 | 66 | |
| 75th | 180 | 78 | 79 | 0.06 | 0.00 | 4 | 28 | 68 | | 3 | 14 | 83 | | 3 | 24 | 72 | |
| 90th | 180 | 73 | 86 | 0.15 | 0.02 | 3 | 30 | 67 | | 3 | 08 | 89 | | 3 | 21 | 77 | |
| **Averages** | | 69 | 78 | 0.09 | 0.02 | 9 | 35 | 56 | | 7 | 18 | 75 | | 6 | 30 | 64 | |
| **Elaboration** | | **Accuracy (%)** | | **Kullback-Leibler** | | **Expected Score (%)** | | | | **Autoscore (%)** | | | | **BERT (%)** | | | |
| % ile | N | AS | BERT | AS | BERT | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 10th | 180 | 41 | 49 | 1.04 | 0.96 | 64 | 36 | 0 | 0 | 21 | 66 | 12 | 1 | 32 | 54 | 13 | 0 |
| 25th | 180 | 42 | 52 | 0.70 | 0.43 | 33 | 63 | 4 | 0 | 8 | 52 | 38 | 2 | 12 | 62 | 22 | 3 |
| 50th | 180 | 31 | 53 | 0.96 | 0.27 | 17 | 63 | 19 | 1 | 7 | 29 | 37 | 26 | 10 | 44 | 37 | 8 |
| 75th | 180 | 23 | 66 | 1.23 | 0.10 | 09 | 48 | 40 | 2 | 4 | 15 | 32 | 48 | 6 | 30 | 58 | 6 |
| 90th | 180 | 20 | 60 | 1.17 | 0.10 | 08 | 44 | 38 | 10 | 6 | 6 | 17 | 71 | 5 | 30 | 60 | 5 |
| **Averages** | | 31 | 56 | 1.02 | 0.37 | 26 | 51 | 20 | 3 | 9 | 34 | 27 | 30 | 13 | 44 | 38 | 5 |
| **Organization** | | **Accuracy (%)** | | **Kullback-Leibler** | | **Expected Score (%)** | | | | **Autoscore (%)** | | | | **BERT (%)** | | | |
| % ile | N | AS | BERT | AS | BERT | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 10th | 180 | 61 | 51 | 0.22 | 0.50 | 56 | 43 | 1 | 0 | 29 | 62 | 9 | 0 | 23 | 58 | 19 | 0 |
| 25th | 180 | 55 | 34 | 0.37 | 0.93 | 26 | 69 | 6 | 0 | 13 | 54 | 32 | 1 | 07 | 38 | 54 | 0 |
| 50th | 180 | 43 | 36 | 0.49 | 0.60 | 15 | 67 | 17 | 1 | 9 | 36 | 43 | 12 | 2 | 31 | 63 | 4 |
| 75th | 180 | 38 | 58 | 0.68 | 0.22 | 8 | 35 | 53 | 3 | 6 | 19 | 36 | 38 | 2 | 20 | 61 | 17 |
| 90th | 180 | 35 | 50 | 0.70 | 0.11 | 7 | 33 | 48 | 12 | 7 | 12 | 19 | 62 | 2 | 19 | 53 | 25 |
| **Averages** | | 46 | 46 | 0.49 | 0.47 | 22 | 49 | 25 | 3 | 13 | 37 | 28 | 23 | 7 | 33 | 50 | 9 |

## Discussion

The purpose of this study was to compare results from a classical essay scoring approach (Autoscore) to a deep learning approach, BERT. The gaming responses considered were: shuffled text (to examine the assumption that word order is learned by neural network models); grammatically correct but nonsense essays (to examine the extent to which grammar is contributing to score); Off-Topic essays (to examine the extent to which essay meaning is contributing to score); and Duplicated essays (to examine the extent to which length is contributing to score, controlling for meaning). Gaming responses were crafted to reflect essay length distributions at the 10th, 25th, 50th, 75th, and 90th percentiles. Six items (one item each for grades 3 through 8) were examined.

Both engines showed adequate performance on engine validation samples, with BERT performing better than Autoscore in exact agreement and QWK, and slightly worse performance on SMD. If filters were used with Autoscore, all Babel essays and duplicate text essays would be captured (and routed for human verification by teachers). For the Off-Topic and Shuffled essays, most responses were assigned the lowest score or low confidence score that would have been routed for human Scoring; however, some items and responses would have received higher scores than desirable from the engine and would not have been routed for review. These results suggest that a more rigorous off-topic filter may be of value to Autoscore. Devising a “shuffled word” filter may not be of value to the program given that these types of responses are very rare.

With regard to performance of the engines without the filters, Autoscore and BERT generally performed poorly on the Shuffled essays and performed worse as essay lengths increased. Autoscore generally outperformed BERT on the Babel essays, although both engines were sensitive to length. Autoscore also outperformed BERT on the Off-Topic essays. Autoscore tended to have higher agreement with the exact score and give scores that were adjacent to the expected score when they were off for the Shuffled, Babel, and Off-Topic essays. BERT outperformed Autoscore on the duplicate text responses, showing less upward bias in its predictions and greater agreement with the expected score.

Thus, BERT’s performance on the gaming essays suggests that the models, at least as implemented in automated essay scoring, are not robust to gaming. This result could be due to the use of word embeddings built upon a very large corpus of text beyond what is typically used in an essay and/or could be due to the dampened use of positional embeddings which are summed with the word embeddings as entries into the model. That BERT performance was impacted by essay length suggests that—at least implicitly—the number of words in the essay is captured in the model, presumably by the contribution of more word embedding vectors as input in longer essays. The fact that BERT performed better than Autoscore on the Duplicated essays suggests that perhaps the contribution of different word embedding vectors may be more important to the model than the same word embedding vector summed with different positional embedding vectors.

This study has three limitations. One limitation is the use of artificially-generated responses, which may not behave like actual examinee responses. A second limitation is that the expected scores were assumed rather than derived from actual human rater scoring, particularly around Off-Topic and Duplicated essays. It is possible that the actual scores assigned by trained raters scoring the item (in the case of Off-Topic essays) or the response (in the case of Duplicated essays) might differ. The third limitation is that the BERT engine was used essentially out of the box, and changes in hyperparameters or the output modelling structure may improve performance on engine on the gaming responses.

The results of this study indicate that deep learning methods like BERT show promise in automated essay scoring because they offer improved overall performance; however, these methods still require unusual and gaming response filters to be used in operational scoring. More work needs to be done to investigate different variations of transformer networks and, just as importantly, to understand how networks can be designed to be robust to gaming and aberrant responses if they are truly to be “language models.”

## References

Alomar, J. (2018). The Illustrated Transformer. http://jalammar.github.io/illustrated-transformer/

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.

Burkhardt, A., & Lottridge, S. (2013, October). *Examining the Impact of Training Samples on Identifying Off-topic Responses in Automated Essay Scoring.* Paper presented at the annual meeting of Northern Rocky Mountain Educational Research Association, Jackson, WY.

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.

Cahill, A., Chodorow, M., & Flor, M. (2018). Developing an e-rater advisory to detect Babel-generated essays. *Journal of Writing Analytics,* 2, 203-224.

Chiu, C-C, Lawson, D., Luo, Y., Tucker, G., Swersky, K., Sutskever, I., & Jaitly, N. (2017). *An Online Sequence-to-Sequence Model for Noisy Speech Recognition.* https://arxiv.org/abs/1706.06428.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Y Harshman, R (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science. 41* (6), 391–407.

Devlin, J., Chang, M-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805.

Esteva, A., Kuprel, B., Novoa, RA., Ko, J., Swetter, SM., Blau, HM., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature,* 542, 115-118.

Henderson, G., & Andrade, A. (2019, April). *Gamification of automated scoring engines.* Paper presented at the National Council on Measurement in Education (NCME). Toronto, CA.

Hewlett Foundation (2012). *Automated Scoring Assessment Prize, Phase 1: Automated Essay Scoring.* https://www.kaggle.com/c/asap-aes.

Hewlett Foundation (2013). *Automated Scoring Assessment Prize, Phase 2: Short Answer Scoring.* https://www.kaggle.com/c/asap-sas.

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering,* 12(02), 145. doi: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(4), 36-46.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9 (8),* 1735-1780.

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

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

Louis, A., & Higgins, D. (2010). Off-topic essay detection using short prompt texts. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 92–95). Los Angeles: Association for Computational Linguistics.

McGraw-Hill Education CTB (2014, December 24). *Smarter Balanced Assessment Consortium Field Test: Automated Scoring Research Studies (in accordance with Smarter Balanced RFP 17)*. Retrieved from: http://www.smarterapp.org/documents/FieldTest_AutomatedScoringResearchStudies.pdf.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*, arXiv:1301.3781.

Page, E. B. (2003). Project Essay Grade: PEG. In *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates.

Pearson and ETS (2015, March 9). *Research Results of PARCC Automated Scoring Proof of Concept Study.* Retrieved from: http://www.parcconline.org/images/Resources/Educatorresources/PARCC_AI_Research_Report.pdf.

Perelman, L. (2014). When 'the state of the art' is counting words. *Assessing Writing, 2*, 104-111.

Pennington, J. S. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* , 1532-1543.

Powers, P. W., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping E-Rater: Challenging the validity of automated essay scoring. *ETS Research Report Series 01-03.*

Rodriguez, P., Jafari, A., & Ormerod, C. (2019). Language models and automated essay scoring. *arXiv preprint*, https://arxiv.org/abs/1909.09482.

Rush, A., (2018). The Annotated Transformer. Retrieved from: https://nlp.seas.harvard.edu/2018/04/03/attention.html on February 15, 2020.

Schnable, T., Labutov, I., Mimno, D., & Joachims, R. (2015, September). Evaluation methods for unsupervised word embeddings. In *Proceedings on Empirical Methods in Natural Language Processing*, 298-307, Lisbon, Portugal.

Shermis, M., & Lottridge, S. (2019, April). *Communicating to the Public About Machine Scoring: What Works, What Doesn't*. Paper presented at the National Conference on Measurement in Education, Toronto, CA.

Shermis, M. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53-76.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). *Rethinking the Inception of Architecture for Computer Vision*. Retrieved from https://www.cvfoundation.org/openaccess/content_cvpr_2016/papers/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L, Gomez, A, Kaiser, L, & Polushkin, I. (2017, June). *Attention is all you need*. Paper presented at the Neural Information Processing Systems (NIPS) Conference, Long Beach, CA.

Weiss, R., Chorowski, J., Navdeep, J., Wu, Y., & Chen, Z. (2017). *Sequence-to-Sequence Models Can Directly Translate Foreign Speech*. https://arxiv.org/abs/1703.08581.

Williams, R., Hinton, G., and Rumelhart, D, (1986). Learning representations by back-propagating errors. *Nature*, 323 (6088), 533-536.

Williamson, D., Xi, X., and Breyer, F.J. (2012). A framework for the evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31(1)*, 2-13.

Wood, S. (2017, April). *Media, the Public, and Automated Scoring*. Paper presented at the National Conference on Measurement in Education, San Antonio, TX.

Yoon, K., Denton, C., Hoang, L, and Rush, A. (2017) Structured attention networks. In *International Conference on Learning Representations*.

Zhang, M., Chen, J., & Chunyi, R. (2016). Evaluating the advisory flags and machine scoring difficulty in the e-rater automated scoring engine. *ETS Research Report Series*.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, 19-27.

# Appendix A: Shuffled Essay Results for Items and Length Percentile Bins

*Table 18. Autoscore and BERT Predictions on Shuffled Essays, by Length Percentile Bin in Conventions*

| | %-ile | Accuracy (%) | | Kullback-Leibler | | Expected | Autoscore (%) | | | BERT (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade | Bin | AS | BERT | AS | BERT | Score | 0 | 1 | 2 | 0 | 1 | 2 |
| 3 | 10th | 100 | 100 | 0.0 | 0.0 | 0 | 100 | 0 | 0 | 100 | 0 | 0 |
| 3 | 25th | 100 | 100 | 0.0 | 0.0 | 0 | 100 | 0 | 0 | 100 | 0 | 0 |
| 3 | 50th | 100 | 100 | 0.0 | 0.0 | 0 | 100 | 0 | 0 | 100 | 0 | 0 |
| 3 | 75th | 87 | 100 | 0.8 | 0.0 | 0 | 87 | 13 | 0 | 100 | 0 | 0 |
| 3 | 90th | 83 | 90 | 1.1 | 0.6 | 0 | 83 | 17 | 0 | 90 | 10 | 0 |
| 4 | 10th | 100 | 100 | 0.0 | 0.0 | 0 | 100 | 0 | 0 | 100 | 0 | 0 |
| 4 | 25th | 100 | 100 | 0.0 | 0.0 | 0 | 100 | 0 | 0 | 100 | 0 | 0 |
| 4 | 50th | 100 | 100 | 0.0 | 0.0 | 0 | 100 | 0 | 0 | 100 | 0 | 0 |
| 4 | 75th | 100 | 100 | 0.0 | 0.0 | 0 | 100 | 0 | 0 | 100 | 0 | 0 |
| 4 | 90th | 100 | 100 | 0.0 | 0.0 | 0 | 100 | 0 | 0 | 100 | 0 | 0 |
| 5 | 10th | 83 | 23 | 1.1 | 6.5 | 0 | 83 | 17 | 0 | 23 | 77 | 0 |
| 5 | 25th | 90 | 7 | 0.6 | 8.3 | 0 | 90 | 10 | 0 | 7 | 93 | 0 |
| 5 | 50th | 90 | 3 | 0.6 | 8.8 | 0 | 90 | 10 | 0 | 3 | 97 | 0 |
| 5 | 75th | 73 | 0 | 1.9 | 9.2 | 0 | 73 | 27 | 0 | 0 | 100 | 0 |
| 5 | 90th | 57 | 0 | 3.3 | 9.2 | 0 | 57 | 43 | 0 | 0 | 100 | 0 |
| 6 | 10th | 87 | 100 | 0.8 | 0.0 | 0 | 87 | 13 | 0 | 100 | 0 | 0 |
| 6 | 25th | 80 | 100 | 1.3 | 0.0 | 0 | 80 | 20 | 0 | 100 | 0 | 0 |
| 6 | 50th | 57 | 100 | 3.3 | 0.0 | 0 | 57 | 43 | 0 | 100 | 0 | 0 |
| 6 | 75th | 50 | 100 | 3.9 | 0.0 | 0 | 50 | 50 | 0 | 100 | 0 | 0 |
| 6 | 90th | 40 | 100 | 4.7 | 0.0 | 0 | 40 | 57 | 3 | 100 | 0 | 0 |
| 7 | 10th | 93 | 43 | 0.4 | 4.5 | 0 | 93 | 7 | 0 | 43 | 57 | 0 |
| 7 | 25th | 63 | 10 | 2.7 | 8.0 | 0 | 63 | 37 | 0 | 10 | 90 | 0 |
| 7 | 50th | 43 | 13 | 4.5 | 7.6 | 0 | 43 | 57 | 0 | 13 | 87 | 0 |
| 7 | 75th | 27 | 0 | 5.9 | 9.2 | 0 | 27 | 67 | 7 | 0 | 100 | 0 |
| 7 | 90th | 20 | 10 | 6.4 | 8.0 | 0 | 20 | 60 | 20 | 10 | 90 | 0 |
| 8 | 10th | 33 | 90 | 5.4 | 0.6 | 0 | 33 | 63 | 3 | 90 | 10 | 0 |
| 8 | 25th | 7 | 80 | 8.2 | 1.3 | 0 | 7 | 90 | 3 | 80 | 20 | 0 |
| 8 | 50th | 13 | 70 | 7.0 | 2.2 | 0 | 13 | 43 | 43 | 70 | 30 | 0 |
| 8 | 75th | 7 | 40 | 7.9 | 4.9 | 0 | 7 | 17 | 77 | 40 | 60 | 0 |
| 8 | 90th | 3 | 20 | 8.8 | 6.6 | 0 | 3 | 0 | 97 | 20 | 73 | 7 |

*Table 19. Autoscore and BERT Predictions on Shuffled Essays, by Length Percentile Bin in Elaboration*

|  | %-ile | Accuracy (%) |  | Kullback-Leibler |  | Expected | Autoscore (%) |  |  |  | BERT (%) |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Grade** | **Bin** | **AS** | **BERT** | **AS** | **BERT** | **Score** | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 3 | 10th | 83 | 87 | 1.1 | 0.8 | 1 | 83 | 17 | 0 | 0 | 87 | 13 | 0 | 0 |
| 3 | 25th | 60 | 57 | 3.0 | 3.3 | 1 | 60 | 40 | 0 | 0 | 57 | 43 | 0 | 0 |
| 3 | 50th | 70 | 40 | 2.2 | 4.9 | 1 | 70 | 30 | 0 | 0 | 40 | 60 | 0 | 0 |
| 3 | 75th | 30 | 20 | 5.5 | 6.5 | 1 | 30 | 60 | 10 | 0 | 20 | 67 | 13 | 0 |
| 3 | 90th | 7 | 10 | 7.7 | 7.3 | 1 | 7 | 60 | 33 | 0 | 10 | 50 | 40 | 0 |
| 4 | 10th | 97 | 90 | 0.2 | 0.6 | 1 | 97 | 3 | 0 | 0 | 90 | 10 | 0 | 0 |
| 4 | 25th | 80 | 90 | 1.3 | 0.6 | 1 | 80 | 20 | 0 | 0 | 90 | 10 | 0 | 0 |
| 4 | 50th | 50 | 83 | 3.9 | 1.1 | 1 | 50 | 50 | 0 | 0 | 83 | 17 | 0 | 0 |
| 4 | 75th | 23 | 73 | 6.5 | 1.9 | 1 | 23 | 77 | 0 | 0 | 73 | 27 | 0 | 0 |
| 4 | 90th | 20 | 63 | 6.6 | 2.7 | 1 | 20 | 73 | 7 | 0 | 63 | 37 | 0 | 0 |
| 5 | 10th | 100 | 100 | 0.0 | 0.0 | 1 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| 5 | 25th | 90 | 80 | 0.6 | 1.3 | 1 | 90 | 10 | 0 | 0 | 80 | 20 | 0 | 0 |
| 5 | 50th | 60 | 43 | 3.0 | 4.5 | 1 | 60 | 40 | 0 | 0 | 43 | 57 | 0 | 0 |
| 5 | 75th | 60 | 50 | 3.0 | 3.9 | 1 | 60 | 40 | 0 | 0 | 50 | 50 | 0 | 0 |
| 5 | 90th | 23 | 27 | 6.2 | 6.2 | 1 | 23 | 63 | 13 | 0 | 27 | 73 | 0 | 0 |
| 6 | 10th | 87 | 93 | 0.8 | 0.4 | 1 | 87 | 13 | 0 | 0 | 93 | 7 | 0 | 0 |
| 6 | 25th | 47 | 57 | 4.2 | 3.3 | 1 | 47 | 53 | 0 | 0 | 57 | 43 | 0 | 0 |
| 6 | 50th | 0 | 7 | 8.9 | 8.2 | 1 | 0 | 90 | 10 | 0 | 7 | 90 | 3 | 0 |
| 6 | 75th | 3 | 3 | 8.1 | 8.2 | 1 | 3 | 63 | 33 | 0 | 3 | 70 | 27 | 0 |
| 6 | 90th | 10 | 3 | 7.3 | 8.0 | 1 | 10 | 20 | 63 | 7 | 3 | 20 | 70 | 7 |
| 7 | 10th | 97 | 100 | 0.2 | 0.0 | 1 | 97 | 3 | 0 | 0 | 100 | 0 | 0 | 0 |
| 7 | 25th | 70 | 100 | 2.2 | 0.0 | 1 | 70 | 30 | 0 | 0 | 100 | 0 | 0 | 0 |
| 7 | 50th | 50 | 97 | 3.9 | 0.2 | 1 | 50 | 50 | 0 | 0 | 97 | 3 | 0 | 0 |
| 7 | 75th | 10 | 100 | 7.8 | 0.0 | 1 | 10 | 87 | 3 | 0 | 100 | 0 | 0 | 0 |
| 7 | 90th | 13 | 87 | 7.4 | 0.8 | 1 | 13 | 83 | 3 | 0 | 87 | 10 | 3 | 0 |
| 8 | 10th | 93 | 100 | 0.4 | 0.0 | 1 | 93 | 7 | 0 | 0 | 100 | 0 | 0 | 0 |
| 8 | 25th | 80 | 100 | 1.3 | 0.0 | 1 | 80 | 20 | 0 | 0 | 100 | 0 | 0 | 0 |
| 8 | 50th | 43 | 97 | 4.3 | 0.2 | 1 | 43 | 50 | 7 | 0 | 97 | 3 | 0 | 0 |
| 8 | 75th | 13 | 83 | 7.2 | 1.1 | 1 | 13 | 70 | 17 | 0 | 83 | 17 | 0 | 0 |
| 8 | 90th | 7 | 83 | 7.7 | 1.0 | 1 | 7 | 63 | 27 | 3 | 83 | 7 | 10 | 0 |

*Table 20. Autoscore and BERT Predictions on Shuffled Essays, by Length Percentile Bin in Organization*

| | %-ile | Accuracy (%) | | Kullback-Leibler | | Expected | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade | Bin | AS | BERT | AS | BERT | Score | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 3 | 10th | 90 | 100 | 0.6 | 0.0 | 1 | 90 | 10 | 0 | 0 | 100 | 0 | 0 | 0 |
| 3 | 25th | 90 | 100 | 0.6 | 0.0 | 1 | 90 | 10 | 0 | 0 | 100 | 0 | 0 | 0 |
| 3 | 50th | 80 | 100 | 1.3 | 0.0 | 1 | 80 | 20 | 0 | 0 | 100 | 0 | 0 | 0 |
| 3 | 75th | 43 | 100 | 4.5 | 0.0 | 1 | 43 | 57 | 0 | 0 | 100 | 0 | 0 | 0 |
| 3 | 90th | 7 | 97 | 8.3 | 0.2 | 1 | 7 | 93 | 0 | 0 | 97 | 3 | 0 | 0 |
| 4 | 10th | 100 | 100 | 0.0 | 0.0 | 1 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| 4 | 25th | 100 | 100 | 0.0 | 0.0 | 1 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| 4 | 50th | 87 | 100 | 0.8 | 0.0 | 1 | 87 | 13 | 0 | 0 | 100 | 0 | 0 | 0 |
| 4 | 75th | 93 | 97 | 0.4 | 0.2 | 1 | 93 | 7 | 0 | 0 | 97 | 3 | 0 | 0 |
| 4 | 90th | 87 | 100 | 0.8 | 0.0 | 1 | 87 | 13 | 0 | 0 | 100 | 0 | 0 | 0 |
| 5 | 10th | 93 | 67 | 0.4 | 2.4 | 1 | 93 | 7 | 0 | 0 | 67 | 33 | 0 | 0 |
| 5 | 25th | 77 | 7 | 1.6 | 8.3 | 1 | 77 | 23 | 0 | 0 | 7 | 93 | 0 | 0 |
| 5 | 50th | 63 | 3 | 2.7 | 8.6 | 1 | 63 | 37 | 0 | 0 | 3 | 93 | 3 | 0 |
| 5 | 75th | 37 | 0 | 5.2 | 8.7 | 1 | 37 | 63 | 0 | 0 | 0 | 80 | 20 | 0 |
| 5 | 90th | 30 | 0 | 5.8 | 8.5 | 1 | 30 | 70 | 0 | 0 | 0 | 60 | 40 | 0 |
| 6 | 10th | 97 | 97 | 0.2 | 0.2 | 1 | 97 | 3 | 0 | 0 | 97 | 3 | 0 | 0 |
| 6 | 25th | 70 | 90 | 2.2 | 0.6 | 1 | 70 | 30 | 0 | 0 | 90 | 10 | 0 | 0 |
| 6 | 50th | 30 | 40 | 5.8 | 4.9 | 1 | 30 | 70 | 0 | 0 | 40 | 60 | 0 | 0 |
| 6 | 75th | 10 | 10 | 8.0 | 8.0 | 1 | 10 | 90 | 0 | 0 | 10 | 90 | 0 | 0 |
| 6 | 90th | 13 | 7 | 7.0 | 8.3 | 1 | 13 | 57 | 30 | 0 | 7 | 93 | 0 | 0 |
| 7 | 10th | 97 | 100 | 0.2 | 0.0 | 1 | 97 | 3 | 0 | 0 | 100 | 0 | 0 | 0 |
| 7 | 25th | 83 | 100 | 1.1 | 0.0 | 1 | 83 | 17 | 0 | 0 | 100 | 0 | 0 | 0 |
| 7 | 50th | 37 | 100 | 5.2 | 0.0 | 1 | 37 | 63 | 0 | 0 | 100 | 0 | 0 | 0 |
| 7 | 75th | 10 | 90 | 8.0 | 0.5 | 1 | 10 | 90 | 0 | 0 | 90 | 7 | 3 | 0 |
| 7 | 90th | 17 | 73 | 6.7 | 1.8 | 1 | 17 | 57 | 27 | 0 | 73 | 23 | 3 | 0 |
| 8 | 10th | 100 | 73 | 0.0 | 1.9 | 1 | 100 | 0 | 0 | 0 | 73 | 27 | 0 | 0 |
| 8 | 25th | 97 | 20 | 0.2 | 6.9 | 1 | 97 | 3 | 0 | 0 | 20 | 80 | 0 | 0 |
| 8 | 50th | 57 | 10 | 3.3 | 7.8 | 1 | 57 | 43 | 0 | 0 | 10 | 87 | 3 | 0 |
| 8 | 75th | 27 | 3 | 6.0 | 8.1 | 1 | 27 | 70 | 3 | 0 | 3 | 33 | 63 | 0 |
| 8 | 90th | 13 | 0 | 7.4 | 8.8 | 1 | 13 | 80 | 7 | 0 | 0 | 13 | 87 | 0 |

## Appendix B: Babel Essay Results for Items and Length Percentile Bins

*Table 21. Autoscore and BERT Predictions on Babel Essays, by Length Percentile Bin in Conventions*

| | %-ile | Accuracy (%) | | Kullback-Leibler | | Expected | Autoscore (%) | | | BERT (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade | Bin | AS | BERT | AS | BERT | Score | 0 | 1 | 2 | 0 | 1 | 2 |
| 3 | 10th | 7 | 0 | 8.35 | 9.21 | 2 | 0 | 93 | 7 | 0 | 100 | 0 |
| 3 | 25th | 7 | 7 | 8.35 | 8.35 | 2 | 0 | 93 | 7 | 0 | 93 | 7 |
| 3 | 50th | 3 | 3 | 8.75 | 8.75 | 2 | 0 | 97 | 3 | 0 | 97 | 3 |
| 3 | 75th | 3 | 10 | 8.75 | 7.96 | 2 | 0 | 97 | 3 | 0 | 90 | 10 |
| 3 | 90th | 0 | 23 | 9.21 | 6.52 | 2 | 0 | 100 | 0 | 0 | 77 | 23 |
| 4 | 10th | 0 | 0 | 9.06 | 9.06 | 2 | 3 | 97 | 0 | 97 | 3 | 0 |
| 4 | 25th | 0 | 0 | 9.21 | 9.21 | 2 | 0 | 100 | 0 | 100 | 0 | 0 |
| 4 | 50th | 0 | 0 | 9.21 | 9.06 | 2 | 0 | 100 | 0 | 97 | 3 | 0 |
| 4 | 75th | 0 | 0 | 9.21 | 8.88 | 2 | 0 | 100 | 0 | 90 | 10 | 0 |
| 4 | 90th | 0 | 0 | 9.21 | 8.81 | 2 | 0 | 100 | 0 | 87 | 13 | 0 |
| 5 | 10th | 87 | 87 | 0.84 | 0.84 | 2 | 0 | 13 | 87 | 0 | 13 | 87 |
| 5 | 25th | 57 | 93 | 3.31 | 0.37 | 2 | 0 | 43 | 57 | 0 | 7 | 93 |
| 5 | 50th | 77 | 100 | 1.61 | 0.00 | 2 | 0 | 23 | 77 | 0 | 0 | 100 |
| 5 | 75th | 80 | 100 | 1.34 | 0.00 | 2 | 0 | 20 | 80 | 0 | 0 | 100 |
| 5 | 90th | 77 | 100 | 1.61 | 0.00 | 2 | 0 | 23 | 77 | 0 | 0 | 100 |
| 6 | 10th | 37 | 47 | 5.17 | 4.22 | 2 | 0 | 63 | 37 | 0 | 53 | 47 |
| 6 | 25th | 37 | 77 | 5.17 | 1.61 | 2 | 0 | 63 | 37 | 0 | 23 | 77 |
| 6 | 50th | 70 | 87 | 2.15 | 0.84 | 2 | 0 | 30 | 70 | 0 | 13 | 87 |
| 6 | 75th | 87 | 93 | 0.84 | 0.37 | 2 | 0 | 13 | 87 | 0 | 7 | 93 |
| 6 | 90th | 100 | 100 | 0.00 | 0.00 | 2 | 0 | 0 | 100 | 0 | 0 | 100 |
| 7 | 10th | 87 | 90 | 0.84 | 0.60 | 2 | 0 | 13 | 87 | 0 | 10 | 90 |
| 7 | 25th | 97 | 100 | 0.16 | 0.00 | 2 | 0 | 3 | 97 | 0 | 0 | 100 |
| 7 | 50th | 100 | 100 | 0.00 | 0.00 | 2 | 0 | 0 | 100 | 0 | 0 | 100 |
| 7 | 75th | 100 | 100 | 0.00 | 0.00 | 2 | 0 | 0 | 100 | 0 | 0 | 100 |
| 7 | 90th | 100 | 100 | 0.00 | 0.00 | 2 | 0 | 0 | 100 | 0 | 0 | 100 |
| 8 | 10th | 100 | 100 | 0.00 | 0.00 | 2 | 0 | 0 | 100 | 0 | 0 | 100 |
| 8 | 25th | 100 | 100 | 0.00 | 0.00 | 2 | 0 | 0 | 100 | 0 | 0 | 100 |
| 8 | 50th | 100 | 100 | 0.00 | 0.00 | 2 | 0 | 0 | 100 | 0 | 0 | 100 |
| 8 | 75th | 100 | 100 | 0.00 | 0.00 | 2 | 0 | 0 | 100 | 0 | 0 | 100 |
| 8 | 90th | 100 | 100 | 0.00 | 0.00 | 2 | 0 | 0 | 100 | 0 | 0 | 100 |

*Table 22. Autoscore and BERT Predictions on Babel Essays, by Length Percentile Bin in Elaboration*

| | | Accuracy (%) | | Kullback-Leibler | | | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade | %-ile Bin | AS | BERT | AS | BERT | Expected Score | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 3 | 10th | 100 | 100 | 0.00 | 0.00 | 1 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| 3 | 25th | 100 | 100 | 0.00 | 0.00 | 1 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| 3 | 50th | 100 | 90 | 0.00 | 0.60 | 1 | 100 | 0 | 0 | 0 | 90 | 10 | 0 | 0 |
| 3 | 75th | 100 | 67 | 0.00 | 2.43 | 1 | 100 | 0 | 0 | 0 | 67 | 33 | 0 | 0 |
| 3 | 90th | 97 | 0 | 0.16 | 9.21 | 1 | 97 | 3 | 0 | 0 | 0 | 100 | 0 | 0 |
| 4 | 10th | 100 | 100 | 0.00 | 0.00 | 1 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| 4 | 25th | 100 | 100 | 0.00 | 0.00 | 1 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| 4 | 50th | 100 | 100 | 0.00 | 0.00 | 1 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| 4 | 75th | 97 | 100 | 0.16 | 0.00 | 1 | 97 | 3 | 0 | 0 | 100 | 0 | 0 | 0 |
| 4 | 90th | 83 | 100 | 1.08 | 0.00 | 1 | 83 | 17 | 0 | 0 | 100 | 0 | 0 | 0 |
| 5 | 10th | 80 | 100 | 1.34 | 0.00 | 1 | 80 | 20 | 0 | 0 | 100 | 0 | 0 | 0 |
| 5 | 25th | 100 | 87 | 0.00 | 0.84 | 1 | 100 | 0 | 0 | 0 | 87 | 13 | 0 | 0 |
| 5 | 50th | 100 | 47 | 0.00 | 4.22 | 1 | 100 | 0 | 0 | 0 | 47 | 53 | 0 | 0 |
| 5 | 75th | 100 | 13 | 0.00 | 7.59 | 1 | 100 | 0 | 0 | 0 | 13 | 87 | 0 | 0 |
| 5 | 90th | 100 | 27 | 0.00 | 6.17 | 1 | 100 | 0 | 0 | 0 | 27 | 73 | 0 | 0 |
| 6 | 10th | 87 | 37 | 0.84 | 5.17 | 1 | 87 | 13 | 0 | 0 | 37 | 63 | 0 | 0 |
| 6 | 25th | 83 | 0 | 1.08 | 9.06 | 1 | 83 | 17 | 0 | 0 | 0 | 97 | 3 | 0 |
| 6 | 50th | 37 | 0 | 5.17 | 8.88 | 1 | 37 | 63 | 0 | 0 | 0 | 10 | 90 | 0 |
| 6 | 75th | 0 | 0 | 9.21 | 8.81 | 1 | 0 | 100 | 0 | 0 | 0 | 0 | 87 | 13 |
| 6 | 90th | 0 | 0 | 9.21 | 8.96 | 1 | 0 | 100 | 0 | 0 | 0 | 0 | 93 | 7 |
| 7 | 10th | 73 | 83 | 1.88 | 1.08 | 1 | 73 | 27 | 0 | 0 | 83 | 17 | 0 | 0 |
| 7 | 25th | 63 | 30 | 2.72 | 5.83 | 1 | 63 | 37 | 0 | 0 | 30 | 70 | 0 | 0 |
| 7 | 50th | 43 | 0 | 4.53 | 8.81 | 1 | 43 | 57 | 0 | 0 | 0 | 87 | 13 | 0 |
| 7 | 75th | 10 | 0 | 7.96 | 8.66 | 1 | 10 | 90 | 0 | 0 | 0 | 77 | 23 | 0 |
| 7 | 90th | 0 | 0 | 9.21 | 8.52 | 1 | 0 | 100 | 0 | 0 | 0 | 57 | 43 | 0 |
| 8 | 10th | 100 | 63 | 0.00 | 2.72 | 1 | 100 | 0 | 0 | 0 | 63 | 37 | 0 | 0 |
| 8 | 25th | 97 | 50 | 0.16 | 3.91 | 1 | 97 | 3 | 0 | 0 | 50 | 50 | 0 | 0 |
| 8 | 50th | 73 | 50 | 1.88 | 3.58 | 1 | 73 | 27 | 0 | 0 | 50 | 30 | 20 | 0 |
| 8 | 75th | 7 | 13 | 8.35 | 7.12 | 1 | 7 | 93 | 0 | 0 | 13 | 20 | 67 | 0 |
| 8 | 90th | 0 | 17 | 9.21 | 6.65 | 1 | 0 | 100 | 0 | 0 | 17 | 43 | 40 | 0 |

*Table 23. Autoscore and BERT Predictions on Babel Essays, by Length Percentile Bin in Organization*

|  | %-ile | Accuracy (%) |  | Kullback-Leibler |  | Expected | Autoscore (%) |  |  |  | BERT (%) |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade | Bin | AS | BERT | AS | BERT | Score | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 3 | 10th | 93 | 100 | 0.37 | 0.00 | 1 | 93 | 7 | 0 | 0 | 100 | 0 | 0 | 0 |
| 3 | 25th | 97 | 90 | 0.16 | 0.60 | 1 | 97 | 3 | 0 | 0 | 90 | 10 | 0 | 0 |
| 3 | 50th | 67 | 10 | 2.43 | 7.96 | 1 | 67 | 33 | 0 | 0 | 10 | 90 | 0 | 0 |
| 3 | 75th | 60 | 0 | 3.01 | 9.06 | 1 | 60 | 40 | 0 | 0 | 0 | 97 | 3 | 0 |
| 3 | 90th | 53 | 0 | 3.61 | 8.21 | 1 | 53 | 47 | 0 | 0 | 0 | 53 | 17 | 30 |
| 4 | 10th | 100 | 100 | 0.00 | 0.00 | 1 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| 4 | 25th | 100 | 97 | 0.00 | 0.16 | 1 | 100 | 0 | 0 | 0 | 97 | 3 | 0 | 0 |
| 4 | 50th | 100 | 67 | 0.00 | 2.43 | 1 | 100 | 0 | 0 | 0 | 67 | 33 | 0 | 0 |
| 4 | 75th | 100 | 33 | 0.00 | 5.50 | 1 | 100 | 0 | 0 | 0 | 33 | 67 | 0 | 0 |
| 4 | 90th | 100 | 17 | 0.00 | 7.22 | 1 | 100 | 0 | 0 | 0 | 17 | 83 | 0 | 0 |
| 5 | 10th | 50 | 100 | 3.91 | 0.00 | 1 | 50 | 50 | 0 | 0 | 100 | 0 | 0 | 0 |
| 5 | 25th | 70 | 63 | 2.15 | 2.72 | 1 | 70 | 30 | 0 | 0 | 63 | 37 | 0 | 0 |
| 5 | 50th | 83 | 0 | 1.08 | 9.21 | 1 | 83 | 17 | 0 | 0 | 0 | 100 | 0 | 0 |
| 5 | 75th | 97 | 0 | 0.16 | 9.21 | 1 | 97 | 3 | 0 | 0 | 0 | 100 | 0 | 0 |
| 5 | 90th | 97 | 0 | 0.16 | 9.21 | 1 | 97 | 3 | 0 | 0 | 0 | 100 | 0 | 0 |
| 6 | 10th | 73 | 0 | 1.88 | 9.21 | 1 | 73 | 27 | 0 | 0 | 0 | 100 | 0 | 0 |
| 6 | 25th | 83 | 0 | 1.08 | 8.63 | 1 | 83 | 17 | 0 | 0 | 0 | 73 | 27 | 0 |
| 6 | 50th | 67 | 0 | 2.43 | 9.21 | 1 | 67 | 33 | 0 | 0 | 0 | 0 | 100 | 0 |
| 6 | 75th | 20 | 0 | 6.87 | 8.71 | 1 | 20 | 80 | 0 | 0 | 0 | 0 | 80 | 20 |
| 6 | 90th | 30 | 0 | 5.83 | 9.21 | 1 | 30 | 70 | 0 | 0 | 0 | 0 | 100 | 0 |
| 7 | 10th | 100 | 47 | 0.00 | 4.22 | 1 | 100 | 0 | 0 | 0 | 47 | 53 | 0 | 0 |
| 7 | 25th | 100 | 3 | 0.00 | 8.75 | 1 | 100 | 0 | 0 | 0 | 3 | 97 | 0 | 0 |
| 7 | 50th | 83 | 0 | 1.08 | 8.81 | 1 | 83 | 17 | 0 | 0 | 0 | 13 | 87 | 0 |
| 7 | 75th | 20 | 0 | 6.87 | 9.21 | 1 | 20 | 80 | 0 | 0 | 0 | 0 | 100 | 0 |
| 7 | 90th | 3 | 0 | 8.75 | 9.21 | 1 | 3 | 97 | 0 | 0 | 0 | 0 | 100 | 0 |
| 8 | 10th | 100 | 33 | 0.00 | 5.50 | 1 | 100 | 0 | 0 | 0 | 33 | 67 | 0 | 0 |
| 8 | 25th | 100 | 0 | 0.00 | 9.21 | 1 | 100 | 0 | 0 | 0 | 0 | 100 | 0 | 0 |
| 8 | 50th | 100 | 0 | 0.00 | 9.21 | 1 | 100 | 0 | 0 | 0 | 0 | 100 | 0 | 0 |
| 8 | 75th | 100 | 0 | 0.00 | 8.52 | 1 | 100 | 0 | 0 | 0 | 0 | 47 | 53 | 0 |
| 8 | 90th | 100 | 0 | 0.00 | 8.57 | 1 | 100 | 0 | 0 | 0 | 0 | 67 | 33 | 0 |

## Appendix C: Off-Topic Essay Results for Items and Length Percentile Bins

*Table 24. Autoscore and BERT Predictions on Off-Topic Essays, by Length Percentile Bin in Conventions*

| | %-ile | Accuracy (%) | | Kullback-Leibler | | Expected (%) | | | Autoscore (%) | | | BERT (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Grade** | **Bin** | **AS** | **BERT** | **AS** | **BERT** | **0** | **1** | **2** | **0** | **1** | **2** | **0** | **1** | **2** |
| 3 | 10th | 53 | 73 | 0.20 | 0.08 | 10 | 50 | 40 | 23 | 63 | 13 | 20 | 57 | 23 |
| 3 | 25th | 57 | 77 | 0.26 | 0.14 | 7 | 20 | 73 | 10 | 50 | 40 | 3 | 43 | 53 |
| 3 | 50th | 40 | 60 | 0.35 | 0.04 | 3 | 40 | 57 | 3 | 80 | 17 | 0 | 47 | 53 |
| 3 | 75th | 53 | 70 | 0.76 | 0.38 | 0 | 33 | 67 | 10 | 57 | 33 | 7 | 37 | 57 |
| 3 | 90th | 57 | 80 | 0.41 | 0.06 | 3 | 30 | 67 | 3 | 73 | 23 | 7 | 43 | 50 |
| 4 | 10th | 30 | 30 | 0.59 | 0.47 | 10 | 50 | 40 | 33 | 67 | 0 | 40 | 53 | 7 |
| 4 | 25th | 13 | 23 | 0.71 | 0.69 | 20 | 23 | 57 | 33 | 63 | 3 | 50 | 47 | 3 |
| 4 | 50th | 50 | 33 | 0.56 | 0.74 | 7 | 60 | 33 | 30 | 70 | 0 | 47 | 50 | 3 |
| 4 | 75th | 40 | 37 | 0.54 | 0.59 | 10 | 50 | 40 | 27 | 73 | 0 | 33 | 67 | 0 |
| 4 | 90th | 33 | 27 | 1.22 | 1.43 | 0 | 40 | 60 | 7 | 93 | 0 | 17 | 70 | 13 |
| 5 | 10th | 63 | 73 | 0.19 | 0.19 | 0 | 43 | 57 | 3 | 53 | 43 | 3 | 30 | 67 |
| 5 | 25th | 63 | 83 | 0.25 | 0.03 | 3 | 17 | 80 | 3 | 47 | 50 | 3 | 27 | 70 |
| 5 | 50th | 73 | 93 | 0.21 | 0.01 | 0 | 23 | 77 | 3 | 37 | 60 | 0 | 17 | 83 |
| 5 | 75th | 70 | 83 | 0.12 | 0.08 | 0 | 30 | 70 | 0 | 53 | 47 | 0 | 13 | 87 |
| 5 | 90th | 60 | 83 | 0.37 | 0.08 | 7 | 13 | 80 | 0 | 47 | 53 | 0 | 10 | 90 |
| 6 | 10th | 57 | 60 | 0.09 | 0.10 | 13 | 43 | 43 | 17 | 23 | 60 | 13 | 23 | 63 |
| 6 | 25th | 67 | 80 | 0.04 | 0.03 | 10 | 33 | 57 | 3 | 40 | 57 | 3 | 33 | 63 |
| 6 | 50th | 80 | 77 | 0.04 | 0.06 | 3 | 53 | 43 | 7 | 40 | 53 | 3 | 37 | 60 |
| 6 | 75th | 60 | 70 | 0.04 | 0.02 | 0 | 43 | 57 | 0 | 30 | 70 | 0 | 33 | 67 |
| 6 | 90th | 70 | 73 | 0.02 | 0.03 | 7 | 30 | 63 | 3 | 27 | 70 | 3 | 23 | 73 |
| 7 | 10th | 57 | 70 | 0.06 | 0.12 | 17 | 30 | 53 | 7 | 27 | 67 | 10 | 13 | 77 |
| 7 | 25th | 57 | 60 | 0.20 | 0.20 | 13 | 40 | 47 | 0 | 30 | 70 | 3 | 20 | 77 |
| 7 | 50th | 70 | 80 | 0.12 | 0.09 | 3 | 43 | 53 | 3 | 20 | 77 | 3 | 23 | 73 |
| 7 | 75th | 53 | 57 | 0.26 | 0.40 | 7 | 40 | 53 | 3 | 10 | 87 | 0 | 7 | 93 |
| 7 | 90th | 77 | 77 | 0.13 | 0.07 | 3 | 27 | 70 | 0 | 10 | 90 | 0 | 17 | 83 |
| 8 | 10th | 80 | 80 | 0.38 | 0.23 | 0 | 20 | 80 | 3 | 0 | 97 | 3 | 7 | 90 |
| 8 | 25th | 90 | 90 | 0.04 | 0.00 | 0 | 17 | 83 | 0 | 7 | 93 | 0 | 13 | 87 |
| 8 | 50th | 90 | 83 | 0.04 | 0.54 | 0 | 17 | 83 | 0 | 7 | 93 | 7 | 0 | 93 |
| 8 | 75th | 80 | 77 | 0.23 | 0.03 | 13 | 20 | 67 | 13 | 0 | 87 | 7 | 17 | 77 |
| 8 | 90th | 93 | 87 | 0.01 | 0.01 | 0 | 27 | 73 | 0 | 20 | 80 | 0 | 20 | 80 |

*Table 25. Autoscore and BERT Predictions on Off-Topic Essays, by Length Percentile Bin in Elaboration*

| | %-ile | Accuracy (%) | | Kullback-Leibler | | Expected | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Grade** | **Bin** | **AS** | **BERT** | **AS** | **BERT** | **Score** | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 3 | 10th | 100 | 97 | 0.00 | 0.16 | 1 | 100 | 0 | 0 | 0 | 97 | 3 | 0 | 0 |
| 3 | 25th | 100 | 90 | 0.00 | 0.60 | 1 | 100 | 0 | 0 | 0 | 90 | 10 | 0 | 0 |
| 3 | 50th | 100 | 87 | 0.00 | 0.84 | 1 | 100 | 0 | 0 | 0 | 87 | 13 | 0 | 0 |
| 3 | 75th | 100 | 77 | 0.00 | 1.61 | 1 | 100 | 0 | 0 | 0 | 77 | 23 | 0 | 0 |
| 3 | 90th | 100 | 73 | 0.00 | 1.88 | 1 | 100 | 0 | 0 | 0 | 73 | 27 | 0 | 0 |
| 4 | 10th | 97 | 90 | 0.16 | 0.60 | 1 | 97 | 3 | 0 | 0 | 90 | 10 | 0 | 0 |
| 4 | 25th | 100 | 97 | 0.00 | 0.16 | 1 | 100 | 0 | 0 | 0 | 97 | 3 | 0 | 0 |
| 4 | 50th | 100 | 90 | 0.00 | 0.60 | 1 | 100 | 0 | 0 | 0 | 90 | 10 | 0 | 0 |
| 4 | 75th | 100 | 83 | 0.00 | 1.08 | 1 | 100 | 0 | 0 | 0 | 83 | 17 | 0 | 0 |
| 4 | 90th | 90 | 63 | 0.60 | 2.72 | 1 | 90 | 10 | 0 | 0 | 63 | 37 | 0 | 0 |
| 5 | 10th | 97 | 83 | 0.16 | 1.08 | 1 | 97 | 3 | 0 | 0 | 83 | 17 | 0 | 0 |
| 5 | 25th | 100 | 93 | 0.00 | 0.32 | 1 | 100 | 0 | 0 | 0 | 93 | 3 | 3 | 0 |
| 5 | 50th | 97 | 73 | 0.16 | 1.88 | 1 | 97 | 3 | 0 | 0 | 73 | 27 | 0 | 0 |
| 5 | 75th | 100 | 90 | 0.00 | 0.60 | 1 | 100 | 0 | 0 | 0 | 90 | 10 | 0 | 0 |
| 5 | 90th | 100 | 77 | 0.00 | 1.61 | 1 | 100 | 0 | 0 | 0 | 77 | 23 | 0 | 0 |
| 6 | 10th | 73 | 40 | 1.88 | 4.47 | 1 | 73 | 27 | 0 | 0 | 40 | 40 | 20 | 0 |
| 6 | 25th | 63 | 37 | 2.61 | 4.81 | 1 | 63 | 33 | 3 | 0 | 37 | 47 | 17 | 0 |
| 6 | 50th | 80 | 30 | 1.34 | 5.45 | 1 | 80 | 20 | 0 | 0 | 30 | 53 | 17 | 0 |
| 6 | 75th | 53 | 13 | 3.61 | 7.16 | 1 | 53 | 47 | 0 | 0 | 13 | 70 | 17 | 0 |
| 6 | 90th | 53 | 13 | 3.61 | 7.01 | 1 | 53 | 47 | 0 | 0 | 13 | 53 | 33 | 0 |
| 7 | 10th | 40 | 37 | 4.85 | 4.90 | 1 | 40 | 60 | 0 | 0 | 37 | 53 | 10 | 0 |
| 7 | 25th | 50 | 37 | 3.91 | 5.04 | 1 | 50 | 50 | 0 | 0 | 37 | 60 | 3 | 0 |
| 7 | 50th | 23 | 27 | 6.52 | 5.71 | 1 | 23 | 77 | 0 | 0 | 27 | 50 | 23 | 0 |
| 7 | 75th | 27 | 27 | 6.17 | 5.74 | 1 | 27 | 73 | 0 | 0 | 27 | 53 | 20 | 0 |
| 7 | 90th | 23 | 20 | 6.52 | 6.46 | 1 | 23 | 77 | 0 | 0 | 20 | 63 | 17 | 0 |
| 8 | 10th | 87 | 77 | 0.76 | 1.61 | 1 | 87 | 10 | 3 | 0 | 77 | 23 | 0 | 0 |
| 8 | 25th | 87 | 60 | 0.84 | 2.90 | 1 | 87 | 13 | 0 | 0 | 60 | 37 | 3 | 0 |
| 8 | 50th | 87 | 73 | 0.84 | 1.88 | 1 | 87 | 13 | 0 | 0 | 73 | 27 | 0 | 0 |
| 8 | 75th | 80 | 80 | 1.34 | 1.22 | 1 | 80 | 20 | 0 | 0 | 80 | 13 | 7 | 0 |
| 8 | 90th | 90 | 90 | 0.60 | 0.60 | 1 | 90 | 10 | 0 | 0 | 90 | 10 | 0 | 0 |

*Table 26. Autoscore and BERT Predictions on Off-Topic Essays, by Length Percentile Bin in Organization*

| | | Accuracy (%) | | Kullback-Leibler | | | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade | %-ile Bin | AS | BERT | AS | BERT | Expected Score | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 3 | 10th | 100 | 83 | 0.00 | 0.91 | 1 | 100 | 0 | 0 | 0 | 83 | 7 | 7 | 3 |
| 3 | 25th | 100 | 60 | 0.00 | 2.83 | 1 | 100 | 0 | 0 | 0 | 60 | 33 | 7 | 0 |
| 3 | 50th | 90 | 83 | 0.60 | 1.08 | 1 | 90 | 10 | 0 | 0 | 83 | 17 | 0 | 0 |
| 3 | 75th | 100 | 47 | 0.00 | 3.86 | 1 | 100 | 0 | 0 | 0 | 47 | 30 | 23 | 0 |
| 3 | 90th | 93 | 73 | 0.37 | 1.69 | 1 | 93 | 7 | 0 | 0 | 73 | 13 | 13 | 0 |
| 4 | 10th | 90 | 73 | 0.60 | 1.70 | 1 | 90 | 10 | 0 | 0 | 73 | 17 | 10 | 0 |
| 4 | 25th | 97 | 70 | 0.16 | 2.05 | 1 | 97 | 3 | 0 | 0 | 70 | 27 | 3 | 0 |
| 4 | 50th | 100 | 60 | 0.00 | 2.83 | 1 | 100 | 0 | 0 | 0 | 60 | 33 | 7 | 0 |
| 4 | 75th | 90 | 53 | 0.60 | 3.49 | 1 | 90 | 10 | 0 | 0 | 53 | 43 | 3 | 0 |
| 4 | 90th | 93 | 47 | 0.37 | 3.92 | 1 | 93 | 7 | 0 | 0 | 47 | 40 | 13 | 0 |
| 5 | 10th | 87 | 60 | 0.84 | 3.01 | 1 | 87 | 13 | 0 | 0 | 60 | 40 | 0 | 0 |
| 5 | 25th | 100 | 63 | 0.00 | 2.55 | 1 | 100 | 0 | 0 | 0 | 63 | 30 | 7 | 0 |
| 5 | 50th | 93 | 37 | 0.37 | 5.17 | 1 | 93 | 7 | 0 | 0 | 37 | 63 | 0 | 0 |
| 5 | 75th | 90 | 40 | 0.60 | 4.85 | 1 | 90 | 10 | 0 | 0 | 40 | 60 | 0 | 0 |
| 5 | 90th | 100 | 27 | 0.00 | 6.17 | 1 | 100 | 0 | 0 | 0 | 27 | 73 | 0 | 0 |
| 6 | 10th | 57 | 23 | 3.31 | 5.83 | 1 | 57 | 43 | 0 | 0 | 23 | 43 | 27 | 7 |
| 6 | 25th | 53 | 7 | 3.49 | 7.51 | 1 | 53 | 43 | 3 | 0 | 7 | 47 | 40 | 7 |
| 6 | 50th | 73 | 10 | 1.88 | 7.37 | 1 | 73 | 27 | 0 | 0 | 10 | 57 | 33 | 0 |
| 6 | 75th | 43 | 3 | 4.53 | 8.09 | 1 | 43 | 57 | 0 | 0 | 3 | 43 | 53 | 0 |
| 6 | 90th | 47 | 0 | 4.22 | 8.39 | 1 | 47 | 53 | 0 | 0 | 0 | 47 | 50 | 3 |
| 7 | 10th | 70 | 30 | 2.15 | 5.36 | 1 | 70 | 30 | 0 | 0 | 30 | 40 | 30 | 0 |
| 7 | 25th | 77 | 23 | 1.51 | 5.99 | 1 | 77 | 20 | 3 | 0 | 23 | 43 | 33 | 0 |
| 7 | 50th | 53 | 20 | 3.61 | 6.32 | 1 | 53 | 47 | 0 | 0 | 20 | 47 | 33 | 0 |
| 7 | 75th | 67 | 10 | 2.43 | 7.42 | 1 | 67 | 33 | 0 | 0 | 10 | 63 | 27 | 0 |
| 7 | 90th | 60 | 10 | 3.01 | 7.39 | 1 | 60 | 40 | 0 | 0 | 10 | 60 | 30 | 0 |
| 8 | 10th | 93 | 47 | 0.37 | 4.10 | 1 | 93 | 7 | 0 | 0 | 47 | 50 | 3 | 0 |
| 8 | 25th | 100 | 43 | 0.00 | 4.41 | 1 | 100 | 0 | 0 | 0 | 43 | 53 | 3 | 0 |
| 8 | 50th | 100 | 27 | 0.00 | 6.17 | 1 | 100 | 0 | 0 | 0 | 27 | 73 | 0 | 0 |
| 8 | 75th | 97 | 33 | 0.16 | 5.37 | 1 | 97 | 3 | 0 | 0 | 33 | 63 | 3 | 0 |
| 8 | 90th | 97 | 20 | 0.16 | 6.87 | 1 | 97 | 3 | 0 | 0 | 20 | 80 | 0 | 0 |

# Appendix D: Duplicated Essay Results for Items and Length Percentile Bins

*Table 27. Autoscore and BERT Predictions on Duplicated Essays, by Length Percentile Bin in Conventions*

| | | Accuracy (%) | | Kullback-Leibler | | Expected (%) | | | Autoscore (%) | | | BERT (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade | %-ile Bin | AS | BERT | AS | BERT | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 |
| 3 | 10th | 57 | 63 | 0.24 | 0.23 | 47 | 20 | 33 | 23 | 50 | 27 | 27 | 50 | 23 |
| 3 | 25th | 50 | 63 | 0.10 | 0.09 | 27 | 43 | 30 | 10 | 47 | 43 | 10 | 53 | 37 |
| 3 | 50th | 60 | 63 | 0.06 | 0.01 | 13 | 57 | 30 | 10 | 43 | 47 | 10 | 53 | 37 |
| 3 | 75th | 83 | 90 | 0.00 | 0.00 | 10 | 30 | 60 | 10 | 27 | 63 | 10 | 27 | 63 |
| 3 | 90th | 73 | 80 | 0.14 | 0.03 | 3 | 37 | 60 | 0 | 17 | 83 | 0 | 37 | 63 |
| 4 | 10th | 63 | 70 | 0.02 | 0.04 | 40 | 53 | 7 | 50 | 43 | 7 | 43 | 43 | 13 |
| 4 | 25th | 53 | 73 | 0.01 | 0.04 | 27 | 47 | 27 | 20 | 53 | 27 | 17 | 60 | 23 |
| 4 | 50th | 67 | 90 | 0.13 | 0.06 | 13 | 63 | 23 | 10 | 43 | 47 | 3 | 73 | 23 |
| 4 | 75th | 67 | 67 | 0.04 | 0.01 | 7 | 57 | 37 | 7 | 43 | 50 | 7 | 63 | 30 |
| 4 | 90th | 47 | 90 | 0.23 | 0.02 | 7 | 57 | 37 | 13 | 23 | 63 | 10 | 47 | 43 |
| 5 | 10th | 53 | 53 | 0.04 | 0.00 | 7 | 47 | 47 | 7 | 33 | 60 | 7 | 47 | 47 |
| 5 | 25th | 77 | 87 | 0.09 | 0.03 | 3 | 27 | 70 | 3 | 10 | 87 | 0 | 27 | 73 |
| 5 | 50th | 83 | 90 | 0.10 | 0.03 | 0 | 23 | 77 | 0 | 7 | 93 | 0 | 13 | 87 |
| 5 | 75th | 87 | 83 | 0.06 | 0.03 | 0 | 23 | 77 | 0 | 10 | 90 | 0 | 13 | 87 |
| 5 | 90th | 80 | 83 | 0.16 | 0.10 | 0 | 23 | 77 | 0 | 3 | 97 | 0 | 7 | 93 |
| 6 | 10th | 67 | 77 | 0.08 | 0.02 | 7 | 53 | 40 | 3 | 37 | 60 | 3 | 50 | 47 |
| 6 | 25th | 77 | 80 | 0.14 | 0.01 | 3 | 37 | 60 | 7 | 13 | 80 | 3 | 30 | 67 |
| 6 | 50th | 53 | 67 | 0.49 | 0.13 | 3 | 43 | 53 | 0 | 3 | 97 | 0 | 23 | 77 |
| 6 | 75th | 53 | 57 | 0.38 | 0.02 | 3 | 40 | 57 | 3 | 3 | 93 | 3 | 30 | 67 |
| 6 | 90th | 63 | 77 | 0.38 | 0.06 | 3 | 40 | 57 | 3 | 3 | 93 | 3 | 23 | 73 |
| 7 | 10th | 57 | 70 | 0.41 | 0.11 | 10 | 37 | 53 | 3 | 3 | 93 | 0 | 37 | 63 |
| 7 | 25th | 80 | 83 | 0.09 | 0.05 | 7 | 20 | 73 | 3 | 7 | 90 | 3 | 10 | 87 |
| 7 | 50th | 87 | 87 | 0.15 | 0.15 | 3 | 13 | 83 | 7 | 0 | 93 | 7 | 0 | 93 |
| 7 | 75th | 90 | 90 | 0.10 | 0.10 | 0 | 10 | 90 | 0 | 0 | 100 | 0 | 0 | 100 |
| 7 | 90th | 87 | 93 | 0.14 | 0.02 | 3 | 13 | 83 | 3 | 0 | 97 | 3 | 7 | 90 |
| 8 | 10th | 47 | 67 | 0.49 | 0.15 | 3 | 57 | 40 | 10 | 10 | 80 | 7 | 30 | 63 |
| 8 | 25th | 60 | 70 | 0.51 | 0.25 | 3 | 37 | 60 | 0 | 0 | 100 | 3 | 7 | 90 |
| 8 | 50th | 67 | 83 | 0.29 | 0.03 | 7 | 27 | 67 | 0 | 3 | 97 | 7 | 17 | 77 |
| 8 | 75th | 90 | 90 | 0.10 | 0.06 | 3 | 7 | 90 | 0 | 0 | 100 | 0 | 13 | 87 |
| 8 | 90th | 90 | 93 | 0.10 | 0.03 | 0 | 10 | 90 | 0 | 0 | 100 | 0 | 3 | 97 |

*Table 28. Autoscore and BERT Predictions on Duplicated Essays, by Length Percentile Bin in Elaboration*

| Grade | %-ile Bin | Accuracy (%) | | Kullback-Leibler | | Expected | | | | Autoscore | | | | BERT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AS | BERT | AS | BERT | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 3 | 10th | 40 | 43 | 0.72 | 0.39 | 77 | 23 | 0 | 0 | 33 | 60 | 7 | 0 | 43 | 53 | 3 | 0 |
| 3 | 25th | 70 | 73 | 0.31 | 0.05 | 23 | 73 | 3 | 0 | 10 | 67 | 23 | 0 | 17 | 73 | 10 | 0 |
| 3 | 50th | 60 | 60 | 1.37 | 1.17 | 37 | 63 | 0 | 0 | 20 | 60 | 20 | 0 | 13 | 70 | 17 | 0 |
| 3 | 75th | 40 | 67 | 2.45 | 0.15 | 13 | 40 | 47 | 0 | 7 | 33 | 27 | 33 | 7 | 20 | 73 | 0 |
| 3 | 90th | 23 | 47 | 1.02 | 0.62 | 7 | 37 | 47 | 10 | 7 | 3 | 23 | 67 | 0 | 3 | 97 | 0 |
| 4 | 10th | 53 | 63 | 0.55 | 0.31 | 77 | 23 | 0 | 0 | 33 | 63 | 3 | 0 | 40 | 60 | 0 | 0 |
| 4 | 25th | 43 | 53 | 0.54 | 0.44 | 43 | 53 | 3 | 0 | 13 | 57 | 30 | 0 | 3 | 87 | 10 | 0 |
| 4 | 50th | 33 | 57 | 1.64 | 0.14 | 23 | 63 | 13 | 0 | 3 | 40 | 37 | 20 | 7 | 67 | 27 | 0 |
| 4 | 75th | 27 | 57 | 2.61 | 0.19 | 13 | 57 | 30 | 0 | 7 | 17 | 43 | 33 | 7 | 33 | 60 | 0 |
| 4 | 90th | 33 | 70 | 1.17 | 0.09 | 10 | 43 | 40 | 7 | 10 | 10 | 23 | 50 | 7 | 40 | 53 | 0 |
| 5 | 10th | 47 | 50 | 0.54 | 0.86 | 73 | 27 | 0 | 0 | 30 | 67 | 3 | 0 | 33 | 57 | 10 | 0 |
| 5 | 25th | 50 | 67 | 0.30 | 0.11 | 30 | 63 | 7 | 0 | 7 | 67 | 27 | 0 | 13 | 70 | 17 | 0 |
| 5 | 50th | 43 | 57 | 0.24 | 0.25 | 10 | 57 | 30 | 3 | 3 | 30 | 57 | 10 | 7 | 30 | 63 | 0 |
| 5 | 75th | 50 | 53 | 0.30 | 0.21 | 13 | 53 | 30 | 3 | 10 | 27 | 43 | 20 | 10 | 30 | 60 | 0 |
| 5 | 90th | 30 | 40 | 0.52 | 0.34 | 10 | 40 | 27 | 23 | 10 | 3 | 23 | 63 | 7 | 40 | 53 | 0 |
| 6 | 10th | 27 | 23 | 3.65 | 4.29 | 30 | 70 | 0 | 0 | 7 | 47 | 47 | 0 | 3 | 43 | 53 | 0 |
| 6 | 25th | 10 | 7 | 1.62 | 2.19 | 23 | 63 | 13 | 0 | 0 | 20 | 70 | 10 | 0 | 20 | 60 | 20 |
| 6 | 50th | 3 | 23 | 6.98 | 3.35 | 7 | 37 | 57 | 0 | 0 | 0 | 20 | 80 | 0 | 0 | 60 | 40 |
| 6 | 75th | 13 | 63 | 1.87 | 0.26 | 0 | 23 | 67 | 10 | 0 | 0 | 10 | 87 | 0 | 3 | 67 | 30 |
| 6 | 90th | 20 | 73 | 1.18 | 0.04 | 0 | 20 | 63 | 17 | 0 | 3 | 10 | 87 | 0 | 10 | 70 | 20 |
| 7 | 10th | 37 | 43 | 1.20 | 0.95 | 57 | 43 | 0 | 0 | 3 | 87 | 10 | 0 | 27 | 60 | 13 | 0 |
| 7 | 25th | 33 | 53 | 3.68 | 1.28 | 43 | 57 | 0 | 0 | 10 | 43 | 47 | 0 | 10 | 73 | 17 | 0 |
| 7 | 50th | 23 | 73 | 2.11 | 0.63 | 10 | 80 | 10 | 0 | 7 | 20 | 53 | 20 | 7 | 67 | 17 | 10 |
| 7 | 75th | 3 | 93 | 4.31 | 0.37 | 0 | 73 | 27 | 0 | 0 | 0 | 53 | 47 | 0 | 70 | 23 | 7 |
| 7 | 90th | 7 | 73 | 6.54 | 0.60 | 10 | 67 | 23 | 0 | 3 | 10 | 10 | 77 | 10 | 57 | 23 | 10 |
| 8 | 10th | 43 | 70 | 0.84 | 0.12 | 70 | 30 | 0 | 0 | 20 | 73 | 0 | 7 | 47 | 53 | 0 | 0 |
| 8 | 25th | 43 | 57 | 2.53 | 1.34 | 33 | 67 | 0 | 0 | 7 | 60 | 33 | 0 | 30 | 50 | 20 | 0 |
| 8 | 50th | 23 | 47 | 2.64 | 0.83 | 17 | 80 | 3 | 0 | 10 | 27 | 37 | 27 | 27 | 33 | 40 | 0 |
| 8 | 75th | 3 | 60 | 5.89 | 0.12 | 17 | 43 | 40 | 0 | 0 | 13 | 17 | 70 | 13 | 23 | 63 | 0 |
| 8 | 90th | 7 | 57 | 2.44 | 0.31 | 13 | 57 | 27 | 3 | 3 | 3 | 10 | 83 | 7 | 30 | 63 | 0 |

*Table 29. Autoscore and BERT Predictions on Duplicated Essays, by Length Percentile Bin in Organization*

| | %-ile | Accuracy (%) | | Kullback-Leibler | | Expected (%) | | | | Autoscore (%) | | | | BERT (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade | Bin | AS | BERT | AS | BERT | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 3 | 10th | 67 | 67 | 0.26 | 0.48 | 67 | 33 | 0 | 0 | 43 | 53 | 3 | 0 | 40 | 53 | 7 | 0 |
| 3 | 25th | 60 | 57 | 0.07 | 0.24 | 27 | 67 | 7 | 0 | 17 | 67 | 17 | 0 | 10 | 63 | 27 | 0 |
| 3 | 50th | 57 | 50 | 1.25 | 2.34 | 23 | 77 | 0 | 0 | 20 | 60 | 17 | 3 | 0 | 70 | 30 | 0 |
| 3 | 75th | 53 | 23 | 1.62 | 5.33 | 17 | 40 | 43 | 0 | 10 | 37 | 30 | 23 | 0 | 30 | 7 | 63 |
| 3 | 90th | 23 | 13 | 1.49 | 2.18 | 7 | 50 | 37 | 7 | 7 | 7 | 13 | 73 | 0 | 10 | 0 | 90 |
| 4 | 10th | 77 | 70 | 0.02 | 0.40 | 73 | 27 | 0 | 0 | 63 | 37 | 0 | 0 | 57 | 37 | 7 | 0 |
| 4 | 25th | 70 | 53 | 0.05 | 0.26 | 50 | 47 | 3 | 0 | 40 | 50 | 10 | 0 | 27 | 53 | 20 | 0 |
| 4 | 50th | 60 | 43 | 0.64 | 0.37 | 33 | 57 | 10 | 0 | 20 | 53 | 17 | 10 | 10 | 60 | 27 | 3 |
| 4 | 75th | 60 | 67 | 0.28 | 0.64 | 20 | 53 | 27 | 0 | 20 | 30 | 47 | 3 | 13 | 40 | 37 | 10 |
| 4 | 90th | 63 | 57 | 0.18 | 0.01 | 13 | 40 | 37 | 10 | 23 | 23 | 27 | 27 | 13 | 33 | 43 | 10 |
| 5 | 10th | 73 | 53 | 0.08 | 0.86 | 47 | 53 | 0 | 0 | 27 | 73 | 0 | 0 | 10 | 80 | 10 | 0 |
| 5 | 25th | 73 | 37 | 0.23 | 1.27 | 17 | 80 | 3 | 0 | 7 | 73 | 20 | 0 | 3 | 40 | 57 | 0 |
| 5 | 50th | 77 | 37 | 0.00 | 1.00 | 0 | 70 | 30 | 0 | 0 | 67 | 33 | 0 | 0 | 13 | 80 | 7 |
| 5 | 75th | 57 | 53 | 0.02 | 0.38 | 7 | 43 | 47 | 3 | 7 | 33 | 57 | 3 | 0 | 13 | 87 | 0 |
| 5 | 90th | 50 | 47 | 0.12 | 0.07 | 3 | 37 | 47 | 13 | 7 | 23 | 40 | 30 | 0 | 33 | 60 | 7 |
| 6 | 10th | 50 | 33 | 0.31 | 0.81 | 37 | 57 | 7 | 0 | 10 | 63 | 27 | 0 | 0 | 57 | 43 | 0 |
| 6 | 25th | 33 | 30 | 0.78 | 0.80 | 17 | 63 | 20 | 0 | 3 | 23 | 70 | 3 | 0 | 23 | 77 | 0 |
| 6 | 50th | 23 | 30 | 1.30 | 0.82 | 3 | 63 | 30 | 3 | 0 | 0 | 67 | 33 | 0 | 7 | 80 | 13 |
| 6 | 75th | 27 | 63 | 1.24 | 0.10 | 0 | 17 | 70 | 13 | 0 | 3 | 17 | 77 | 0 | 10 | 60 | 30 |
| 6 | 90th | 33 | 67 | 0.91 | 0.10 | 0 | 23 | 53 | 23 | 0 | 7 | 7 | 87 | 0 | 13 | 43 | 43 |
| 7 | 10th | 40 | 30 | 1.47 | 2.91 | 47 | 53 | 0 | 0 | 3 | 80 | 17 | 0 | 7 | 57 | 37 | 0 |
| 7 | 25th | 33 | 7 | 4.54 | 7.97 | 20 | 80 | 0 | 0 | 7 | 37 | 57 | 0 | 3 | 7 | 90 | 0 |
| 7 | 50th | 20 | 37 | 2.01 | 0.94 | 7 | 63 | 30 | 0 | 7 | 3 | 70 | 20 | 3 | 3 | 93 | 0 |
| 7 | 75th | 17 | 73 | 2.16 | 0.21 | 0 | 30 | 67 | 3 | 0 | 0 | 23 | 77 | 0 | 7 | 93 | 0 |
| 7 | 90th | 17 | 57 | 1.16 | 0.37 | 7 | 30 | 50 | 13 | 0 | 7 | 13 | 80 | 0 | 13 | 87 | 0 |
| 8 | 10th | 57 | 53 | 0.60 | 0.86 | 63 | 37 | 0 | 0 | 27 | 67 | 7 | 0 | 23 | 67 | 10 | 0 |
| 8 | 25th | 60 | 23 | 1.46 | 4.65 | 23 | 77 | 0 | 0 | 3 | 77 | 20 | 0 | 0 | 43 | 57 | 0 |
| 8 | 50th | 20 | 20 | 1.56 | 1.73 | 23 | 73 | 3 | 0 | 7 | 33 | 53 | 7 | 0 | 33 | 67 | 0 |
| 8 | 75th | 17 | 67 | 3.64 | 0.09 | 7 | 27 | 67 | 0 | 0 | 13 | 40 | 47 | 0 | 20 | 80 | 0 |
| 8 | 90th | 23 | 60 | 1.43 | 0.20 | 10 | 17 | 67 | 7 | 3 | 7 | 17 | 73 | 0 | 13 | 87 | 0 |