DSGA 1016 Final Project Report

Qiwenjing Jiang (qj336@nyu.edu)

Erqian Wang (ew1708@nyu.edu)

Kristine Zeng (yz4792@nyu.edu)

Jialing Li (jl9716@nyu.edu)

Abstract

In this research, Automated Essay Scoring (AES) systems utilizing DeBERTa, DistilBERT, and GPT-3.5 are implemented to evaluate essays written by 8th-12th grade English language learners. These systems are trained to assess essays across six aspects - cohesion, syntax, vocabulary, phraseology, grammar, and conventions. Additional textual features are added to evaluate and compare the scoring decisions in depth. By analyzing the differences and correlations between human and machine scores, we aim to reveal how computational models can emulate or diverge from human cognitive patterns in language assessment. This comparison not only highlights the potential and limitations of AES systems in understanding complex human judgments but also contributes to efforts aimed at refining these systems to better align with human cognition.

Keywords: Automated Essay Scoring; DeBERTa; DistilBERT; GPT-3.5; Cognitive Science; Natural Language Processing

Introduction

Cognitive load theory(Sweller, Ayres, & Kalyuga, 2011) is a theory that explores human cognitive architecture—specifically sensory memory, working memory and long-term memory. The theory provides a framework for understanding how information is processed and learned. As indicated in multiple studies, grammatical complexity is closely associated with readers' cognitive load.(Mikk, 2008; Schluroff, 1982)

In this research, we aim to leverage the theory to analyze how features in student essays affect the cognitive load of evaluators which in turn influences their grading decisions. In specific, we use features like sentence complexity and lexical diversity and compare human gradings with those generated by NLP models. Through the comparison, we wish to test the hypothesis that certain textual features impose different levels of cognitive load, which leads to variance in grading.

Methods and Models

Data

We employed a dataset provided by Vanderbilt and The Learning Agency Lab from Kaggle. It includes 3911 essays written by 8th–12th grade English language learners and their scores on 6 measures: cohesion, syntax, vocabulary, phraseology, grammar, and conventions. The scores range from 1.0 to 5.0 with increments of 0.5.

We first split the data into 80/20 train-test datasets. Then, we added the following features to the test dataset to enhance our understanding of how NLP models understand the grading task:

- **word count** number of words in an essay
- **sentence count** number of sentences in an essay
- **lexical diversity** ratio of unique words to the total number of words in an essay
- **average sentence length** average number of words per sentence
- **spelling error** number of words misspelled, this was identified using Python package pyspellchecker
- **rare word count** number of unique rare words used, this was calculated by identifying the 15,000 most common words in brown corpus of Python package nltk and selecting words in the essay that are spelled correctly and not included in the 15,000 common words.

Data Analysis Methods

Model Comparison To assess and contrast the overall performance of the models, we calculate and compare their means and standard deviations. Additionally, we employ the Root Mean Square Deviation (RMSE) to evaluate how closely the models' grading aligns with human assessments. The RMSE provides insights into which model most accurately mirrors the human grading system.

Feature Significance To understand feature influence on model accuracy, we utilized hypothesis testing in the following steps:

- Calculate the absolute difference in grading between the predictions made by each model and the human grading
- For each feature, the code selects the top 50 essays where the model performed best (smallest difference between predicted and actual grades) and the bottom 50 essays where the model performed worst (largest difference).

- For each model, perform statistical hypothesis tests (t-tests) to compare the distributions of specific features between the top 50 and bottom 50 essays for each model.

This analysis determines whether there are statistically significant differences in the added features, which may influence the grading accuracy of the models.

DeBERTa

We utilized the DeBERTaV3 model, a cutting-edge transformer architecture known for its efficiency and efficacy in NLP tasks. The model was trained to predict all six critical metrics of essay evaluation. Each essay was represented as an input to the model, and the outputs were the predicted scores across these six measures.

Training process

- **Tokenization** Essays were tokenized using the tokenizer associated with `DeBERTa-v3-base`, converting texts into input IDs and attention masks necessary for transformer models.
- **Model architecture** The core of our model was the DeBERTa transformer, supplemented by a mean pooling layer and a linear output layer to map the pooled representation to the scores of the six metrics.
- **Training loop** We employed a custom training loop that included gradient accumulation to handle larger batch sizes effectively, given GPU memory constraints. The loop used a SmoothL1Loss, which is less sensitive to outliers than mean squared error.

Inference The model inference involved preparing and tokenizing test data, loading the best-performing model weights for each cross-validation fold, and aggregating predictions from multiple folds to prevent overfitting.

DistilBERT

We utilized DistilBERT for training on essays, aiming to predict scores related to six linguistic dimensions: cohesion, syntax, vocabulary, phraseology, grammar, and conventions.

- **Data preparation** The corpus was tokenized using the distilBERT tokenizer, which converted texts into input IDs and attention masks necessary for transformer models. During the tokenization process, each input essay was truncated to a maximum of 500 tokens with padding. The tokenized data was separated into training, validation and testing set. Data loaders were created for each set.
- **Model architecture** A DistilBERT model for sequence classification was configured, and the classifier was modified to a linear layer that generates 6 continuous values, representing different aspects of linguistic scores.
- **Metrics**
  - **Custom Accuracy**:
    * The custom accuracy function was set up along with the MSE loss to evaluate the accuracy of predictions by determining if the absolute difference between predicted and actual scores for each linguistic dimension falls within a specified threshold, set at 0.25. It provided a measure of how well the model's predictions align with the actual scores across all linguistic dimensions, offering insights into the model's performance in capturing nuanced linguistic features.
  - **Mean Squared Error (MSE) Loss**:
    * MSE loss was utilized to quantify the average squared difference between predicted and actual values. By penalizing larger deviations more heavily, MSE loss enabled the model to make predictions that closely match the actual scores across all linguistic dimensions, facilitating effective training and evaluation of the model's performance.
- **Training loop** Model training loop was established with custom accuracy evaluation while employing Mean Squared Error Loss. An optimizer and scheduler were set up, and the model was trained and validated over thirty epochs.
- **Inference** The testing set was prepared during the data preparation step. The trained DistilBERT model was applied to the testing test and generated predictions for linguistic scores across six dimensions. Results are summarized in a DataFrame for comparison with actual values.

GPT-3.5

We also developed an automated English essay evaluation model with GPT-3.5 from OpenAI. We accessed GPT in notebook through OpenAI API, and sent prompts to OpenAI's chat completion endpoint through openai.ChatCompletion.create() function. Specifically, we were using "gpt-3.5-turbo" to process all requests. Different from BERT models, we provided the essays in their original human-readable format (without processing) to GPT, and prompted the model to evaluate them based on cohesion, syntax, vocabulary, phraseology, grammar, and conventions. The model's response would be parsed and the scores would be extracted. Since GPT-3.5 could not be directly trained based on the train set, we adjusted prompts and explored two ways to have GPT evaluate the test essays: scoring without grading samples and scoring with grading samples.

GPT Scoring Without Grading Samples With GPT-3.5's advanced natural language understanding, we assumed that it is capable of evaluating the cohesion, syntax, vocabulary, phraseology, grammar, and conventions of English essays on its own, while grading criteria being held consistently with the same prompt. Thus, in this method, we input each student's essay to GPT-3.5 and prompted it to grade each aspect on a scale of 1.0 to 5.0, in 0.5 increments. With our initial phrasings of the prompt, we noticed that GPT-3.5 gave

pretty low scores. To meet the high school standard, after a few tests, we refined and finalized our prompt to

"Grade the following English essay based on high school standards (8th-12th grade). Evaluate 'cohesion', 'syntax', 'vocabulary', 'phraseology', 'grammar', 'conventions' individually, scoring each from 1.0 to 5.0 in 0.5 increments. Grade leniently with a 5.0 score being achievable. Return only the numerical scores in a list: {essay}"

This prompt resulted in more lenient and reasonable scores.

GPT Scoring With Grading Samples We also developed a GPT scoring model given grading samples. Here, we prompted GPT to score each essay on the same six aspects, and provided it with 5 sample essays with scorings from the train set as reference. These samples included essays with high, average, and low scores. The prompt was structured as follows:

"Grade the following high school English essay (8th-12th grade). Evaluate 'cohesion', 'syntax', 'vocabulary', 'phraseology', 'grammar', 'conventions' individually, scoring each from 1.0 to 5.0 in 0.5 increments. A score of 5.0 should be attainable even if the essay is not perfect. Avoid very low scores unless absolutely necessary. The grading samples below should be taken as references. Return only the numerical scores in a list: {essay}

Grading Samples and scores in ['cohesion', 'syntax', 'vocabulary', 'phraseology', 'grammar', 'conventions'] format:
Essay 1: ...
scores: [4.0, 5.0, 5.0, 5.0, 5.0, 4.5]
Essay 2: ...
Scores: [4.5, 3.0, 4.5, 4.0, 4.0, 4.0]
Essay 3: ...
Scores: [3.0, 2.5, 2.0, 2.0, 2.5, 3.0]
Essay 4: ...
Scores: [3.5, 2.5, 3.0, 3.0, 2.5, 2.5]
Essay 5: ...
Scores: [2.0, 2.5, 2.0, 2.0, 1.5, 2.0]"

This approach had GPT review and analyze the scoring criterion and aimed to replicate the standard for scoring.

Results

Model Comparison

To understand the general performances of models on the grading task, we first calculated mean and standard deviation of each feature.

In addition, we adopted RMSE as the evaluation metric. For each model, we compare the 6 feature gradings with those of the original data.

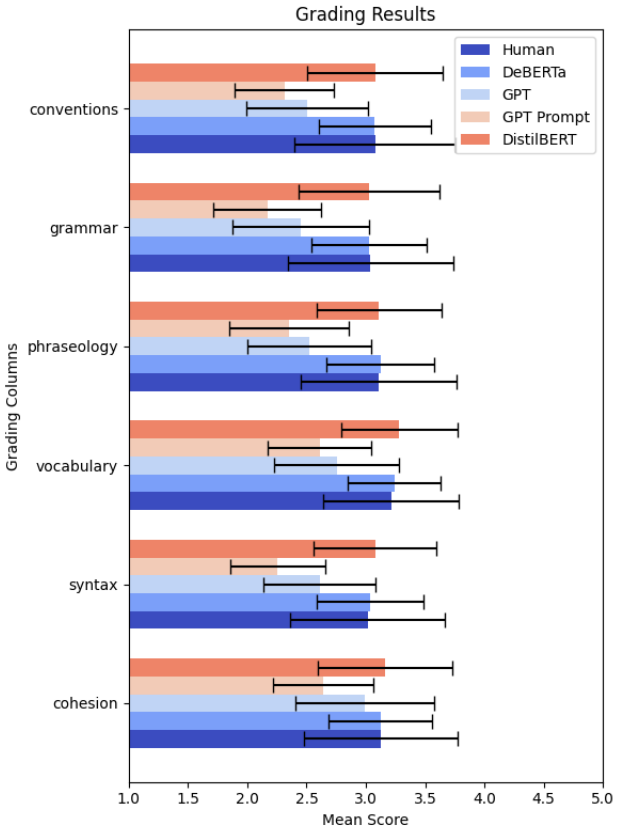

Figure 1: Scoring Mean and Standard Deviation

Model	Cohesion	Syntax	Vocab	Phrase	Grammar	Convention
DeBERTa	0.492	0.450	0.427	0.457	0.474	0.447
GPT	0.636	0.750	0.725	0.877	0.934	0.820
GPT Prompt	0.783	1.000	0.830	0.998	1.101	0.998
DistilBERT	0.639	0.602	0.549	0.618	0.696	0.620

Table 1: RMSE Scores

From Table 1, we can identify DeBERTa model has the leading performance, followed by DistilBERT. GPT models have less ideal performance compared to DeBERTa and DistilBERT.

DeBERTa

The RMSE suggests that DeBERTa performs with great accuracy across all measures, with the lowest error observed in vocabulary scoring and the highest in cohesion and grammar.

Hypothesis Testing Hypothesis tests were conducted to determine whether specific textual features could significantly influence DeBERTa's scoring discrepancies between essays. Significant differences were observed in:

- **Syntax with `avg_sentence_length`:** Essays where DeBERTa's syntax scores closely matched human ratings featured significantly shorter average sentence lengths (mean: 26.48) compared to those where DeBERTa's assessments diverged significantly (mean: 36.23). The t-test statistic was 2.25 with a p-value of 0.03, suggesting that essays with simpler, shorter sentences tend to be scored more consistently between DeBERTa and humans.
- **Vocabulary with `word_count`:** Vocabulary scoring discrepancies between DeBERTa and human ratings were significantly associated with the essay's word count. Essays with more accurate vocabulary scores had a lower average word count (mean: 393.9) compared to those with larger discrepancies (mean: 517.3). The t-test yielded a statistic of 3.33 with a p-value less than 0.001.

The test results indicate that certain textual complexities, such as longer sentences and higher word counts, are associated with larger discrepancies in model scoring.

DistilBERT

Compared to the RMSE scores of DeBERTa across the six assessed aspects, DistilBERT exhibits a slightly inferior performance. However, it still achieves commendable accuracy across all measures.

Hypothesis Testing Hypothesis tests were also conducted to determine whether specific textual features could significantly influence distilBERT scoring discrepancies between essays. We can observe significant differences in the following parts:

- **Vocabulary with `word_count`:** A positive t-statistic of 3.62 with a p-value of 0.0 indicates that essays with higher word counts demonstrate a stronger correlation with human grading accuracy.
- **Vocabulary with `lexical_diversity`:** Negative t-statistic of -2.14 and a p-value of 0.03, indicating lower lexical diversity in top-performing essays.
- **Vocabulary with `rare_words`:** A positive t-statistic of 2.37 with a p-value of 0.02 indicates that a higher usage of rare words correlates with better grading alignment between the model and human graders.
- **Phraseology with `word_count`:** A positive t-statistic of 2.05 and a p-value of 0.04, consistent with results observed in the vocabulary category, suggest that longer essays tend to align more closely with human grading. This indicates that both the model and human graders may favor or more accurately assess essays with greater length.
- **Convention with `Spelling_errors`:** A negative t-statistic of -2.13 with a p-value of 0.04 indicates that the grading of essays with fewer spelling errors tends to align more closely with human grading. This suggests that both the model and human graders consistently rate essays with fewer spelling mistakes more favorably.

Summary

- The hypothesis testing results show that **word count** significantly impacts grading across multiple linguistic aspects, with longer essays generally aligning better with human grading.
- **Lexical simplicity** and the correct use of **rare words** positively affect model performance, while **spelling accuracy** is crucial for high grading consistency.
- No significant features were identified in the categories of cohesion and syntax, which may suggest that these aspects are either uniformly handled by the model or are not captured effectively by the current feature set.

GPT-3.5

From Figure 1, we can see that GPT-3.5 models tend to give lower scores on all measures compared to DeBERTa and DistilBERT. However, both GPT (without grading samples) and GPT Prompt (with grading samples) demonstrate smaller standard deviations relatively, indicating more consistent gradings. This suggests that GPT-3.5 models apply grading criteria more uniformly across essays compared to the other models.

Evaluating the RMSE scores shown in Table 1, we can tell that GPT-3.5 models score very differently from humans, resulting in high RMSE values. For cohesion scores, GPT and GPT Prompt achieve RMSE scores of 0.636 and 0.783, respectively. These scores are the lowest errors among the six measures for GPT and GPT Prompt and are relatively closer to that of DistilBERT (0.639). However, for Grammar, GPT and GPT Prompt produce the highest error with RMSE values of 0.934 and 1.101, suggesting that GPT-3.5 models are stricter when evaluating grammatical correctness. Unlike what we expected, GPT (without grading samples) generally outperforms GPT Prompt (with grading samples), as indicated by smaller errors in all grading columns

Hypothesis Testing As we conduct hypothesis testing on GPT-3.5 models to evaluate whether the grading significantly differs for certain measures based on specific textual features, we notice that lexical diversity, spelling errors, average sentence length, and sentence count largely influence the grading performance of GPT-3.5 models in most measures. Textual features significantly affect whether GPT-3.5 models grade similarly to human raters across all six grading measures.

GPT (Without Grading Samples)

- **Syntax with `word_count`, `sentence_count`, `rare_words`:** The positive t-statistic values of these three textual features—3.96, 3.25, and 3.32, respectively—and p-values of 0.0 each indicate that higher word and sentence counts, as well as more frequent use of rare words, are correlated with model output scores closer to human grading.

- **Syntax, with `avg_sentence_length`:** A negative t-statistic value of -3.47 and p-value of 0.0 indicate that longer sentence length is associated with scoring more different from humans, which means that GPT model and human value the use of long and short sentences very differently regarding Grammar grading.
- **Syntax with `lexical_diversity`:** A negative t-statistic value of -2.11 and p-value of 0.04 suggest that, surprisingly, higher lexical diversity correlates with scores more different from human grading. However, a relatively larger p-value compared to that of the previous textual features implies that this correlation is less statistically significant.
- **Phraseology with `rare_words`:** A positive t-statistic value of 2.14 and p-value of 0.03 suggest that more common use of rare words is associated with scores closer to human grading.
- **Grammar with `lexical_diversity`:** A higher lexical diversity is correlated with scores similar to human grading, suggested by the positive t-statistic value of 2.78 and p-value of 0.01.
- **Convention with `avg_sentence_length`:** Similar to the previous point, a negative t-statistic value of -2.75 and p-value of 0.01 suggest that larger average sentence length is correlated with scoring more different from human grading on the Convention measure.
- **Convention with `spelling_errors`:** A negative t-statistic value of -2.38 and p-values of 0.02 suggest that fewer spelling errors are associated with scores closer to human rating.

GPT Prompt(With Grading Samples)

- **Cohesion, Syntax, and Phraseology with `sentence_count`:** All positive t-statistic values of 2.14, 3.09, 2.03, and p-values of 0.03, 0, 0.05, indicate that more sentences are correlated with model scoring closer to human scoring on Cohesion, Syntax, and Phraseology.
- **Syntax, Phraseology, Convention with `avg_sentence_length`:** All negative t-statistic values of -2.7, -3.66, -2.92, and p-values of 0.01, 0.0, 0.01, show that longer average sentence length correlates with scores more different from human grading.
- **Syntax, Grammar, Convention, with `spelling_errors`:** All negative t-statistic values of -2.08, -2.06, -3.11 and p values of 0.04, 0.04, 0.0 suggest that fewer spelling errors correspond with scores closer to human grading.
- **Vocabulary, Grammar, Convention with `lexical_diversity`:** All positive t-statistic values of 2.35, 3.94, 3.23 and p-values of 0.02, 0.0, 0.0 indicate that higher lexical diversity is associated with scores similar to human grading.

Summary Hypothesis testing illustrates that textual features strongly impact whether GPT models would perform similar to human in essay grading. Most correlations are intuitively straightforward. Generally, both GPT (without grading samples) and GPT Prompt (with grading samples) evaluate essays with fewer spelling errors, more sentence and word count, more frequent use of rare words, and more complex lexical diversity in a manner similar to human grading. And these are the more technically objective characteristics of a high-quality essay - long and comprehensive with diverse language use and minimal errors. Regarding the more subjective features like average sentence length, GPT models and human grade differently, which is also intuitive as it is common for different human graders as well - some would value short and clear expressions while others prefer long sentences that could demonstrate proficiency in Phraseology, Grammar, and Syntax. Overall, GPT models share more common grading criteria with humans on objective or technically relative aspects of outstanding essays.

Discussion

DeBERTa

Simpler sentences generally reduce cognitive load and are easier to process. Humans are likely to find them clearer, possibly reflecting a cognitive bias towards simplicity and clarity in communication. DeBERTa's alignment with human scoring in these instances could reflect its effectiveness in mimicking this aspect of human cognitive processing, where simplicity is associated with comprehensibility and correctness.

Shorter texts might be less demanding in terms of information processing, allowing both humans and models to focus more on individual word choice and usage. Thus, DeBERTa's performance might mirror a human-like processing efficiency, where fewer words allow for more focused and accurate assessment of vocabulary.

These findings underscore that while DeBERTa can effectively approximate human scoring in certain contexts, its performance varies significantly based on textual characteristics that also influence human cognitive processing. This variability is particularly evident in how the model handles complexity and information density, which are critical factors in human reading and comprehension. Understanding these patterns helps in refining the model's algorithms to better align with human cognitive behaviors, enhancing both its accuracy and its utility as a tool for educational assessment.

DistilBERT

Our results have shown that while DistilBERT slightly underperforms compared to DeBERTa across the six evaluated linguistic aspects, it nonetheless maintains respectable

accuracy across all measures. Notably, both models excel in assessing vocabulary, as evidenced by their lowest RMSE scores in this category. This proficiency underscores the models' capability to effectively evaluate word usage, which is critical in grading essays where vocabulary richness and appropriateness are key metrics.

Conversely, both models struggle with grammar and cohesion, areas that consistently present the highest RMSE scores. This pattern suggests inherent challenges in automated systems' ability to interpret and assess more complex linguistic constructs, which may involve nuanced grammatical relationships and the logical flow of ideas. Such difficulties highlight potential limitations in current NLP technologies and underscore the necessity for advancements in understanding contextual and syntactical nuances within text.

Hypothesis testing further elucidates how specific textual features impact the grading discrepancies between DistilBERT and human grading. Our findings reveal significant correlations, particularly in vocabulary assessment, where essays with higher word counts align more closely with human grading, suggesting a model bias towards longer texts. Additionally, essays with lower lexical diversity and a higher use of rare words also tend to receive grades that more closely mirror human assessment, indicating that certain types of word usage may either confound or enhance the model's accuracy.

Moreover, the correlation of longer essays with more accurate grading in phraseology and the influence of spelling accuracy in conventions both point to a broader trend: both DistilBERT and human graders may favor essays that adhere more strictly to conventional writing norms, such as longer lengths and fewer spelling errors.

GPT-3.5

Overall, GPT-3.5 models tend to grade the essays more harshly across all six aspects - Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Convention - compared to DeBERTa and DistilBERT models. The grading is also largely different from human grading, resulting in high RMSE scores. Since we cannot technically train GPT-3.5 like DeBERTa and DistilBERT to fit high school grading standards through the Chat Completetion endpoint, but can only prompt it to grade the essays based on high school standards and provide a few sample gradings as reference, the overall grading standard is set by GPT itself. Thus, the rubric might not strictly follow 8th - 12th grade standards, and points deduction might be more harshly applied. However, it is worth noting that the two GPT-3.5 models we employed both grade with more consistency, as shown by smaller standard deviations of the scores. Also, we discovered through hypothesis testing that GPT models hold similar grading criteria with humans on objective or technically relative aspects of high-quality essays, such as more complex lexical diversity, fewer spelling errors, and more frequent use of rare words. Well-written essays generally exhibit much smaller differences in scores between GPT and human graders.

Since one of our models is not given grading samples, and the other model is only provided with 5 grading samples, they are almost entirely AI-based English essay evaluation models. These strict grading standards of GPT may be attributed to extensive training on large amounts of language data, which equips it with a thorough understanding of grammar, syntax, and structure. This makes the models highly sensitive to structural and grammatical errors. As a result, minor issues that might be overlooked by human graders would still likely be detected by GPT and result in point deductions, leading to the overall harsher grading. However, unlike human graders, these models lack the flexibility to freely adjust their grading standards to align with students' levels. Human graders can consider the context, student progress, and the goals of writing during essay assessment, all of which are measures that GPT models cannot fully integrate into their grading processes

Conclusion

In conclusion, our analysis of the DeBERTa, DistilBERT, and GPT-3.5 models offers significant insights into the capabilities and limitations of automated essay grading systems. DeBERTa and DistilBERT exhibit proficiency in assessing vocabulary, mirroring human-like cognitive processing by emphasizing simplicity and clarity, which enhances comprehensibility. However, both models encounter challenges with more complex linguistic constructs like grammar and cohesion, reflecting inherent difficulties in automated systems' understanding of nuanced grammatical relationships and the logical flow of ideas. This highlights a crucial area for future enhancements in NLP technologies.

GPT-3.5, while grading more harshly and differently from human assessors, shows a consistent application of grading criteria, particularly in detecting structural and grammatical issues due to its extensive training on language data. Nevertheless, its strict grading standards underscore the need for flexibility in automated grading systems to better mimic human graders who consider context and individual student progress in their assessments.

Overall, our study reveals that while automated essay grading systems can approximate human scoring in certain contexts, their efficacy is closely tied to the textual characteristics and complexity. Understanding these nuances and integrating them into the model's algorithms is essential for improving their accuracy and utility in educational settings. Future advancements should focus on enhancing the models' ability to interpret complex syntactical and contextual

nuances, aiming to reduce grading discrepancies and better align with human cognitive processes.

References

Mikk, J. (2008, May). Sentence length for revealing the cognitive load reversal effect in text comprehension. *Educational Studies*, *34*(2), 119–127. doi: 10.1080/03055690701811164

Schluroff, M. (1982, Sep). Pupil responses to grammatical complexity of sentences. *Brain and Language*, *17*(1), 133–145. doi: 10.1016/0093-934x(82)90010-4

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer New York.