Using Linguistic Features to Predict the Response Process Complexity Associated with Answering Clinical MCQs

Victoria Yaneva[1] **Daniel Jurich**[1] **Le An Ha**[2] **Peter Baldwin**[1]

[1]National Board of Medical Examiners, Philadelphia, USA
{vyaneva, djurich, pbaldwin}@nbme.org
[2]University of Wolverhamton, UK
ha.l.a@wlv.ac.uk

Abstract

This study examines the relationship between the linguistic characteristics of a test item and the complexity of the response process required to answer it correctly. Using data from a large-scale medical licensing exam, clustering methods identified items that were similar with respect to their relative difficulty and relative response-time intensiveness to create *low response process complexity* and *high response process complexity* item classes. Interpretable models were used to investigate the linguistic features that best differentiated between these classes from a descriptive and predictive framework. Results suggest that nuanced features such as the number of ambiguous medical terms help explain response process complexity beyond superficial item characteristics such as word count. Yet, although linguistic features carry signal relevant to response process complexity, the classification of individual items remains challenging.

1 Introduction

The success of high-stakes exams, such as those used in licensing, certification, and college admission, depends on the use of items (test questions) that meet stringent quality criteria. To provide useful information about examinee ability, good items must be neither too difficult, nor too easy for the intended test-takers. Furthermore, the timing demands of items should be such that different exam forms seen by different test-takers should entail similar times to complete. Nevertheless, while an extreme difficulty or mean response time can indicate that an item is not functioning correctly, within these extremes variability in difficulty and item response time is expected. For good items, it is hoped that this variability simply reflects the breadth and depth of the relevant exam content.

The interaction between item *difficulty* (as measured by the proportion of examinees who respond correctly) and *time intensiveness* (as measured by the average time examinees spend answering) can help quantify the complexity of the response process associated with an item. This is valuable, since the more we know about the way examinees think about the problem presented in an item, the better we can evaluate exam validity. Although easier items usually require less time than difficult items, the interaction between these two item properties is not strictly linear – examinees may spend very little time responding to certain difficult items and, likewise, examinees may spend a great deal of time on items that are relatively easy. The idea of response process complexity is best illustrated with items that have similar difficulty but different mean response times. In such cases, one item may require the formation of a complex cognitive model of the problem and thus take a long time, while another item with a similar level of difficulty may require factual knowledge that few examinees recall (or that many recall incorrectly) and thus take a short time on average. The interaction between item difficulty and time intensity can therefore provide valuable information about the complexity of the response process demanded by an item, which, we argue, can be further explained by examining the linguistic properties of the item.

In this paper, we use a data-driven approach to capture the interaction between item difficulty and response time within a pool of 18,961 multiple-choice items from a high-stakes medical exam, where each item was answered by 335 examinees on average. For our data, this resulted in the definition of two clusters, one of which consisted of items that are relatively easy and less time-intensive, and another one which consisted of items that are relatively difficult and/or time-intensive. For the purposes of this study, we name these two clusters *low-complexity* class and *high-complexity* class, respectively. The use of the term *response process*

Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 223–232

A 16-year-old boy is brought to the emergency department because of a 2-day history of fever, nausea, vomiting, headache, chills, and fatigue. He has not had any sick contacts. He underwent splenectomy for traumatic injury at the age of 13 years. He has no other history of serious illness and takes no medications. He appears ill. His temperature is 39.2°C (102.5°F), pulse is 130/min, respirations are 14/min, and blood pressure is 110/60 mm Hg. On pulmonary examination, scattered crackles are heard bilaterally. Abdominal shows a well-healed midline scar and mild, diffuse tenderness to palpation. Which of the following is the most appropriate next step in management?
(A) Antibiotic therapy (B) Antiemetic therapy
(C) CT scan of the chest (D) X-ray of the abdomen
(E) Reassurance

Table 1: An example of a practice item

complexity here is not based on an operational definition of this construct, which would require extensive research on its own, but rather, as a succinct label that summarises the differences between the two classes along the interaction of empirical item difficulty and item time intensiveness.

Studying the linguistic characteristics of these two categories may help test developers gain a more nuanced understanding of how cognitively complex items differ from those with a straightforward solution. Provided that strong relationships are found, such insight can also be used to guide item writers or inform innovative automated item generation algorithms when seeking to create high- or low-complexity items. For this reason, our goal is not to train a black-box model to predict item complexity; instead, our goal is to isolate interpretable relationships between item text and item complexity that can inform our understanding of the response process and provide better item-writing strategies.

In addition to its utility for improving high-stakes exams, the problem of modeling response process complexity is interesting from an NLP perspective because it requires the modeling of cognitive processes beyond reading comprehension. This is especially relevant for the data used here because, as we explain in Section 3 below, the items in our bank assess expert-level clinical knowledge and are written to a common reading level using standardized language.

Contributions: i) We use unsupervised clustering to define classes of high and low response-process complexity from a large sample of items and test-takers in a high-stakes medical exam; ii) the study provides empirical evidence that linguistic characteristics carry signal relevant to an item's response process complexity; iii) the most predictive features are identified through several feature selection methods and their potential relationship to response process complexity is discussed; iv) the errors made by the model and their implications for predicting response process complexity are analysed.

2 Related Work

This section discusses related work on the topics of modeling item difficulty and response time.

Most NLP studies modeling the difficulty of test questions for humans have been conducted in the domain of reading comprehension, where the readability of reading passages is associated with the difficulty of their corresponding comprehension questions (Huang et al., 2017; Beinborn et al., 2015; Loukina et al., 2016). For other exams, taxonomies representing knowledge dimensions and cognitive processes involved in the completion of a test task have been used to predict the difficulty of short-answer questions (Padó, 2017) and identify skills required to answer school science questions (Nadeem and Ostendorf, 2017). Difficulty prediction has also been explored in the context of evaluating automatically generated questions (Alsubait et al., 2013; Ha and Yaneva, 2018; Kurdi, 2020; Kurdi et al., 2020) through measures such as question-answer similarity.

Response time prediction has mainly been explored in the field of educational testing using predictors such as item presentation position (Parshall et al., 1994), item content category (Parshall et al., 1994; Smith, 2000), the presence of a figure (Smith, 2000; Swanson et al., 2001), and item difficulty and discrimination (Halkitis et al., 1996; Smith, 2000). The only text-related feature explored in these studies was *word count*, and it was shown to have a very limited predictive power in most domains.

Several studies have explored the prediction of item difficulty and response time in the context of clinical multiple choice questions (MCQs). Ha et al. (2019) propose a large number of linguis-

tic features and embeddings for modeling item difficulty. The results show that the full model outperforms several baselines with a statistically significant improvement, however, its practical significance for successfully predicting item difficulty remains limited, confirming the challenging nature of the problem. Continuations of this study include the use of transfer learning to predict difficulty and response time (Xue et al., 2020), as well as using predicted difficulty for filtering out items that are too easy or too difficult for the intended examinee population (Yaneva et al., 2020). Baldwin et al. (2020) used a broad range of linguistic features and embeddings (similar to those in Ha et al. (2019)) to predict item response time, showing that a wide range of linguistic predictors at various levels of linguistic processing were all relevant to response-time prediction. The predicted response times were then used in a subsequent experiment to improve fairness by reducing the time intensity variance of exam forms.

3 Data

The data[1] used in this study comprises 18,961 Step 2 Clinical Knowledge items from the United States Medical Licensing Examination (USMLE®), a large-scale high-stakes medical assessment. All items were MCQs. An example practice item[2] is given in Table 1. The exam comprises several one-hour testing blocks with 40 items per block. All items test medical knowledge and are written by experienced item-writers following guidelines intended to produce items that vary in their difficulty and response times only due to differences in the medical content they assess. These guidelines stipulate that item writers adhere to a standard structure and avoid excessive verbosity, extraneous material not needed to answer the item, information designed to mislead the test-taker, and grammatical cues (e.g., correct answers that are more specific than the other options). All items were administered between 2010 and 2015 as pretest items and presented alongside scored items on operational exams. Examinees were medical students from accredited US and Canadian medical schools taking the exam for the first time and had no way of knowing which items were pretest items and which were scored. On average, each item was attempted by 335 examinees (SD = 156.8).

3.1 Identifying items with high and low response process complexity

We base our definition of the two classes of items on empirical *item difficulty* and *time intensity*. Item difficulty is measured by the proportion of examinees who answered the item correctly, a metric commonly referred to by the educational testing community as *p-value* and calculated as follows:

$$P_i = \frac{\sum_{n=1}^{N} U_n}{N},$$

where P_i is the p-value for item i, U_n is the 0-1 score (incorrect-correct) on item i earned by examinee n, and N is the total number of examinees in the sample. Thus, difficulty measured in this way ranges from 0 to 1 and higher values correspond to easier items.

Time intensity is found by taking the arithmetic mean response time, measured in seconds, across all examinees who attempted a given item. This includes all time spent on the item from the moment it is presented on the screen until the examinee moves to the next item, as well as any revisits.

To assign items to classes, p-value and mean response time are rescaled such that each variable has a mean of 0 and a standard deviation of 1. Moreover, we use two quantitative methods to categorize items and retain only those items where there was agreement between the two methods.

Method 1: Items were classified by applying a K-means clustering algorithm via the `kmeans` function in Python's `Scikit-learn` (Pedregosa et al., 2011). K-means is an unsupervised data classification technique that discovers patterns in the data by assigning instances to a pre-defined number of classes (Wagstaff et al., 2001). This approach also allows us to evaluate the plausibility of categorizing items into more than two complexity classes, or whether the items fail to show any meaningful separation along the interaction of p-value and duration (one class). Results suggest that two classes best fit these data and identified 11,067 items as low complexity and 7,894 items as high complexity[3].

[1] The data cannot be made available due to exam security considerations.

[2] Source: https://www.usmle.org/pdfs/step-2-ck/2020_Step2CK_SampleItems.pdf

[3] We also experimented with hierarchical clustering, which led to similar results. The hierarchical clustering dendrogram suggested that there are meaningful distances between two clusters in the data, and much smaller distances between a higher number of more fine-grained clusters.

Method 2: Any item with a rescaled p-value greater than its rescaled mean response time – indicating that the item is relatively easier than it is time-consuming – is classified as low-complexity (11,682 items). Likewise, the remaining items, which had rescaled p-values less than their rescaled mean response times, were assigned to the high-complexity class (7,279 items). Put another way, if an item takes less time than we would expect given its difficulty, the item is classified as *low response process complexity* and if it takes more time than we would expect, it is classified as *high response process complexity*.

The two methods achieved strong agreement, with only 673 (3.5%) items being assigned to different classes across methods. These discrepant items are excluded, leaving a total of 18,288 items for further analysis: 11,038 low-complexity items and 7,250 high-complexity ones. Figure 1 shows the class assignment, p-value, and mean response time for each item.

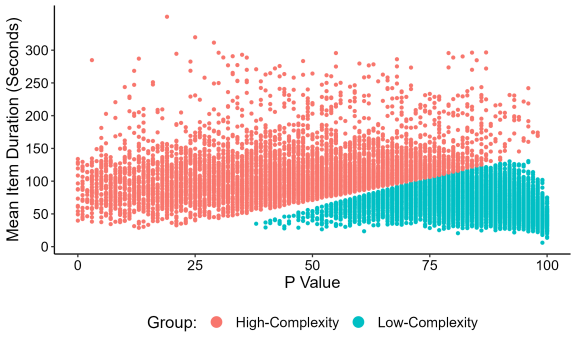

Figure 1: Class assignment by p-value and response time for each item. Note that discrepant items were excluded, as illustrated by the gap between the two class distributions.

As can be seen from the figure, the class of *low-complexity* items was dense and homogenous compared to the *high-complexity class*, meaning that it contained a large number of easy items whose response times were always below 125 seconds. The high-complexity class on the other hand was highly heterogeneous, with items whose response times and p-values spanned almost the entire scale.

4 Features

We use a set of interpretable linguistic features, many of which were previously used for predicting item difficulty (Ha et al., 2019) and response time (Baldwin et al., 2020) in the domain of clinical MCQs. These features were extracted using code made available by Ha et al. (2019) and to these, we add several predictors specifically related to the medical content of the items, as well as standard item metadata.

4.1 Linguistic features

As noted, this study replicates the feature extraction procedure described and made available by Ha et al. (2019). Approximately 90 linguistic features were extracted from each item's text (the full item including answer options) and are summarized in Table 2. They span several levels of linguistic processing including surface lexical and syntactic features, semantic features that account for ambiguity, and cognitively motivated features that capture properties such as imageability and familiarity. Common readability formulae are used to account for surface reading difficulty. The organization of ideas in the text is captured through text cohesion features that measure the number and types of connective words within an item. Finally, word frequency features (including threshold frequencies) measure the extent to which items utilize frequent vocabulary.

Combinations of these features have the potential to capture different aspects of item content that are relevant to response complexity. For example, medical terms can be expected to have lower absolute frequencies and familiarity ratings, among other characteristics, and combinations of these features may suggest a higher density of medical terms and specialized language in some items compared to others. Another example is the temporal organization of the information about the patient history and symptoms described in the item and captured by temporal connectives, where it is reasonable to expect that more temporally intricate cases would require higher response process complexity to solve. Similarly, a high number of causal connectives would indicate a higher complexity of causal relationships among the events that led to the patient seeing a doctor, which may also be associated with higher cognitive demands.

4.2 Clinical content features

This group of features relates to the *medical content of the items* by mapping terms and phrases in the text to medical concepts contained in the Unified Medical Language System (UMLS) Metathesaurus (Schuyler et al., 1993) using Metamap (Aronson, 2001). The number of UMLS terms that appear in an item may indicate the amount of medical content

Group	N	Summary of features	Resources
Lexical	5	Word Count, Content word count, Content word count without stop-words, Average word length in syllables, Complex word count	
Syntactic	29	POS count, Phrase count (for each POS), Type count, Comma count, Average phrase length, Negation, Type-token ratio, Average sentence length, Average depth of tree, Clause count (relative, conditional), Average number of words before the main verb, Passive-active ratio, Proportion active VPs, Proportion passive VPs, Agentless passive count	Stanford NLP Parser (Manning et al., 2014)
Semantic	11	Polysemic word count, Average senses for: content words, nouns, verbs, adjectives, auxiliary verbs, adverbs; Average noun/verb distance to WordNet root, Average noun-and-verb distance to WordNet root, Answer words in WordNet ratio	WordNet (Miller, 1995)
Readability	7	Flesch Reading Ease, Flesch-Kincaid grade level, Automated Readability Index, Gunning Fog, Coleman Liau, SMOG, SMOG Index	See Dubay (2004) for definitions
Cognitive	14	Absolute values, ratios, and ratings for Concreteness, Imageability, Familiarity, Age of acquisition, Meaningfulness (Colorado norms), Meaningfulness (Paivio norms)	MRC Psycholinguistic Database (Coltheart, 1981)
Frequency	10	Average frequency (relative, absolute and rank) for all words and for content words; Threshold frequencies for words not in the first 2,000/3,000/4,000/5,000 most common words	British National Corpus (Leech et al., 2014)
Cohesion	5	Counts of Temporal, Causal, Additive connectives and All connectives; Referential pronoun count	

Table 2: Linguistic features extracted for each item following Ha et al. (2019)

the item contains (note that a given term found in the items can refer to multiple UMLS concepts).

First, we ask: *how many of the words and phrases in the items are medical terms?* This information is captured by *UMLS Terms Count*, indicating the number of terms in an item that appear in the UMLS wherein each instance of a given term contributes to the total count, as well as *UMLS Distinct Terms Count:* the number of terms in an item that appear in the UMLS wherein multiple instances of a given term contribute only once to the total count. The same kinds of counts are done for medical phrases – *UMLS Phrases Count* refers to the number of phrases in an item. For example, Metamap maps *'ocular complications of myasthenia gravis'* to two phrases: the noun phrase *'ocular complications'* and the prepositional phrase *'of myasthenia gravis'* (Aronson, 2001).

Next, we introduce features that measure the ambiguity of medical terms within the items. These include *Average Number of Competing UMLS Concepts Per Term Count*, which captures the average number of UMLS concepts that a term could be referring to, averaged for all terms in an item, and weighted by the number of times Metamap returns the term. A similar version of this feature but without weighting by the number of times Metamap returns the term is *Average Number of UMLS Concepts Per Term Count*. This metric is then computed at the level of sentences and items, resulting in: *Average Number of UMLS Concepts per Sentence*, which measures the medical ambiguity of sentences and *UMLS Concept Count*, which measures item medical ambiguity through the total number of UMLS concepts all terms in an item could refer to. Finally, *UMLS concept incidence* refers to the number of UMLS concepts per 1000 words.

4.3 Standard Item Features

This group of features refers to metadata describing item content. *Presence of an image* is a binary categorical variable indicating whether the item includes an image such an X-ray or an MRI that needs to be examined. Another variable is *Content category*, which describes 18 generic topic categories such as "Cardiovascular", "Gastrointestinal", "Behavioral Health", 'Immune System", and so on. Another variable, *Physician Task* describes tasks required by the item, e.g., determine a diagnosis, choose the correct medicine, apply foundational science concepts, and others. Finally, we also include the *Year* the item was administered as a predictor (2010 - 2015) to account for potential changes in response process complexity and examinee samples over time.

4.4 Classification

This section describes three baseline models (Section 4.5), the training of classifiers using the full feature set (Section 4.6), and the feature selection procedures (Section 4.7).

	Logistic regression			Random forests		
	Precision	**Recall**	**Weighted F1**	**Precision**	**Recall**	**Weighted F1**
Majority class	0.37	0.6	0.46	0.37	0.6	0.46
Word count	0.57	0.6	0.48	0.56	0.6	0.54
Standard item features	0.62	0.63	0.59	0.59	0.59	0.59
Full feature set	0.64	0.65	0.63	0.68	0.68	**0.67**
Selected linguistic features	0.63	0.65	0.63	0.67	0.68	**0.66**

Table 3: Weighted F1 scores for different models on the test set

4.5 Baseline Models

Three classification baselines were computed to benchmark the predictive benefit given by linguistics features over standard item characteristics:

Majority Class Baseline: Since the low-complexity class contains a higher number of items, it is more likely that an item would be correctly predicted as belonging to this class.

Word Count: This baseline examines the possibility that response process complexity is simply a function of item length.

Standard Item Features: This baseline comprises *Word count, Presence of an image, Content category, Physician task* and *Year*. This model reflects the standard item characteristics that most testing organizations would routinely store.

4.6 Full feature models

After scaling the features, two models were fit using Python's `scikit-learn` library and the full set of features: a logistic regression model and a random forests one (400 trees). Twenty percent of the data (3,658 items) were used as a test set.

4.7 Feature selection

Feature selection was undertaken to better understand which features were most strongly associated with class differences. The selection process utilized three distinct strategies, where the final set of selected features comprises only those features retained by all three methods. After applying feature selection to the training set, the predictive performance of the selected features is evaluated on the test set and compared to the performance of the full feature set and the baseline models outlined above.

Embedded methods: The first method is LASSO regularized regression wherein the coefficients of variables that have low contributions towards the classification performance are shrunk to zero by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value. We use the LassoCV algorithm with 100-fold cross validation and maximum iterations set to 5,000.

Wrapper methods: We next apply recursive feature elimination, performed using two different classification algorithms: random forests classifier (400 trees, step = 5) and gradient boosting classifier (Friedman, 2002) (default parameters, step = 5).

The final set of selected linguistic features comprised 57 features that were retained by all three strategies. These features and their evaluation are discussed in sections 5 and 7.

5 Results

Table 3 presents the classification results for the baselines, the full feature set, and the selected features for both logistic regression and random forests. Results are reported using a weighted F1 score, which is a classification accuracy measure based on the mean between the precision and recall after adjusting for class imbalance.

The linguistic and clinical content features improve predictive accuracy above the baselines, yielding a higher F1 score than the strongest baseline (.67 compared to .59). The reduced feature set does not lead to a meaningful performance drop compared to the full feature set, suggesting that no signal was lost due to feature elimination.

Figure 2 reports the eight best-performing features: *UMLS phrases count, Unique word count, Polysemic word count, Average noun phrase length, Automated readability index, Prepositional phrases, UMLS distinct terms count,* and *Concreteness ratio*.

6 Error analysis

The output of the selected-features prediction model was analyzed further in order to get insight into this model's performance. As could be expected, the majority class of low-complexity items was predicted more accurately than the high-complexity class, as shown by the confusion matrix in Table 4. An interesting observation was made during a follow-up classification experiment, which showed that this effect remained when using

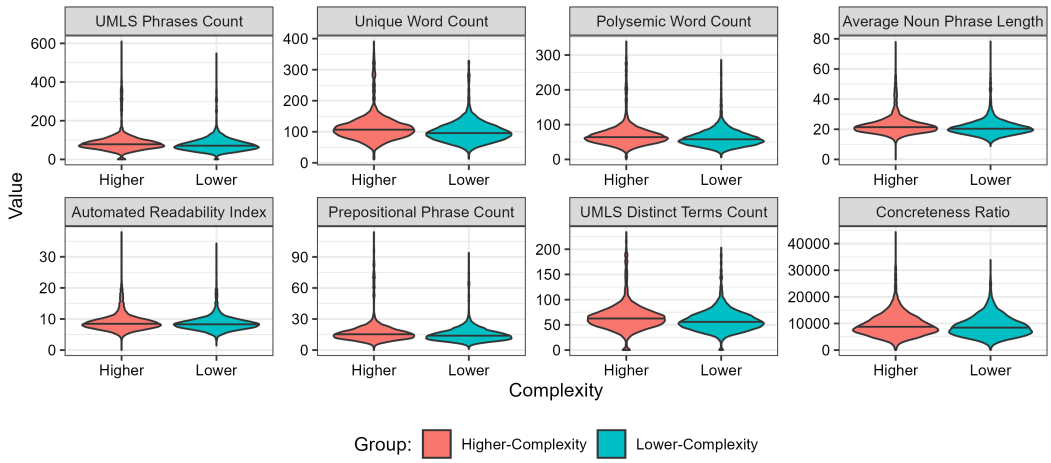

Figure 2: Distributions and median values for the top eight features by group.

balanced classes[4]. This shows that the success in predicting this class cannot be attributed solely to its prevalence but potentially also to its high homogeneity compared to the high-complexity class.

	High Complexity	Low Complexity
High Complexity	617	828
Low Complexity	332	1881

Table 4: Confusion matrix for the results from the selected features model using random forests (F1 = 0.66)

Next, we plot the model errors across the two classes of low-complexity and high-complexity items, as shown in Figure 3. Notably, items with average response times below 150 seconds were predicted as low-complexity most of the time, with minimal consideration of their p-value. This shows that what the model effectively learned was to distinguish between items with long and short mean response times, which overpowered its ability to predict the p-value parameter. This finding is consistent with previous work, where response times in Baldwin et al. (2020) were predicted more successfully than p-value using a similar set of linguistic features in Ha et al. (2019). Finally, analysis of the feature distributions across these four classes revealed no unexpected patterns.

7 Discussion

The results presented in the previous section lead to three main findings: i) the linguistic characteristics of the items carry signal relevant to response process complexity; ii) no individual features stand out as strong predictors, and iii) the most important features were those related to syntax and semantics.

The first of these findings relates to the fact that the linguistic characteristics of the items carry signal that is predictive of response process complexity, revealing that the problems posed by low-complexity and high-complexity items are described using slightly different language. While this signal outperformed several baselines, the overall low predictive utility of the models suggests that there are other factors, yet to be captured, that have a significant effect on response process complexity.

The retention of 56 features indicates that individual linguistic predictors provide a weak classification signal but, taken together, they complement each other in a way that ultimately provides a higher accuracy. The fact that there are many predictive features with none standing out is also a positive evaluation outcome for item writing quality, as it shows that the response process complexity associated with an item is not distributed along a small number of linguistic parameters.

The most important features that helped with classification were those related to syntax and semantics (Figure 2). The poor performance of the *Word Count* baseline suggests that differences in response process complexity cannot be explained solely by item length and that more complex linguistic features capture some of the nuance in the response process. As can be seen in Figure 2, high-complexity items contain a slightly higher number of UMLS phrases and (distinct) medical terms, as well as a higher number of unique words. These features suggest high-complexity items re-

[4]Classes were balanced using the `balanced_subample` setting of the `class_weight` parameter in Scikit-learn's `RandomForrestClassifier`

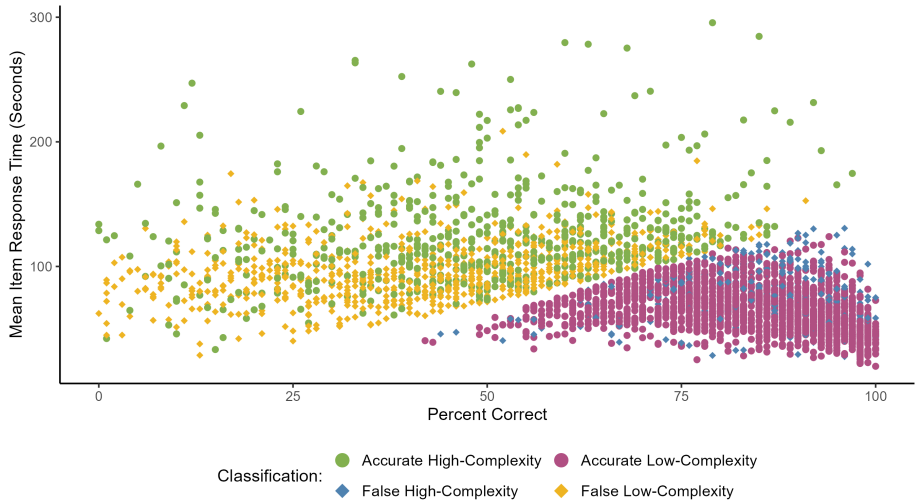

Figure 3: Error distribution for the two classes

peat words less frequently and may contain a higher concentration of new information and specialized terminology than low-complexity items. The individual phrases in high-complexity items are also slightly longer, which naturally influences readability metrics that are based on word and sentence length, such as the *Automated Readability Index* (higher values are indicative of a more complex text). Prepositional phrases were also identified as more important than other phrase types in distinguishing between response process complexity. Prepositional phrases often serve as modifiers of the primary noun phrase and the higher number of prepositional phrases in the high-complexity items suggests the use of more specific descriptions (e.g., "small cell carcinoma of the ovary" instead of just "small cell carcinoma"). The words contained in the high-complexity items also have slightly higher concreteness levels, providing another indication that they may contain more terms, as terms tend to be more concrete than common words. Finally, the words contained in the high-complexity items also tend to have more possible meanings, as indicated by the polysemous word count variable, which results in higher complexity owing to disambiguation efforts. Overall, these features indicate that the language used in the low-complexity items is less ambiguous and descriptive, and potentially contains fewer medical terms.

One limitation of the study is the fact that it treats item difficulty and time intensiveness as independent variables. This may not always be the case, as examinees do employ strategies to optimize their time. Given finite time limits, examinees may ignore time intensive items if they believe the time needed for such items can be better utilized attempting other, less time intensive items. Therefore, the relationship between difficulty and response time and their association with item text would differ for exams that do not impose strict time limits.

When using data-driven approaches to defining item classes, our data did not lend itself to a categorization that would allow investigating high difficulty/low response time items and vice-versa. While the approach taken in this paper has a higher ecological validity, studying such cases in the future may lead to a greater understanding of various aspects of response process complexity and their relationship to item text. Other future work includes exploration of potential item position effects.

8 Conclusion

The experiments presented in this paper are, to the best of our knowledge, the first investigation of the relationship between item text and response process complexity. The results showed that such a relationship exists. To the extent that items were written as clearly and as concisely as possible, the findings suggest that high-complexity medical items generally include longer phrases, more medical terms, and more specific descriptions.

While the models outperformed several baselines, they required a large number of features to do so and the predictive utility remained low. Ultimately, this shows the challenging nature of modeling response process complexity using interpretable models and the lack of a straightforward way to manipulate this item property.

References

Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. 2013. A similarity-based theory of controlling mcq difficulty. In *e-Learning and e-Technologies in Education (ICEEE), 2013 Second International Conference on*, pages 283–288. IEEE.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Peter Baldwin, Victoria Yaneva, Janet Mee, Brian E Clauser, and Le An Ha. 2020. Using natural language processing to predict item response times and improve test construction. *Journal of Educational Measurement*.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2015. Candidate evaluation strategies for improved difficulty prediction of language tests. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

William H. Dubay. 2004. *The Principles of Readability*. Impact Information.

Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378.

Le An Ha and Victoria Yaneva. 2018. Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 389–398.

Le An Ha, Victoria Yaneva, Peter Balwin, and Janet Mee. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Perry N Halkitis et al. 1996. Estimating testing time: The effects of item characteristics on response latency.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question difficulty prediction for reading problems in standard tests. In *AAAI*, pages 1352–1359.

Ghader Kurdi. 2020. *Generation and mining of medical, case-based multiple choice questions*. Ph.D. thesis, PhD thesis, University of Manchester.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Anastassia Loukina, Su-Youn Yoon, Jennifer Sakano, Youhua Wei, and Kathy Sheehan. 2016. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3245–3253.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Farah Nadeem and Mari Ostendorf. 2017. Language based mapping of science assessment items to skills. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 319–326.

Ulrike Padó. 2017. Question difficulty–how to estimate without norming, how to use for automated grading. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10.

Cynthia G Parshall et al. 1994. Response latency: An investigation into determinants of item-level timing.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. 1993. The umls metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217.

Russell Winsor Smith. 2000. An exploratory analysis of item parameters and characteristics that influence item level response time.

David B Swanson, Susan M Case, Douglas R Ripkey, Brian E Clauser, and Matthew C Holtman. 2001. Relationships among item characteristics, examine characteristics, and response times on usmle step 1. *Academic Medicine*, 76(10):S114–S116.

Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. 2001. Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584.

Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 193–197.

Victoria Yaneva, Le An Ha, Peter Baldwin, and Janet Mee. 2020. Predicting item survival for multiple choice questions in a high-stakes medical exam. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6812–6818.