Using natural language processing to predict item response times and improve test construction

Using Natural Language Processing to Predict Item Response Times and Improve Test Construction

Peter Baldwin, Victoria Yaneva, Janet Mee, and Brian E. Clauser
National Board of Medical Examiners
Le An Ha
University of Wolverhampton

In this article, it is shown how item text can be represented by (a) 113 features quantifying the text's linguistic characteristics, (b) 16 measures of the extent to which an information-retrieval-based automatic question-answering system finds an item challenging, and (c) through dense word representations (word embeddings). Using a random forests algorithm, these data then are used to train a prediction model for item response times and predicted response times then are used to assemble test forms. Using empirical data from the United States Medical Licensing Examination, we show that timing demands are more consistent across these specially assembled forms than across forms comprising randomly-selected items. Because an exam's timing conditions affect examinee performance, this result has implications for exam fairness whenever examinees are compared with each other or against a common standard.

The time needed to respond to test questions has implications both for the logistics of exam administration and for fairness. When tests are administered in commercial test centers, time limits must be established so that seat time can be scheduled in advance. More generous time limits increase the cost of the administration; however, less generous time limits may differentially affect examinees. Moreover, these differential effects may be more pronounced when different forms of an examination are not parallel with respect to the time demands they place on examinees. For these reasons, response time has been a longstanding interest of measurement specialists, which has led to various test assembly strategies that are designed to meet timing objectives.

In the context of computer adaptive testing (CAT), efforts to tailor item sets so that they are equally speeded have been proposed by several authors (e.g., Fan, Wang, Chang, & Douglas, 2012; van der Linden, 2005, 2009; van der Linden, Scrams, & Schnipke, 1999). Generally, this has been accomplished by modeling the labor required by test questions and the speed at which examinees labor. The approach proposed by van der Linden (2009) for use in adaptive testing settings has also been extended to include fixed-form construction (van der Linden, 2011). Scrams and Smith (2010) also recommended an approach that could be used for constructing fixed form tests.

These approaches may be useful, but they require that item response times are known prior to the administration. This requirement will be satisfied in some cases; however, more often some or all of the items within a timed section of the test will have unknown response times. This occurs because of the need to collect data on new items. When these new items are presented alongside scored items, variability in the time required to complete the new items may affect performance on the scored items. Even when new items are piloted in separate unscored sections variability in the time requirements may affect the estimated item parameters making them less useful in future test construction.

If item response times could be accurately predicted, the predictions could be used during test assembly to reduce this uncertainty, which would create greater consistency in the work demands across test forms and/or separately timed sections. To this end, this paper presents two related studies. The first study identifies several sources of collateral data based on item text and evaluates their utility for predicting mean examinee response times on test items. In the second study we evaluate the practical usefulness of the prediction models for test construction.

Background

In 1953, the state of the art for collecting latency data was to have a proctor or an assistant hold up signs displaying elapsed time, which examinees then could record next to their answers (Ebel, 1953). Although it was reasonably effective, this method had the inconvenient requirements that the displayed times be updated every 10 seconds and that examinees interrupt their test taking after each item to record the elapsed time. Readers will not be surprised to learn that this strategy was not widely adopted. It wasn't until response time data could be collected automatically, reliably, and routinely through computer-based testing that response-time research received broad attention.

In addition to item selection in the context of CAT and form construction, which were mentioned above, response-time data have been used in many applications since becoming commonly available. These include: the identification of aberrant response patterns (Fitzgerald, Fremer, & Maynes, 2005; Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; Miller, 1986; van der Linden & van Krimpen-Stoop, 2003); the modeling of examinee motivation (Wise, 2017; Wise & DeMars, 2006; Wise & Kong,

2005); improvements in scoring (Tatsuoka & Tatsuoka, 1979; Thissen, 1983; Wang & Hanson, 2005); the identification of item features that affect response time (Halkitis, Jones, & Pradhan, 1996; Smith, 2000); and the study of differential time intensiveness across examinee subgroups (Zenisky & Baldwin, 2006). There has also been some research into *predicting* response times using collateral item data; however, work in this area has been relatively modest (e.g., Halkitis et al., 1996; Parshall, Mittelholtz, & Miller, 1994; Smith, 2000) and by and large these studies have been preoccupied with identifying effective predictors rather than using predicted values to improve practice.

To date, researchers have focused on a narrow range of predictors: item presentation position (Parshall et al., 1994), item content and/or cognitive classification (Parshall et al., 1994; Smith, 2000), the presence of a figure (Smith, 2000; Swanson, Case, Ripkey, Clauser, & Holtman, 2001), item difficulty and discrimination (Halkitis et al., 1996; Smith, 2000), and word count (Halkitis et al., 1996; Smith, 2000; Swanson, et al., 2001). Of these previously studied predictors, only content/cognitive classification, word count, and the presence of a figure are applicable for improving test construction. As will be seen below, for our purposes here we excluded item difficulty and discrimination from consideration because we took the view that if data were collected such that difficulty and discrimination could be estimated, so too could response time thereby making the problem of predicting response time irrelevant. For this reason, the present study focuses on item collateral data that can be known prior to pretesting. We evaluated 113 linguistic features, 16 measures related to information-retrieval-based, automatic-question-answering technology, and the representation of items as dense vectors (word embeddings). Since linguistic features, information retrieval, and word embeddings may be unfamiliar to some readers the following sections provide a brief introduction to these topics.

Linguistic Features

Our decision to examine linguistic features was motivated by the hypothesis that items that are more linguistically complex would take more time to complete. Linguistic complexity is often referred to as *readability* (De Clercq & Hoste, 2016; DuBay, 2004; Flesch, 1948). The evolution of readability research has led to three distinct method types. Benjamin (2012) describes these as *traditional methods,*

methods inspired by cognitive science, and *methods based on the use of statistical language-modelling tools*.

Traditional methods were developed in the earliest stages of readability research, which focused on predicting linguistic complexity through vocabulary measures (e.g., Thorndike & Lorge, 1944) and specially calibrated formulae (e.g., Flesch, 1948; Gunning, 1952). The most commonly used features in these readability formulae are word length, sentence length, and vocabulary measures from word lists (e.g., the number of words from the text that are also in a list of "easy" words).

The second category, characterized by *cognitive methods*, was motivated by the idea that reading comprehension is determined by more sophisticated factors than those used by traditional methods. These include, for example, propositions and inferences (Kintsch, 1995), text coherence (Graesser, McNamara, Louwerse, & Cai, 2004), and word properties such as word concreteness, word familiarity, word imageability (the ease by which a mental image of a word is formed), and age of acquisition (the age at which a word is typically encountered) (Coltheart, 1981).

Finally, the methods based on statistical language-modelling emerged from developments in Natural Language Processing (NLP) that allowed faster and more robust extraction of features such as word frequency, part of speech, and word and sentence length. These advancements also included new algorithms for predicting text difficulty based on advances in the field of machine learning. In addition to calculating previously identified features more quickly, NLP capitalizes on novel features such as word or character sequences, syntactic complexity (e.g., the depth of syntax trees), the ambiguity of words based on machine-readable ontologies such as WordNet (Pedersen, Patwardhan, & Michelizzi, 2004), and the diversity of text represented by the ratio of number of unique words to the total number of words.

Information Retrieval and Question Answering

Our use of information from the field of automatic question answering is motivated by the literature showing that examinees generally take more time to answer more difficult questions (e.g., Swanson, et al., 2001). We hypothesized that items that are more difficult for automated systems may also be more difficult for humans and so, by extension, may be associated with longer response times.

Question answering represents a very challenging problem for automatic systems because in many cases it requires reasoning over facts and the more complex the questions, the lower the performance of such systems. Nevertheless, question-answering systems have been shown to perform well for relatively simple reading-comprehension questions or questions derived from various texts—e.g., the Stanford Question Answering Dataset (Rajpurkar, Zhang, Lopyrev, & Liang, 2016). A popular method for obtaining answers for such questions features a technique called information retrieval (Manning, Raghavan & Schütze, 2010). The functioning of an information-retrieval-based question-answering system is outlined in Gupta and Gupta (2012), where they describe a system that takes a user's query as input (e.g., "Who won the first Nobel Prize in Literature?") and returns an answer based on a pool of web documents. This is achieved by first filtering out irrelevant documents and paragraphs based on whether or not they contain keywords from the query. After that, the system searches for relevant sentences within the documents. This relevance is based on a measure of semantic similarity between the query and the identified sentences. Based on this measure, the candidate phrases are ranked and the candidate answer that has the highest semantic similarity with the query is proposed as the correct answer. As will be shown below, we apply a similar technique to multiple-choice questions as a potential way to quantify aspects of item difficulty. Specifically, we use the stem and each answer option as a query to search a database of text related to the content of our items. The premise behind this approach is two-fold: (a) that concepts occurring in both the stem and key are more likely to occur together in documents than concepts from the stem and the distractors; and (b) that a potentially higher similarity between the stem and the key would in turn result in documents with a higher similarity to the stem-and-key query. To the extent that this is true, the success of the automatic question-answering system will be related to these measures and, in this way, such measures are related to the item difficulty for the system.

Word Embeddings

While the extraction of linguistic features and their use as variables in machine-learning models has been the predominant research paradigm in NLP for several decades, approaches related to advances in *deep learning* (a specific type of machine learning that uses Artificial Neural Network algorithms

(Hassoun, 1995)) have become increasingly popular. This shift was made possible due to the accumulation of large amounts of digital-text data (e.g., billions of words on Google) and increases in computational power through parallel processing. In this new approach, words (and sometimes other linguistic units) are represented as dense vectors of real numbers called *embeddings*, which are "learned" using neural networks. The main linguistic theory behind the concept of word embeddings is the theory of distributional semantics, which specifies that words are defined through the company they keep with other words. Therefore, the vector representations of individual words are derived through modeling hundreds of thousands of co-occurrences of a given word in various contexts. Word embeddings capture and quantify a wide variety of information about words and longer units (including the quantification of some semantic relations) using vectors with a smaller number of dimensions compared to the vast number of dimensions derived through simple co-occurrence matrices.

Study I

This study focuses on predicting response time using the item characteristics described in the previous section. Specifically, we extract these features from an empirical dataset and then compare the effectiveness of different modeling approaches for building a prediction model. Study II will then provide a practical evaluation of the utility of using those predictions in test construction. The crucial feature of the characteristics selected for predicting response times is their availability prior to the first administration of the item.

Method

Data. A set of over 12,000 items that were administered across four years of the United States Medical Licensing Examination (USMLE®) Step 2 Clinical Knowledge (CK) exam were used in this study. Step 2 CK is part of a three-step examination sequence taken by physicians with an MD degree seeking licensure in the United States. All items were being pretested for the first time; response time and other performance information was unavailable prior to this administration. These items were selected because prior, unpublished research indicated that the distribution of response times for pretest items—some of which do not function well enough to be used for scoring—may be different than that for scored

items. Using this data set ensured that the evaluation of our prediction procedures was being carried out under realistic conditions. The sensitivity of scores from Step 2 CK to varying time limits has been studied by Harik et al. (2018). The evidence reported in that paper suggests that examinee performance can be affected by the amount of available time, so increasing the uniformity of the time demand across forms of the examination is a matter of practical importance.

Each Step 2 CK test form contains approximately 320 multiple-choice items that are administered in eight one-hour separately timed sections or blocks. Once a given block is completed, it cannot be revisited. A proportion of the items on each test form are unscored and reserved for pretesting of new items or other experimental uses. Most often, these items are presented alongside scored exam content; however, entire blocks are occasionally designated as experimental. Examinees are not told which items are scored and which are not. Examinees are randomly assigned to test forms and, for each examinee, items are randomly ordered within each block. In this way, mean response times for each item are directly comparable—i.e., the effects of examinee sample and item position are largely accounted for by the randomization.

All items were designed to test clinical knowledge and were written by experienced item-writers following a set of guidelines stipulating adherence to a standard structure. Items were also professionally edited and reviewed by content experts to ensure adherence to guidelines and accuracy of content. On average, each item was answered by 328 examinees (SD = 67). For this study we used item responses from examinees from US and Canadian medical schools taking the exam for the first time. The items were relatively lengthy, with the average number of words per item equal to 155 (SD = 71). In terms of difficulty, the mean p-value of the items in our sample was .69 (SD = .24).

Feature extraction. As mentioned above, in addition to using the presence of a figure to model mean response time, three general types of features—linguistic, features based on automated question answering, and word embeddings—were utilized.

Linguistic features. Linguistic features refer to linguistic characteristics of an item's text that can be represented numerically. We identified and extracted 113 such features. Specific examples are

provided in the following paragraphs. A complete list of the features and their definitions is available in Appendix A.

Lexical features. Lexical features, as the name implies, quantify characteristics of a text based on its constituent words. These features include counts (raw and/or normalized by the number of words or number of sentences in an item) for content words, nouns, verbs, adjectives, adverbs, and numerals; type-token ratios, which quantify language diversity as the ratio of the number of unique words used to the number of words used; measures of lexical complexity such as the average word length in syllables and the number of words having more than three syllables; the prevalence of medical concepts based on the Unified Medical Language System Meta-thesaurus (UMLS; Schuyler, Hole, Tuttle, & Sherertz, 1993); and, finally, we considered word usage commonality as represented by the average word frequency within the British National Corpus (Leech, 1992) as well as whether or not each word is among the 2000, 3000, 4000, or 5000 most frequently used words in English based on the same resource.

Syntactic features. Syntactic features refer to the characteristics of sentence structure and the rules governing it. Here, these features were measured by the average length of sentences, the depth of their syntactic trees, the presence of negated words, and the counts and average lengths of specific types of phrases (e.g., noun phrases, verb phrases, and so on). Because certain syntactic constructions such as the passive voice increase text complexity, the proportion of verb phrases that were passive was also included.

Semantic ambiguity features. Items that contain more ambiguous constructions may require more time to read and understand. To examine this, we modeled the ambiguity of words using the WordNet ontology (Miller, 1995), wherein words are represented through their semantic relations to other words, which allows information about the relatedness of their meanings to be extracted. This information was then used to quantify semantic ambiguity.

Cohesion features. Cohesion refers to the way various pieces of information within the text are organized to form a meaningful whole. These features were extracted by counting the overall number of connective words such as additive connectives (e.g., “and”, “also”) and temporal connectives (e.g.,

"next", "finally", "soon"). Both raw counts and counts normalized by the number of words in each item were used.

Readability features. In order to capture the overall linguistic difficulty of the items, we applied a set of widely used readability formulae. Readability formulae use surface text characteristics such as word length and sentence length to measure the difficulty of specific types of text with respect to a reference population. Since no such formulae exist for medical multiple-choice questions specifically, we included the performance of a wide range of formulae developed for various types of texts and readerships. These include the Flesch Reading Ease (Flesch, 1948); Flesch Kincaid Grade Level (Kincaid, Fishburne, Rogers, & Chissom, 1975); Automated Readability Index (ARI; Smith & Senter, 1967); and several others.

Cognitively-motivated features. While useful, the above readability formulae have been criticized for relying on surface complexity measures and failing to capture deeper levels of lexical complexity such as how abstract a word is or how familiar it is to the reader. Therefore, we included a set of cognitively-motivated features derived from the MRC Psycholinguistic Database (Coltheart, 1981). These features are based on human ratings of characteristics such as *imageability*, *familiarity*, and *age of acquisition*.

Information retrieval features. Information retrieval features come from the field of automatic question answering. The automatic question-answering system we developed followed the approach presented in Clark et al. (2018), which requires the use of a corpus of relevant text. For this purpose, our corpus comprised abstracts of medical articles contained in the MEDLINE database[1], which was queried using keywords—similar to the way one searches for information with a search engine. In our case, each query is based on the stem of a multiple-choice question together with one of the options and the five most relevant abstracts from the MEDLINE database are identified. Each query is then assigned various retrieval scores summarizing the relevance of the matched abstracts, which in turn enables the system to assign a score to each stem-and-answer-option combination (in addition to using all words, these queries were repeated using *nouns only*, and *nouns + verbs + adjectives*). To the extent that the information

[1] https://www.nlm.nih.gov/bsd/medline.html

retrieval scores for the *stem + correct answer* are higher than those for the *stem + options*, the question answering system will have more confidence in its answer choice and the item will be deemed relatively easy. More details about this approach can be found in Ha and Yaneva (2019).

Embeddings. Word embeddings are word vector representations produced using specific language modelling techniques that capture a wide variety of information including semantics. As noted above, the main linguistic theory behind the concept of word embeddings is that words are defined through the company they keep with other words; therefore, the vector representations of individual words are derived through modeling hundreds of thousands of co-occurrences of a given word in various contexts.

We experimented with two types of pre-trained embeddings: Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013) and ELMo (Peters et al., 2018); however, for our analyses, both types performed with similar accuracy so only the results using the Word2Vec embeddings are reported. Word2Vec embeddings represent 300-dimensional word vectors, derived through the co-occurrence of each word in large collections of texts.

Model Building. The linguistic features, information-retrieval-based features, and word embeddings may carry valuable information for different aspects of text processing that potentially can be used to predict item response times; however, a large number of potential predictors can pose a problem for standard modeling techniques like regression where the risk overfitting can be high. Because of this concern, we took a different approach to model building than earlier researchers who relied on regression alone to predict response time (e.g., Halkitis et al., 1996; Smith, 2000). Exploiting these features required using algorithms to train prediction models capable of coping with a large number of possible predictors. In such circumstances, the choice of algorithm is usually determined empirically through testing. Feature selection can be performed through stepwise addition or removal of features, through selecting the top N features that have the strongest relationship with the dependent variable, or as part of the model itself (embedded methods). We experimented with a number of the machine learning algorithms implemented in the WEKA toolkit (Hall et al., 2009) using our full set of features. (An introduction to data mining in

the context of educational measurement can be found in Sinharay, 2016; a recent application can be found in Man, Harring, & Sinharay, 2019.) All models were evaluated using a 10-fold cross validation, where data are split into ten folds or parts. Each model is then trained on nine folds and tested on the one remaining fold. By repeating the process ten times, each fold becomes a test fold once and the results are cross validated.

The best performance was achieved using an ensemble decision-tree algorithm called random forests (Breiman, 2001). Random forests combines the predictive power of a number of "weaker" algorithms called decision trees (Quinlan, 1986). Decision trees are a type of predictive model that predicts a target value (represented as the leaves of a tree) based on several input variables (represented as the branches of the tree). In random forests, a random subset of features and data points is selected for the construction of a decision tree (hence, the name "random forests") and, after a large number of decision trees have been trained in this manner using random data and feature sub-samples, their predictions are averaged and taken as the prediction for the target variable. Using information from multiple decision-tree models, each using a random sub-sample of all features, reduces the variance of the predictions, while preserving, to a certain extent, the low bias nature of individual decision trees.

In addition to the model based on the random forests approach, five linear regression models were estimated. These linear models show what can be achieved without relying on more sophisticated methods—or, more specifically, the value added by using random forests in this context. The first of these linear models used the full set of features to predict response time and the remaining four models were based on one predictor each—*average word frequency*, *presence of an image*, *word count*, and *unique word count*. The single-predictor models were intended to act as baselines for comparison with the full model and were selected based on prior work on response time prediction (presence of an image, word count) or their established relation to reading difficulty (average word frequency, unique word count)

Results

Table 1 reports the Pearson product-moment correlation coefficient between the observed and predicted mean response times, as well as Root Mean Squared Deviation (RMSD; in seconds) for the two

models trained on the full feature set and each of the baseline models. We first explore four individual baselines using linear regression (the 'Individual Baselines' section), then combine these baselines using both linear regression and random forests (the 'Combined Baselines' section), and, finally, we test the full set of predictors using both linear regression and random forests (the 'All Features' section). Using the full feature set, correlation coefficients were .69 (RMSD = 23.50) for the random forests algorithm and .59 (RMSD = 26.04) for the linear regression. These models significantly outperform the baselines; the two best performing individual baselines were *word count* and *unique word count*, which had correlations of .37 (RMSD = 29.75) and .42 (RMSD = 29.16), respectively. The best combined baseline (random forests) had a correlation of .55 (RMSD = 27.00).

Table 1
Comparison of Prediction Models

	Models	Pearson's *r*	RMSD (seconds)
Individual baselines (linear regression)	Average word frequency	.18	31.57
	Presence of an image	.20	31.47
	Word count	.37	29.75
	Unique word count	.42	29.16
Combined baselines	Linear regression	.47	28.22
	Random forests	.55	27.00
*All features**	Linear regression	.59	26.04
	Random forests	**.69**	**23.50**

* The full feature set includes the 16 IR features, 113 linguistic features, as well as W2V embeddings of 300 dimensions.

To better understand the contributions of individual feature types, separate random forest models were also trained for each type. Results, reported in Table 2, show that word embeddings and linguistic features performed similarly (both $r = .67$); however, no single feature type performed quite as well as the combination of all features, which suggests that despite their similar performance, the feature types each captured some unique information.

Table 2
Results for Random Forest Models Based on Individual Feature Types

Feature type	*r*	RMSD (seconds)
Information retrieval (16)	.55	26.79
Linguistic (113)	.67	23.98
Embeddings	.67	23.90

Next, we used Lasso regression to identify the most important linguistic features for our sample. The result suggested that the number of unique words has the highest predictive power and was among the features with positive coefficients, followed by the presence of an image and an IR query containing the sum of the retrieval indices of the nouns from the stem and options. Strong predictors with negative coefficients were the number of additive connectives, the number of numbers, and the number of words that were not among the 5000 most commonly used words. The top 10 features with positive and negative coefficients can be found in Appendix B.

Finally, we explored optimizing the models through a greedy forward or backward search through the space of feature subsets. The results from these optimized models, which were based on a subset of selected features, are presented in Table 3.

Table 3
Results for the Optimized (Selected Features) Models

Model	*r*	RMSD (seconds)
Linear regression	.55	26.81
Random forests	.70	23.15

Overall, *slightly* better performance was achieved using the optimized feature set (i.e., selected features from the full feature set) to train a random forests model (r = .70, RMSD = 23.15). Like the model using the full feature set, this model significantly outperformed the baseline models; however, its improvement over the model using the full feature set was not significant.

The results presented in the previous paragraphs indicate that using the identified features allows for substantially better prediction of response times than that provided by the simpler approaches based on word count or presence of an image that have been suggested by previous researchers. It also appears that the random forests algorithm may be superior to simple regression as a means of aggregating the information from a large number of features. The question that remains is whether these predicted values result in a practical improvement when used in test construction. This question is addressed in study II.

Study II

Method

This second study evaluates the practical utility of the predictions obtained from study I. Previous authors have recognized the importance of building test forms that are parallel with respect to their time demands (e.g., Fan, Wang, Chang, & Douglas, 2012; Scrams & Smith, 2010; van der Linden, 2005, 2009; van der Linden, 2011; van der Linden, Scrams, & Schnipke, 1999). This may be done to ensure that scores from different test forms are not affected by the imposed time limits or to ensure that the effects of the time limits are equivalent across forms. These approaches require that the response time for each item be known in advance. In contrast, the present study focuses on building equivalent forms when the response times are unknown. To evaluate the usefulness of the predicted response times described in study I, we used these predictions to build test forms intended to meet certain timing targets and then compared the empirical response times for these forms to these targets. For this evaluation three approaches to test assembly were compared. In the first condition (denoted *random*), which represents what happens when predicted response times are not used, items were randomly assigned to blocks (i.e., timed sections); in the second condition (denoted *simple model*), which represents what can be accomplished using a single easily-obtained covariate and a simple linear model, items were assigned based on predicted mean response times based on word count alone; and in the third condition (denoted *full model*), assignment was based on the full prediction model described in study I (using the optimized feature set). Aside from being superficial and straightforward, word count is a measure of the quantity of text; thus, the extent to which the full model outperforms word count can be attributed to text characteristics beyond simply quantity. Numerous conditions were considered related to the size of the pool, number of blocks, and the number of items per block and each of these conditions was replicated to evaluate the stability of the results across samples.

Item block assembly. An explanation of the various study conditions, replications, and so on is provided below but first we present the general case of a single replication and a set of conditions to make clear the process by which item blocks were assembled. First, an item pool is defined by selecting, at

random, a set of P items from the full set of 12,123 pretest items used in this study. Let μ be the mean of the mean response times for whatever *unselected* items remain from the full set. Since this mean is based on the unselected items for a given assembly problem, it represents the expected mean response time for pretest items that, in practice, could be known a priori based on historical data. Let the target block time, τ, be given by $\tau = I\mu$, where I is the number of items per block (specifically, the number of items per block with unknown latencies).

From the selected item pool of P items, a subset of K items then is selected, also at random, where K equals the total number of items needed to populate some specified number of item blocks, B, each with some specified number of items, I (i.e., $K = BI$, the product of the total number of blocks and the number of items per block). These K items are then used to populate, at random, the given set of B item blocks. These random blocks are used for the random assignment baseline condition (*random*) and as the starting condition for the two prediction model conditions. The number of items in the pool (P), the number of blocks (B), and the number of items in each block (I) are varied as study conditions. Blocks for the two prediction model conditions were assembled such that the differences between the sum of their predicted mean response times and the target block time, τ, was minimized. (The optimization procedure used to accomplish this is described in Appendix C.)

Evaluation criteria. For each block b, whether assembled using predicted values or at random, there will be a corresponding sum of *empirical* mean response times $\sum_{i=1}^{I} t_{bi}$. This sum will deviate from its block target time, τ, by some amount $d_b = \left(\sum_{i=1}^{I} t_{bi} \right) - \tau$ and the resultant distribution of d across all blocks and replications for a given set of study conditions is treated as the sampling distribution for the prediction error. The means for these distributions are of some interest; however, of greater interest are the variances of these distributions—the smaller the variance, the greater the consistency of time needed across block, which is the objective of the procedure.

Study conditions. Results are reported as a function of three study conditions: *pool size*, *number of blocks*, and *number of items per block*. For the *pool size* condition, the number of blocks and number of items per block were held constant at 20 blocks and 20 items per block. Pool sizes included 400, 500, 600, 700, 800, 900, and 1000. For the *number of blocks* condition, pool size was held constant at 1000 and number of items per block was kept constant at 20. Five conditions were tested: 10, 20, 30, 40, and 50 blocks. For the *number of items per block* condition, pool size was again held constant at 1000 and this time number of blocks was held constant at 20. As with the *number of blocks* condition, five conditions were tested: 10, 20, 30, 40, and 50 items per block. These study conditions are summarized in Table 4. A fully crossed design would have created 7 × 5 × 5 = 175 conditions; however, half of these conditions were simply not possible (e.g., 30 blocks at 30 items per block *requires* a pool size of at least 30 × 30 = 600 items thereby making the 400 and 500 item pool size conditions impossible) and, moreover, results were highly stable across pool size and number of blocks conditions and highly predictable across the number of items per block conditions. In this way, the reported conditions generalize sufficiently to tell the full story.

Table 4
Conditions for Evaluation Study II

Study condition	Study condition values		
	Pool size	*Number of blocks*	*Items per block*
Pool size	400, 500, …, 1000	20	20
Number of blocks	1000	10, 20, …, 50	20
Items per block	1000	20	10, 20, …, 50

Results

Means and standard deviations, in seconds, for the distributions of discrepancies, d, between the sum of empirical mean response times, $\sum_{i=1}^{I} t_i$, and the target sum of mean response times, τ, are reported in Appendix D. As noted, the means are not of great interest here and generally they were not significantly different from zero for the *random* and *full model* conditions. Even in the case of the *simple*

model (word count) condition, the greatest mean d across all conditions was 23 seconds, which, while not zero, was only .12 standard deviations from zero.

The more relevant findings were those relating to the variability of the discrepancies. Standard deviations for the *random* item selection method are reduced by about 14 seconds on average when items are selected using the simple model and by 42 seconds on average when the full model is used. A permutation test reveals that these reductions were significant at the .05 level except once: the difference between the random and simple model methods for the pool size = 1000, number of blocks = 50, number of items per block = 20 condition, which is a somewhat more challenging condition to optimize because it requires the use of the entire pool (50 × 20 = 1000).

Standard deviation results are also shown graphically in Figures 1-3 and it can be seen that the differences in standard deviation across item selection methods is reasonably consistent across all conditions. Moreover, the standard deviations themselves are relatively unaffected by pool size or number of blocks; however, they are positively related to number of items per block. These positive trends are expected because the variance of a sum (the sum of the uncorrelated mean response times of the items in a block in this case) equals the sum of the variances and thus the greater the number of items, the greater the variance. This relationship is sometimes referred to as Bienaymé's equality (Loeve, 1977) and the standard deviations predicted by this equality are also shown in Figure 3 by "X." For the item-level variance (the variances that are summed in Bienaymé's formula), the variance of the observed mean response times was used, which should correspond to the *random* selection method's results, which it does very closely.

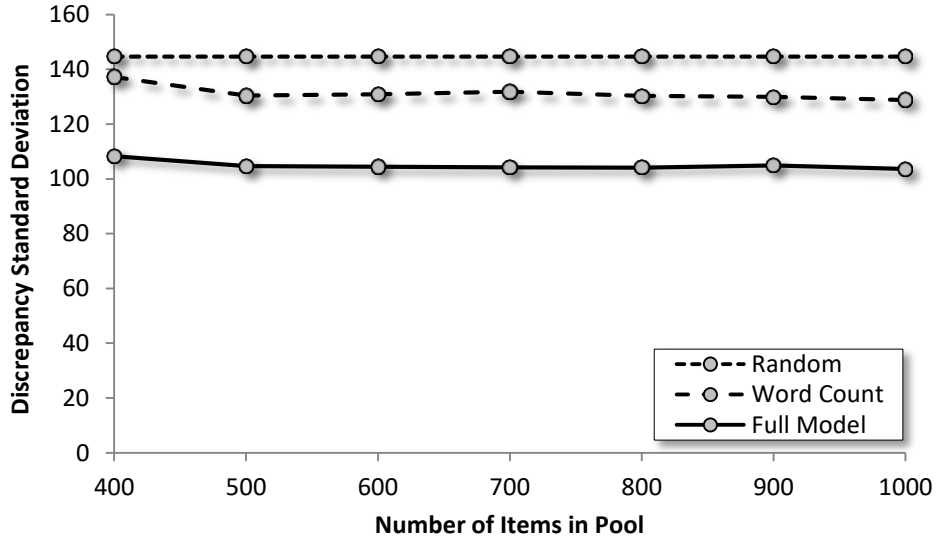

Figure 1. Standard deviations for the distributions of discrepancies, d *, between the sum of empirical mean response times,* $\sum_{i=1}^{I} t_i$ *, and the target sum of mean response times,* τ *, in seconds for seven* pool sizes. Number of blocks *and* items per block *were both fixed at 20.*

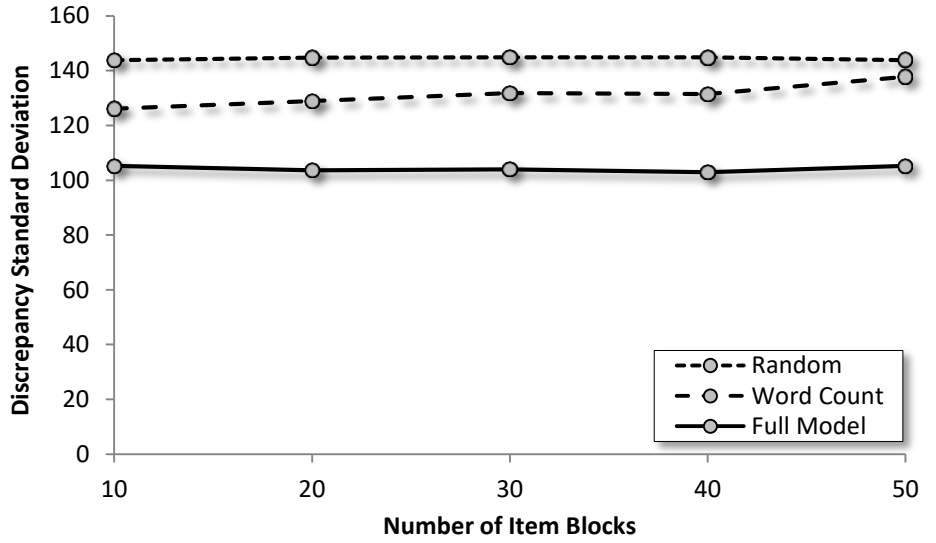

Figure 2. Standard deviations for the distributions of discrepancies, d *, between the sum of empirical mean response times,* $\sum_{i=1}^{I} t_i$ *, and the target sum of mean response times,* τ *, in seconds for five* number of blocks *conditions.* Pool size *was fixed at 1000 items and* items per block *was fixed at 20.*

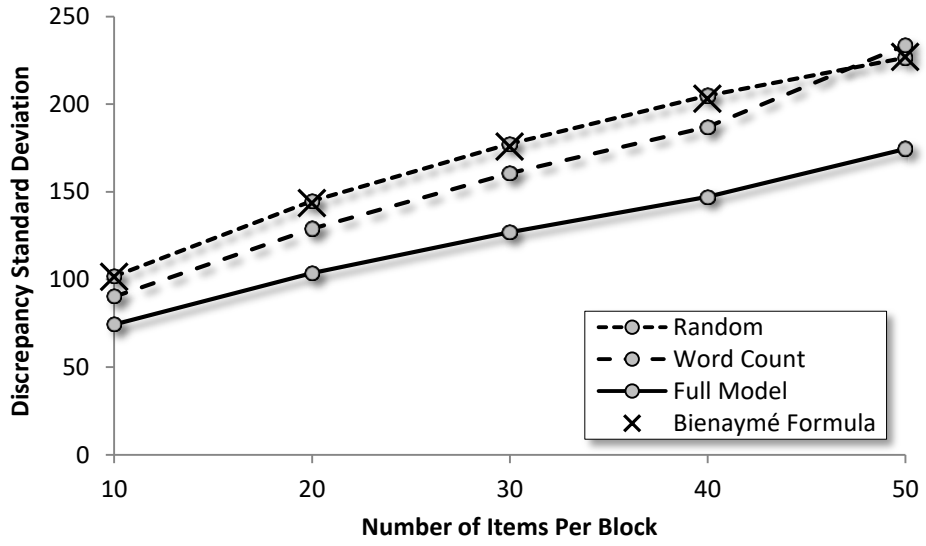

Figure 3. Standard deviations for the distributions of discrepancies, d *, between the sum of empirical mean response times,* $\sum_{i=1}^{I} t_i$ *, and the target sum of mean response times,* τ *, in seconds for five* items-per-block *conditions.* Pool size *was fixed at 1000 items and* number of blocks *was fixed at 20. Also shown (indicated by the* **X***) are the expected standard deviations for the* random *item selection method given that the variance of a sum of item mean response times should equal the sum of the variances of the item mean response times. This relationship is sometimes called* Bienaymé's equality *(Loeve, 1977) and explains why the variance increases as a function of number of items per block.*

Again, the standard deviations presented in Figures 1-3 (and also reported in Appendix D) represent the variability around the mean for the time demand for a block of items. Fairness requires that the time demands should be equal for examinees completing putatively parallel blocks of items. Eliminating this variability completely is not possible when tests are constructed using items for which empirical response time data are unavailable; however, the results suggest that the proposed procedure can substantially reduce the variability across blocks. This could have practically important consequences for the comparability of scores across test forms.

Discussion

The purpose of this study was threefold: (a) to identify and describe item features that covary with response time, (b) to show how the set of features together can be productively used to build a prediction model for response time, and (c) to demonstrate the potential practical value of using these

predictions to improve test assembly when items are included that have not been administered previously. To this end, we showed how item text could be represented by (a) 113 features quantifying the text's linguistic characteristics; (b) 16 measures of the extent to which an information-retrieval-based automatic question answering system finds an item challenging; and (c) through dense word representations (word embeddings). We then showed how these data could be used to train a model that predicts item response times, which then could be used to assemble test forms. In study II, we found that these predictions could reduce the variability in the sum of expected response times across blocks of items.

In this second study, a consistent story across a range of pool sizes, number of item blocks, and number of items per block was observed: the expected variability in the time needed to complete a block of items could be substantially reduced by using *predicted* response times during block assembly and using the full-model predictions reduced this variability the most. On the USMLE®, the operational exam from which our data were taken, blocks typically contain approximately 40 items. To provide a concrete example, consider the condition in which 10 of these items have unknown response times. If these 10 items are selected at random, 20.4% of the item blocks will differ in their sum of expected response times by more than three minutes; but this frequency is reduced to 8.6% by using the full-model predictions to select items. Three minutes is a substantial amount of time for a 60-minute testing block on this exam, which suggests that using the predictions can substantially improve fairness.

Although these results suggest that the procedures we described provide a practical improvement in test construction that can be accomplished using information (i.e., item text) already available prior to any pretesting, it is worth considering the extent to which these results generalize to other settings. This study used test material from a medical licensing examination. Other settings could have lexical features that are more or less strongly related to the response time requirements of items. This is, of course, an empirical question and beyond the scope of the present study. Nonetheless, it is reasonable to speculate, for example, that this approach may be less useful in a mathematics test in which items may have relatively little text and that it may be more useful in a reading-comprehension test in which the complexity of the text is systematically manipulated to create more (and less) difficult items. The tenfold

cross validation used in the present study supports the view that the predictions made in this study would be similarly useful when applied to other test items like the ones used in this study. It remains to be seen what the limits are of "like the ones used in this study." Thus, while we might reasonably expect these limits to include other medical items with different content, such as those used in other components of the USMLE® or medical certification tests, we also might expect relatively homogeneous item subsets to benefit from set-specific prediction models. At present, we recommend that test developers evaluate these procedures in their own specific contexts.

Another factor potentially limiting the generalizability of the results reported in this paper is that content constraints and item difficulty were ignored in optimizing block time based on the predicted item response times. This may be appropriate in circumstances such as the Step 2 CK examination (used in this study) where the predicted response times are for unscored pretest items. In such circumstances, pretest item selection may not be subject to many—if any—content constraints, information targets, or other assembly objectives (i.e., the constraints may apply only to scored items). By contrast, if predicted response times were being used to construct *scored* test forms for a testing program that did not pretest items, the additional assembly constraints (e.g., content constraints) could limit the usefulness of the proposed procedure. The extent to which usefulness is affected would depend on the relative importance and stringency of the timing objectives and whatever additional assembly objectives existed as well as the size and composition of the item pool used for test construction.

More generally, in this study the target expected time was known and the goal simply was to meet it as closely and as consistently as possible; however, it has been suggested elsewhere (e.g., van der Linden, 2011) that a more nuanced approach to mitigating speededness would be to estimate the *distribution* of (test-level) response times and the risk of running out of time using the lognormal model described in van der Linden (2006). Nevertheless, the discrepancies between the target and empirical expected block times are a very interpretable and direct measure of the utility of the predicted response times, which seemed appropriate for this stage of our research. Now that proof-of-concept has been

demonstrated, one area for future research is the use of predicted response times within a response-time model.

The potential limitations of this study aside, the reported results suggest that the proposed approach has considerable potential to improve test construction. Producing the collateral item data used in this study does not require any special data collection effort and can be extracted at the point at which an item is written. For this reason, they can be used with relatively little effort and cost. In this study, we showed how varied and valuable these data can be and our findings suggest that further research in this area could be profitable.

Appendix A

All included linguistic features and their definitions.

Table A1
Lexical Features

Feature	Raw count	Normalized by word count	Normalized by sentence count
Number of words	√		
Number of nouns	√	√	
Number of verbs	√	√	
Number of adjectives	√	√	
Number of adverb	√	√	
Number of number	√	√	
Number of comma	√	√	
Number of complex words (words with more than 3 syllables)	√	√	
Number of unique words	√	√	
Number of content words	√	√	
Number of content words excluding stopwords (e.g., is, the, an)	√		
Word length in syllables		√	
Number of UMLS* concepts	√	√	√
Number of words that refer to multiple UMLS* concepts		√	

*Unified Medical Language System Meta-thesaurus

Table A2
Syntactic Features

Feature	Raw count	Normalized by word count	Normalized by sentence count	Other
Average length in words for sentences				√
Average depth of syntax tree for sentences				√
Number of negated words	√	√		
Number of negated words, stem	√			
Number of negated words, lead-in	√			
Number of main noun phrases	√	√		
Average noun phrase length in words for main noun phrases				√
Number of noun phrases (main noun phrases *and* embedded noun phrases)	√	√		
Average noun phrase length in words for all noun phrases (both main and embedded)				√
Number of prepositional phrases	√	√	√	
Number of verb phrases	√	√		
Proportion of verb phrases that are active-voice verb phrases				√
Proportion of verb phrases that are passive-voice verb phrases				√
Ratio of active- to passive-voice verb phrases				√
Number of words occurring before the main verb			√	
Number of agentless passive-voice constructions	√			
Number of relative clauses	√	√		
Proportion of clauses that are relative clauses				√
Number of conditional clauses	√	√		

Table A3
Semantic Ambiguity Features

Feature	Raw count	Average	Normalized by word count
Number of polysemic words	√		√
Number of word senses for content words		√	
Number of word senses for nouns		√	
Number of word senses for verbs		√	
Number of word senses for non-auxiliary verbs		√	
Number of word senses for adjectives		√	
Number of word senses for adverbs		√	
Distance between words and their roots for nouns		√	
Distance between words and their roots for verbs		√	
Distance between words and their roots for nouns and verbs		√	
Words in answer that are also in WordNet	√		

Note. All features based on WordNet ontology (Miller, 1995).

Table A4
Readability Indices

Feature	Source
Flesch Reading Ease	Flesch, 1948
Flesch Kincaid Grade Level	Kincaid *et al*., 1975
Automated Readability Index	Smith and Senter, 1967
Gunning Fog	Gunning, 1952
Coleman-Liau	in DuBay, 2004
SMOG	McLaughlin, 1969
SMOG Index	McLaughlin, 1969

Table A5
Word Frequency Features

Feature	Raw count	Average	Normalized by word count
Word frequency for all words		√	
Word frequency relative to word frequency in item text for all words		√	
Word frequency rank for all words		√	
Word frequency for content words only		√	
Word frequency relative to word frequency in item text for content words		√	
Word frequency rank for content words		√	
Number of words not in the 2,000 most frequently used words	√		√
Number of words not in the 3,000 most frequently used words	√		√
Number of words not in the 4,000 most frequently used words	√		√
Number of words not in the 5,000 most frequently used words	√		√

Note. Frequencies based on the British National Corpus (BNC; Leech, 1992).

Table A6
Cognitively Motivated Features

Feature	Average*	Corrected average†	Ratio‡	Proportion absent§
Word imageability	√	√	√	√
Word familiarity	√	√	√	√
Word concreteness	√	√	√	√
Word age of acquisition	√	√	√	√
Word meaningfulness (Colorado)		√		
Word meaningfulness (Paivio)		√		

Note. All feature ratings from the MRC Psycholinguistic Database (Coltheart, 1981).
* Average rating after assigning *unrated* words a rating of 0.
† Average rating of the rated words only.
‡ Sum of ratings of the rated words only multiplied by the ratio of number of rated words to the total number of words in the item.
§ Number of *unrated* content words divided by the total number of content words in the item.

Table A7
Text Cohesion Features

Feature	Raw count	Normalized by word count
Number of connectives	√	√
Number of connectives – additive	√	√
Number of connectives - temporal	√	√
Number of connectives - causal	√	√
Number of referential pronouns	√	√

Appendix B

Table B1
The 10 Most Important Positively and Negatively Related Features

Rank from most to least important	Features with positive coefficients	Features with negative coefficients
1	Number of unique words, raw count (12.42)	Number of additive connectives, raw count (- 6.09)
2	Presence of an image (6.8)	Number of numbers, raw count (-4.06)
3	IR query containing the sum of the retrieval indices of the nouns from the stem and options (4.43)	Number of words not in the list of the 5000 most common words, raw count (-3.23)
4	Number of numbers, normalized by word count (3.95)	Number of referential pronouns, normalized by word count (-2.85)
5	Number of conditional clauses, normalized by word count (3.82)	Number of negated words, normalized by word count (-2.49)
6	Concreteness ratio (2.62)	Number of relative clauses, normalized by word count (-2.36)
7	Number of nouns, normalized by word count (2.42)	Number of nouns, raw count (-2.00)
8	Number of additive connectives, normalized by word count (2.33)	Word imageability, average (-1.87)
9	Number of adjectives, raw count (2.32)	IR query containing the maximum retrieval score from all words in the stem (-1.74)
10	Average depth of syntax tree for sentences (2.27)	Word meaningfulness (Paivio), corrected average (-1.21)

Note. Based on Lasso regression estimation of feature importance. Standardized coefficient values are given in parentheses.

Appendix C

Optimizing block assembly using predicted mean response times in this context is a relatively modest problem. Mixed-integer programming strategies for test assembly have been described by van der Linden (2006), Diao and van der Linden (2011), and elsewhere. We began the analysis using this approach, but subsequently adopted a simpler procedure that produced equally good results much more quickly. This simpler procedure was sufficient because our block assembly problem optimized only a single feature, response time; however, this simplicity was not merely a contrivance of our study design: pretest items are often not subject to stringent block-specific content specifications, information targets, and so on. This simpler approach was accomplished using two substitution techniques. Because they are parts of the same overall block assembly problem, let's designate these two substitution techniques stage 1 and stage 2.

As noted in the main text, first an item pool is defined by selecting, at random, a set of P items from the full set of 12,123 pretest items used in this study. Let μ be the mean of the mean response times for whatever *unselected* items remain from the full set and let the target block time, τ, be given by $\tau = I\mu$, where I is the number of items per block (specifically, the number of items per block with unknown latencies).

From the selected item pool of P items, a subset of K items then is selected, also at random, where K equals the total number of items needed to populate some specified number of item blocks, B, each with some specified number of items, I (i.e., $K = BI$, the product of the total number of blocks and the number of items per block). These K items are then used to populate, at random, the given set of B item blocks.

After selecting and assigning these K items from the pool of P items, some $P - K$ items remain. In stage 1, each of these remaining $P - K$ unselected items is, one-at-a-time, exchanged on a trial basis with each of the selected K items. Whichever trial exchange results in the greatest reduction in the squared difference between the relevant block's sum of mean *predicted* response times and the target

block time, $\left(\tau - \sum_{i=1}^{I} s_{bi}\right)^2$ or $\left(\tau - \sum_{i=1}^{I} f_{bi}\right)^2$, then is adopted, where s_{bi} and f_{bi} are the predicted response times for item i in block b for the simple and full model conditions, respectively. Stage 1 is repeated until there are no unselected items that reduce this squared difference by more than .001 squared seconds.

In stage 2, the updated set of $P - K$ *unselected* items are set aside altogether and each *selected* item in the now modified blocks is swapped, one-at-a-time, on a trial basis with every other selected item in each of the other blocks. The squared differences between τ and each of the two blocks' sum of mean predicted response times is calculated and then these two squared differences are summed. Here, whichever trial swap results in the greatest reduction in this sum of squared differences is adopted. As with stage 1, stage 2 was repeated until there were no swaps that reduced this sum of squared differences by more than .001 squared seconds.

This two-stage optimization process performed well. Setting aside the conditions that required the use of all the items in the pool (and which therefore did not benefit from stage 1 of the block assembly process), the *greatest* mean absolute discrepancy between the sum of predicted mean response times and τ was .10 seconds across all remaining conditions and replications comprising a total of 52,000 item blocks. These discrepancies were so small compared with the prediction errors that they can be ignored for our purposes. That is, errors due to suboptimal form assembly effects were inconsequential in this context.

Appendix D

Means and standard deviations, in seconds, for the distributions of discrepancies, d, between the sum of empirical mean response times, $\sum_{i=1}^{I} t_i$, and the target sum of mean response times, τ, are reported in Tables D1, D2, and D3 for the pool size, number of blocks, and number of items per block study conditions, respectively. Each table shows results for all three of the item selection methods: *random*, *simple model* (word count), and *full model*.

Table D1
The Means and Standard Deviations for the Distribution of Discrepancies for the Pool Size Condition

Pool Size	*Item Selection Method*		
	Random	Simple Model (word count)	Full Model
400	0 (-4.9, 5.7); 145	0 (-4.6, 5.4); 137	0 (-3.6, 4.4); 108
500	0 (-4.9, 5.7); 145	16 (11.6, 21.2); 130	-2 (-6.3, 1.4); 105
600	0 (-4.9, 5.8); 145	15 (9.8, 19.4); 131	-3 (-6.6, 1.1); 104
700	0 (-5.0, 5.7); 145	15 (10.4, 20.1); 132	-3 (-7.0, .7); 104
800	0 (-5.0, 5.7); 145	16 (10.9, 20.5); 130	-3 (-7.1, .5); 104
900	0 (-4.9, 5.7); 145	17 (11.8, 21.4); 130	-3 (-7.1, .7); 105
1000	0 (-4.9, 5.7); 145	15 (10.3, 19.8); 129	-4 (-7.4, .3); 104

Note. Means and standard deviations for the distribution of discrepancies, d, between the sum of empirical mean response times, $\sum_{i=1}^{I} t_i$, and the target sum of mean response times, τ, in seconds for seven pool sizes. Number of blocks and items per block were both fixed at 20. Also reported (parenthetically) are bootstrapped 95% confidence intervals for the means and standard deviations, which are reported in each cell following the semicolon: *Mean* (*Confidence Interval*); *SD.*

Table D2
The Means and Standard Deviations for the Distribution of Discrepancies for the Number of Blocks Condition

Number of Blocks	*Item Selection Method*		
	Random	Simple Model (word count)	Full Model
10	2 (-5.3, 9.6); 144	20 (13.3, 26.4); 126	0 (-5.2, 5.8); 105
20	0 (-4.9, 5.7); 145	15 (10.3, 19.8); 129	-4 (-7.4, .3); 104
30	1 (-3.8, 4.9); 145	12 (8.0, 15.9); 132	-6 (-8.8, -2.5); 104
40	-1 (-4.4, 3.1); 145	14 (10.1, 17.0); 131	-3 (-6.1, -.7); 103
50	0 (-3.3, 3.4); 144	0 (-3.1, 3.3); 138	0 (-2.4, 2.5); 105

Note. Means and standard deviations of the distribution of discrepancies, d, between the sum of empirical mean response times, $\sum_{i=1}^{I} t_i$, and the target sum of mean response times, τ, in seconds for five number-of-blocks conditions. Items per block was fixed at 20 and pool size was fixed at 1000. Also reported (parenthetically) are bootstrapped 95% confidence intervals for the means and standard deviations, which are reported in each cell following the semicolon: *Mean* (*Confidence Interval*); *SD*.

Table D3
The Means and Standard Deviations for the Distribution of Discrepancies for the Number of Items per Block Condition

Number of Items Per Block	*Item Selection Method*		
	Random	Simple Model (word count)	Full Model
10	1 (-2.7, 4.8); 102	9 (5.6, 12.2); 90	1 (-1.9, 3.6); 74
20	0 (-4.9, 5.7); 145	15 (10.3, 19.8); 129	-4 (-7.4, .3); 104
30	1 (-5.8, 7.3); 177	20 (14.4, 26.3); 161	-6 (-10.8, -1.5); 127
40	-1 (-8.8, 6.2); 205	23 (15.6, 29.4); 187	-7 (-12.0, -1.1); 147
50	0 (-8.1, 8.5); 227	0 (-8.4, 8.8); 234	0 (-6.2, 6.6); 175

Note. Means and standard deviations of the distribution of discrepancies, d, between the sum of empirical mean response times, $\sum_{i=1}^{I} t_i$, and the target sum of mean response times, τ, in seconds for five items- per- block conditions. Number of blocks was fixed at 20 and pool size was fixed at 1000. Also reported (parenthetically) are bootstrapped 95% confidence intervals for the means and standard deviations, which are reported in each cell following the semicolon: *Mean* (*Confidence Interval*); *SD*.

References

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*(1), 63-88.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology, 33*(4), 497-505.

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement*, *35*(5), 398-409.

De Clercq, O., & Hoste, V. (2016). All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, *42*(3), 457-490.

Dubay, W. H. (2004). The principles of readability. Impact Information. *Costa Mesa, CA.*

Ebel, R. L. (1953). The use of item response time measurements in the construction of educational achievement tests. *Educational and Psychological Measurement*, *13*(3), 391-401.

Fan, Z., Wang, C., Chang, H. H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*(5), 655-670.

Fitzgerald, C., Fremer, J., & Maynes, D. (2005, October). *Test security issues and best practices in detecting test fraud.* Paper presented at the Innovations in Testing Research Conference, October 7, 2005, Victoria, B.C.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221-233.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193-202.

Gunning, R. (1952). *The technique of clear writing.* McGraw-Hill, New York.

Gupta, P., & Gupta, V. (2012). A survey of text question answering techniques. *International Journal of Computer Applications*, *53*(4), 1-8.

Ha, L.A., & Yaneva, V. (2019) Automatic Question Answering for Medical MCQs: Can It Go Further than Information Retrieval? *Proceedings of the 12th Conference on Advances in Natural Language Processing* (RANLP 2019), Varna, Bulgaria.

Halkitis, P. N., Jones, J. P., & Pradhan, J. (1996, April).*Estimating testing time: The effects of item characteristics on response latency.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.

Harik, P., Clauser, B. E., Grabovsky, I., Baldwin, P., Margolis, M. J., Bucak, D., ... & Haist, S. (2018). A Comparison of Experimental and Observational Approaches to Assessing the Effects of Time Constraints in a Medical Licensing Examination. *Journal of Educational Measurement*, *55*(2), 308-327.

Hassoun, M. H. (1995). *Fundamentals of artificial neural networks*. MIT press.

J. P. Kincaid, R. P. Fishburn, R. L. Rogers, and B. S. Chissom. *Derivation of new readability formulas for navy enlisted personnel.* Technical Report Research Branch Report 8-75, Millington, Tenn, Naval Air Station, 1975.

Kintsch, W. (1995). How readers construct situation models for stories: The role of syntactic cues and causal inferences. *Coherence in spontaneous text*, 139-160.

Leech, G. N. (1992). 100 million words of English: the British National Corpus (BNC).

Loeve, M. (1977). *Graduate texts in mathematics: Vol. 45. Probability theory* (4th ed.). Berlin, Germany: Springer.

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, *16*(1), 100-103.

Marianti, S., Fox, J. P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*(6), 426-451.

Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of Reading*, 12(8), 639-646.

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

Miller, M. D. (1986). Time allocation and patterns of item response. *Journal of Educational Measurement*, *23*(2), 147-156.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Parshall, C. G., Mittelholtz, D. J., & Miller, T. R. (1994, April). *Response latency: An investigation into determinants of item-level timing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004, May). WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004* (pp. 38-41). Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250.*

Schuyler, P. L., Hole, W. T., Tuttle, M. S., & Sherertz, D. D. (1993). The UMLS Metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, *81*(2), 217-222.

Scrams, D. J., & Smith, R. W. (2010, May). *A practical approach to balancing time demands across test forms.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Smith R. W. (2000, April). *An exploratory analysis of item parameters and characteristics that influence item response time.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Smith, E. A., & Senter, R. J. (1967). Automated readability index. *Technical Report AMRLTR-66-220*, University of Cincinnati, Cincinnati, Ohio.

Swanson, D. B., Case, S. M., Ripkey, D. R., Clauser, B. E., & Holtman, M. C. (2001). Relationships among item characteristics, Examinee characteristics, and response times on the USMLE® Step 1. *Academic Medicine, 79* (10 October Supplement), S114-S116.

Tatsuoka, K. K., & Tatsuoka, M. M. (1979). A model for incorporating response-time data in scoring achievement tests. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 236– 256). Minneapolis : University of Minnesota, Department of Psychology, Psychometric Methods Program .

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.),*New horizons in testing: Latent trait test theory and computerized adaptive testing*. (pp. 179– 203) New York : Academic Press.

Thorndike, E. L., & Lorge, I. (1944). The teacher's word book of 30,000 words. Oxford, England: Bureau of Publications, Teachers Co.

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, *42*(3), 283-302.

van der Linden, W. J. (2006). *Linear models for optimal test design*. Springer Science & Business Media.

van der Linden, W. J. (2009). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement, 33*, 25–41.

van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement, 48*(1), 44-60.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 195-210.

van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251-265.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29,* 323– 339.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36,* 52-61.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19-38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183.

Zenisky, A. L., & Baldwin, P. (2006, April). *Using response time data in test development and validation: Research with beginning computer users*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.