

# Machine Learning–Driven Language Assessment

Burr Settles and Geoffrey T. LaFlair

Duolingo  
Pittsburgh, PA USA  
{burr, geoff}@duolingo.com

Masato Hagiwara\*

Octanove Labs  
Seattle, WA USA  
masato@octanove.com

## Abstract

We describe a method for rapidly creating language proficiency assessments, and provide experimental evidence that such tests can be valid, reliable, and secure. Our approach is the first to use machine learning and natural language processing to induce proficiency scales based on a given standard, and then use linguistic models to estimate item difficulty directly for computer-adaptive testing. This alleviates the need for expensive pilot testing with human subjects. We used these methods to develop an online proficiency exam called the Duolingo English Test, and demonstrate that its scores align significantly with other high-stakes English assessments. Furthermore, our approach produces test scores that are highly reliable, while generating item banks large enough to satisfy security requirements.

## 1 Introduction

Language proficiency testing is an increasingly important part of global society. The need to demonstrate language skills—often through standardized testing—is now required in many situations for access to higher education, immigration, and employment opportunities. However, standardized tests are cumbersome to create and maintain. Lane et al. (2016) and the *Standards for Educational and Psychological Testing* (AERA et al., 2014) describe many of the procedures and requirements for planning, creating, revising, administering, analyzing, and reporting on high-stakes tests and their development.

In practice, test items are often first written by subject matter experts, and then “pilot tested” with a large number of human subjects for psy-

chometric analysis. This labor-intensive process often restricts the number of items that can feasibly be created, which in turn poses a threat to security: Items may be copied and leaked, or simply used too often (Cau, 2015; Dudley et al., 2016). Security can be enhanced through *computer-adaptive testing* (CAT), by which a subset of items are administered in a personalized way (based on examinees’ performance on previous items). Because the item sequences are essentially unique for each session, there is no single test form to obtain and circulate (Wainer, 2000), but these security benefits only hold if the item bank is large enough to reduce item exposure (Way, 1998). This further increases the burden on item writers, and also requires significantly more item pilot testing.

For the case of language assessment, we tackle both of these development bottlenecks using machine learning (ML) and natural language processing (NLP). In particular, we propose the use of test item formats that can be automatically created, graded, and psychometrically analyzed using ML/NLP techniques. This solves the “cold start” problem in language test development, by relaxing manual item creation requirements and alleviating the need for human pilot testing altogether.

In the pages that follow, we first summarize the important concepts from language testing and psychometrics (§2), and then describe our ML/NLP methods to learn proficiency scales for both words (§3) and long-form passages (§4). We then present evidence for the validity, reliability, and security of our approach using results from the Duolingo English Test, an online, operational English proficiency assessment developed using these methods (§5). After summarizing other related work (§6), we conclude with a discussion of limitations and future directions (§7).

\* Research conducted at Duolingo.

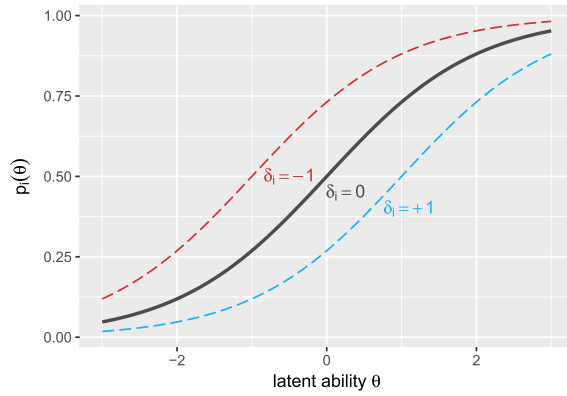


Figure 1: The Rasch model IRF, showing the probability of a correct response  $p_i(\theta)$  for three test item difficulties  $\delta_i$ , across examinee ability level  $\theta$ .

## 2 Background

Here we provide an overview of relevant language testing concepts, and connect them to work in machine learning and natural language processing.

### 2.1 Item Response Theory (IRT)

In psychometrics, *item response theory* (IRT) is a paradigm for designing and scoring measures of ability and other cognitive variables (Lord, 1980). IRT forms the basis for most modern high-stakes standardized tests, and generally assumes:

1. An examinee's response to a test item is modeled by an *item response function* (IRF);
2. There is a unidimensional latent *ability* for each examinee, denoted  $\theta$ ;
3. Test items are locally independent.

In this work we use a simple logistic IRF, also known as the *Rasch model* (Rasch, 1993). This expresses the probability  $p_i(\theta)$  of a correct response to test item  $i$  as a function of the difference between the item *difficulty* parameter  $\delta_i$  and the examinee's *ability* parameter  $\theta$ :

$$p_i(\theta) = \frac{1}{1 + \exp(\delta_i - \theta)}. \quad (1)$$

The response pattern from equation (1) is shown in Figure 1. As with most IRFs,  $p_i(\theta)$  monotonically increases with examinee ability  $\theta$ , and decreases with item difficulty  $\delta_i$ .

In typical standardized test development, items are first created and then “pilot tested” with human subjects. These pilot tests produce many (examinee, item) pairs that are graded correct

or incorrect, and the next step is to estimate  $\theta$  and  $\delta_i$  parameters empirically from these grades. The reader may recognize the Rasch model as equivalent to binary logistic regression for predicting whether an examinee will answer item  $i$  correctly (where  $\theta$  represents a weight for the “examinee feature,”  $-\delta_i$  represents a weight for the “item feature,” and the bias/intercept weight is zero). Once parameters are estimated,  $\theta$ s for the pilot population can be discarded, and  $\delta_i$ s are used to estimate  $\theta$  for a future examinee, which ultimately determines his or her test score.

We focus on the Rasch model because item difficulty  $\delta_i$  and examinee ability  $\theta$  are interpreted on the same scale. Whereas other IRT models exist to generalize the Rasch model in various ways (e.g., by accounting for item discrimination or examinee guessing), the additional parameters make them more difficult to estimate correctly (Linacre, 2014). Our goal in this work is to estimate item parameters using ML/NLP (rather than traditional item piloting), and a Rasch-like model gives us a straightforward and elegant way to do this.

### 2.2 Computer-Adaptive Testing (CAT)

Given a bank of test items and their associated  $\delta_i$ s, one can use CAT techniques to efficiently administer and score tests. CATs have been shown to both shorten tests (Weiss and Kingsbury, 1984) and provide uniformly precise scores for most examinees, by giving harder items to subjects of higher ability and easier items to those of lower ability (Thissen and Mislevy, 2000).

Assuming test item independence, the conditional probability of an item response sequence  $\mathbf{r} = \langle r_1, r_2, \dots, r_t \rangle$  given  $\theta$  is the product of all the item-specific IRF probabilities:

$$p(\mathbf{r}|\theta) = \prod_{i=1}^t p_i(\theta)^{r_i} (1 - p_i(\theta))^{1-r_i}, \quad (2)$$

where  $r_i$  denotes the graded response to item  $i$  (i.e.,  $r_i = 1$  if correct,  $r_i = 0$  if incorrect).

The goal of a CAT is to estimate a new examinee's  $\theta$  as precisely as possible with as few items as possible. The precision of  $\theta$  depends on the items in  $\mathbf{r}$ : Examinees are best evaluated by items where  $\delta_i \approx \theta$ . However, because the true value of  $\theta$  is unknown (this is, after all, the reason for testing!), we use an iterative adaptive algorithm. First, make a “provisional”

CEFR	Level Description	Scale
C2	Proficient / Mastery	100
C1	Advanced / Effective	80
B2	Upper Intermediate / Vantage	60
B1	Intermediate / Threshold	40
A2	Elementary / Waystage	20
A1	Beginner / Breakthrough	0

Table 1: The Common European Framework of Reference (CEFR) levels and our corresponding test scale.

estimate  $\hat{\theta}_t \propto \arg\max_{\theta} p(\mathbf{r}_t|\theta)$  by maximizing the likelihood of observed responses up to point  $t$ . Then, select the next item difficulty based on a “utility” function of the current estimate  $\delta_{t+1} = f(\hat{\theta}_t)$ . This process repeats until reaching some stopping criterion, and the final  $\hat{\theta}_t$  determines the test score. Conceptually, CAT methods are analogous to *active learning* in the ML/NLP literature (Settles, 2012), which aims to minimize the effort required to train accurate classifiers by adaptively selecting instances for labeling. For more discussion on CAT administration and scoring, see Segall (2005).

### 2.3 The Common European Framework of Reference (CEFR)

The *Common European Framework of Reference (CEFR)* is an international standard for describing the proficiency of foreign-language learners (Council of Europe, 2001). Our goal is to create a test integrating reading, writing, listening, and speaking skills into a single overall score that corresponds to CEFR-derived ability. To that end, we designed a 100-point scoring system aligned to the CEFR levels, as shown in Table 1.

By its nature, the CEFR is a descriptive (not prescriptive) proficiency framework. That is, it describes what kinds of activities a learner should be able to do—and competencies they should have—at each level, but provides little guidance on what specific aspects of language (e.g., vocabulary) are needed to accomplish them. This helps the CEFR achieve its goal of applying broadly across languages, but also presents a challenge for curriculum and assessment development for any particular language. It is a coarse description of potential target domains—tasks, contexts, and conditions associated with language use (Bachman and Palmer, 2010; Kane, 2013)—that can be

sampled from in order to create language curricula or assessments. As a result, it is left to the developers to define and operationalize constructs based on the CEFR, targeting a subset of the activities and competences that it describes.

Such work can be seen in recent efforts undertaken by linguists to profile the vocabulary and grammar linked to each CEFR level for specific languages (particularly English). We leverage these lines of research to create labeled data sets, and train ML/NLP models that project item difficulty onto our CEFR-derived scale.

### 2.4 Test Construct and Item Formats

Our aim is to develop a test of general English language proficiency. According to the CEFR global descriptors, this means the ability to understand written and spoken language from varying topics, genres, and linguistic complexity, and to write or speak on a variety of topics and for a variety of purposes (Council of Europe, 2001).

We operationalize part of this construct using five item formats from the language testing literature. These are summarized in Table 2 and collectively assess reading, writing, listening, and speaking skills. Note that these items may not require examinees to perform all the linguistic tasks relevant to a given CEFR level (as is true with any language test), but they serve as strong proxies for the underlying skills. These formats were selected because they can be automatically generated and graded at scale, and have decades of research demonstrating their ability to predict linguistic competence.

Two of the formats assess vocabulary breadth, known as *yes/no* vocabulary tests (Figure 2). These both follow the same convention but vary in modality (text vs. audio), allowing us to measure both written and spoken vocabulary. For these items, the examinee must select, from among text or audio stimuli, which are real English words and which are English-like pseudowords (morphologically and phonologically plausible, but have no meaning in English). These items target a foundational linguistic competency of the CEFR, namely, the written and spoken vocabulary required to meet communication needs across CEFR levels (Milton, 2010). Test takers who do well on these tasks have a broader lexical inventory, allowing for performance in a variety of language use situations. Poor performance on these tasks indicates a more basic inventory.

Item Format	Scale Model	Skills	References
Yes/No (text)	Vocab (§3)	L,R,W	Zimmerman et al. (1977); Staehr (2008); Milton (2010)
Yes/No (audio)	Vocab (§3)	L,S	Milton et al. (2010); Milton (2010)
C-Test	Passage (§4)	R,W	Klein-Braley (1997); Reichert et al. (2010); Khodadady (2014)
Dictation	Passage (§4)	L,W	Bradlow and Bent (2002, 2008)
Elicited Speech	Passage (§4)	R,S	Vinther (2002); Jessop et al. (2007); Van Moere (2012)

Table 2: Summary of language assessment item formats in this work. For each format, we indicate the machine-learned scale model used to predict item difficulty  $\delta_i$ , the linguistic skills it is known to predict (L = listening, R = reading, S = speaking, W = writing), and some of the supporting evidence from the literature.

(a) Yes/No (text)

(b) Yes/No (audio)

Figure 2: Example test item formats that use the vocabulary scale model to estimate difficulty.

The other three item formats come out of the integrative language testing tradition (Alderson et al., 1995), which requires examinees to draw on a variety of language skills (e.g., grammar, discourse) and abilities (e.g., reading, writing) in order to respond correctly. Example screenshots of these item formats are shown in Figure 4.

The *c-test* format is a measure of reading ability (and to some extent, writing). These items contain passages of text in which some of the words have been “damaged” (by deleting the second half of every other word), and examinees must complete the passage by filling in missing letters from the damaged words. The characteristics of the damaged words and their relationship to the text ranges from those requiring lexical, phrasal, clausal, and discourse-level comprehension in order to respond correctly. These items indicate how well test takers can process texts of varied abstractness and complexity versus shorter more concrete texts, and have been shown to reliably

predict other measures of CEFR level (Reichert et al., 2010).

The *dictation* task taps into both listening and writing skills by having examinees transcribe an audio recording. In order to respond successfully, examinees must parse individual words and understand their grammatical relationships prior to typing what they hear. This targets the linguistic demands required for overall listening comprehension as described in the CEFR. The writing portion of the dictation task measures examinee knowledge of orthography and grammar (markers of writing ability at the A1/A2 level), and to some extent meaning. The *elicited speech* task taps into reading and speaking skills by requiring examinees to say a sentence out loud. Test takers must be able to process the input (e.g., orthography and grammatical structure) and are evaluated on their fluency, accuracy, and ability to use complex language orally (Van Moere, 2012). This task targets sentence-level language skills that incorporate simple-to-complex components of both the reading and speaking “can-do” statements in the CEFR framework. Furthermore, both the dictation and elicited speech tasks also measure working memory capacity in the language, which is regarded as shifting from lexical competence to structure and pragmatics somewhere in the B1/B2 range (Westhoff, 2007).

### 3 The Vocabulary Scale

For the experiments in this section, a panel of linguistics PhDs with ESL teaching experience first compiled a CEFR vocabulary wordlist, synthesizing previous work on assessing active English language vocabulary knowledge (e.g., Capel, 2010, 2012; Cambridge English, 2012). This standard-setting step produced an inventory of 6,823 English words labeled by CEFR level, mostly in the B1/B2 range (■■■■). We did not conduct

any formal annotator agreement studies, and the inventory does include duplicate entries for types at different CEFR levels (e.g., for words with multiple senses). We used this labeled wordlist to train a vocabulary scale model that assigns  $\delta_i$  scores to each yes/no test item (Figure 2).

### 3.1 Features

Culligan (2015) found character length and corpus frequency to significantly predict word difficulty, according IRT analyses of multiple vocabulary tests (including the yes/no format). This makes them promising features for our CEFR-based vocabulary scale model.

Although character length is straightforward, corpus frequencies only exist for *real* English words. For our purposes, however, the model must also make predictions for English-like *pseudowords*, since our CAT approach to yes/no items requires examinees to distinguish between words and pseudowords drawn from a similar CEFR-based scale range. As a proxy for frequency, we trained a character-level Markov chain language model on the OpenSubtitles corpus<sup>1</sup> using modified Kneser-Ney smoothing (Heafield et al., 2013). We then use the log-likelihood of a word (or pseudoword) under this model as a feature.

We also use the *Fisher score* of a word under the language model to generate more nuanced orthographic features. The Fisher score  $\nabla x$  of word  $x$  is a vector representing the gradient of its log-likelihood under the language model, parameterized by  $\mathbf{m}$ :  $\nabla x = \frac{\partial}{\partial \mathbf{m}} \log p(x|\mathbf{m})$ . These features are conceptually similar to trigrams weighted by *tf-idf* (Elkan, 2005), and are inspired by previous work leveraging information from generative sequence models to improve discriminative classifiers (Jaakkola and Haussler, 1999).

### 3.2 Models

We consider two regression approaches to model the CEFR-based vocabulary scale: *linear* and *weighted-softmax*. Let  $y_x$  be the CEFR level of word  $x$ , and  $\delta(y_x)$  be the 100-point scale value corresponding to that level from Table 1.

For the linear approach, we treat the difficulty of a word as  $\delta_x = \delta(y_x)$ , and learn a linear function with weights  $\mathbf{w}$  on the features of  $x$  directly. For weighted-softmax, we train a six-way multinomial

<sup>1</sup>We found movie subtitle counts (Lison and Tiedemann, 2016) to be more correlated with the expert CEFR judgments than other language domains (e.g., Wikipedia or newswire).

Vocabulary Scale Model	$r_{ALL}$	$r_{XV}$
Linear regression	.98	.30
w/o character length	.98	.31
w/o log-likelihood	.98	.34
w/o Fisher score	.38	.38
Weighted-softmax regression	.90	.56
w/o character length	.91	.56
w/o log-likelihood	.89	.51
w/o Fisher score	.46	.46

Table 3: Vocabulary scale model evaluations.

$\approx \delta$	English Words	Pseudowords
90	loft, proceedings	fortheric, retray
70	brutal, informally	insequent, vasera
50	delicious, unfairly	anage, compatively
30	into, rabbit	knoce, thace
10	egg, mother	cload, eut

Table 4: Example words and pseudowords, rated for difficulty by the weighted-softmax vocabulary model.

regression (MaxEnt) classifier to predict CEFR level, and treat difficulty  $\delta_x = \sum_y \delta(y)p(y|x, \mathbf{w})$  as a weighted sum over the posterior  $p(y|x, \mathbf{w})$ .

### 3.3 Experiments

Experimental results are shown in Table 3. We report Pearson’s  $r$  between predictions and expert CEFR judgments as an evaluation measure. The  $r_{ALL}$  results train and evaluate using the same data; this is how models are usually analyzed in the applied linguistics literature, and provides a sense of how well the model captures word difficulty for *real* English words. The  $r_{XV}$  results use 10-fold cross-validation; this is how models are usually evaluated in the ML/NLP literature, and gives us a sense of how well it generalizes to English-like *pseudowords* (as well as English words beyond the expert CEFR wordlist).

Both models have a strong, positive relationship with expert human judgments ( $r_{ALL} \geq .90$ ), although they generalize to unseen words less well ( $r_{XV} \leq .60$ ). Linear regression appears to drastically overfit compared to weighted-softmax, since it reconstructs the training data almost perfectly while explaining little of the variance among cross-validated labels. The feature ablations also reveal that Fisher score features are

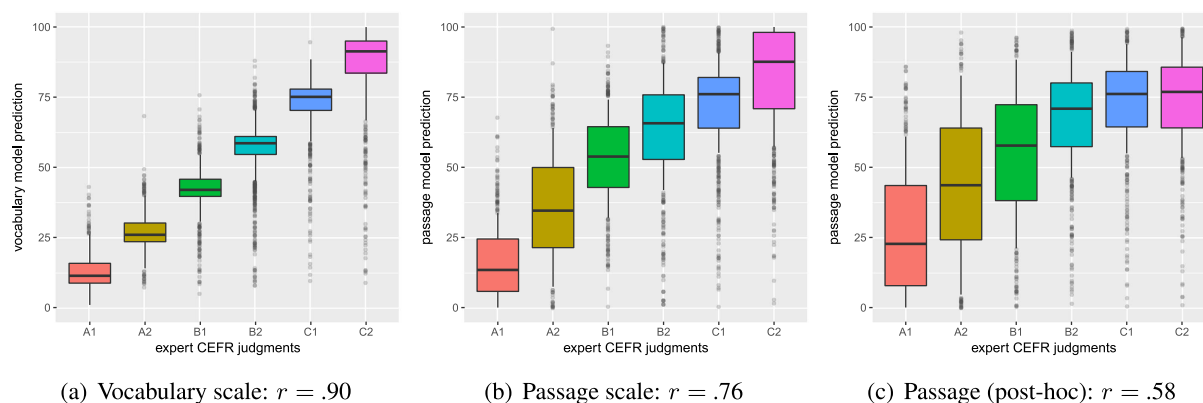


Figure 3: Boxplots and correlation coefficients evaluating our machine-learned proficiency scale models. (a) Results for the weighted-softmax vocabulary model ( $n = 6,823$ ). (b) Cross-validation results for the weighted-softmax passage model ( $n = 3,049$ ). (c) Results applying the trained passage model, post-hoc, to a novel set of “blind” texts written by ESL experts at targeted CEFR levels ( $n = 2,349$ ).

the most important, while character length has little impact (possibly because length is implicitly captured by all the Fisher score features).

Sample predictions from the weighted-softmax vocabulary scale model are shown in Table 4. The more advanced words (higher  $\delta$ ) are rarer and mostly have Greco-Latin etymologies, whereas the more basic words are common and mostly have Anglo-Saxon origins. These properties appear to hold for non-existent pseudowords (e.g., ‘cloud’ seems more Anglo-Saxon and more common than ‘fortheric’ would be). Although we did not conduct any formal analysis of pseudoword difficulty, these illustrations suggest that the model captures qualitative subtleties of the English lexicon, as they relate to CEFR level.

Boxplots visualizing the relationship between our learned scale and expert judgments are shown in Figure 3(a). Qualitative error analysis reveals that the majority of mis-classifications are in fact under-predictions simply due to polysemy. For example: ‘*a just cause*’ (C1) vs. ‘*I just left*’ ( $\delta = 24$ ), and ‘*to part ways*’ (C2) vs. ‘*part of the way*’ ( $\delta = 11$ ). Because these more basic word senses do exist, our correlation estimates may be on the conservative side. Thus, using these predicted word difficulties to construct yes/no items (as we do later in §5) seems justified.

## 4 The Passage Scale


For the experiments in this section, we leverage a variety of corpora gleaned from online sources, and use combined regression and ranking techniques to train longer-form passage scale models.

Type the missing letters to complete the text below

Other nations have also been successful when hosting the tournament. Sweden (runners-up i n 1958), Chile ( t h i r d place i n 1962 ), Korea Republic ( fourth p l a c e in 2002 ), and Mexico ( quarter-finals in 1970 and 1986 ) a h have t h best r e s when s e r as h o . So far, South Africa (2010) was the only host nation to fail to advance beyond the first round.

(a) C-Test


Type in the English statement that you hear

 He see himself in the mirror|

Number of replays left: 2

(b) Dictation

Click the microphone and say:

 Click to stop

“My telephone number has changed permanently.”

3.1 SEC

(c) Elicited Speech

Figure 4: Example test item formats that use the passage scale model to estimate difficulty.

These models can be used to predict difficulty for c-test, dictation, and elicited speech items (Figure 4).

In contrast to vocabulary, little to no work has been done to profile CEFR text or discourse features for English, and only a handful of “CEFR-labeled” documents are even available for model training. Thus, we take a *semi-supervised learning*



approach (Zhu and Goldberg, 2009), first by learning to rank passages by overall difficulty, and then by propagating CEFR levels from a small number of labeled texts to many more unlabeled texts that have similar linguistic features.

#### 4.1 Features

Average word length and sentence length have long been used to predict text difficulty, and in fact measures based solely on these features have been shown to correlate ( $r = .91$ ) with comprehension in reading tests (DuBay, 2006). Inspired by our vocabulary model experiments, we also trained a word-level unigram language model to produce log-likelihood and Fisher score features (which is similar to a bag of words weighted by *tf-idf*).

#### 4.2 Corpora

We gathered an initial training corpus from on-line English language self-study Web sites (e.g., free test preparation resources for popular English proficiency exams). These consist of reference phrases and texts from reading comprehension exercises, all organized by CEFR level. We segmented these documents and assigned documents’ CEFR labels to each paragraph. This resulted in 3,049 CEFR-labeled passages, containing very few A1 texts, and a peak at the C1 level (■■■■■). We refer to this corpus as CEFR.

Due to the small size of the CEFR corpus and its uncertain provenance, we also downloaded pairs of articles from English Wikipedia<sup>2</sup> that had also been rewritten for Simple English<sup>3</sup> (an alternate version that targets children and adult English learners). Although the CEFR alignment for these articles is unknown, we hypothesize that the levels for texts on the English site should be higher than those on the Simple English site; thus by comparing these article pairs a model can learn features related to passage difficulty, and therefore the CEFR level (in addition to expanding topical coverage beyond those represented in CEFR). This corpus includes 3,730 article pairs resulting in 18,085 paragraphs (from both versions combined). We refer to this corpus as WIKI.

We also downloaded thousands of English sentences from Tatoeba,<sup>4</sup> a free, crowd-sourced database of self-study resources for language learners. We refer to this corpus as TATOEBEA.

<sup>2</sup><https://en.wikipedia.org>.

<sup>3</sup><https://simple.wikipedia.org>.

<sup>4</sup><https://tatoeba.org>.

Passage Ranking Model	AUC <sub>CEFR</sub>	AUC <sub>WIKI</sub>
Linear (rank) regression	.85	.75
w/o characters per word	.85	.72
w/o words per sentence	.84	.75
w/o log-likelihood	.85	.76
w/o Fisher score	.79	.84

Table 5: Passage ranking model evaluations.

#### 4.3 Ranking Experiments

To rank passages for difficulty, we use a linear approach similar to that of Sculley (2010). Let  $\mathbf{x}$  be the feature vector for a text with CEFR label  $y$ . A standard linear regression can learn a weight vector  $\mathbf{w}$  such that  $\delta(y) \approx \mathbf{x}^\top \mathbf{w}$ . Given a pair of texts, one can learn to rank by ‘‘synthesizing’’ a label and feature vector representing the difference between them:  $[\delta(y_1) - \delta(y_2)] \approx [\mathbf{x}_1 - \mathbf{x}_2]^\top \mathbf{w}$ . The resulting  $\mathbf{w}$  can still be applied to single texts (i.e., by subtracting the  $\mathbf{0}$  vector) in order to score them for ranking. Although the resulting predictions are not explicitly calibrated (e.g., to our CEFR-based scale), they should still capture an overall ranking of textual sophistication. This also allows us to combine the CEFR and WIKI corpora for training, since relative difficulty for the latter is known (even if precise CEFR levels are not).

To train ranking models, we sample 1% of paragraph pairs from CEFR (up to 92,964 instances), and combine this with the cross of all paragraphs in English  $\times$  Simple English versions of the same article from WIKI (up to 25,438 instances). We fix  $\delta(y) = 25$  for Simple English and  $\delta(y) = 75$  for English in the WIKI pairs, under a working assumption that (on average) the former are at the A2/B1 level, and the latter B2/C1.

Results using cross-validation are shown in Table 5. For each fold, we train using pairs from the training partition and evaluate using individual instance scores on the test partition. We report the AUC, or area under the ROC curve (Fawcett, 2006), which is a common ranking metric for classification tasks. Ablation results show that Fisher score features (i.e., weighted bag of words) again have the strongest effect, although they improve ranking for the CEFR subset while harming WIKI. We posit that this is because WIKI is topically balanced (all articles have an analog from both versions of the site), so word and sentence length alone are in fact good discriminators. The CEFR results indicate

$\approx \delta$	Candidate Item Text
90	A related problem for aerobic organisms is oxidative stress. Here, processes including oxidative phosphorylation and the formation of disulfide bonds during protein folding produce reactive oxygen species such as hydrogen peroxide. These damaging oxidants are removed by antioxidant metabolites such as glutathione, and enzymes such as catalases and peroxidases.
50	In 1948, Harry Truman ran for a second term as President against Thomas Dewey. He was the underdog and everyone thought he would lose. The Chicago Tribune published a newspaper on the night of the election with the headline ‘‘Dewey Defeats Truman.’’ To everyone’s surprise, Truman actually won.
10	Minneapolis is a city in Minnesota. It is next to St. Paul, Minnesota. St. Paul and Minneapolis are called the ‘‘Twin Cities’’ because they are right next to each other. Minneapolis is the biggest city in Minnesota with about 370,000 people. People who live here enjoy the lakes, parks, and river. The Mississippi River runs through the city.

Table 6: Example WIKI paragraphs, rated for predicted difficulty by the weighted-softmax passage model.

that 85% of the time, the model correctly ranks a more difficult passage above a simpler one (with respect to CEFR level).<sup>5</sup>

#### 4.4 Scaling Experiments

Given a text ranking model, we now present experiments with the following algorithm for propagating CEFR levels from labeled texts to unlabeled ones for semi-supervised training:

1. Score all individual passages in CEFR, WIKI, and TATOEB A (using the ranking model);
2. For each labeled instance in CEFR, propagate its CEFR level to the five most similarly ranked neighbors in WIKI and TATOEB A;
3. Combine the label-propagated passages from WIKI and TATOEB A with CEFR;
4. Balance class labels by sampling up to 5,000 passages per CEFR level (30,000 total);
5. Train a passage scale model using the resulting CEFR-aligned texts.

Cross-validation results for this procedure are shown in Table 7. The weighted-softmax regression has a much stronger positive relationship with CEFR labels than simple linear regression. Furthermore, the label-propagated WIKI and TATOEB A supplements offer small but statistically significant improvements over training on CEFR texts alone. Since these supplemental passages also expand the feature set more than tenfold (i.e., by

<sup>5</sup>AUC is also the effect size of the Wilcoxon rank-sum test, which represents the probability that a randomly chosen text from WIKI English will be ranked higher than Simple English. For CEFR, Table 5 reports macro-averaged AUC over the five ordinal breakpoints between CEFR levels.

Passage Scale Model	$r_{\text{cefr}}$
Weighted-softmax regression	.76
w/o TATOEB A propagations	.75
w/o WIKI propagations	.74
w/o label-balancing	.72
Linear regression	.13

Table 7: Passage scale model evaluations.

increasing the model vocabulary for Fisher score features), we claim this also helps the model generalize better to unseen texts in new domains.

Boxplots illustrating the positive relationship between scale model predictions and CEFR labels are shown in Figure 3(b). This, while strong, may also be a conservative correlation estimate, since we propagate CEFR document labels down to paragraphs for training and evaluation and this likely introduces noise (e.g., C1-level articles may well contain A2-level paragraphs).

Example predictions from the WIKI corpus are shown in Table 6. We can see that the C-level text ( $\delta \approx 90$ ) is rather academic, with complex sentence structures and specialized jargon. On the other hand, the A-level text ( $\delta \approx 10$ ) is more accessible, with short sentences, few embedded clauses, and concrete vocabulary. The B-level text ( $\delta \approx 50$ ) is in between, discussing a political topic using basic grammar, but some colloquial vocabulary (e.g., ‘underdog’ and ‘headline’).

#### 4.5 Post-Hoc Validation Experiment

The results from §4.3 and §4.4 are encouraging. However, they are based on data gathered from the Internet, of varied provenance, using possibly noisy labels. Therefore, one might question whether the resulting scale model correlates well with more trusted human judgments.



To answer this question, we had a panel of four experts—PhDs and graduate students in linguistics with ESL teaching experience—compose roughly 400 new texts targeting each of the six CEFR levels (2,349 total). These were ultimately converted into c-test items for our operational English test experiments (§5), but because they were developed independently from the passage scale model, they are also suitable as a “blind” test set for validating our approach. Each passage was written by one expert, and vetted by another (with the two negotiating the final CEFR label in the case of any disagreement).

Boxplots illustrating the relationship between the passage scale model predictions and expert judgments are shown in Figure 3(c), which shows a moderately strong, positive relationship. The flattening at the C1/C2 level is not surprising, since the distinction here is very fine-grained, and can be difficult even for trained experts to distinguish or produce (Isbell, 2017). They may also be dependent on genre or register (e.g., textbooks), thus the model may have been looking for features in some of these expert-written passages that were missing for non-textbook-like writing samples.

## 5 Duolingo English Test Results

The Duolingo English Test<sup>6</sup> is an accessible, online, computer-adaptive English assessment initially created using the methods proposed in this paper. In this section, we first briefly describe how the test was developed, administered, and scored (§5.1). Then, we use data logged from many thousands of operational tests to show that our approach can satisfy industry standards for psychometric properties (§5.2), criterion validity (§5.3), reliability (§5.4), and test item security (§5.5).

### 5.1 Test Construction and Administration

Drawing on the five formats discussed in §2.4, we automatically generated a large bank of more than 25,000 test items. These items are indexed into eleven bins for each format, such that each bin corresponds to a predicted difficulty range on our 100-point scale (0–5, 6–15, . . . , 96–100).

The CAT administration algorithm chooses the first item format to use at random, and then cycles through them to determine the format for each subsequent item (i.e., all five formats have

equal representation). Each session begins with a “calibration” phase, where the first item is sampled from the first two difficulty bins, the second item from the next two, and so on. After the first four items, we use the methods from §2.2 to iteratively estimate a provisional test score, select the difficulty  $\delta_i$  of the next item, and sample randomly from the corresponding bin for the next format. This process repeats until the test exceeds 25 items or 40 minutes in length, whichever comes first. Note that because item difficulties ( $\delta_i$ s) are on our 100-point CEFR-based scale, so are the resulting test scores ( $\theta$ s). See Appendix A.1 for more details on test administration.

For the yes/no formats, we used the vocabulary scale model (§3) to estimate  $\delta_x$  for all words in an English dictionary, plus 10,000 pseudowords.<sup>7</sup> These predictions were binned by  $\delta_x$  estimate, and test items created by sampling both dictionaries from the same bin (each item also contains at least 15% words and 15% pseudowords). Item difficulty  $\delta_i = \bar{\delta}_x$  is the mean difficulty of all words/pseudowords  $x \in i$  used as stimuli.

For the c-test format, we combined the expert-written passages from §4.5 with paragraphs extracted from other English-language sources, including the wiki corpus and English-language literature.<sup>8</sup> We followed standard procedure (Klein-Braley, 1997) to automatically generate c-test items from these paragraphs. For the dictation and elicited speech formats, we used sentence-level candidate texts from WIKI, TATOEBE, English Universal Dependencies,<sup>9</sup> as well as custom-written sentences. All passages were then manually reviewed for grammaticality (making corrections where necessary) or filtered for inappropriate content. We used the passage scale model (§4) to estimate  $\delta_i$  for these items directly from raw text.

For items requiring audio (i.e., audio yes/no and elicited speech items), we contracted four native English-speaking voice actors (two male, two female) with experience voicing ESL instructional materials. Each item format also has its own stat-

<sup>7</sup>We trained a character-level LSTM RNN (Graves, 2014) on an English dictionary to produce pseudowords, and then filtered out any real English words. Remaining candidates were manually reviewed and filtered if they were deemed too similar to real words, or were otherwise inappropriate.

<sup>8</sup><https://www.wikibooks.org>.

<sup>9</sup><http://universaldependencies.org>.

<sup>6</sup><https://englishtest.duolingo.com>.

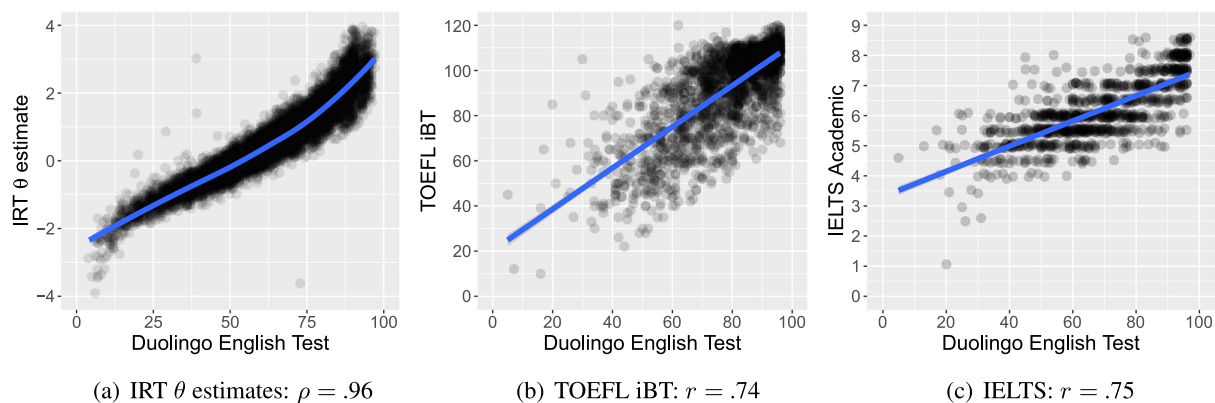


Figure 5: Scatterplots and correlation coefficients showing how Duolingo English Test scores, based on our ML/NLP scale models, relate to other English proficiency measures. (a) Our test score rankings are nearly identical to those of traditional IRT  $\theta$  estimates fit to real test session data ( $n = 21,351$ ). (b–c) Our test scores correlate significantly with other high-stakes English assessments such as TOEFL iBT ( $n = 2,319$ ) and IELTS ( $n = 991$ ).

istical grading procedure using ML/NLP. See Appendix A.2 for more details.

## 5.2 Confirmatory IRT Analysis

Recall that the traditional approach to CAT development is to first create a bank of items, then pilot test them extensively with human subjects, and finally use IRT analysis to estimate item  $\delta_i$  and examinee  $\theta$  parameters from pilot data. What is the relationship between test scores based on our machine-learned CEFR-derived scales and such pilot-tested ability estimates? A strong relationship between our scores and  $\theta$  estimates based on IRT analysis of real test sessions would provide evidence that our approach is valid as an alternative form of pilot testing.

To investigate this, we analyzed 524,921 ⟨examinee, item⟩ pairs from 21,351 of the tests administered during the 2018 calendar year, and fit a Rasch model to the observed response data post-hoc.<sup>10</sup> Figure 5(a) shows the relationship between our test scores and more traditional “pilot-tested” IRT  $\theta$  estimates. The Spearman rank correlation is positive and very strong ( $\rho = .96$ ), indicating that scores using our method produce rankings nearly identical to what traditional IRT-based human pilot testing would provide.

<sup>10</sup>Because the test is adaptive, most items are rarely administered (§5.5). Thus, we limit this analysis to items with  $>15$  observations to be statistically sound. We also omit sessions that went unscored due to evidence of rule-breaking (§A.1).

## 5.3 Relationship with Other English Language Assessments

One source of *criterion validity* evidence for our method is the relationship between these test scores and other measures of English proficiency. A strong correlation between our scores and other major English assessments would suggest that our approach is well-suited for assessing language proficiency for people who want to study or work in and English-language environment. For this, we compare our results with two other high-stakes English tests: TOEFL iBT<sup>11</sup> and IELTS.<sup>12</sup>

After completing our test online, we asked examinees to submit official scores from other tests (if available). This resulted in a large collection of recent parallel scores to compare against. The relationships between our test scores with TOEFL and IELTS are shown in Figures 5(b) and 5(c), respectively. Correlation coefficients between language tests are generally expected to be in the .5–.7 range (Alderson et al., 1995), so our scores correlate very well with both tests ( $r > .7$ ). Our relationship with TOEFL and IELTS appears, in fact, to be on par with their published relationship with each other ( $r = .73$ ,  $n = 1,153$ ), which is also based on self-reported data (ETS, 2010).

## 5.4 Score Reliability

Another aspect of test validity is the *reliability* or overall consistency of its scores (Murphy

<sup>11</sup><https://www.ets.org/toefl>.

<sup>12</sup><https://www.ielts.org/>.

Reliability Measure	<i>n</i>	Estimate
Internal consistency	9,309	.96
Test-retest	526	.80

Table 8: Test score reliability estimates.

Security Measure	Mean	Median
Item exposure rate	.10%	.08%
Test overlap rate	.43%	<.01%

Table 9: Test item bank security measures.

and Davidshofer, 2004). Reliability coefficient estimates for our test are shown in Table 8. Importantly, these are high enough to be considered appropriate for high-stakes use.

*Internal consistency* measures the extent to which items in the test measure the same underlying construct. For CATs, this is usually done using the “split half” method: randomly split the item bank in two, score both halves separately, and then compute the correlation between half-scores, adjusting for test length (Sireci et al., 1991). The reliability estimate is well above .9, the threshold for tests “intended for individual diagnostic, employment, academic placement, or other important purposes” (DeVellis, 2011).

*Test-retest* reliability measures the consistency of people’s scores if they take the test multiple times. We consider all examinees who took the test twice within a 30-day window (any longer may reflect actual learning gains, rather than measurement error) and correlate the first score with the second. Such coefficients range from .8–.9 for standardized tests using identical forms, and .8 is considered sufficient for high-stakes CATs, since adaptively administered items are distinct between sessions (Nitko and Brookhart, 2011).

### 5.5 Item Bank Security

Due to the adaptive nature of CATs, they are usually considered to be more secure than fixed-form exams, so long as the item bank is sufficiently large (Wainer, 2000). Two measures for quantifying the security of an item bank are the *item exposure rate* (Way, 1998) and *test overlap rate* (Chen et al., 2003). We report the mean and median values for these measures in Table 9.

The exposure rate of an item is the proportion of tests in which it is administered; the average item

exposure rate for our test is .10% (or one in every 1,000 tests). While few tests publish exposure rates for us to compare against, ours is well below the 20% (one in five tests) limit recommended for unrestricted continuous testing (Way, 1998). The test overlap rate is the proportion of items that are shared between any two randomly-chosen test sessions. The mean overlap for our test is .43% (and the median below .01%), which is well below the 11–14% range reported for other operational CATs like the GRE<sup>13</sup> (Stocking, 1994). These results suggest that our proposed methods are able to create very large item banks that are quite secure, without compromising the validity or reliability of resulting test scores.

## 6 Related Work

There has been little to no work using ML/NLP to drive end-to-end language test development as we do here. To our knowledge, the only other example is Hoshino and Nakagawa (2010), who used a support vector machine to estimate the difficulty of cloze<sup>14</sup> items for a computer-adaptive test. However, the test did not contain any other item formats, and it was not intended as an integrated measure of general language ability.

Instead, most related work has leveraged ML/NLP to predict test item difficulty from operational test logs. This has been applied with some success to cloze (Mostow and Jang, 2012), vocabulary (Susanti et al., 2016), listening comprehension (Loukina et al., 2016), and grammar exercises (Perez-Beltrachini et al., 2012). However, these studies all use multiple-choice formats where difficulty is largely mediated by the choice of distractors. The work of Beinborn et al. (2014) is perhaps most relevant to our own; they used ML/NLP to predict c-test difficulty at the word-gap level, using both macro-features (e.g., paragraph difficulty as we do) as well as micro-features (e.g., frequency, polysemy, or cognateness for each gap word). These models performed on par with human experts at predicting failure rates for English language students living in Germany.

Another area of related work is in predicting text difficulty (or readability) more generally. Napoles

<sup>13</sup><https://www.ets.org/gre>.

<sup>14</sup>Cloze tests and c-tests are similar, both stemming from the “reduced redundancy” approach to language assessment (Lin et al., 2008). The cloze items in the related work cited here contain a single deleted word with four multiple-choice options for filling in the blank.

and Dredze (2010) trained classifiers to discriminate between English and Simple English Wikipedia, and Vajjala et al. (2016) applied English readability models to a variety of Web texts (including English and Simple English Wikipedia). Both of these used linear classifiers with features similar to ours from §4.

Recently, more efforts have gone into using ML/NLP to align texts to specific proficiency frameworks like the CEFR. However, this work mostly focuses on languages other than English (e.g., Curto et al., 2015; Sung et al., 2015; Volodina et al., 2016; Vajjala and Rama, 2018). A notable exception is Xia et al. (2016), who trained classifiers to predict CEFR levels for reading passages from a suite of Cambridge English<sup>15</sup> exams, targeted at learners from A2–C2. In addition to lexical and language model features like ours (§4), they showed additional gains from explicit discourse and syntax features.

The relationship between test item difficulty and linguistic structure has also been investigated in the language testing literature, both to evaluate the validity of item types (Brown, 1989; Abraham and Chapelle, 1992; Freedle and Kostin, 1993, 1999) and to establish what features impact difficulty so as to inform test development (Nissan et al., 1995; Kostin, 2004). These studies have leveraged both correlational and regression analyses to examine the relationship between passage difficulty and linguistic features such as passage length, word length and frequency, negations, rhetorical organization, dialogue utterance pattern (question-question, statement-question), and so on.

## 7 Discussion and Future Work

We have presented a method for developing computer-adaptive language tests, driven by machine learning and natural language processing. This allowed us to rapidly develop an initial version of the Duolingo English Test for the experiments reported here, using ML/NLP to directly estimate item difficulties for a large item bank in lieu of expensive pilot testing with human subjects. This test correlates significantly with other high-stakes English assessments, and satisfies industry standards for score reliability and test security. To our knowledge, we are the

first to propose language test development in this way.

The strong relationship between scores based on ML/NLP estimates of item difficulty and the IRT estimates from operational data provides evidence that our approach—using items’ linguistic characteristics to predict difficulty, a priori to any test administration—is a viable form of test development. Furthermore, traditional pilot analyses produce inherently *norm-referenced* scores (i.e., relative to the test-taking population), whereas it can be argued that our method yields *criterion-referenced* scores (i.e., indicative of a given standard, in our case the CEFR). This is another conceptual advantage of our method. However, further research is necessary for confirmation.

We were able to achieve these results using simple linear models and relatively straightforward lexical and language model feature engineering. Future work could incorporate richer syntactic and discourse features, as others have done (§6). Furthermore, other indices such as narrativity, word concreteness, topical coherence, etc., have also been shown to predict text difficulty and comprehension (McNamara et al., 2011). A wealth of recent advances in neural NLP that may also be effective in this work.

Other future work involves better understanding how our large, automatically-generated item bank behaves with respect to the intended construct. Detecting differential item functioning (DIF)—the extent to which people of equal ability but different subgroups, such as gender or age, have (un)equal probability of success on test items—is an important direction for establishing the fairness of our test. While most assessments focus on demographics for DIF analyses, online administration means we must also ensure that technology differences (e.g., screen resolution or Internet speed) do not affect item functioning, either.

It is also likely that the five item formats presented in this work over-index on language *reception* skills rather than *production* (i.e., writing and speaking). In fact, we hypothesize that the “clipping” observed to the right in plots from Figure 5 can be attributed to this: Despite being highly correlated, the CAT as presented here may over estimate overall English ability relative to tests with more open-ended writing and speaking exercises. In the time since the present experiments were conducted, we have updated the Duolingo

<sup>15</sup><https://www.cambridgeenglish.org>.

English Test to include such writing and speaking sections, which are automatically graded and combined with the CAT portion. The test–retest reliability for these improved scores is .85, and correlation with TOEFL and IELTS are .77 and .78, respectively (also, the “clipping” effect disappears). We continue to conduct research on the quality of the interpretations and uses of Duolingo English Test scores; interested readers are able to find the latest ongoing research at <https://go.duolingo.com/dettechnicalmanual>.

Finally, in some sense what we have proposed here is partly a solution to the “cold start” problem facing language test developers: How does one estimate item difficulty without any response data to begin with? Once a test is in production, however, one can leverage the operational data to further refine these models. It is exciting to think that such analyses of examinees’ response patterns (e.g., topical characteristics, register types, and pragmatic uses of language in the texts) can tell us more about the underlying proficiency scale, which in turn can contribute back to the theory of frameworks like the CEFR.

## Acknowledgments

We would like to thank Micheline Chalhoub-Deville, Steven Sireci, Bryan Smith, and Alina von Davier for their input on this work, as well as Klinton Bicknell, Erin Gustafson, Stephen Mayhew, Will Monroe, and the *TACL* editors and reviewers for suggestions that improved this paper. Others who have contributed in various ways to research about our test to date include Cynthia M. Berger, Connor Brem, Ramsey Cardwell, Angela DiCostanzo, Andre Horie, Jennifer Lake, Yena Park, and Kevin Yancey.

## References

- R. G. Abraham and C. A. Chapelle. 1992. The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal*, 76(4):468–479.
- AERA, APA, and NCME. 2014. *Standards for Educational and Psychological Testing*.
- J. C. Alderson, C. Clapham, and D. Wall. 1995. *Language Test Construction and Evaluation*, Cambridge University Press.
- D. Andrich. 1978. A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573.
- L. Bachman and A. Palmer. 2010. *Language Assessment in Practice*. Oxford University Press.
- L. Beinborn, T. Zesch, and I. Gurevych. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, 2:517–530.
- A. R. Bradlow and T. Bent. 2002. The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America*, 112:272–284.
- A. R. Bradlow and T. Bent. 2008. Perceptual adaptation to non-native speech. *Cognition*, 106:707–729.
- J. D. Brown. 1989. Cloze item difficulty. *JALT Journal*, 11:46–67.
- Cambridge English. 2012. Preliminary wordlist.
- A. Capel. 2010. A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1.
- A. Capel. 2012. Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, 3.
- S. Cau. 2015. TOEFL questions, answers leaked in China. *Global Times*.
- S. Chen, R. D. Ankenmann, and J. A. Spray. 2003. Exploring the relationship between item exposure rate and item overlap rate in computerized adaptive testing. *Journal of Educational Measurement*, 40:129–145.
- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.
- B. Culligan. 2015. A comparison of three test formats to assess word difficulty. *Language Testing*, 32(4):503–520.
- P. Curto, N. J. Mamede, and J. Baptista. 2015. Automatic text difficulty classifier-assisting the selection of adequate reading materials for European Portuguese teaching. In *Proceedings*

- of the International Conference on Computer Supported Education, pages 36–44.
- P. T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstien. 2005. A tutorial on the cross-entropy method. *Annals of Operations Research*, 34:19–67.
- R. F. DeVellis. 2011. *Scale Development: Theory and Applications*, Number 26 in Applied Social Research Methods. SAGE Publications.
- W. H. DuBay. 2006. *Smart Language: Readers, Readability, and the Grading of Text*, Impact Information.
- R. Dudley, S. Stecklow, A. Harney, and I. J. Liu. 2016. As SAT was hit by security breaches, College Board went ahead with tests that had leaked. *Reuters Investigates*.
- C. Elkan. 2005. Deriving TF-IDF as a Fisher kernel. In M. Consens and G. Navarro, editors, *String Processing and Information Retrieval*, volume 3772 of *Lecture Notes in Computer Science*, pages 295–300. Springer.
- ETS. 2010. Linking TOEFL iBT scores to IELTS scores - A research report. ETS TOEFL Report.
- T. Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.
- R. Freedle and I. Kostin. 1993. The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items. ETS Research Report 93-13.
- R. Freedle and I. Kostin. 1999. Does the text matter in a multiple-choice test of comprehension? the case for the construct validity of TOEFL’s minitalks. *Language Testing*, 16(1):2–32.
- A. Graves. 2014. Generating sequences with recurrent neural networks. *arXiv*, 1308.0850v5 [cs.NE].
- K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 690–696.
- A. Hoshino and H. Nakagawa. 2010. Predicting the difficulty of multiple-choice close questions for computer-adaptive testing. *Research in Computing Science*, 46:279–292.
- D. Isbell. 2017. Assessing C2 writing ability on the Certificate of English Language Proficiency: Rater and examinee age effects. *Assessing Writing*, 34:37–49.
- T. Jaakkola and D. Haussler. 1999. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems (NIPS)*, volume 11, pages 487–493.
- L. Jessop, W. Suzuki, and Y. Tomita. 2007. Elicited imitation in second language acquisition research. *Canadian Modern Language Review*, 64(1):215–238.
- M.T. Kane. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50:1–73.
- E. Khodadady. 2014. Construct validity of C-tests: A factorial approach. *Journal of Language Teaching and Research*, 5(6):1353–1362.
- C. Klein-Braley. 1997. C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, 14(1):47–84.
- I. Kostin. 2004. Exploring item characteristics that are related to the difficulty of TOEFL dialogue items. ETS Research Report 04-11.
- S. Lane, M. R. Raymond, and S. M. Downing, editors. 2016. *Handbook of Test Development*, 2nd edition. Routledge.
- W. Y. Lin, H. C. Yuan, and H. P. Feng. 2008. Language reduced redundancy tests: A re-examination of cloze test and c-test. *Journal of Pan-Pacific Association of Applied Linguistics*, 12(1):61–79.
- J. M. Linacre. 2014. 3PL, Rasch, quality-control and science. *Rasch Measurement Transactions*, 27(4):1441–1444.
- P. Lison and J. Tiedemann. 2016. OpenSubtitles-2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 923–929.
- F. M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*, Routledge.



- A. Loukina, S. Y. Yoon, J. Sakano, Y. Wei, and K. Sheehan. 2016. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 3245–3253.
- G. N. Masters. 1982. A Rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174.
- D. S. McNamara, A. C. Graesser, Z. Cai, and J. Kulikowich. 2011. Coh-Metrix easability components: Aligning text difficulty with theories of text comprehension. In *Annual Meeting of the American Educational Research Association (AERA)*.
- J. Milton. 2010. The development of vocabulary breadth across the CEFR levels. In I. Bartning, M. Martin, and I. Vedder, editors, *Communicative Proficiency and Linguistic Development: Intersections Between SLA and Language Testing Research*, volume 1 of *EuroSLA Monograph Series*, pages 211–232. EuroSLA.
- J. Milton, J. Wade, and N. Hopkins. 2010. Aural word recognition and oral competence in English as a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, and M. Torreblanca-López, editors, *Insights Into Non-Native Vocabulary Teaching and Learning*, volume 52, pages 83–98. Multilingual Matters.
- J. Mostow and H. Jang. 2012. Generating diagnostic multiple choice comprehension cloze questions. In *Proceedings of the Workshop on Building Educational Applications Using NLP*, pages 136–146.
- K. R. Murphy and C. O. Davidshofer. 2004. *Psychological Testing: Principles and Applications*, Pearson.
- C. Napoles and M. Dredze. 2010. Learning Simple Wikipedia: A cogitation in ascertaining abecedarian language. In *Proceedings of the Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, pages 42–50.
- S. Nissan, F. DeVincenzi, and K. L. Tang. 1995. An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension. ETS Research Report 95-37.
- A. J. Nitko and S. Brookhart. 2011. *Educational Assessment of Students*. Merrill.
- L. Perez-Beltrachini, C. Gardent, and G. Kruszewski. 2012. Generating grammar exercises. In *Proceedings of the Workshop on Building Educational Applications Using NLP*, pages 147–156.
- G. Rasch. 1993. *Probabilistic Models for Some Intelligence and Attainment Tests*, MESA Press.
- M. Reichert, U. Keller, and R. Martin. 2010. The C-test, the TCF and the CEFR: A validation study. In *The C-Test: Contributions from Current Research*, pages 205–231. Peter Lang.
- D. Sculley. 2010. Combined regression and ranking. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD)*, pages 979–988.
- D. O. Segall. 2005. Computerized adaptive testing. In K. Kempf-Leonard, editor, *Encyclopedia of Social Measurement*. Elsevier.
- B. Settles. 2012. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool.
- S. G. Sireci, D. Thissen, and H. Wainer. 1991. On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3):237–247.
- L. S. Staehr. 2008. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36:139–152.
- M. L. Stocking. 1994. Three practical issues for modern adaptive testing item pools. ETS Research Report 94-5.
- Y. T. Sung, W. C. Lin, S. B. Dyson, K. E. Change, and Y. C. Chen. 2015. Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, 99(2):371–391.
- Y. Susanti, H. Nishikawa, T. Tokunaga, and H. Obari. 2016. Item difficulty analysis of english vocabulary questions. In *Proceedings of the International Conference on Computer Supported Education*, pages 267–274.
- D. Thissen and R. J. Mislevy. 2000. Testing algorithms. In H. Wainer, editor, *Computerized Adaptive Testing: A Primer*. Routledge.

- S. Vajjala, D. Meurers, A. Eitel, and K. Scheiter. 2016. Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 38–48.
- S. Vajjala and T. Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153.
- A. Van Moere. 2012. A psycholinguistic approach to oral assessment. *Language Testing*, 29:325–344.
- T. Vinther. 2002. Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, 12(1):54–73.
- E. Volodina, I. Pilán, and D. Alfter. 2016. Classification of Swedish learner essays by CEFR levels. In *Proceedings of EUROCALL*, pages 456–461.
- H. Wainer. 2000. *Computerized Adaptive Testing: A Primer*. Routledge.
- W. D. Way. 1998. Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, 17(4):17–27.
- D. J. Weiss and G. G. Kingsbury. 1984. Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21:361–375.
- G. Westhoff. 2007. Challenges and opportunities of the CEFR for reimagining foreign language pedagogy. *The Modern Language Journal*, 91(4):676–679.
- M. Xia, E. Kochmar, and T. Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the Workshop on Building Educational Applications Using NLP*, pages 12–22.
- X. Zhu and A. B. Goldberg. 2009. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool.
- J. Zimmerman, P. K. Broder, J. J. Shaughnessy, and B. J. Underwood. 1977. A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. *Intelligence*, 1(1):5–31.

## A Appendices

### A.1 Test Administration Details

Tests are administered remotely via Web browser at <https://englishtest.duolingo.com>. Examinees are required to have a stable Internet connection and a device with a working microphone and front-facing camera. Each test session is recorded and reviewed by human proctors before scores are released. Prohibited behaviors include:

- Interacting with another person in the room
- Using headphones or earbuds
- Disabling the microphone or camera
- Moving out of frame of the camera
- Looking off screen
- Accessing any outside material or devices
- Leaving the Web browser

Any of these behaviors constitutes rule-breaking; such sessions do not have their scores released, and are omitted from the analyses in this paper.

### A.2 Item Grading Details

The item formats in this work (Table 2) are not multiple-choice or true/false. This means responses may not be simply “correct” or “incorrect,” and require more nuanced grading procedures. While partial credit IRT models do exist (Andrich, 1978; Masters, 1982), we chose instead to generalize the binary Rasch framework to incorporate “soft” (probabilistic) responses.

The maximum-likelihood estimation (MLE) estimate used to score the test (or select the next item) is based on the log-likelihood function:

$$LL(\hat{\theta}_t) = \log \prod_{i=1}^t p_i(\hat{\theta}_t)^{r_i} (1 - p_i(\hat{\theta}_t))^{1-r_i},$$

which follows directly from equation (2). Note that maximizing this is equivalent to minimizing *cross-entropy* (de Boer et al., 2005), a measure

of disagreement between two probability distributions. As a result,  $r_i$  can just as easily be a probabilistic response ( $0 \leq r_i \leq 1$ ) as a binary one ( $r_i \in \{0, 1\}$ ). In other words, this MLE optimization seeks to find  $\hat{\theta}_t$  such that the IRF prediction  $p_i(\hat{\theta}_t)$  is most similar to each probabilistic response  $r_i$ . We believe the flexibility of this generalized Rasch-like framework helps us reduce test administration time above and beyond a binary-response CAT, since each item's grade summarizes multiple facets of the examinee's performance on that item. To use this generalization, however, we must specify a probabilistic grading procedure for each item format. Since an entire separate manuscript can be devoted to this topic, we simply summarize our approaches here.

The yes/no vocabulary format (Figure 2) is traditionally graded using the sensitivity index  $d'$ —a measure of separation between signal (word) and noise (pseudoword) distributions from signal detection theory (Zimmerman et al., 1977). This index is isomorphic with the AUC (Fawcett, 2006), which we use as the graded response  $r_i$ . This can be interpreted as “the probability that the examinee can discriminate between English words and pseudowords at level  $\delta_i$ .”

C-test items (Figure 4(a)) are graded using a weighted average of the correctly filled word-gaps, such that each gap's weight is proportional to its length in characters. Thus,  $r_i$  can be interpreted

as “the proportion of this passage the examinee understood, such that longer gaps are weighted more heavily.” (We also experimented with other grading schemes, but this yielded the highest test score reliability in preliminary work.)

The dictation (Figure 4(b)) and elicited speech (Figure 4(c)) items are graded using logistic regression classifiers. We align the examinee's submission (written for dictation; transcribed using automatic speech recognition for elicited speech) to the expected reference text, and extract features representing the differences in the alignment (e.g., string edit distance,  $n$ -grams of insertion/substitution/deletion patterns at both the word and character level, and so on). These models were trained on aggregate human judgments of correctness and intelligibility for tens of thousands of test item submissions (stratified by  $\delta_i$ ) collected during preliminary work. Each item received  $\geq 15$  independent binary judgments from fluent English speakers via Amazon Mechanical Turk,<sup>16</sup> which were then averaged to produce “soft” (probabilistic) training labels. Thus  $r_i$  can be interpreted as “the probability that a random English speaker would find this transcription/utterance to be faithful, intelligible, and accurate.” For the dictation grader, the correlation between human labels and model predictions is  $r = .86$  (10-fold cross-validation). Correlation for the elicited speech grader is  $r = .61$  (10-fold cross-validation).

<sup>16</sup><https://www.mturk.com/>.