Jump-Starting Item Parameters for Adaptive Language Tests

Arya D. McCarthy[1*], **Kevin P. Yancey**[2], **Geoffrey T. LaFlair**[2],
Jesse Egbert[3], **Manqian Liao**[2], **and Burr Settles**[2]

[1]Johns Hopkins University
[2]Duolingo
[3]Northern Arizona University

Abstract

A challenge in designing high-stakes language assessments is calibrating the test item difficulties, either a priori or from limited pilot test data. While prior work has addressed 'cold start' estimation of item difficulties without piloting, we devise a multi-task generalized linear model with BERT features to *jump-start* these estimates, rapidly improving their quality with as few as 500 test-takers and a small sample of item exposures (≈6 each) from a large item bank (≈4,000 items). Our joint model provides a principled way to compare test-taker proficiency, item difficulty, and language proficiency frameworks like the Common European Framework of Reference (CEFR). This also enables new item difficulty estimates without piloting them first, which in turn limits item exposure and thus enhances test security. Finally, using operational data from the Duolingo English Test, a high-stakes English proficiency test, we find that difficulty estimates derived using this method correlate strongly with lexico-grammatical features that correlate with reading complexity.

1 Introduction

High-stakes language assessment demands high reliability, validity, and security (AERA et al., 2014). These goals are at odds with each other during the test design process. Large-scale pilot testing of new **items** to accurately measure their psychometric characteristics (e.g., difficulty and discrimination) risks that those items will be copied and leaked (Cao, 2015; Dudley, 2016). Computer-adaptive tests, which more precisely score test-takers by selecting items of appropriate difficulty on-the-fly, exacerbate this conflict: their item banks must be large enough to cover all proficiency levels while ensuring the security of test items.

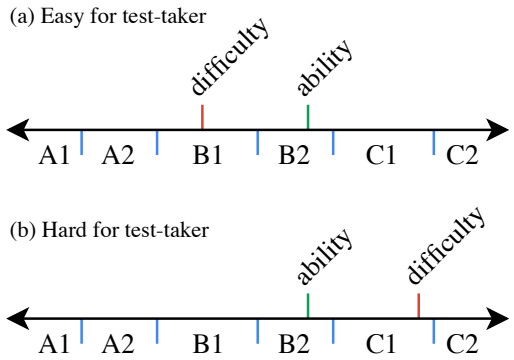


Figure 1: Model-estimated test-taker ability, item difficulty, and CEFR cutpoints all exist on the same real line and can be meaningfully compared (§3.3).

During the life cycle of the test, compromised or frequently shown items must be retired. Introducing newly written items to replace or grow the item bank usually entails new rounds of piloting. In this paper, we present a principled method to **jump start** new items without extensive piloting. A small amount of pilot data on older items (as few as 500 test-takers, each seeing 6 test items on average) is sufficient for calibrating new items' difficulty and discrimination parameters.

The closest work to ours (Settles et al., 2020) describes automatic estimation of item difficulty from linguistic features by supervised learning of an external difficulty scale: the **Common European Framework of Reference** (CEFR; Council of Europe, 2001). We recognize this as **weak supervision** from a related task (Pino et al., 2019; Ruder et al., 2019; Wang et al., 2019), and we seek to perform *direct* supervision of item difficulty from test-takers. Settles et al. (2020) provide no mechanism for this. Conversely, traditional **item response theory** (IRT; Lord, 1980; Hambleton et al., 1991) provides such mechanisms but is hamstrung by a *cold start problem* that Settles et al. (2020) address.

[*]Research conducted during an internship at Duolingo.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 883–899
November 7–11, 2021. ©2021 Association for Computational Linguistics

In this work, we fuse and generalize conventional IRT with the work of Settles et al. (2020), remedying limitations of each. A single model estimates difficulty *a priori* to guide piloting; permits incremental learning from test-takers; and jump-starts difficulty estimates of newly written items, with similar quality as on observed items.

We therefore make the following contributions:

1. We design (§3) and test (§4) a principled probabilistic model that incorporates arbitrary linguistic features of passage-based items. As illustrated in Figure 1, it measures item difficulty, test-taker ability, and CEFR level on a common logit scale (§3.3).
2. We show how BERT-derived passage embeddings (Devlin et al., 2019) facilitate strong generalization to new test items (**jump-starting**; §5) on a high-stakes English proficiency test. We outperform Settles et al. (2020) in all measures of difficulty estimation on test-taker response data while requiring *much less* data than traditional IRT models.
3. We provide linguistic validation of these difficulty estimates (§7): they correlate *strongly* with lexico-grammatical features known to characterize reading complexity, which helps to interpret test-takers' skills.

2 Background: Item Response Theory

Item response theory is the basis for most modern high-stakes standardized tests. It jointly assesses each person's ability and the difficulty of each item in the item bank. The major distinction between IRT and its predecessors, such as classical test theory, is that IRT assumes the item characteristics (e.g., difficulty) to be independent from the person or sample characteristics (e.g., the person ability distribution). IRT models relate the probability of answering correctly to a person's latent scalar ability θ through an item response function. The simplest such model is the **Rasch model** (Rasch, 1960), a special case of logistic regression with one parameter per test-taker p (their ability $\theta_p \in \mathbb{R}$) and one parameter per test item i (its difficulty $b_i \in \mathbb{R}$):

$$p(y = 1 \mid p, i) = \sigma(\theta_p - b_i). \quad (1)$$

where σ is the sigmoid function.[1] Extensions of IRT allow polytomous (rather than dichotomous) response variables (e.g., Masters, 1982; Andrich, 1978; Eckes, 2011) and integrate temporal knowledge by Bayesian knowledge tracing (Khajah et al., 2014). In §5, we compare our proposed model to an extension of the Rasch model called **2PL**:

$$p(y = 1 \mid p, i) = \sigma(a_i \cdot (\theta_p - b_i)), \quad (2)$$

where $a_i \in \mathbb{R}_{>0}$ is the item's discrimination parameter (Hambleton et al., 1991), governing the slope of the sigmoid function. Items with low discrimination are less sensitive to test-taker ability.

Decomposing items into linguistic skills We do not believe that the test items represent independent, atomic skills for a test-taker to master. Instead, each item may amalgamate several skills—perhaps mastery of certain linguistic attributes—that are shared across test items. The student's performance will depend on their attainment of each skill to different degrees.

One of the best-known item-explanatory IRT models is the **linear logistic test model** (LLTM; Fischer, 1973) that decomposes the item difficulty parameter of the Rasch model into a linear combination of features extracted from each item.

The Rasch model is a special case whose features are indicator functions—only one of which will be active per test item. In our LLTM, the total number of active features is higher, the total number of features is lower, and they are not orthogonal.[2] A single feature pertains to multiple items, providing regularization (i.e., sharing feature weights amongst test items) and thereby increasing robustness. The features can characterize unseen items as well, which enables generalization. This is crucial to our jump-starting process.

Cold starts and jump-starts In traditional test development processes, one wants to pinpoint item parameters by pilot testing on test-takers whose ability is near the item's true difficulty. But without initial difficulty estimates, computer-adaptive tests cannot choose the most informative items to present to pilot testers. This is costly for large item banks and a security risk for high-stakes tests.

This **cold start problem** is mitigated by discriminative, feature-based difficulty estimation (Beinborn et al., 2014; Settles et al., 2020) according to an external standard like the CEFR. However,

[1] Note that we use the standard variable names from IRT.

[2] The same motivation exists in language modeling: representing each word in a vocabulary V as a dense embedding of dimensionality $d \ll |V|$ (vs. a one-hot vector) reduces orthogonality to enable parameter-sharing among word types (Bengio et al., 2003; Turian et al., 2010; Mikolov et al., 2010).

once test administration yields enough test-taker responses, the item parameters should be directly calibrated on these responses.

Further, these test-taker responses can inform the difficulty of *new* items that grow an existing item bank, akin to establishing a prior that can be refined via Bayesian updating as test-taker responses are collected (Raina et al., 2006). We refer to this as **jump-starting** the parameters of the new items.

3 Approach and Formalism

We extend the 2PL model in two key ways. First, while (2) uses a single parameter for each item's difficulty, like Fischer (1973) we decompose item difficulty into a weighted sum of features (§3.1). We decompose item discrimination the same way. Second, we integrate CEFR-labeled data for indirect supervision of passage difficulty (§3.2) within an ordinal–logistic regression multi-task model (§3.3).

3.1 Student Modeling (TEST-TAKER)

Enumerate exam sessions $p \in \{1, \ldots, P\}$ and items $i \in \{1, \ldots, I\}$. The 2PL model models the probability of the test-taker in exam session p responding to item i correctly as

$$p(Y_{p,i} = 1; \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}) = \sigma\left(a_i(\theta_p - b_i)\right) = \frac{\exp(a_i(\theta_p - b_i))}{1 + \exp(a_i(\theta_p - b_i))}, \quad (3)$$

The higher the test-taker's ability, the higher the probability of a correct response.

To model our test-takers' responses, we generalize 2PL into an LLTM by decomposing the item discrimination a_i and item difficulty b_i into multiple "skills". Extending (3), we model the probability as

$$p(Y_{p,i} = 1; \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{v}) = \sigma\left((\boldsymbol{w}^\intercal \boldsymbol{\phi}(i))\,(\theta_p - \boldsymbol{v}^\intercal \boldsymbol{\phi}(i))\right), \quad (4)$$

where $\boldsymbol{w} \in \mathbb{R}^K$ and $\boldsymbol{v} \in \mathbb{R}^K$ are weight vectors and $\boldsymbol{\phi}$ is a vector of K **feature functions** used to extract features from an item. In LLTMs, these are typically "skills" associated with the item, but the log-linear formulation elegantly allows for arbitrary numerical features to be incorporated.

3.2 CEFR Level Modeling (CEFR)

The Common European Framework of Reference for Language defines guidelines for the language proficiency of non-native language learners. Its six levels of proficiency are (from lowest to highest) A1, A2, B1, B2, C1, and C2. Because CEFR categories are ordered, we treat predicting a passage's CEFR level z_i as an ordinal regression problem.

To do this, we define a generalized linear model, which generalizes linear regression and the log-linear models common in statistical natural language processing (e.g. the one defined in §3.1), for ordinal regression. It computes a logit with a function that is linear in the parameters, then transforms the logit into the mean of the distribution function. This transformation requires an invertible *link function*. For linear regression, this is the identity function, and for logistic and softmax regression it is the logit function. In our ordinal regression case, we choose the *logistic cumulative link* (McCullagh, 1980; Agresti, 2010; Pedregosa et al., 2017). Applying this, we define the probability of level z as

$$p(Z_i = z; \boldsymbol{\lambda}, \boldsymbol{b}) = \begin{cases} \sigma(\xi_1) & z = 1 \\ \sigma(\xi_z) - \sigma(\xi_{z-1}) & 1 < z < C \\ 1 - \sigma(\xi_{C-1}) & z = C, \end{cases} \quad (5)$$

where a learnable, sorted vector $\boldsymbol{\lambda}$ of $C - 1$ cutpoints divides the difficulty scale into C levels according to $\xi_z \triangleq \lambda_z - b_i, \ \forall z$. The level z is determined by which cutpoints the logit falls between.

As in the student modeling, we compute difficulty as $b_i = \boldsymbol{v}^\intercal \boldsymbol{\phi}(i)$. Similarly, we must jointly learn the passage difficulties and the values that contextualize them—in this case, $\boldsymbol{\lambda}$ instead of $\boldsymbol{\theta}$. In other words, the CEFR model's parameters are $\boldsymbol{\lambda}$ and $\boldsymbol{v}$. We have intentionally designed this model to share structure and parameters with the linear logistic test model. (We elaborate on this in §3.3.) What differs is the classification task and the consequent manner in which the classification likelihood is computed from features.

3.3 Multitasking on One Line (JOINT)

How are we able to directly rate a test-taker's ability or a passage's difficulty in terms of its CEFR level? Recall that in both the ordinal CEFR model and the binary student model, we decompose the difficulty of item i as $b_i = \boldsymbol{v}^\intercal \boldsymbol{\phi}(i)$. Our computation therefore depends on a vector $\boldsymbol{v}$ of weights that govern the (relative) contribution of each passage feature. These weights are shared between the two prediction tasks, so parameter estimation on one task can hone the other's estimate.

These b values exist on the same real line, as do the CEFR cutpoints $\boldsymbol{\lambda}$ and the test-taker ability

Figure 2: In this work, we focus on C-test items like the one above. Modeling the test-taker's ability to fill each blank correctly depends on the difficulty of the test item.

θ values. The three are thus directly comparable (see Figure 1). We can reason about the relative difficulty of a passage for a test-taker (from the difference in logits), as well as their CEFR level (from the cutpoints the ability lies between or beyond).

We cannot perform this direct comparison if we follow Settles et al. (2020) and use nominal CEFR classification (e.g., with a softmax function atop some $\boldsymbol{v}$-weighted feature extraction instead of ordinal regression)—there would be one logit per class for each passage (i.e., in $\mathbb{R}^6$)[3] instead of a single logit per passage (i.e., in $\mathbb{R}$).

We tune the item discrimination weights $\boldsymbol{w}$, item difficulty weights $\boldsymbol{v}$, and CEFR cutpoints $\boldsymbol{\lambda}$ via maximum likelihood estimation (MLE) to predict both passage difficulty with respect to the CEFR scale and the correctness of test-taker responses.

3.4 Feature Design

For passage featurization, we use a Transformer-based passage representation using BERT (Devlin et al., 2019), which is able to implicitly represent linguistic content of its input text (Tenney et al., 2019; Ettinger, 2020). For $k \in \{1, 2, \ldots, 768\}$, let

$$\phi_k(i) \triangleq \text{BERT}(\text{text}(i))_k,$$

where the function BERT extracts a 768-dimensional embedding vector to represent each passage of text.[4]

4 Experimental Setup

We train our jump-start model (which we hereafter refer to as BERT-LLTM) on either CEFR-labeled data, test-taker responses from a high-stakes English proficiency test, or both. Specifics of learning and reproducibility details are given in Appendix B. We then compare several measures of model fit, both on held-out test-taker exam sessions and fully held-out items. The latter represents adding new items to an already-available test. §5 shows that we are able to jump-start the difficulty estimates for these items, outperforming three baselines, including Settles et al. (2020).

4.1 Data

CEFR-labeled Dataset We use 3,826 English passages automatically extracted and labeled with CEFR levels ( ) by two subject-matter experts (SME) with education and experience in applied linguistics/language teaching, with an average weighted κ between annotator pairs of 0.599. Disagreements beyond one score-point were adjudicated by the lead of the labeling study. Two instances of each passage are created in the final dataset: one instance with each label when there is disagreement by one CEFR level between annotators, and two instances with the agreed-upon CEFR level otherwise. We hold out 10 % of the passages for evaluation.

Test-taker Response Dataset We also use a much larger collection of test-taker response data from the Duolingo English Test, a high-stakes, adaptive English proficiency test. We collected 100,495 exam sessions over a 14-month period, extracting C-test item responses (Klein-Braley, 1984; Khodadady, 2014; Reichert et al., 2010) for this study. Each session contained test-taker responses to 5–8 distinct items drawn from an item bank of 4,151 unique C-test items. The extreme sparsity would pose a challenge for accurately estimating item parameters via a traditional IRT model, with 1–3 parameters for each test item (Linacre, 2014).

[3] More precisely, these are confined to the simplex Δ^5.

[4] Other passage features can be included as well. We found that word frequency quantile features had minimal effect.

A C-test item is similar to a cloze-deletion item (Taylor, 1953), except that words are only partially omitted. Specifically, for each partially omitted (i.e., "damaged") word in the passage, the first half of the word is provided, as is the number of missing characters (see Figure 2). In our dataset, we treat each damaged word (i.e., sub-item) as a separate instance of the same item, which was either answered correctly or not.[5]

The proficiency test we consider contains other item types besides C-tests. We use these to compute a gold-standard ability estimate for each exam session. This is the score on other item types in the exam, ignoring the C-test items. As a high-stakes exam, these ability estimates are highly accurate, with a test–retest reliability of 0.92. We use this gold standard as the ability parameters $\boldsymbol{\theta}$ when training the 2PL IRT baseline and our BERT-LLTM model. It is also needed for some metrics (§4.2). We consider joint estimation in Appendix C.

Finally, we held out 10 % of sessions for evaluation and used the other 90 % of sessions for training. Evaluation on the test-taker response dataset requires additional care: the fact that multiple item responses share the same exam sessions or items violates the i.i.d. assumption underlying supervised machine learning evaluation practices. In particular, items that occur in the training data set will also occur in the test dataset, but with responses from different test-takers.

Item-Split Dataset To ensure that our model generalizes well on new, unseen items, we created an additional split of the data described above. In the *item-split dataset*, we randomly sample 3 % of items, and we hold out all sessions where at least one of those items were administered (which is roughly 20 % of sessions). The remaining sessions are used for training. In the evaluation phase of the item-split experiments, we only evaluate results on the held-out items, not entire sessions.

[5]In effect, this is how we dichotomized the response data, which is required for the logistic models upon which IRT and LLTM models are usually based. We also considered continuous IRT models (Deonovic et al., 2020) in Appendix A, but these are much less commonly used and our experiments (not reported here) demonstrated that they fit the data very poorly. Our approach has the limitation that it does not consider the differences in difficulty among damaged words within the same passage; it effectively models the average difficulty among those damaged words. We may get additional precision by modeling the difficulty at the word level, but we leave this to a future work.

4.2 Metrics

We measure performance on the CEFR data with these two measures:

Pearson's r The coefficient of determination between the model's logit-scale difficulty estimate and the CEFR label (using A1 = 0, A2 = 1, etc.).

Spearman's ρ The rank correlation between the model's logit-scale difficulty estimate and its CEFR label. This can capture nonlinear relationships in the data.

When evaluating the models on the test-taker response dataset, we predict the test-taker's score on each item using the gold-standard abilities and the learned item parameters. We also compute an aggregate score of each test-taker's performance on the C-test items (referred to as the item-type score), by using the learned item parameters and the test-taker's item scores on C-test items. We then evaluate the following metrics:

Item Mean Score / Predicted Score Pearson's r The correlation between the item's mean score across all sessions, and the item's mean predicted score across all sessions. This relates the item's difficulty to the item's predicted difficulty (both relative to the audience each item was administered to).

Cross-Entropy The cross-entropy between each sub-item score and the model's predicted probability of getting a sub-item of that item correct.

Residual standard deviation A residual is the difference between the item score predicted by the model for a given test-taker, and the test-taker's actual item score. In a well-calibrated model, the standard deviation of residuals across all items in the test dataset is small.

Item-Type / Total Pearson's r The Pearson correlation between the item-type score and the gold-standard ability estimates. This metric is common in assessment research (Furr, 2017).

Test–retest Reliability The correlation of item-type scores among all pairs of exam sessions in the test dataset taken by the same test-taker within 30 days of each other. Like above, this reliability estimate is common in assessment research (Furr, 2017).

4.3 Baselines

We compare our model to three baselines:

ALL-SAME. Here, we fix all item difficulties at 0.0 and all item discriminations at 1.0. This baseline shows what is trivially attainable on the test-taker response metrics.

2PL-IRT. In this model, item parameters are estimated using (2), where the test-taker ability parameters $\boldsymbol{\theta}$ are fixed to the gold-standard ability estimates. This is a considered a strong baseline in assessment research. By design, this model cannot predict test-taker performance on new items before those items have undergone pilot testing.

SETTLES-ET-AL. We reproduce the features and model design of the work closest to ours, Settles et al. (2020), but train on the same CEFR dataset as our other models, without data augmentation.[6] By its design, their model cannot use test-taker data.

5 Results and Interpretation

Here we study how well our model captures test-takers' ability to answer C-test items based upon the passage's difficulty, and how well it contextualizes this with the CEFR scale.

5.1 Comparison to baselines

We present CEFR-labeling and test-taker modeling metrics in Table 1. Our joint model's item parameters are better calibrated, more consistent, and lead to a more reliable test than the estimates from Settles et al. (2020). It outperforms SETTLES-ET-AL on all five measures of test-taker modeling: relative improvements in cross-entropy (40 %), residual standard deviation (15 %), item score r (44 %), item total r (14 %), and test–retest reliability (17 %). CEFR prediction slightly improves.

The joint model also performs very similarly to the standard 2PL-IRT, matching or beating it on 4 out of 5 metrics, despite having far fewer free parameters. However, as we'll see in §5.2–5.3, 2PL-IRT is incapable on unseen items and plummets when less test-taker response data is available.[7]

Finally, we note that while JOINT and 2PL outperform ALL-SAME on all measures, SETTLES-ET-AL has a higher cross-entropy than this trivial baseline. This only reflects that SETTLES-ET-AL cannot estimate item discrimination parameters $\boldsymbol{a}$; its cross-entropy can be altered by arbitrarily changing the fixed, shared discrimination parameter.

5.2 Model ablation experiments

How valuable is multitasking? We find that having both sources of data available leads to strong test-taker modeling while contextualizing difficulties with the CEFR scale. Using only test-taker data produces small improvements in three test-taker modeling metrics compared to the JOINT model, but unsurprisingly yields a dramatic decrease in CEFR labeling performance.

Similarly, the model trained on only CEFR data improves slightly on CEFR modeling but drops substantially on test-taker modeling. The performance becomes similar to SETTLES-ET-AL, which also does not use test-taker data. With robust featurization in our log-linear model, no pilot testing is necessary to get accurate item difficulty estimates. We thus provide an alternative solution to the cold start problem, which is able to improve as test-taker data becomes available.

5.3 Jump-starting new items

As a high-stakes test evolves, we must introduce new items with new parameters to be estimated. Pilot testing is resource-intensive, so we wish to expedite it by jump-starting item parameters with initial estimates from our model, requiring less pilot data. We use the item-split dataset described above to simulate this scenario and evaluate our model's generalization to new items (see "Jump-starting New Items" in Table 1).

On this data split, we again achieve better item mean r, cross-entropy, and residual standard deviation than SETTLES-ET-AL. (Recall that 2PL-IRT cannot even be used in this scenario.) As before, training on only test-taker data produces further gains, at the cost of CEFR labeling performance. In all, this indicates that our principled joint model effectively jump-starts the item a and b parameters.

5.4 Data ablation experiments

In §5.2, we note that the model is viable for a cold start and improves as test-taker responses are recorded. How quickly can we expect this improvement? We measure this by recursively removing half the test-taker data, fitting a model on the remainder, and evaluating its performance.

[6] Settles et al. (2020) used additional datasets in a semi-supervised label propagation scheme akin to the Yarowsky algorithm (Yarowsky, 1995).

[7] This occurs because 2PL's parameters are per-item, akin to one-hot encoding. When an item isn't included in training (the essence of the jump-start scenario), 2PL does not have any parameters to represent it. It thus cannot make predictions for items unseen in training.)

Model	CEFR-labeling Metrics		Test-taker Response Metrics				
	Pearson's r	Spearman's ρ	Item Mean Score / Pred. Pearson's r	Cross-entropy	Residual st. dev.	Item-Type / Total Pearson's r	Test–Retest Reliability
Baselines							
ALL-SAME	0.00	0.00	0.32	0.73	0.21	0.41	0.18
2PL-IRT	N/A†	N/A†	**0.74**	0.56	**0.16**	0.74	0.62
SETTLES-ET-AL	0.81	0.75	0.43	0.91	0.20	0.66	0.53
BERT-LLTM							
CEFR OBJECTIVE	**0.84**	**0.77**	0.45	0.88	0.28	0.70	0.51
TEST-TAKER OBJECTIVE	0.73	0.49	0.66	**0.55**	**0.16**	**0.75**	**0.63**
JOINT OBJECTIVE	0.82	0.76	0.62	**0.55**	0.17	**0.75**	0.62
*Jump-starting New Items**							
BERT-LLTM							
TEST-TAKER OBJECTIVE	0.74	0.56	0.59	0.53	0.16	N/A‡	N/A‡
JOINT OBJECTIVE	0.80	0.73	0.52	0.54	0.17	N/A‡	N/A‡

* These experiments are run on a different split of the test-taker dataset, so they have an additional source of variance when comparing to the models above.
† 2PL IRT can only estimate difficulties for passages with test-taker data, so they cannot be evaluated on the CEFR-labeling task.
‡ Since only held-out items are evaluated in these experiments, we cannot score all items in the session and thus cannot compute these metrics.

Table 1: Metrics for each training objective. Lower rows are a different data split and not directly comparable.

In Figure 3 we see that even with only 452 exam sessions, we still reach near-optimum performance. By comparison, 2PL-IRT model performance degraded even with 64,000 sessions, and needed at least 8,000 sessions to even beat the trivial ALL-SAME baseline. Similar trends emerge when examining other test-taker response metrics. This is expected, given that the 2PL-IRT model generally requires 200–400 test-taker responses per item (Henning, 1987).

Notably, at 452 sessions with an average of 6.15 items per session, *there are fewer total responses than the number of items in the item bank*. In this reduced dataset, each item in the item bank was viewed 0.67 times, on average. This paucity of data shows that we can achieve high calibration and reliability while preserving test security. Similar trends occur for test–retest reliability and item mean Pearson's r (Appendix D).

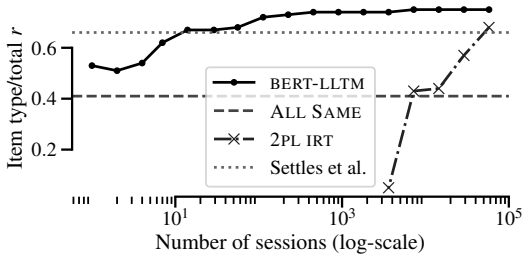


Figure 3: Item type/total r as the number of pilot sessions is reduced. Only 500 exam sessions are needed for near-optimal results. Other metrics trend similarly.

6 Recipe for a Language Proficiency Test

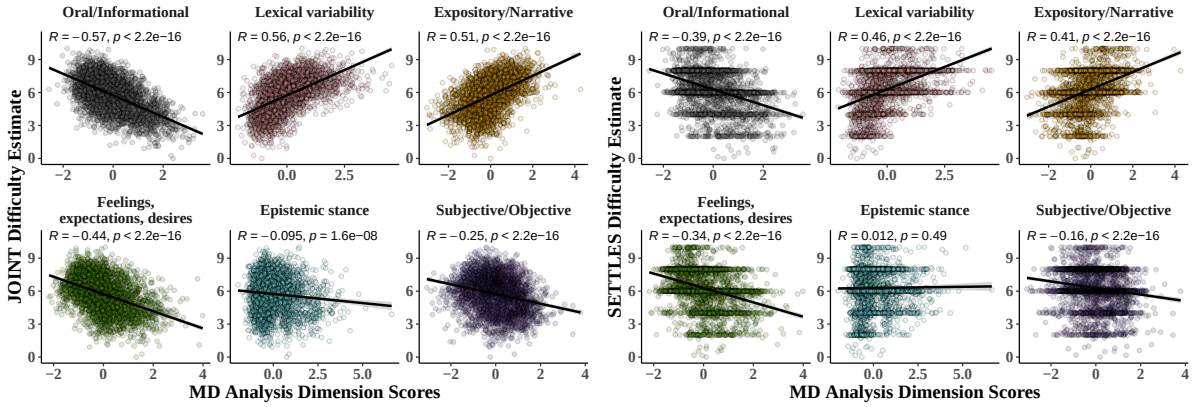
The better calibration and reliability of JOINT over baselines, its aptitude for cold start scenarios, and the ability to jump-start new item parameters suggest a "user guide" for building and growing a new English proficiency exam for a cold start.

1. Begin from a cold start, using item difficulties learned by ordinal regression on CEFR-labeled passages to accelerate piloting.
2. As the test is administered, incrementally improve the estimates using the JOINT objective, adjusting the test-taker response objective's weight depending upon to the amount of data collected. (We found that as few as 500 short exam sessions were sufficient to reach near-optimal performance for this model).
3. Eventually, the feature-based operational-only objective will outgrow the CEFR data. If CEFR levels of test-takers and items are not important to the test construct, the CEFR-labeled data may eventually be retired.
4. Introduce new test items, accelerating their piloting by jump-starting difficulty and discrimination estimates with the model.

This approach enables continuous improvement of item difficulty and discrimination parameters. Importantly, our model can elegantly handle training from any of these data combinations available.

7 Linguistic Interpretation

Here we triangulate between the BERT features and linguistic features of the text, examining whether our model generates theoretically valid, *linguistically meaningful* difficulty estimates.



(a) Training BERT-LLTM with the JOINT objective

(b) Training from Settles et al. (2020)

Figure 4: Relating model-derived β estimates to our six identified dimensions of functional language; 3 dimensions show significant, strong correlation ($|r| > 0.5$) in our model vs. none from Settles et al. (2020).

Our approach, **multi-dimensional (MD) analysis** (Biber, 1988; Biber and Conrad, 2019), was developed in applied linguistics to comprehensively describe linguistic variation in a corpus. This is a process that employs exploratory factor analysis to reduce a large set of computationally identified linguistic features in passages down to a smaller set of interpretable "dimensions" for which each passage receives a dimension score (see Biber and Conrad (2019) for complete details). The interpretation of each dimension is based on linguistic theory, the results of previous MD analysis literature, and a qualitative analysis of the texts with relatively high and low dimension scores.

Based on these sources of evidence, we developed a functional interpretation for each of the six resulting factors:

D1: Oral/involved vs. Literate/informational
D2: Lexical variability
D3: Expository vs. Narrative style
D4: Personal feelings, expectations, and desires
D5: Epistemic stance expression
D6: Subjective vs. Objective description

Details of our MD analysis are in Appendix E.

Figure 4 illustrates the relationship between the final difficulty estimates from our models and the MD-based linguistic dimension scores for each passage. The BERT-LLTM (JOINT) model (Figure 4a) shows moderately strong relationships between the difficulty estimates and the dimension scores on the first four of the six dimensions. In contrast, the Settles et al. (2020) baseline (Figure 4b) showed moderate to weak relationships, and were consistently lower than those of our model. This indicates that the BERT-LLTM difficulty estimates do a better job of honing in on pertinent linguistic features. For example, as passages are estimated to be more difficult, they become more informational, have greater lexical variability, and become more expository and objective.

While privacy and test integrity prevents the release of in-use test items, we present some retired test items with their model predictions in Table 2, to allow a qualitative analysis. Encouragingly, it is apparent that less clausal, more phrasal and lexically varied items are scored as more difficult by the BERT-LLTM. This aligns with the quantitative findings of MD analysis dimensions above.

8 Related Work

The challenge of estimating passage difficulty is widely considered, with roots in the readability research (Flesch, 1943; Chall and Dale, 1995). Recent approaches leverage techniques from natural language processing (Beinborn et al., 2014, 2015; Loukina et al., 2016; Settles et al., 2020). Rather than using explicit, interpretable linguistic features of the passage, our features (§3.4) are drawn from a large, pre-trained neural network's representation of the passage.

Pre-trained distributed representations of text are in widespread use in natural language processing; recent work leverages these large models for assessment research. Ha et al. (2020) provides an empirical study, comparing a battery of pre-trained models on their ability to predict test-taker proficiency from short-answer and multiple choice questions on a high-stakes medical exam. In a related

Passage	SME CEFR level	Predicted CEFR level	Predicted difficulty
Tara: Do you want to go to the museum today? *Billy: No, I don't like the museum very much. I want to go to the movie theater. Tara: I don't like any of the movies at the movie theater.* Billy: OK, we'll go to the café. Tara: OK!	A2	A2	−6.75
Since water is so important, you might wonder if you're drinking enough. *There is no magic amount of water that kids need to drink every day. Usually, kids like to drink something with meals and should definitely drink when they are thirsty. But when it's warm out or you're exercising, you'll need more.* Be sure to drink some extra water when you're out in warm weather, especially while playing sports or exercising.	B1/B2	B2	−4.04
The same as all eight of Connecticut's counties, there is no county government and no county seat. *In Connecticut, towns are responsible for all local government activities, including fire and rescue, snow removal and schools. In a few cases, neighboring towns will share some resources (e.g., water, gas, etc.).* New London County is only a group of towns on a map. It has no governmental authority.	B2	B2	−2.98
Ariel University, formerly the College of Judea and Samaria, is the major Israeli institution of higher education in the West Bank. *With close to 13,000 students, it is Israel's largest public college. The college was accredited in 1994 and awards bachelor's degrees in arts, sciences, technology, architecture and physical therapy.* The school's current temporary status is that of a "university institution" conferred by the Israel Defense Forces, but it remains without university accreditation.	B2	B2	−2.40
The basic operation of a telephone involves sound waves being converted into electrical signals. *These signals can then be sent over long distances from a device transmitting these signals at one end to a device receiving them at another. The original telephone system involved direct connections between two locations or parties.* However, this was rapidly changed to a more flexible system where a central office would direct calls towards an intended receiver.	C1	C1	−1.06

Table 2: Five retired items from the Duolingo English Test (https://englishtest.duolingo.com), with their expert-annotated CEFR level and model predictions. The italicized sentences are damaged in the C-test.

vein, Xue et al. (2020) explore transfer learning for predicting item difficulty and response time, again using large pre-trained models. None of these use our joint probability model to improve learning and contextualize the scores of test-taker ability and item difficulty or relate deep features to linguistically measurable attributes of the passages; nor do they provide ablations showing our minimal need for test-taker data.

9 Conclusion

We have introduced a multitask model BERT-LLTM to estimate item parameters for adaptive language tests, and demonstrated that it improves upon the performance of Settles et al. (2020) by incorporating test-taker response data, and matches the strong 2PL-IRT baseline on most metrics. Furthermore, we showed that 3,000 pilot item administrations were sufficient for good performance with a large (>4,000) item bank, whereas the standard 2PL-IRT model required 200 times as many administrations to achieve similar performance (see Figure D.2). Finally, we showed that model's item parameter estimates generalize well even for items that have not yet been piloted, whereas a 2PL-IRT model cannot generalize to unseen items at all.

These results have significant applications in the area of high-stakes language proficiency testing. In addition to addressing the *cold start* problem described by Settles et al. (2020), models like BERT-LLTM can be used to *jump-start* new items by providing good initial parameter estimates. This reduces the number of item exposures needed during piloting and increases the feasibility of maintaining large item banks, both of which are crucial for maintaining the security of a high-stakes test.

Furthermore, the difficulty estimates produced by our model are much more strongly correlated with four of six dimensions of functional language derived from co-occurring lexico-grammatical features, suggesting both that the model is keying into these features (see Tenney et al., 2019) and that these linguistic features are related to the true difficulty of the passages. This illuminates the linguistic abilities of the test-takers.

In the future, we plan to show that BERT-LLTM's parameter estimates can be used as Bayesian priors for a 2PL-IRT model to achieve even more accurate results. Furthermore, we can expand to other item types. Finally, we will explore modeling the word-level difficulty and discrimination of damaged words in the passage.

References

AERA, APA, and NCME. 2014. *Standards for Educational and Psychological Testing*.

Alan Agresti. 2010. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons.

David Andrich. 1978. A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573. Publisher: Springer.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, 2:517–530.

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

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155.

D Biber. 1988. *Variation across speech and writing*. Cambridge University Press.

D Biber, B Gray, and K Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45:5–35.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*, 2nd edition. Cambridge University Press.

P Billingsley. 2012. Probability and measures, Wiley series in probability and statistics.

Siqi Cao. 2015. TOEFL questions, answers leaked in China: reviewer. In *Global Times*.

Raymond B. Cattell. 1966. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276. PMID: 26828106.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale–Chall readability formula*. Brookline Books.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Council of Europe. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Benjamin Deonovic, Maria Bolsinova, Timo Bechger, and Gunter Maris. 2020. A Rasch model and rating system for continuous responses collected in large-scale learning systems. *Frontiers in Psychology*, 11:3520.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Renee Dudley. 2016. 'Massive' breach exposes hundreds of questions for upcoming SAT exams. In *Reuters Investigates*.

T. Eckes. 2011. Item banking for C-tests: A polytomous Rasch modeling approach.

Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Gerhard H. Fischer. 1973. The linear logistic test model as an instrument in educational research. *Acta psychologica*, 37(6):359–374.

Rudolf Flesch. 1943. Marks of readable style; a study in adult education. *Teachers College Contributions to Education*.

R Michael Furr. 2017. *Psychometrics: an introduction*. SAGE publications.

Stuart Geman, Elie Bienenstock, and René Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58.

Le An Ha, Victoria Yaneva, Polina Harik, Ravi Pandian, Amy Morales, and Brian Clauser. 2020. Automated prediction of examinee proficiency from short-answer questions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 893–903, Barcelona, Spain (Online). International Committee on Computational Linguistics.

R.K. Hambleton, H. Swaminathan, and H.J. Rogers. 1991. *Fundamentals of Item Response Theory*. Measurement Methods for the So. SAGE Publications.

Grant Henning. 1987. *A guide to language testing*. Heinle & Heinle.

D.J. Jackson and D.L. Tweed. 1980. Note on the squared multiple correlation as a lower bound to communality. *Psychometrika*, 45:281–284.

Henry F. Kaiser. 1974. An index of factorial simplicity. *Psychometrika*, 39(1):31–36.

Mohammad Khajah, Yun Huang, José P. González-Brenes, Michael Mozer, and Peter Brusilovsky. 2014. Integrating knowledge tracing and item response theory: A tale of two frameworks. In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization co-located with the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP2014), Aalborg, Denmark, July 7-11, 2014*, volume 1181 of *CEUR Workshop Proceedings*. CEUR-WS.org.

E. Khodadady. 2014. Construct validity of C-tests: A factorial approach. *Journal of Language Teaching and Research*, 5(6):1353–1362.

Christine Klein-Braley. 1984. Advance prediction of difficulty with C-tests. *Practice and Problems in Language Testing*.

J.M. Linacre. 2014. 3PL, Rasch, quality-control and science. *Rasch Measurement Transactions*, 27(4):1441–1444.

Gabriel Loaiza-Ganem and John P Cunningham. 2019. The continuous Bernoulli: fixing a pervasive error in variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Frederic M Lord. 1980. *Applications of item response theory to practical testing problems*. Routledge.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

Geoff N. Masters. 1982. A Rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174.

Peter McCullagh. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048. ISCA.

Hans Müller. 1987. A rasch model for continuous ratings. *Psychometrika*, 52(2):165–181.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. 2017. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18(55):1–35.

Juan Pino, Liezl Puzon, Jiatao Gu, Xutai Ma, Arya D. McCarthy, and Deepak Gopinath. 2019. Harnessing Indirect Training Data for End-to-End Automatic Speech Translation: Tricks of the Trade. In *16th International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong. Zenodo.

Rajat Raina, Andrew Y. Ng, and Daphne Koller. 2006. Constructing informative priors using transfer learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 713–720, New York, NY, USA. Association for Computing Machinery.

G. Rasch. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Studies in mathematical psychology. Danmarks Paedagogiske Institut.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Wilson L. Taylor. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden. Association for Computational Linguistics.

Norman D. Verhelst. 2019. *Exponential Family Models for Continuous Responses*, pages 135–160. Springer International Publishing, Cham.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy. Association for Computational Linguistics.

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

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

A Additional Information on Modeling Fractional Responses

In §3, we describe how to score binary items. However, the C-test item contains several "sub-item" that must be scored. How can we integrate this into our model? Regressing to the fraction of blanks answered correctly is not well-founded probabilistically (Loaiza-Ganem and Cunningham, 2019), and rounding to 0 or 1 loses information.

One option is modeling a *continuous Bernoulli distribution*. This distribution, long known to the psychometric community (Müller, 1987; Verhelst, 2019), has recently been explored in machine learning (Loaiza-Ganem and Cunningham, 2019). Its support is $[0, 1]$ instead of $\{0, 1\}$, and its unnormalized probability density can be expressed in the same form as a function of the natural parameter:

$$p(y \mid \eta) \propto \exp(\eta x).$$

The normalizing constant must then account for this continuous support. However, Monte Carlo estimates of cross-entropy for continuous variables can be difficult to interpret because differential entropy can become negative. Additionally, we found this to yield poor empirical results.

We suspect that the continuous Bernoulli distribution fits the data poorly because it is designed for "U-shaped" distributions, i.e., convex density functions. By contrast, our item-score data is approximately normally distributed.

A second issue with the continuous Bernoulli distribution is that probability mass is wasted. The possible scores on a given C-test item are quantized, based on the number of damaged letters N. To handle this, in §4 we model an item as N separate items that share item parameters. A more sophisticated model could separately parameterize each blank, perhaps using information about which words are damaged.

Another option for fractional responses is dyadic expansion (Billingsley, 2012; Deonovic et al., 2020), recursively split the interval $[0, 1]$ and predicting one bit for each split. This resembles predicting a fixed-precision binary fraction, bit by bit.

B Additional Information on Parameter Estimation and Reproducibility

We train parameters using AdamW (Loshchilov and Hutter, 2019), a stochastic gradient method, to maximize the log-joint likelihood $\sum_{(p,i)} \log p(y_{p,i}, z_i \mid \theta, \boldsymbol{w}, \boldsymbol{v})$ over all test-taker responses in the test-taker data and all tagged passages in the CEFR data. This objective combines (3) and (5). If either y or z is not observed, we marginalize it out.

B.1 2PL-IRT details

Our 2PL implementation follows the text of Embretson and Reise (2013). It amounts to a three-step process which we summarize here.

1. For each item, we fit a logistic regression model to the observed data and the gold-standard ability estimates θ. This yields an intercept and slope for each item.

2. We use empirical Bayes shrinkage to obtain the shrunken logistic regression coefficient, in order to avoid the issue that small sample sizes (i.e., small number of respondents) may yield inaccurate logistic regression coefficient estimates.

3. These shrunken parameters are transformed to the 2PL discrimination or difficulty scale.

B.2 Reproducibility details

We train for 1,000 epochs using batching and no regularization. A search over coefficients in $\{1v1, 1v2, 1v4, 1v8\}$ to govern the relative importance of the two training objectives (CEFR vs test-taker) found that 1v8 yielded the best compromise between the two training objectives. The initial parameters $\in \mathbb{R}^{768}$ are all drawn from $\mathcal{U}(-\sqrt{k}, \sqrt{k})$, where k is the number of features, except for the ordered CEFR cutpoints, which we arbitrarily initialize as $[-2.5, -1.5, -0.5, 0.5, 1.5]$. For AdamW, we use a learning rate of 0.01 and averaging coefficients $(0.9, 0.999)$.

For passage representation, we use the publicly available, pre-trained XLM-RoBERTa (Conneau et al., 2020), a variant of BERT which removes the standard BERT architecture's next sentence prediction objective, alters the training regimen, and is trained on over 100 languages' text. While we do not evaluate cross- or multi-lingual performance, XLM-RoBERTa's strong performance in downstream discriminative tasks across languages suggests an avenue for robustly generalizing our difficulty and proficiency estimation to other high-resource languages.

We implement our model in the PyTorch differentiable computation library (Paszke et al., 2019).

Training and validation on a CPU is extremely fast, completing an entire experiment in about an hour on a single 2.9 GHz Intel i9 GHz processor.

Why not fine-tune BERT? Arguably, a tighter model fit (i.e., a lower conditional cross-entropy on training data) could be achieved by discriminative fine-tuning of the BERT parameters, rather than using BERT as a static feature extractor. The model then becomes log-*nonlinear* in the learnable parameters $\boldsymbol{\theta}$, which now include BERT's parameters. This introduces a host of local optima, making the model *non-identifiable*, unlike our log-linear formulation with regularization. This translates into higher-variance estimates of test-taker ability (Geman et al., 1992), subject to the parameter initialization and the vagaries of sampling during optimization. High-stakes proficiency testing demands reliable, consistent estimates, so we prefer the identifiability of our log-linear model.

Additionally, there are other approaches for using BERT to represent the passage suggested in the semantic similarity literature. We could average the BERT token embeddings instead of using the `[CLS]` embedding, or we could use a variant more tailored to fixed-length representation (Reimers and Gurevych, 2019). We leave these comparisons to future work.

C Additional Information on Training with Free θ Parameters

The experiments described in §4 rely on a gold-standard test-taker ability estimated from test-taker performance on non–C-test items. In other words, we evaluate in the case that a new item type is being piloted in the context of an high-reliability test where the test-takers' abilities are known. Generally, there are a number of ways we could acquire high-quality estimates of the pilot test-taker's ability. Our setup would thus be applicable in many "cold start" situations.

When you cannot acquire the pilot test-taker's ability directly, you can infer the abilities by not fixing $\boldsymbol{\theta}$ in the model, instead learning the parameter vector. For completeness, we train a model with this setup. The results are given in Table C.1. While the BERT-LLTM did worse, the hyperparameters were not tuned for this vastly larger number of parameters. Also, in such a piloting scenario, one would generally administer much more than 6 pilot items to each test-taker, so that each θ could be more accurately inferred. Nonetheless, these are better than either the Rasch model or 2PL could produce in this case, both of which would require hundreds of test-taker responses per item (Henning, 1987).

Metric	Value
Item Mean Score / Predicted r	0.62
Cross-entropy	0.57
Residual st.dev.	0.18
Item type/total score r	0.72
Test–retest reliability	0.59

Table C.1: Model performance when test-taker ability $\boldsymbol{\theta}$ is learned instead of gold-standard.

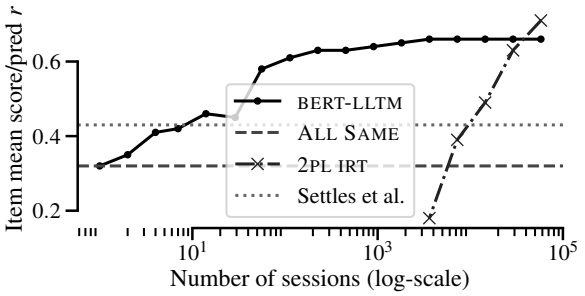


Figure D.1: Mean item score Pearson's r as the number of training sessions is reduced shows that as few as 500 exam sessions are necessary.

D Additional Information on Data Ablation

In Figure 3, we reported the item type total correlation as the number of exam sessions used in training was reduced. Here we show the item mean score/prediction correlation (Figure D.1) and the test-retest reliability (Figure D.2). The trend is similar to what was reported in §5.4: with 452 exam sessions, each item is administered during piloting fewer than one time on average to achieve a strong model fit.

E Additional Information on Multi-Dimensional Analysis

Here we provide the linguistic variables from our multidimensional analysis in §7, an overview of how the linguistic features were extracted, and a summary of the factor analysis. The variables themselves are in Table E.1 and their loadings are in Table E.2.

Linguistic feature extraction We used the Biber tagger to annotate the lexico-grammatical

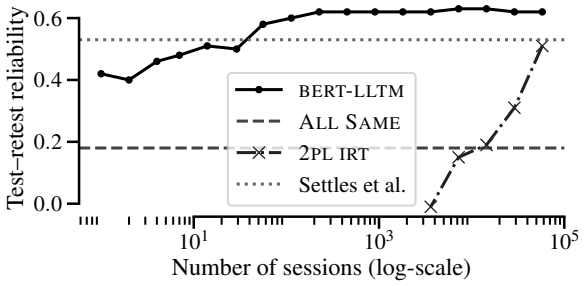


Figure D.2: Test–retest reliability as the number of training sessions is reduced shows that with substantially fewer training samples, high reliabilty is maintained.

features of the C-test passages. The Biber Tagger has been developed and revised by Douglas Biber over the past 30 years. The current version has both probabilistic and rule-based components and uses multiple large-scale dictionaries. After tagging, we calculated normalized rates of occurrence for more than 150 lexico-grammatical features[8].

We pruned redundant features, those with extremely low mean rates of occurrence, and those with many zeros from this starting list. This resulted in a set of 80 linguistic variables. We further pruned with the squared multiple correlation method (Jackson and Tweed, 1980), resulting in the 43 final linguistic features given in Table E.1.

Factor analysis Factor analysis, a technique related to principal components analysis, models data as a linear combination of latent factors. In our factor analysis, we found an acceptable Kaiser–Meyer–Olkin index (Kaiser, 1974) of 0.63 out of 1.0, suggesting the appropriateness of our 43 variables. To determine the optimal number of factors, we examined a scree plot (Cattell, 1966), which suggested a six-factor solution, and these six factors accounted for 45% of the variance in the dataset. We include the factor (structure) matrix for the 43 variables in Table E.2.

Functional interpretation of factors We computed factor scores for every C-test passage on each of the six factors in the final factor solution, using the regression method. We used a liberal factor loading cutoff of 0.20 and, for purposes of our interpretations, assigned a linguistic variable to the factor on which it loaded strongest. This concluded the quantitative portion of the factor analysis, and we turned our attention to qualitatively interpreting the six factors. We set out to functionally interpret each of the six factors as underlying dimensions of linguistic variation.

[8]This tagger achieves accuracy levels comparable to other taggers of its time (see Biber et al., 2011) and analyzes a larger set of lexico-grammatical features than most other taggers, including a wide range of semantic categories for words (e.g., nouns: animate, cognitive, concrete, technical, quantity, place, group, abstract) and lexico-grammatical features (e.g., *that*-complement clauses controlled by stance nouns, such as in *The claim that I would even do that is ridiculous.*).

Abbreviation	Feature
adv	Adverb (not inc. emphatics, hedges, amplifiers, downtoners, time/place adverbs)
advl_stance_all	Stance adverbs
all_def_art	Definite articles
all_indef_art	Indefinite articles
conj_advl	Adverbial – conjuncts
conj_all	All conjunctions
contract	Contractions
emphatic	Emphatics
infinitive	Infinitive verbs
jj_attr	Attributive adjective
jj_pred	Predicative adjective
mod_poss	Modals of possibility
mod_pred	Modal of prediction
nn_abstact	Abstract nouns
nn_all	Total nouns
nn_common	Common nouns
nn_concrete	Concrete nouns
nn_nom	Nominalizations
nn_place	Place nouns
nn_premod	Pre-modifying nouns (noun-noun sequences)
nn_proper	Proper nouns
passive_short	Agentless passive verbs
prep	Prepositions
pro_1	First person pronouns
pro_2	Second person pronouns
pro_3	Third person pronoun (except 'it')
pro_it	Pronoun "it"
th_stance_all	'That' complement clauses controlled by stance verbs
th_vb	'That' complement clause controlled by all verbs
that_del	'That' deletion
to_stance_all	'To' complement clauses controlled by stance verbs
tt_ratio	Type/token ratio (first 100 words)
vb_act	Activity verbs
vb_be	Verb "be" (uninflected present tense, verb and auxiliary)
vb_comm	Communication verbs
vb_mental	Mental verbs
vb_past	Past tense verbs
vb_present	Verb (uninflected present, imperative & third person)
vb_progress	Verbs – present progressive
wh_cls	Wh- clauses
wh_ques	Wh- questions
word_count	Total words
word_length	Average word length

Table E.1: The forty-three linguistic variables included in final factor analysis

	D1	D2	D3	D4	D5	D6
adv	0.142		-0.137	0.104		0.464
advl_stance_all	0.195	-0.13	-0.131			0.339
all_def_art	-0.247	0.263		-0.297	-0.127	-0.443
all_indef_art					-0.188	
conj_advl						0.257
conj_all	0.267	-0.166	-0.202	0.235		0.379
contract	0.565	-0.236	-0.196	0.289	0.207	0.277
emphatic	0.211					0.17
infinitive	0.18	-0.144	-0.118	0.662	0.103	0.185
jj_attr	-0.334	0.246	0.303	-0.349		-0.168
jj_pred	0.199	-0.146				0.284
mod_poss	0.262	-0.119	0.152			0.241
mod_pred	0.454	-0.173	-0.106	0.266	0.113	0.18
nn_abstact	-0.127		0.316	-0.112		
nn_all	-0.688	0.364	0.557	-0.573	-0.282	-0.657
nn_common	-0.519	0.234	0.669	-0.315	-0.35	-0.34
nn_concrete			0.177		-0.259	
nn_nom			0.165	-0.189		
nn_place				-0.114	-0.116	-0.244
nn_premod	-0.337	0.131	0.559	-0.208	-0.199	-0.239
nn_proper	-0.31	0.354	-0.199	-0.303		-0.422
passive_short	-0.229	0.128		-0.167		-0.11
prep	-0.417	0.376	0.11	-0.429	-0.159	-0.538
pro_1	0.571	-0.277	-0.212	0.326	0.229	0.2
pro_2	0.472	-0.162		0.266		0.189
pro_3		-0.218	-0.448	0.275	0.156	0.275
pro_it	0.256	-0.104				0.107
th_stance_all	0.271		-0.121	0.303	0.858	0.232
th_vb	0.172			0.158	0.539	0.188
that_del	0.279	-0.123	-0.161	0.182	0.452	0.176
to_stance_all	0.212	-0.149	-0.119	0.651	0.188	0.103
tt_ratio	-0.313	0.931	0.117	-0.185	0.117	-0.105
vb_act	0.219	-0.247	-0.225	0.379	-0.126	0.118
vb_be	0.29					0.207
vb_comm			-0.107	0.108	0.241	0.142
vb_mental	0.402	-0.232	-0.21	0.53	0.407	0.268
vb_past	-0.349		-0.542		0.156	-0.115
vb_present	0.762	-0.444	0.154	0.362	0.108	0.497
vb_progress	0.117			0.221	0.117	0.156
wh_cls	0.168			0.146		0.124
wh_ques	0.257					
word_count	-0.32	0.921	0.149	-0.216		-0.231
word_length	-0.604	0.448	0.488	-0.457		-0.186

Table E.2: Factor (structure) matrix; loadings $>$ .10