Item Difficulty Modeling for a Sentence Reading Efficiency Task with Language Model Simulations

Stanford CS224N Custom Project

Wanjing Anya Ma
Graduate School of Education
Stanford University
wanjingm@stanford.edu

Abstract

Silent reading is the most common form of reading, and has been found to be a better indicator of reading comprehension compared to oral reading fluency (ORF) for students in 2nd grade and higher. Our previous research conducted a case study to demonstrate how to generate and evaluate parallel test forms of this task with language model simulations Zelikman et al. (2023). Although our previous work shows promising results on using language models to predict the item parameters of this SRE task, the accuracy for each individual item estimate remains unstable and highly depends on which students included in the simulation set. The goal of this project is to explore the generalization ability of LM simulations to new students and new items. We first replicated the item response simulator with a smaller language model Phi-2. We then compared the response time predictions obtained by the LM simulation data and by the real data. The results indicate the response time is more attributed to the person parameters than the item parameters, but more complex neural network model doesn't help with increasing the generalizability of the model compared with the simple linear regression model. We conclude item difficulty modeling of completely new items is still a challenging topic even with complex LM simulations and extensive previous response data.

1 Key Information to include

- TA mentor: Yann Dubois
- External collaborators (if you have any): Eric Zelikman (Eric and I collaborated on the LLM simulation part of this project in the winter quarter, and I worked independently on item generation, data collection, simulation response analysis, and neural network construction.)
- External Mentors: Jason Yeatman (my advisor)
- Sharing project: NA

2 Introduction

Silent reading is the most common form of reading, and has been found to be a better indicator of reading comprehension compared to oral reading fluency (ORF) for students in 2nd grade and higher Kim et al. (2012). Our previous research have developed and validated an online 3-minute timed measure to assess sentence reading efficiency (SRE) Tran et al., where students are asked to tell the sentence either TRUE or FALSE as fast and as accurate as they can 1. We also conducted a case study to demonstrate how to generate and evaluate parallel test forms of this task with language model (LM) simulations Zelikman et al. (2023). Although our previous work shows promising results on using language models to predict the item parameters of this SRE task, the accuracy for each individual item estimate remains unstable and highly depends on which students included in the simulation set.

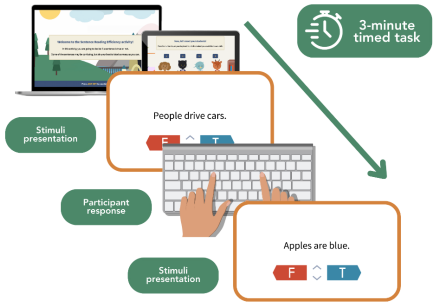

Figure 1: textbfSentence Reading Efficiency Task

This is a shared problem for language model simulations across many tasks that the results highly depend on an aggregated score, and large repetitions may be extremely expensive to achieve. The task of this project is to predict item difficulty of an unseen sentence reading efficiency item. Language models (eg. GPT4) can be used to generate huge amount of candidate test items, but it is just the first step of a test development. The goal of this project is to explore the generalization ability of LM simulations to new students and new items. Specifically, we ask two questions:

1. Person features vs. item features: Which types of features are better predictors of response time in the SRE task?
2. How well do the item difficulty parameters from LM simulations generalize to new students in the SRE task?

If we have a better understanding of the generalization ability of the LM simulations and obtain more accurate item difficulty parameter estimates, we can further develop computerized adaptive test to make the assessment more efficient (although it is not the scope of this project).

3 Related Work

Item difficulty modeling in psychometrics Item difficulty modeling becomes an increasingly important topic in psychometric community, especially with the need of larger item bank for more adaptive and personalized assessments. Belov et al. (2024) explores the use neural network of response patterns to calibrate item difficulty of 3PL with small sample. There are serveal psychometric models suitable for the SRE task. For example Kara et al. (2020) parameterizes the oral reading fluency to both person level and item level. The speed–accuracy response model proposed by Maris and van der Maas Maris and Van der Maas (2012) and improved by van Rijn van Rijn and Ali (2018) is a popular psychometric model suitable for modeling the item and person parameters when there is a time constraint in a measurement.

Item response simulation using LMs Building simulators/agents to answer questions becomes increasingly popular with the development of language models. In educational applications, Srivastava and Goodman (2021) fine-tuned an LM to predict the probability of a student answering a question correctly in an adaptive setting. Our previous work **?** fine-tuned an LM to simulate student's responses to sentence reading efficiency tasks and uses the simulator to calibrate new test items and generate parallel tests.

4 Approach

This task has three major components:

Language Model Simulations. Previous work has demonstrated fine-tuning language models with existing student responses can generate and evaluate high-quality test forms for measuring sentence reading efficiency (SRE) Zelikman et al. (2023). However, the major limitation of previous work was that the cost of fine-tuning was extremely expensive and could not be generalized for a research-based

education application with limited resources. To address this, we changed the language model from LLaMA 2 Touvron et al. (2023) to Phi-2 Microsoft Research (2024), which largely reduced the allocated compute resources and training time. This also allowed us to increases number and diversity of student responses in the training. In addition, because the trained reward model also has a smaller size, we had the capacity to increase number of students used in the simulation (from 100 to 250) for predicting the item difficulty for each new item. This improvement builds a foundation for a more stable item difficulty modeling.

Efficient Item Calibration. In psychometrics, obtaining a stable item parameter (i.e. calibration) for a new item usually requires at least 200 responses from a representative population. This calibration process is extremely expensive when there is large number of item items and when each student has limited test time to respond new items that don't contribute to scoring. A classic way to calibrate new items in a real testing scenario is to design a test that has a mix of calibrated items and new items. However, one of the constraints in the SRE task is it is a 3-minute timed measure, so all the calibrated items have to be given in a fixed order for scoring purpose. To calibrate GPT-generated new items, followed by a short break after the standard 3-minute SRE measure, we we create another 90s block to test new items only. We split the 700 new items to 50 test bundles with balanced item difficulty based on LM item simulations. Each student will be randomly assigned 5 test bundles from a total of 50, each consisting of a maximum of 70 items they can respond to within 90 seconds. This sampling strategy has two advantages: 1. For each new student, we have their ability estimate from the standard SRE block with calibrated items only, which can be used as a prior for item difficulty modeling. 2. Items in the same bundle will be guaranteed to be seen by the same students, allowing us to learn the relative difficulty between items.

Generalization Analysis. In a psychometric model, the probability of a correct response or the expected response time distribution depends on both item parameters and person parameters. In the SRE task, we use students' response patterns to previous items as person parameters. However, the item parameters for new items are often limited to basic linguistic features such as length and readability scores.

Fortunately, we can use language model (LM) simulations to generate simulated responses from previous students to new items. With these simulated responses, we can train a neural network to predict responses using only person parameters and limited item parameters. This approach serves as a generalization test to assess how well LM-generated responses can be used as training data to predict real student responses.

5 Experiments

This section contains the following.

5.1 Data

There are two parts of data in this project. First, to fine-tune the language model and create simulations for predicting the new 700 test items, we have collected more than 4000 K-12 student responses on a 3-minute standard SRE task (shown in 1) over the past two years (Tran et al.). Second, to collect the ground true parameters for these 700 new items from about 1500 K-12 students with in total of more than 36000 unique item responses. For each response, we collect the response accuracy (correct/incorrect) and response time in milliseconds. Based on our existing exclusion criteria, we excluded participants whose median response time is faster than 1000 ms AND with response accuracy lower than 0.6. We also excluded responses with extreme response time: faster than 500ms and longer than 20,000ms. After exclusion criteria, there are in total of 33544 unique student responses and each new item has been seen by around 50 real students and around 250 LM simulated students.

5.2 Evaluation method

This project is more exploratory than achieving a certain benchmark metric. Here we define three checkpoints to evaluate our work. First, our previous work shows the aggregated item parameter (median response time) based on LM simulation results correlates with actual median response time

around 0.6 for both true sentences and false sentence. In this project, we change to a smaller model but with more diverse training data, we should expect similar or better correlation. Second, to evaluate the neural network model we use to predict the response time, we use Pearson correlation and mean squared error (MSE) between predict and actual responses. We expect the correlation should be relatively higher and the MSE be relatively lower for the neural network model with response patterns as embedding than the simple linear regression with aggregated features. Third, we qualitatively check which items are more or less generalizable by looking through the list of items and their associated correlation between predicted and actual response.

5.3 Experimental details

Item response simulator. The fine-tuning part follows the procedures of our previous work (Zelikman et al., 2023): we selected a constant learning rate of 1e5 with a batch size of 32 sets of items, including a random sample of up to 30 student item-response pairs (fewer only if the student responded to fewer items). We set both classifier dropout and hidden dropout rates as 0.2, and adopted flash attention 2 implementation(Dao, 2023).

GPT4 generated sentences. We used openAI API to generate new items with GPT4. We let GPT4 see a list of calibrated items, and we then prompt: "For each sentence below, generate 10 unique sentences that follow the following rules: 1. vary by content and semantics. 2. keep the similar length and readability. 3. safe and appropriate for children to read. 4. similar number of true and false sentences.". We set temperature as 1 to engage GPT4 to generate diverse items, used 5,000 tokens each time, and set presence penalty to 0.2. We iterated this process until we have 1000 unique true sentence and 1,000 unique false sentence. We then asked four human experts in the reading assessment domain to exclude sentences that have potential biased or harm for children to read.

LM simulated responses. We used our trained item response simulator to simulate previous students to respond new items and predict their response time log scale and probability of answer this item correct. We then used the simulated responses to investigate two person parameters: median response time and total score from a standard SRE task (total correct - total incorrect within 3 minutes), as well as two item parameters: number of words per sentence and traditional readability metric Flesh Kincaid.

Neural Network Model. We then constructed a more complex training data with more complex person parameters. Since everyone took the same standard SRE task with a maximum of 130 items, for each student, we added 130 correct/incorrect features and 130 response time features in log scale. The matrix is very sparse because students only responded to a subset of items given the time limit, and the missing is not by random. Therefore, we imputed the response time missing values by filling in 10, which corresponds to a very long response time around 20,000 ms. We also imputed the correct/incorrect missing values as 0, meaning incorrect for all items that run out of time to answer.

We then combined these 260 features with 2 item features (length and flesh kincaid) as the training data, and used simulated response time as the predicted variable. We split the training data to 0.7 training, 0.15 evaluation, and 0.15 testing. We designed a simple feedforward neural network with three layers (from 268 to 64, from 64 to 32, and from 32 to 1), and set the dropout rate to be 0.5. We experimented several different settings about learning rate and number of epochs, and decided to use 0.001 learning rate and 20 epochs for best performance. We applied the best model to the test set within the simulation data and the real data (with new participants and real responses on these new items). Please note, we didn't not train our model in any of the real data, because we are curious about how these LM simulated data can generalize the prediction of real student response.

5.4 Results

2 suggests that with a smaller model Phi-2 and less training time in general, the predictions are almost equivalent as what we have found with LLaMA 2 in our previous work.

Based on comparison shown in Table 1., person parameters that leverage previous responses on calibrated items are stronger predictors to item response time than item parameters (length and flesh kincaid). This highlights, in the SRE task, the observed response time is more attributed by person level than the item level. Among all the aggregated features, the median response time on previous

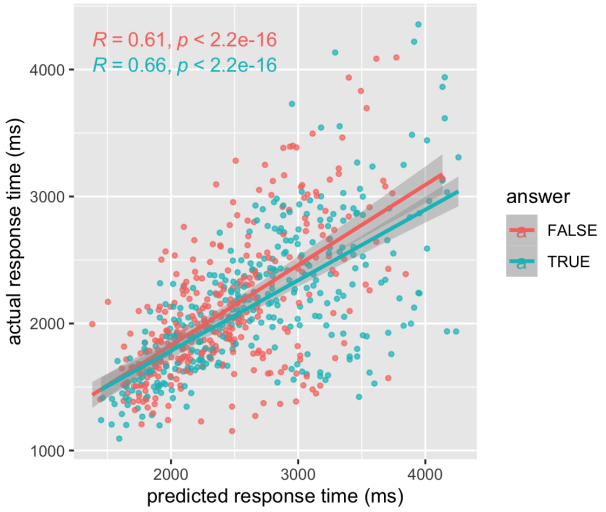

Figure 2: Response Time Prediction

	Predictors	Simulation data Correlation	Simulation data MSE	Real data Correlation	Real data MSE
item features	length	0.408	/	0.287	/
	flash kincaid	0.176	/	0.12	/
peson features	previous total score	-0.636	/	-0.437	/
	previous median rt	0.747	/	0.484	
ML models (predicted vs. actual)	Linear Regression: flash kincaid + length + previous total score + previous median rt	0.87	0.201	0.578	0.478
	NN: flash kincaid + length + previous rt patterns + previous correct patterns (262 features)	0.9	0.176	0.57	0.48

Table 1: Feature comparison and generalization ability between simulation data and its application to real data

calibrated items is most correlated with the response time on new items (r = 0.747 in simulation data). By comparing with the baseline of using linear regression model, the neural network does not add too much value in adding more complex prior response features. However, given the improvement of MSE, we found that the neural network can help to match the predicted response time to actual response time more accurately in terms of the absolute value.

By comparing the prediction performance of the neural network model between simulation data 3 and real data 4, we found that the current LM simulated item difficulty modeling still faces the challenge of generalization, which will be discussed in the next section.

6 Analysis

To understand if certain items are more easily predicted than the others, we plotted the correlations for each single item between their predicted and actual response time, shown in Figure 5 5. First, there is no difference in terms of generalizability between true and false sentence. Second, there is no trend on either harder or easier items are more predicted than the others. For example, both sentences "A triangle has three sides." and "Beds love to swim." only have correlation around 0.1, which means the simulation and real data respond completely differently on them. The sentence "A kitten has a tail." has the highest correlation 0.83.

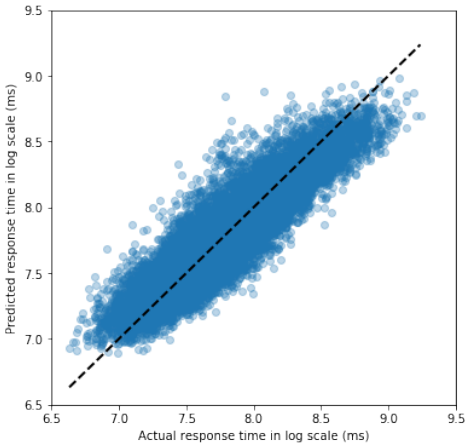

Figure 3: prediction within simulation data

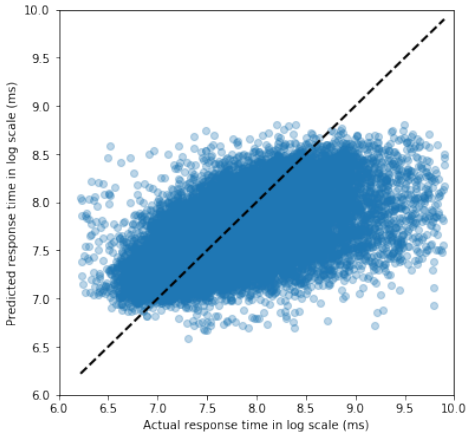

Figure 4: prediction to apply real data

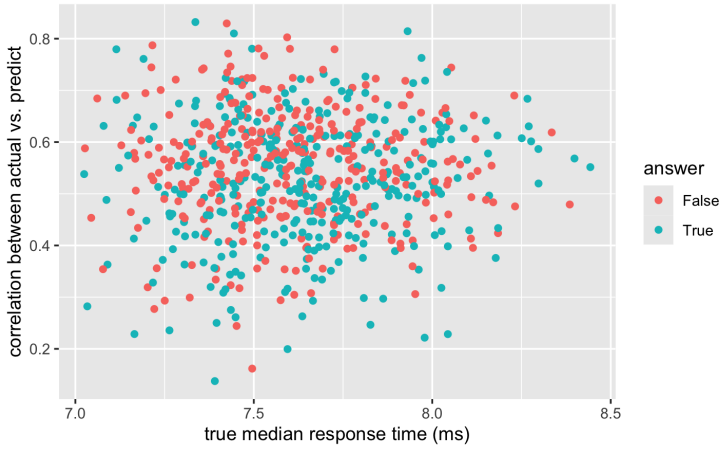

Figure 5: Correlation between predict and actual response time across items

One exciting insight from this project is understanding the aggregated scores (total score and median response time) from student previous responses in the standard SRE task are sufficient to represent the person parameters. This is meaningful for us to calibrate new items with confidence as long as we keep students taking the SRE task along with piloting new items.

7 Conclusion

In this project, with language model simulations, we present a novel approach by combining both person parameters and item parameters to predict the response time of new items in a sentence reading efficiency task. The results indicate in this reading task, the response time is more attributed to the person parameters than the item parameters, but more complex neural network model doesn't help with increasing the generalizability of the model compared with the simple linear regression model. We conclude item difficulty modeling of completely new items is still a challenging topic even with complex LM simulations and extensive previous response data. There are several limitations in this project. First, we included all response patterns from the standard SRE task to fit the neural network, but we could decrease number of items and investigate the performance of the model with smaller number of previous responses. Second, due to the time limit, we didn't incorporate complex psychometric modeling such as speed-accuracy trade-off modeling to model the relationship between response accuracy and response time.

8 Ethics Statement

This work involves significant ethical considerations, particularly in the handling of student data to ensure privacy. We have used only d-identifiable data, adhering to strict usage standards and IRB guidelines. Moreover, before presenting any test item generated by language models, especially to children, we conduct multiple review and analysis rounds to confirm its appropriateness. Our approach aims to minimize the burden of test creation, relying on humans primarily as final reviewers and experts, thus reducing the number of items they need to assess. Most importantly, we emphasize the importance of incorporating actual student responses into the simulation loop to predict the difficulty of each new test item, rather than relying solely on language models.

References

Dmitry I Belov, Oliver Lüdtke, and Esther Ulitzsch. 2024. Likelihood-free estimation of irt models in small samples: A neural networks approach.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Yusuf Kara, Akihito Kamata, Cornelis Potgieter, and Joseph FT Nese. 2020. Estimating model-based oral reading fluency: A bayesian approach. *Educational and Psychological Measurement*, 80(5):847–869.

Young-Suk Kim, Richard K Wagner, and Danielle Lopez. 2012. Developmental relations between reading fluency and reading comprehension: A longitudinal study from grade 1 to grade 2. *Journal of experimental child psychology*, 113(1):93–111.

Gunter Maris and Han Van der Maas. 2012. Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, 77(4):615–633.

Microsoft Research. 2024. Phi-2: The future of small language models. Technical report, Microsoft. `https://phi2.io/`.

Peter W van Rijn and Usama S Ali. 2018. A generalized speed–accuracy response model for dichotomous items. *Psychometrika*, 83(1):109–131.

Megha Srivastava and Noah Goodman. 2021. Question generation for adaptive education. *arXiv preprint arXiv:2106.04262*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. Technical Report 2307.09288, arXiv. `https://arxiv.org/abs/2307.09288`.

Jasmine Elizabeth Tran, Jason Yeatman, Amy Burkhardt, Wanjing Anya Ma, Jamie Mitchell, Maya Yablonski, Liesbeth Gijbels, Carrie Townley-Flores, and Adam Richie-Halford. Development and validation of a rapid online sentence reading efficiency assessment.

Eric Zelikman, Wanjing Ma, Jasmine Tran, Diyi Yang, Jason Yeatman, and Nick Haber. 2023. Generating and evaluating tests for k-12 students with language model simulations: A case study on sentence reading efficiency. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2190–2205, Singapore. Association for Computational Linguistics.