Word Discriminations for Vocabulary Inventory Prediction

Frankie Robertson
University of Jyväskylä
frankie@robertson.name

Abstract

The aim of vocabulary inventory prediction is to predict a learner's whole vocabulary based on a limited sample of query words. This paper approaches the problem starting from the 2-parameter Item Response Theory (IRT) model, giving each word in the vocabulary a difficulty and discrimination parameter. The discrimination parameter is evaluated on the sub-problem of question item selection, familiar from the fields of Computerised Adaptive Testing (CAT) and active learning. Next, the effect of the discrimination parameter on prediction performance is examined, both in a binary classification setting, and in an information retrieval setting. Performance is compared with baselines based on word frequency. A number of different generalisation scenarios are examined, including generalising word difficulty and discrimination using word embeddings with a predictor network and testing on out-of-dataset data.

1 Introduction

Given a small sample of words, how well can we predict whether a learner knows some out-of-sample word? This is the task of vocabulary inventory prediction. A clear motivation for the topic is to enable quicker and more precise placement testing. For example, a 40 word self-assessed word knowledge quiz used as a benchmark in this paper is quick enough that an L2 learner returning to a language learning app after a long break, in which they may have either forgotten a lot or had a lot of extra exposure to their target language, can be placed again quickly without excessive disruption.

This paper addresses the following research questions:

1. What are the empirical differences in performances between difficulty parameters produced by estimation of Item Response Theory (IRT) models and those based on word frequency in terms of their application to vocabulary inventory prediction?
2. How well can the IRT parameters of difficulty and discrimination be regressed based on word embeddings?
3. Which approaches from the field of Computerised Adaptive Testing (CAT) help to select good items to query? Does the addition of a discrimination parameter help with question selection?
4. Does the addition of a discrimination parameter help with the final prediction step?

2 Related Work

Milton (2009) refers to the common assumption when quantifying vocabulary acquisition that words are learnt in approximately descending order of frequency as the frequency assumption. It has been used in the field of reading research, for example in estimating vocabulary size, but can also provide a simple baseline for the task of vocabulary inventory prediction.

Avdiu et al. (2019) approached the problem through feature engineering, taking frequency profiles of different genres and associating learners with them according to their responses. They used a large section of the data for training, without testing a scenario in which learned data is to be generalised to new learners with less data available, as in this paper

Item Response Theory (IRT) (Tatsuoka et al., 1968; Baker, 2001) is widely used to determine item difficulties and examinee ability in academic assessments. A key drawback of traditional IRT is that the actual content of the items is ignored. Instead, items are only understood in terms of their responses. This leaves no possibility of generalising item parameters to unseen items. Recent work

Proceedings of Recent Advances in Natural Language Processing, pages 1188–1195
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_134

has begun to generalise difficulty scores based on representations based on items' textual content using deep neural networks. For example Benedetto et al. (2021) first fit an IRT model on questions from a cloud technology certification exam, before training a transformer model to regress the resulting difficulty scores, allowing generalisation to new questions without a pre-testing stage.

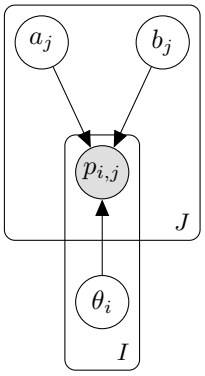
Ehara (2019) approaches the problem of vocabulary prediction by fitting a Rasch (1960) model, equivalent to a 1-parameter logistic IRT model. The problem was modelled such that an equivalent neural network was constructed which included features based on Glove (Pennington et al., 2014) word embeddings. As with Avdiu et al. (2019), a single stage of training was performed so that the ability of the learners was learnt simultaneously with the weights of the prediction network. This network did not beat a word frequency and logistic regression baseline. In this paper, a 2-parameter logistic IRT model is fitted as an initial step, before proceeding to generalise these parameters using a word embedding based regressor.

Computerised Adaptive Testing (CAT) (Lord, 1977; Wainer, 2000) has not been widely applied to the task of vocabulary inventory estimation. A CAT system selects questions based on a examinee's previous answers in order to converge on an accurate ability estimate faster. Related, but outside of CAT/IRT setting, Ehara et al. (2014a) builds graphs made from a combination of multiple corpora combined and apply label propagation to find a fixed set of queries to in-effect give a more accurate ability estimate than choosing at random. Restricting ourselves to the adaptive setting, the main prior art is the website http://testyourvocab.com/, which uses CAT to estimate vocabulary size based on word frequencies. To the best of the author's knowledge, there is no prior work attempting to quantify how accurate the ability estimates obtained when applying CAT to the problem of vocabulary inventory estimation are.

3 Method

3.1 Datasets

Three datasets are used in this paper. The first, SVD12K, is due to Ehara et al. (2012) and contains 12 000 words rated on a 5-point scale by 16 learners of English, most of whom have Japanese as their native language. Following Ehara et al. (2014b), the first learner is discarded due to lower quality data. The learners in SVD12K were all students of the University of Tokyo and we speculate that it is quite possible they have all learnt English for similar purposes, i.e. academic usage, and may have even attended the same English classes.



$$a_j \sim \mathcal{N}(1.2, 0.25)$$
$$b_j \sim \mathcal{N}(0, 1)$$
$$\theta_i \sim \mathcal{N}(0, 1)$$

Figure 1: Plate diagram showing the Bayesian network corresponding to the 2-parameter logistic IRT model.

The other two datasets are used as additional test sets, so as to see how well the techniques generalise beyond the potentially rather narrow distribution of SVD12K. Both of the two extra datasets are constructed such that they should be mainly composed of learners with Japanese as their L1, i.e. testing of generalisation beyond learner L1 is not considered here. Ehara (2018) introduce EVKD1, a dataset consisting of responses to a 100 word 4-way multiple choice test given to 100 participants, administered using a Japanese crowdsourcing platform. Respondents were asked to choose the correct definition of a word given in a context sentence. The final dataset is a section of responses to the website TestYourVocab[1] limited to responses from 2018 by participants who selected their country as "Japan". This dataset has a different selection of responses for each person.

3.2 Fitting an IRT Model

Given a matrix of responses $r_{i,j}$ indexed by items i and respondents j, an IRT model predicts latent features of the items and respondents. Respondents are assigned abilities θ_i, while in 2-parameter IRT models, items are assigned difficulties a_j and discriminations b_j. Typically we predict binomial responses based on an Item Characteristic Curve

[1] Obtained by direct request from the owner of http://testyourvocab.com/.

(ICC) like so:

$$ICC_j(\theta) = \left(1 + e^{-a_j(\theta - b_j)}\right)^{-1}$$
$$P(r_{i,j}|\theta_i, a_j, b_j) = ICC_j(\theta_i)$$
$$Q(r_{i,j}|\theta_i, a_j, b_j) = 1 - ICC_j(\theta_i)$$

The data of Ehara et al. (2012) is rated on a 5-point scale, suggesting a graded IRT model. A typical formulation may try and learn separate difficulty and discrimination parameters per item-level pair, significantly increasing the number of parameters to be learnt. In order to reduce the amount of data necessary to fit the IRT model, we learn only one difficulty discrimination per item and create fixed global offsets $l_{1...4} \geq 0$ to create offset difficulties for the thresholds. We then model:

$$P(r^\star_{i,j} \geq k|\theta_i, a_j, b_j) = ICC_j(\theta_i - \textstyle\sum_{s=k}^{4} l_s)$$

And note that:

$$P(r^\star_{i,j} = k) = P(r^\star_{i,j} \geq k) - P(r^\star_{i,j} \geq k+1)$$
$$P(r^\star_{i,j} \geq 1) = 1$$
$$P(r^\star_{i,j} \geq 6) = 0$$

We estimate the Maximum A Posteriori (MAP) with Stan (Carpenter et al., 2017). The priors are illustrated alongside Figure 1. After fitting the model we revert to considering the binomial case by defining $P(r_{i,j}) := P(r^\star_{i,j} = 5)$.

3.3 Frequencies as a Difficulty Baseline

A simple frequency baseline for difficulty was constructed based on the word frequencies of the wordfreq (Speer et al., 2018) library. The wordfreq library incorporates frequencies from multiple corpora of different registers, ensuring balanced coverage by taking equal contributions from each register after removing outliers. Internally, wordfreq stores log frequencies on an 800 point scale. These are first negated and then standardized according to their mean and standard deviation based on the words in the SVD12K dataset so that they lie in the same range as the IRT difficulties.

To the best of the author's knowledge, given good frequency data, this baseline has not yet been significantly surpassed on this task in the setting where there are only a small number responses available from the learner, making it effectively state-of-the-art.

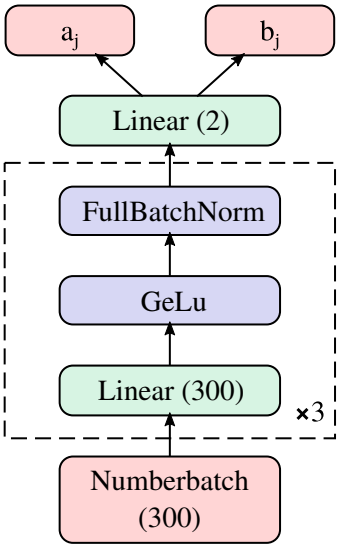


Figure 2: The architecture of the IRT item parameter regressor network.

3.4 Generalising IRT Item Parameters

In order to generalise the difficulty and discrimination parameters beyond the words present at IRT model estimation time, a Multi-Layer Perceptron (MLP) was trained as a regressor for both parameters. Words are input to the network as Numberbatch 19.08 (Speer et al., 2017) embeddings. These 300 dimensional embeddings, based on lemmas rather than word forms, are constructed by combining multiple distributional word embeddings with information from the ConceptNet lexical knowledge graph. They were chosen because most vocabulary tests are either based on lemmas or word families rather than word forms, and because they have performed well in previous studies.

The architecture shown in Figure 2 was implemented using PyTorch (Paszke et al., 2019). The GeLu activation function (Hendrycks and Gimpel, 2016) and BatchNorm (Ioffe and Szegedy, 2015) are used as non-linearities. Since full batch training is used here, the BatchNorm damping parameter, which is intended to stabilise random variations in minibatches, is not used. The Adam optimizer (Kingma and Ba, 2015) was used with a learning rate of 0.003. Training was performed for 50 iterations and the best iteration on the validation set created by 1:11 validation:train split was chosen.

3.5 Computerised Adaptive Testing

The aim of Computerised Adaptive Testing (CAT) (Lord, 1977; Wainer, 2000) is to estimate a learner's ability parameter θ as accurately as possi-

ble with as few queries as possible. Key parts of a CAT system are initialisation, next item selection, and θ estimation. After initialisation, the system repeatedly queries a new item from the learner and re-estimates θ^* until a termination condition. Here, we terminate after having made 40 queries, and always initialise θ^* to be 0.

Next item selection rules are typically formulated as choosing an next item so as to maximise some measure of merit. Here we consider the maximisation of Fisher information introduced to the field of CAT by Lord (1977), and denoted as MaxInfo. For the 2-parameter logistic IRT model the Fisher information is defined as:

$$I_j(\theta) = a_j^2 ICC_{a_j,b_j}(\theta)(1 - ICC_{a_j,b_j}(\theta))$$

An alternative next item selection rule is due to Urry (1970), and denoted as such, and simply picks questions close to the current estimate of θ. Note that this is equivalent to the max entropy heuristic in active learning, which queries the data point about which the current version of the classifier is most uncertain.

There are two approaches for estimating θ^*. The first, denoted *Full-ICC*, starts from a binomial IRT model introduced in Section 3.2 and incomplete response data $U = \{u_j | j \in J, u_j \in \{0, 1\}\}$. We then obtain θ^* by maximum likelihood estimation:

$$\mathcal{L}(\theta) = \prod_{u_j \in U} P(r_{i,j}|\theta, a_j, b_j)^{u_j} \times Q(r_{i,j}|\theta, a_j, b_j)^{(1-u_j)}$$
$$\theta^* = \arg\max_{\theta} \mathcal{L}(\theta)$$

The second, denoted *Difficulty Only*, ignores the discriminations of the items, which is equivalent to setting all $a_j = 1$. Substituting the resulting ICC expressions into the likelihood reveals an equivalence with logistic regression. Namely, after fitting a logistic regression model on the responses U, we get a model with coefficient m and intercept c. We then find that $\theta^* = \frac{-c}{m}$.

In early iterations, there may only be positive or negative responses. In this case we apply the method of Dodd (1990), which averages the previous theta estimate with either the maximum or minimum item difficulty value depending on the direction in which θ^* would otherwise diverge.

As a non-CAT baseline, there is stratified random selection, denoted Rand. In order to guarantee a reasonable range of item difficulties are asked, strata for the words are created by ordering by frequency and splitting into 5 equal sized strata. The random selection procedure then chooses 40 items randomly, taking equally from each stratum.

The catsim Python library (De Rizzo Meneghetti and Aquino Junior, 2017) is used for the implementations of all CAT techniques.

3.6 Evaluation

The vocabulary inventory prediction task can be viewed as a binary classification problem. The Receiver Operator Characteristic (ROC) curve plots the recall of the positive class against the recall of the negative class by varying the classifier threshold. Statistics based on ROC curve, such as Area Under ROC (AUROC) enjoy the key advantage of threshold invariance. On the other hand, we typically do have to pick some threshold and for this reason, a metric based on a default threshold of 0.5 is given: Matthews Correlation Coefficient (MCC). The second angle on the problem is that of known and unknown word retrieval. In this case Average Precision (AP) acts as a threshold invariant measure of retrieval performance. We consider AP+ and AP- for measuring retrieval performance from the two classes of known and unknown respectively.

AUROC does not change significantly based on exact ability estimate of the learner due to its lack of a fixed threshold. Here, we use it only to explore different ways the difficulty parameter can be obtained and the effect of including the discrimination parameter. Being based on a fixed threshold, MCC is highly sensitive to the actual ability estimate, and so it gives a more realistic picture of performance practically. The metrics AP+ and AP- are used to measure an upper bound on the performance on the retrieval tasks.

Intuitively, we can see low values of discrimination as reflecting a degree of uncertainty about a word's true difficulty. The information retrieval perspective is particularly relevant here since the presence of the discrimination parameter means that, for example in unknown word retrieval, words that are highly discriminating but less difficult could be returned earlier than words with low discrimination that are more difficult, potentially improving performance.

4 Experiments

We first evaluate how well the item/word parameters from the IRT model can be regressed with

Gen	Param	MAE	Norm
Words	Diff	0.595	0.495
	Discrim	0.148	1.365
Both	Diff	0.608	0.506
	Discrim	0.147	1.356

Table 1: Table containing the both the raw Mean Absolute Error (MAE) and the MAE normalised by the true standard deviation of difficulties and discriminations as predicted in two word generalisation scenarios.

the chosen architecture. Next we move on to consider how well various CAT approaches can estimate learners' abilities. Finally, the results for the final task of vocabulary inventory prediction are presented, first cross validating on the SVD12K dataset and then training on the whole SVD12K dataset and testing on the extra datasets.

Four generalisation scenarios are considered across experiments:

Gen-None No generalisation; The IRT model is fitted on the same data as the test data.

Gen-Word Generalising only to new words; 3-fold cross validation is performed on words, with the IRT model being fitted on ⅔ training words, before fitting the MLP on the results to predict the out-of-vocabulary ⅓ of words.

Gen-Respondent Generalising only to new learners; 3-fold cross validation is performed on participants, with the IRT model being fitted on ⅔ participants, from which the item parameters are used as-is on the out-of-sample ⅓ of participants.

Gen-Both Generalising to new words and learners; 9-fold cross validation is performed, consisting of the product of 3-fold cross validation on participants with 3-fold cross validation on words.

4.1 Predicting Item Parameters

Table 1 gives the results evaluating the performance of the IRT parameter regressor. When looking at the results normalised by true standard deviation, it is clear that the parameter of discrimination is more difficult to predict. The lower error in predicting difficulties in the Gen-Words scenario suggests that the more accurate IRT predictions made with more data do indeed provide an easier target for the network to fit. However, the actual errors are quite close, and the generalisation scenarios tend to give similar results, so for this reason Gen-Words is not considered further in the later results.

Gen	Estimator	Next Item	MAE	Norm
Both	Full ICC	Rand	1.204	1.136
		Urry	1.117	1.053
		Max-Info	1.110	1.047
	Difficulty only	Rand	1.087	1.025
		Urry	1.037	0.978
		Max-Info	1.114	1.051
Resp.	Full ICC	Rand	1.334	1.258
		Urry	1.150	1.084
		Max-Info	1.199	1.131
	Difficulty only	Rand	1.280	1.207
		Urry	1.068	1.007
		Max-Info	1.211	1.142
None	Full ICC	Rand	1.372	1.294
		Urry	1.105	1.042
		Max-Info	1.395	1.316
	Difficulty only	Rand	1.249	1.178
		Urry	1.233	1.163
		Max-Info	1.338	1.262

Table 2: Table showing the raw MAE and MAE normalized by standard deviation of estimated difficulties after 40 questions versus true difficulties.

4.2 θ-estimation

We now turn to the matter of how well θ is estimated using different approaches. The results are shown in Table 2.

For both next item selection methods and θ-estimation methods, including the discrimination parameter seemed to decrease performance. Noteworthy is that the best overall score is obtained by difficulty-based CAT for the Gen-Both and Gen-Resp, with this setting in the Gen-Both scenario outperforming the others, showing that the regressed word difficulties perform well for this task. For the Gen-None scenario, including the full ICC when estimating θ appeared to help. It may be that having non-regressed discrimination values based on responses from more respondents helped in this case.

However, since discriminations appear to not be generally useful for finding θ in any generalisation scenario, they are not used further in the next section and the Urry (1970) next item rule is used together with the difficulty only θ estimator.

4.3 Vocabulary Inventory Prediction

We now evaluate the final task of vocabulary inventory prediction. Table 3 shows the results on this task using the metrics introduced in Section 3.6. The experiments compare the use of dif-

Gen	Diff.	Dis.	ROC	MCC	AP+	AP-
Both	Resp.	Off	0.829	0.398	0.848	0.722
		On	0.827	0.398	0.845	0.720
	Freq	Off	0.805	0.260	0.842	0.659
		On	0.799	0.260	0.839	0.656
Resp.	Resp.	Off	0.882	0.492	0.879	0.786
		On	0.882	0.492	0.880	0.786
	Freq	Off	0.805	0.346	0.842	0.659
		On	0.805	0.346	0.843	0.658
None	Resp.	Off	0.915	0.598	0.933	0.840
		On	0.918	0.598	0.934	0.843
	Freq	Off	0.805	0.359	0.842	0.659
		On	0.811	0.359	0.845	0.666

Table 3: Table showing results on the SVD12K dataset in different generalisation settings given different choices of source difficulty parameter and whether to include the discrimination parameter in predictions.

Gen	Diff.	Dis.	ROC	MCC	AP+	AP-
Freq	Freq	Off	0.690	0.228	0.711	0.650
Pred	Resp.	Off	0.658	0.195	0.676	0.614
		On	0.656	0.205	0.676	0.608
	Freq	On	0.680	0.228	0.704	0.648
Mix	Resp.	Off	0.670	0.261	0.711	0.625
		On	0.677	0.280	0.718	0.625
	Freq	On	0.687	0.262	0.715	0.654

Table 4: Table showing results on the EVKD1 dataset of different choices of source difficulty parameter and whether to include the discrimination parameter in predictions.

ficulties from IRT versus the wordfreq baseline, and whether or not the discrimination parameter is used for prediction. The idea behind using the discrimination parameter in prediction is that highly discriminating words may receive more confident scores even when they're further from ability estimate than a nearer lowly discriminating word since the discrimination parameter acts as a measure of certainty of the item's difficulty.

From the results we can see that using difficulties based on word frequencies reduces performance across the board. The inclusion of the discrimination parameter in most cases does not seem to make too much of a change, slightly decreasing performance in the Gen-Both scenario, and making very little difference for Gen-Respondent. Although there is a small improvement in the Gen-None case, this reflects the IRT model's goodness of fit, rather than how well the values generalise.

Gen.	Diff.	Dis.	ROC	MCC	AP+	AP-
Freq	Freq	Off	0.895	0.612	0.886	0.903
Pred	Resp.	Off	0.843	0.516	0.835	0.844
		On	0.837	0.516	0.826	0.839
	Freq	On	0.888	0.612	0.880	0.896
Mix	Resp.	Off	0.878	0.609	0.871	0.877
		On	0.876	0.609	0.872	0.875
	Freq	On	0.893	0.612	0.886	0.901

Table 5: Table showing results on the TestYourVocab dataset of different choices of source difficulty parameter and whether to include the discrimination parameter in predictions.

4.4 Generalising Vocabulary Inventory Prediction

We now turn to a scenario in which all the data from the SVD12K dataset is used for training, equivalent to the Gen-None scenario, but the resulting item parameters are tested on external datasets. We test on the EVKD1 data set and TestYourVocab dataset introduced in Section 3.1. The results are given in Tables 4 & 5.

Since the EVKD1 set is a 4-way multiple choice test, we account for correct answers by guessing by using an item response curve with a guessing probability of 0.25, similar to the 3-parameter logistic IRT model:

$$ICC_{a_j,b_j}(\theta) = 0.25 + \frac{0.75}{1 + e^{-a_j(\theta_i - b_j)}}$$

Since there is a limited number of training words available in these datasets, in these experiments, no CAT is used, and instead the difficulty parameter is estimated based on 40 words taken at regular intervals from the frequency ranked list. There are three generalisation scenarios: *Freq*, where only frequency data is used; *Pred*, where only predictions from the generalisation model are used; and *Mix*, where item parameters are used directly from the IRT model fitted on SVD12K where possible, falling back to predictions when items available in SVD12K. Other variations are as in Section 4.3. For both datasets, frequency based difficulties outperform difficulties estimated from SVD12K, suggesting these do not generalise well to other datasets. The inclusion of the discrimination parameter appears to have a consistent small negative effect across all these experiments.

5 Discussion

We now summarise and discuss some of the main results of the experiments. Firstly, the discrimination parameter does not appear to help with query item selection, however it remains somewhat inconclusive whether it can help with estimating the learner ability θ since this was the best configuration in the Gen-None case. It may be that with sufficiently high quality estimates of the discrimination values, using this for θ-estimation would help more. The approach which appeared best overall in this case however, and which was used for later experiments on the SVD12K dataset ignored the discrimination parameter altogether for both steps of the CAT stage.

The difficulty parameter generalises reasonably well, while the discrimination parameter generalises quite poorly when regressed using a MLP based on Numberbatch representations of the word items. Since item difficulty here is closely related to frequency, it seems quite possible that a lot of the generalisation is happening based on frequency information encoded in the word embeddings. When considering how well both parameters generalised, we should note that only one type of word embedding and regressor was tried, and others may generalise this parameter better.

The regressed difficulties perform better than the frequency data on in-dataset data, while performing worse on out-of-dataset data. Given all datasets contained mostly Japanese learners of English, this suggests that both the IRT parameter and the MLP generalising may have over fitted on narrow attributes of the particular cohort of University of Tokyo students making up SVD12K. Conversely we see that that high quality, balanced word frequency data generalises rather well.

Usage of the discrimination parameter for vocabulary inventory prediction was largely inconclusive, with some evidence against it. In many cases, it appeared to decrease performance on metrics such as AUROC, however some tasks showed a promising but insignificant boost in AP-.

It is unclear exactly why the discrimination parameter failed to provide significant improvements in either next-item selection, θ-estimation or vocabulary inventory prediction. It is possible that the amount of response data was not sufficient either in terms of the number of respondents, or in terms of representing a diverse range of abilities, to obtain accurate word discrimination estimates. Apart from simply finding and integrating more vocabulary knowledge data, one direction for future work is trying to find corpus derived measures which correlate with word discrimination, analogously to the negative correlation between word frequency and word difficulty. This would also effectively address the failure to generalise the word discriminations parameter to out of vocabulary words.

We hope the methods of evaluating the different sub-tasks of the vocabulary inventory prediction task in the settings demonstrated here can help establish practices for evaluating this task more throughly. We also hope that the framing given here inspires others to tackle the problem in the challenging, but more broadly applicable setting of vocabulary inventory prediction having a small, limited number of queries.

The code to replicate all experiments is made available at `https://github.com/frankier/vocabirt`.

References

Drilon Avdiu, Vanessa Bui, and Klára Ptačinová Klimčíková. 2019. Predicting learner knowledge of individual words using machine learning. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 1–9, Turku, Finland. LiU Electronic Press.

F. Baker. 2001. *The basics of item response theory*, 2nd edition.

Luca Benedetto, Giovanni Aradelli, Paolo Cremonesi, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2021. On the application of transformers for estimating the difficulty of multiple-choice questions from text. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–157.

B. Carpenter, A. Gelman, M. Hoffman, D. Lee, Ben Goodrich, M. Betancourt, Marcus A. Brubaker, J. Guo, P. Li, and Allen B. Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1–32.

Douglas De Rizzo Meneghetti and Plinio Thomaz Aquino Junior. 2017. Application and Simulation of Computerized Adaptive Tests Through the Package catsim. *arXiv e-prints*, page arXiv:1707.03012.

B. G. Dodd. 1990. The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement*, 14:355 – 366.

Yo Ehara. 2018. Building an english vocabulary knowledge dataset of japanese english-as-a-second-language learners using crowdsourcing. In *Proceedings of LREC 2018*, Miyazaki, Japan.

Yo Ehara. 2019. Neural rasch model: How do word embeddings adjust word difficulty? In *PACLING*.

Yo Ehara, Yusuke Miyao, H. Oiwa, Issei Sato, and H. Nakagawa. 2014a. Formalizing word sampling for vocabulary prediction as graph-based active learning. In *EMNLP*.

Yo Ehara, Yusuke Miyao, Hidekazu Oiwa, Issei Sato, and Hiroshi Nakagawa. 2014b. Formalizing word sampling for vocabulary prediction as graph-based active learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1374–1384.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining words in the minds of second language learners: Learner-specific word difficulty. In *Proceedings of COLING 2012*, page 799814, Mumbai, India. The COLING 2012 Organizing Committee.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv: Learning*.

S. Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv*, abs/1502.03167.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

F. Lord. 1977. A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, 1:100 – 95.

James Milton. 2009. *Measuring second language vocabulary acquisition*.

Adam Paszke, S. Gross, Francisco Massa, A. Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Z. Lin, N. Gimelshein, L. Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach 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 *NeurIPS*.

Jeffrey Pennington, R. Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

George Rasch. 1960. Probabilistic models for some intelligence and attainment tests: Danish institute for educational research. *Denmark Paedogiska, Copenhagen*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. pages 4444–4451.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2.2.

M. Tatsuoka, F. Lord, M. R. Novick, and A. Birnbaum. 1968. Statistical theories of mental test scores. *Journal of the American Statistical Association*, 66:651.

V. W. Urry. 1970. *A Monte-Carlo investigation of logistic mental test models.* Ph.D. thesis, Purdue University.

H. Wainer. 2000. Computerized adaptive testing: A primer.