Optimal Ranking of Test Items using the Rasch Model

Divyanshu Vats
Sailthru Inc.

Andrew S. Lan
Rice University

Christoph Studer
Cornell University

Richard G. Baraniuk
Rice University

***Abstract*—We study the problem of ranking test items, i.e., the ordering of items according to the *amount of information* they provide on the latent trait of the respondents. We focus on educational applications, where instructors are interested in ranking questions so as to select a small set of informative questions in order to efficiently assess the students' understanding on the course material. Using the Rasch model for modeling student responses, we prove that the simple algorithm of sorting the item level parameters of the Rasch model is optimal in the setting where the goal is to maximize the entropy of the student responses. We demonstrate the optimality of the sorting algorithm using both theoretical results and using empirical results on several real-world datasets. Furthermore, we also demonstrate how the sorting algorithm can be used in a *batch adaptive* manner for predicting unobserved student responses.**

I. Introduction

Tests and surveys that consist of tens to hundreds of items remain to be among the most effective ways to probe the latent traits/opinions of respondents [1]. Depending on the application, the latent trait can refer to the ability of a student in answering questions in a test or refer to the psychological type of a human subject when participating in a survey. Over the past few decades, there have been several advances in designing tests and surveys that can be used to effectively measure the latent trait of a respondent [2]–[5]. A practical application of these computerized adaptive testing (CAT) techniques are seen in standardized tests such as the SAT, the GRE, and the GMAT.

Despite the prevalence of CATs and other test design strategies, there has been very limited theory in trying to understand the performance of various test design strategies. For example, consider the problem of trying to select q items from a test bank of Q items to design a test for N students in a class. If we assume that all students will receive the same test, a common item selection strategy is to select the items that are neither too difficult nor too hard. This strategy for item selection can be formalized using the Rasch model [6], where each item is associated with a difficulty parameter and the item selection strategy simply selects items that are close to the mean difficulty parameter. Although this simple algorithm is well known there is—to the best of our knowledge—no theoretical analysis in the literature that quantifies the performance of the algorithm.

Our main contribution in this paper is to quantify the performance of the simple algorithm that sorts the item difficulty parameters of the Rasch model to rank items. Our analysis reveals that the simple sorting algorithm, which we refer to as Rank-Rasch, is equivalent to the optimal algorithm (which is not possible to implement in practice) based on maximizing the entropy of student responses when the number of students N is large. We supplement the theoretical results with empirical results on real-world data. Our results have important consequences in online settings, such as Massive open online courses (MOOCs)[1], where several thousands of students take courses simultaneously on an on-line platform [7], [8]. Our theoretical results indicate that since a large number of students are taking a course, instructors can use Rank-Rasch to select items on a test. This strategy has potential to increase the efficiency of item selection when an instructor is managing a course with thousands of students. Although our analysis is limited to the setting where all students receive the same test, we also show that Rank-Rasch is compatible with adaptive testing by grouping items into batches and grouping students with similar abilities [9].

The rest of the paper is organized as follows. Section II reviews the Rasch model. Section III reviews the Rank-Rasch algorithm. Section IV presents our main results on the optimality of Rank-Rasch. Section V presents experimental results on real educational data. We conclude in Section VI.

II. The Rasch Model

We study the ranking of items in an educational setting, where items correspond to questions and the respondents correspond to students. To mathematically characterize the students' responses to items, we make use of the classical Rasch model [6]. We analyze the responses of N students to Q items. Our main objective is to rank the Q items so that an instructor can easily select the top subset of items for a test. The Rasch model relies on two sets of parameters:

- Difficulty parameters $\mu_1, \dots, \mu_Q$ associated with the Q items. If $\mu_i > \mu_{i'}$, then item i is considered to be more difficult than item i'.
- Ability parameters $a_1, \dots, a_N$ associated with the N students. If $a_j > a_{j'}$, then student j is considered to have a better chance of answering the items correctly when compared to student j'.

Let $Y_{i,j}$ be a binary-valued random variable that represents the correct/incorrect $(1/0)$ response of student j to item i.

[1] See, e.g., coursera.org, edX.org, and udacity.com.

Using the item difficulty parameter μ_i and the student ability parameter a_j, the response $Y_{i,j}$ is characterized as a Bernoulli random variable satisfying

$$P(Y_{i,j} = 1|\mu_i, a_j) = \frac{1}{1 + \exp(-(a_j - \mu_i))}\,. \quad (1)$$

The Rasch model not only allows for a predictive model for student responses, but also allows for an interpretable model to understand the item difficulties and the student abilities. Define the vectors $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_Q]^T$ and $\mathbf{a} = [a_1, \ldots, a_N]^T$. Under the Rasch model, we also have the following joint distribution:

$$P(\{Y_{i,j} = y_{i,j}, i = 1, \ldots, Q\}|\boldsymbol{\mu}, \mathbf{a}) = \prod_{i=1}^{Q} P(Y_{i,j} = y_{i,j}|\boldsymbol{\mu}, \mathbf{a})\,, \quad (2)$$

$$P(Y_{i,j} = y_{i,j}|\boldsymbol{\mu}, \mathbf{a}) = P(Y_{i,j} = y_{i,j}|a_i, \mu_j)\,, \quad (3)$$

where $y_{i,j}$ is the response observed from student j answering item i. In words, (2) implies that when conditioned on all the difficulty and ability parameters, the responses to the items are independent of each other for a given student. Equation (3) implies that when conditioned on all the difficulty and ability parameters, $Y_{i,j}$ only depends on the difficulty of item i and the ability of student j.

III. Ranking Items using the Rasch Model

In this section, we review the simple ranking algorithm based on the Rasch model. Given Q items, we aim to rank the items so that the items at the top of the list are more informative than the items at the bottom. Informally, an item i is more informative than item j if responding to item i provides for information about a student's ability parameter than responding to item j. See Section IV for a formal definition of item informativeness.

In addition to using the Rasch model in (1), we make the following two assumptions:

(A1) The difficulty parameters $\mu_1, \ldots, \mu_Q$ are known.

(A2) The ability parameters $a_1, \ldots, a_N$ are assumed to be independent and identically distributed samples drawn from a random variable with mean zero and probability density function $f(a)$, which is symmetric around 0, i.e., $f(a) = f(-a)$, and non-increasing for $a \geq 0$.

Assumption (A1) can be easily satisfied by using data obtained from prior offerings of a course. In particular, given some data on a given number of students answering the Q items, we can infer the difficulty parameters using standard maximum-likelihood estimation algorithms [10], [11]. Assumption (A2) makes the Rasch model identifiable and enables an analysis of the optimality of our ranking algorithm in Section IV. To rank items using the Rasch model, we simply sort the items based on the absolute value of the difficulty parameters:

Ranking Item using the Rasch Model (Rank-Rasch)
Find a ranking $i_1, i_2, \ldots, i_Q$ s.t. $|\mu_{i_1}| \leq \cdots \leq |\mu_{i_Q}|$; ties are borken arbitrarily.

In words, Rank-Rasch selects items whose difficulty parameters closely match the *mean ability parameters* over all the students in a course. Rank-Rasch is commonly used as a heuristic for initializing adaptive testing systems [2]. Despite the simplicity of Rank-Rasch, we show in Section IV that Rank-Rasch performs as well as an algorithm that uses information about the *unknown* ability parameter of each student in order to select items that minimize the error in estimating the ability parameters of all the students.

IV. Optimality of Rank-Rasch

In this section, we present our main result regarding the optimality of Rank-Rasch. Section IV-A presents an optimal algorithm, which is impossible to implement in practice, based on maximizing mean student entropy. Section IV-B presents our result on comparing the performance of the Rank-Rasch and the optimal algorithm.

A. Item Ranking using Entropy

In order to study the item-ranking problem, we use the notion of entropy of the student responses, as it is characterizes how informative a question is on estimating the latent abilities of the students. More specifically, the entropy of a binary random variable X is given by

$$H(X) = -p_0 \log(p_0) - (1 - p_0) \log(1 - p_0), \quad (4)$$

where $p_0 = p(X = 0)$ and $\log$ is the natural logarithm. Informally, the entropy measures the uncertainty of the random variable X. When ranking two items, it is preferable to select an item whose responses are more uncertain than the other across different students. The reason is that if all the N students submit their responses to the more uncertain item, then their responses to the remaining items can be better predicted than if the students submit their responses to the less uncertain item. In this way, we can say that the most uncertain item (i.e., the item with the highest entropy) is also the most informative item.

The entropy of an item i is the entropy of the random vector $\mathbf{y}_i = [y_{i,1}, \ldots, y_{i,N}]^T$, which corresponds to the responses submitted by the N students to the item i. Using the Rasch model in (1), the definition of entropy in (4), and the conditional independence relationship in (2), the entropy of item i can be written as

$$S_i = \sum_{j=1}^{N} H(Y_{i,j}; \mu_i, a_j)\,, \quad (5)$$

$$= \sum_{j=1}^{N} \left[\log(1 + e^{(a_j - \mu_i)}) - \frac{a_j - \mu_i}{1 + e^{-(a_j - \mu_i)}}\right], \quad (6)$$

where μ_i is the difficulty of item i, a_j is the ability of student j, and $H(Y_{i,j}|\mu_i, a_j)$ is the entropy of $Y_{i,j}$ given μ_i

and a_j. Recall that we want to rank items in such a way that the items with highest entropy are at the top of the ranking. Therefore, a natural criterion for item ranking is to sort E_i for $i = 1, \ldots, Q$. This leads to the following Optimal-Rank algorithm:

Optimal-Rank

Step 1. Compute S_i, defined in (5), for $i = 1, \ldots, Q$.
Step 2. Find a ranking $i_1, i_2, \ldots, i_Q$ such that $S_{i_1} \geq S_{i_2} \geq \cdots \geq S_{i_Q}$; ties are broken arbitrarily.

We note that Optimal-Rank cannot be implemented in practice as computing S_i requires knowledge of the *unknown* ability parameters of each student. On the other hand, Rank-Rasch only depends on the known difficulty parameter, and can hence easily be implemented in practice.

B. Main Theoretical Result

Having shown that the optimal item ranking algorithm, which selects items with the highest entropy first, is practically infeasible, we now characterize the performance of Rank-Rasch. For notational simplicity, we assume that the item difficulty parameters are sorted so that

$$|\mu_1| < \cdots < |\mu_Q| \,. \tag{7}$$

From (7) it is clear that Rank-Rasch will rank the items as $1, 2, \ldots, Q$. We assume that there are no ties in (7) so that Rank-Rasch identifies a unique ranking. Our main result is given as follows.

Theorem 1. *Suppose the responses of N students to the Q items are modeled using the Rasch model (1). Assume that the independence conditions in (2)–(3) holds, (A1)–(A2) hold with $m = 0$, the items are sorted as in (7), and Rank-Rasch outputs the ranking σ over items $1, 2, \ldots, Q$. Then, $\mathbb{P}(\sigma = \sigma^{opt}) \geq 1 - 2Q \exp(-N E_{\min}^2 / 2 \log 2)$, where*

$$E_{\min} = \min_{i=1,\ldots,Q-1} \mathbb{E}_a[H(Y_i; a, \mu_i) - H(Y_{i+1}; a, \mu_{i+1}))] \,, \tag{8}$$

where σ^{opt} is the output of Optimal-Rank, and the expectation $\mathbb{E}_a[\,\cdot\,]$ is with respect to the distribution of the ability parameters defined in (A2).

Theorem 1 characterizes the probability that the ranking computed by Rank-Rasch is equal to the ranking computed by Optimal-Rank. The proof of Theorem 1 is given in the Appendix. The main idea of the proof is to first use concentration results to approximate (5) and then to show that Assumption (A2) ensures that Rank-Rasch solves the approximated problem with high probability. Furthermore, it is clear that if Q and $E_{\min}$ are fixed as N increases, then $\lim_{N\to\infty} \mathbb{P}(\sigma = \sigma^{\text{opt}}) = 1$. We now make some additional remarks regarding Theorem 1.

Remark 1. A simple calculation shows that to guarantee that $\mathbb{P}(\sigma \neq \sigma^{\text{opt}}) \leq \epsilon$, the number of students N needs to be at least $\log(2Q\log(2)/\epsilon)/(2E_{\min}^2)$. Thus, $E_{\min}$ plays an important role in determining the performance of Rank-Rasch. In particular, a smaller value of $E_{\min}$ requires a larger number of students for optimal item selection using Rank-Rasch. Furthermore, since $E_{\min}$ can be numerically approximated, we can easily quantify how hard the problem of ranking items is for a particular set of items.

Remark 2. Figure 1 illustrates the dependence of the performance of Rank-Rasch on the problem parameters, namely the number of students N, the number of items Q, and the quantity $E_{\min}$ defined in (8). In particular, Figure 1 studies the probability that $\sigma = \sigma^{\text{opt}}$ as N, Q, and $E_{\min}$ are varied. The ability parameters are sampled from a standard normal distribution. Each pixel in the figure corresponds to the empirical probability of accurate item selection computed over 100 trials. Brighter regions correspond to the probability being close to one and the darker regions correspond to the probability being close zero. As a consequence of Theorem 1, smaller values of $E_{\min}$ require a larger number of students for accurate item selection.

Remark 3. The main assumption in Theorem 1 involves the distribution of the ability parameters in (A2). In particular, we assume that the density $f(a)$ of the ability parameters is symmetric around 0 and non-increasing for $a \geq 0$. This includes several common unimodal distributions including the Gaussian, Laplacian, and uniform distributions.

Remark 4. We note that our optimality criterion was to maximize the entropy of the responses. Alternatively, we could have used other metrics to define the notion of an optimal ranking. For example, metrics based on the Fisher information matrix are popular in the adaptive testing literature [2], [3], [5]. Although we do not show the computations here, under similar conditions as in Theorem 1, Rank-Rasch is also nearly optimal under the criterion for maximizing the Fisher information.

V. Real Educational Data Experiments

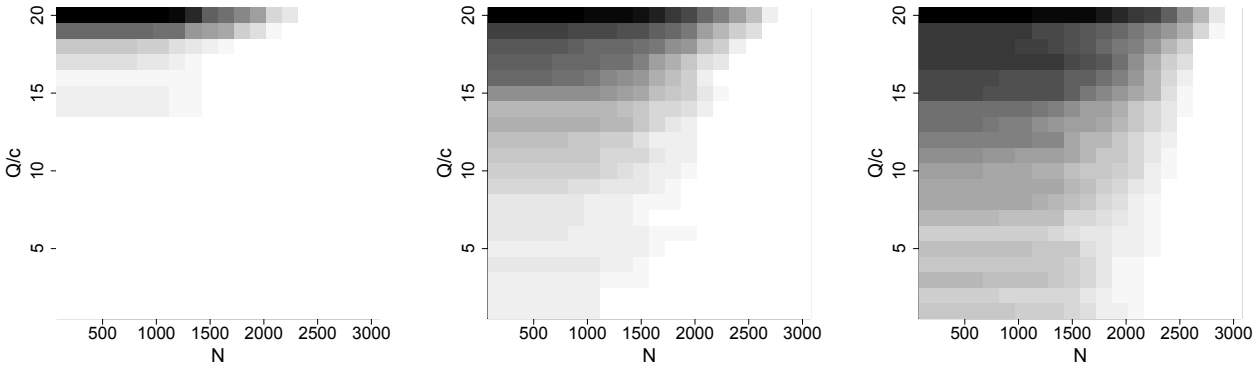
In this section, we verify the optimality of Rank-Rasch using real educational datasets. Section V-A describes the datasets. Section V-B presents empirical results for Rank-Rasch. Section V-C presents empirical results of a batch-adaptive version of Rank-Rasch.

A. Datasets

We use six educational datasets that consist of binary-valued (correct/incorrect) graded student responses to illustrate the performance of Rank-Rasch. Table 1 summarizes these datasets. A brief description of the datasets is as follows:

MT: A dataset from a high-school algebra test conducted on Amazon's Mechanical Turk.

UD_1 and UD_2: Two datasets from a high school admission test that consists of questions from physics, chemistry, mathematics, and biology. The test is scored so that that students receive $+3$ points for a correct response



(a) $c = 1$, $E_{\min} = 0.03$. (b) $c = 2$, $E_{\min} = 0.015$. (c) $c = 3$, $E_{\min} = 0.01$.

Fig. 1. Empirical probability illustrating the space of problem parameters over which the item ranking is optimal for (a) $Q = 10$ items, (b) $Q = 20$ items, and (c) $Q = 30$ items. Brighter (darker) regions indicate that the probability of the item ranking being optimal is close to one (zero).

TABLE I
DESCRIPTION OF THE EDUCATIONAL DATASETS. N IS THE NUMBER OF STUDENTS, Q IS THE NUMBER OF ITEMS, AND % MISSING IS PERCENTAGE OF MISSING DATA FROM THE NQ DATA POINTS

Dataset	N	Q	% missing
MT	99	34	0.00
UD_1	1714	60	39.3
UD_2	1567	60	29.3
CT_1	53	82	21.3
CT_2	97	203	0.00
ET	3193	438	93.3

(mapped to $y_{i,j} = 1$), -1 points for an incorrect response (mapped to $y_{i,j} = 0$), and 0 points for not providing any response (mapped to $y_{i,j}$ being unobserved).

CT_1 and CT_2: Datasets from a university course on introduction to signal processing and another course on introduction to computer engineering.

ET: A dataset from a testing company which gives questions to programmers and then grades their responses to assess their employability for software engineering jobs.

For all the datasets, we used all the items to learn the Rasch model parameters.

B. Performance of Rank-Rasch

We now evaluate the performance of Rank-Rasch. We denote each dataset by a $Q \times N'$ matrix $\mathbf{Y}$. We split the dataset into a training set $\mathbf{Y}^{\text{train}}$ and a testing set $\mathbf{Y}^{\text{test}}$, where $\mathbf{Y}^{\text{test}}$ is of size $Q \times N$ and $N = \lfloor N'/2 \rfloor$. We treat $\mathbf{Y}^{\text{train}}$ as data from a prior offering of a course and use this data to estimate the difficulty parameters of the items and the probability distribution of the students' ability parameters $f(a)$.

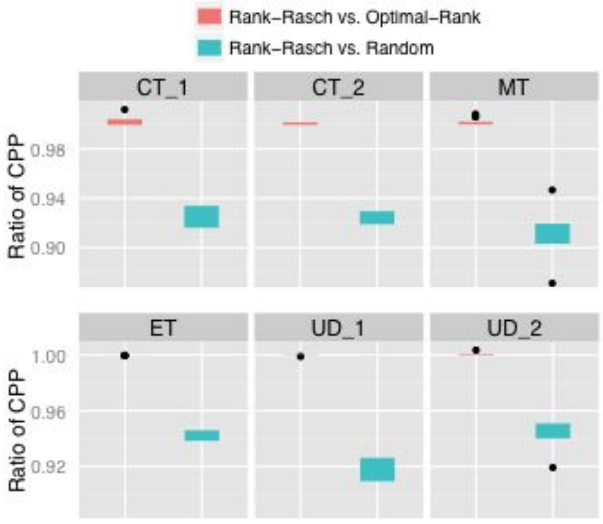


Fig. 2. Boxplot showing the ratios of the cumulative prediction performance (CPP) between different algorithms for six different education datasets computed using 25 different splits of the data. The ratios correspond to the Optimal-Rank CPP over the Rank-Rasch CPP and the Random CPP over the Rank-Rasch CPP, respectively. We clearly see that (i) Rank-Rasch outperforms randomly ranking the items, and (ii) Rank-Rasch has nearly the same performance as Optimal-Rank, which is infeasible in practice.

We evaluate the performance of ranking on $\mathbf{Y}^{\text{test}}$ by computing the cumulative prediction performance (CPP). In particular, if $\mathcal{I}_q$ represents the top q items in a ranking, we evaluate the prediction performance on the items $\mathcal{I}_q^c$ using the responses to $\mathcal{I}_q$ for each $q = 1, \ldots, Q-1$. In this way, we compute the CPP score as follows:

$$\text{CPP} = \frac{1}{N(Q-1)} \sum_{\ell=1}^{Q-1} \frac{1}{Q-\ell} \sum_{i=\ell+1}^{Q} \sum_{j=1}^{N} I(y_{i,j}^{\text{test}} = \widehat{y}_{i,j}^{\text{test}}),$$

$$\widehat{y}_{i,j}^{\text{test}} = I\left(\frac{1}{1 + \exp(-(\widehat{a}_j^i - \mu_i))} > 0.5 \right), \tag{9}$$

where $I(x = y) = 1$ if $x = y$ and 0 otherwise, $I(x > y) = 1$ if $x > y$ and 0 otherwise, and $\widehat{a}_j^i$ is the ability parameter of student j computed using the responses from the top i responses in a ranking. Note that we have assumed that the ranking is $1, 2, \ldots, Q$ to compute (9). Informally, the CPP measures how well the students' responses to items at the bottom of the ranking can be predicted from their responses to the items at the top of the ranking. It is clear that $0 \leq$ CPP ≤ 1 and a larger CPP corresponds to a better ranking.

Since we have shown that Rank-Rasch is nearly optimal for large enough N in Theorem 1, we compare Rank-Rasch to the optimal ranking algorithm Optimal-Rank outlined in Section IV-A. However, recall that Optimal-Rank is impractical as it requires knowledge about the unknown ability parameters of each student. To approximate Optimal-Rank, we estimate the ability parameter for each student using $\mathbf{Y}^{\text{test}}$. In addition to Optimal-Rank, we also compare Rank-Rasch to an algorithm that randomly ranks the Q items.

Figure 3 shows the boxplot of the ratios of CPPs between different algorithms over 25 random splits of the six datasets. In the first plot, we show the ratio between the Optimal-Rank CPP and the Rank-Rasch CPP. In the second plot, we show the ratio between the Random CPP and the Rank-Rasch CPP.

Remark 5. As expected by Theorem 1, Rank-Rasch has similar performance to that of Optimal-Rank, with Optimal-Rank performing slightly better in some instances. This is seen by observing that the ratio of the Optimal-Rank CPP and Rank-Rasch CPP mostly lies in the interval $[1, 1 + \epsilon]$, with ϵ being at most 0.013. Furthermore, Rank-Rasch performs significantly better than the Random algorithm. This result verifies that Rank-Rasch is very effective in ranking test items.

Remark 6. We observe that the variation in the difference between Rank-Rasch and Optimal-Rank is much smaller for the larger datasets (ET, UD_1, and UD_2) when compared to the smaller datasets (CT_1, CT_2, and MT). This suggests that Rank-Rasch is better suited in settings where the ranking of items needs to be performed for a large number of students.

C. Adaptive Version of Rank-Rasch

In this section, we show that Rank-Rasch can be applied adaptively in real-world testing scenarios, although the item ranking is produced in a non-adaptive manner. The key idea is to apply Rank-Rasch to *batches* of items and *groups* of students, and iteratively update the item rankings and student groupings after each batch of questions.

As described in Section V-B, we split the dataset into a training set $\mathbf{Y}^{\text{train}}$ to estimate the probability distribution of the students' ability parameters $f(a)$, and a testing set $\mathbf{Y}^{\text{test}}$ to test the performance of Rank-Rasch. Using the estimate of $f(a)$, we generate an initial ranking of the items. Next, we select the first batch of items to be the ones at the top of the initial ranking using the Rank-Rasch algorithm. Once students respond to these items, we obtain estimates of each student's ability parameter using their responses to these items in the testing set, and group them by putting students with similar ability parameters into the same group. Next, we update the item rankings for each group of students individually using the estimates of the ability parameters of students in that group, and select the next batch of items for each group using these updated rankings. We repeat the above process until we have used all the items in the testing set.

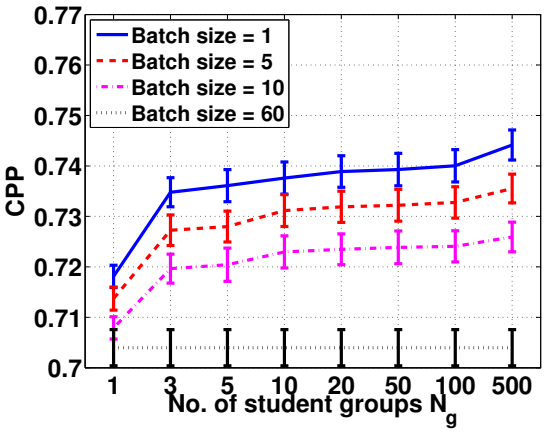


Fig. 3. Plot of the cumulative prediction performance (CPP) versus the number of student groups N_g with different item batch sizes for the batch-adaptive Rank-Rasch approach on the UD_2 dataset. The prediction performance increases as N_g increases and batch size decreases. Note that the curve with batch size $= 60$ corresponds to applying Rank-Rasch in a fully non-adaptive manner, while the point with $N_g = 500$ on the curve with batch size $= 1$ corresponds to applying Rank-Rasch in a fully adaptive manner.

Figure 4 shows the CPP versus the number of student groups $N_g \in \{1, 3, 5, 10, 20, 50, 100, 500\}$ for various values of batch sizes over 25 random splits of the UD_2 dataset.

Remark 7. We observe that smaller batch sizes and larger number of student groups lead to higher CPP values (better prediction performance on the unobserved student responses). Note that in Figure 4, the curve with batch size equal to 60 corresponds to applying Rank-Rasch in a fully non-adaptive manner, while the data point with the number of student groups equal to 500 on the curve with batch size equal to 1 corresponds to applying Rank-Rasch in a fully adaptive manner [3]–[5], as in every batch one new item is selected and each student belongs to their own group.

Remark 8. Compared to the Rank-Rasch algorithm applied in a fully non-adaptive manner, Rank-Rasch applied in a batch-adaptive manner utilizes the differences between groups of students (weak, average and strong) to obtain a better estimate of the students' ability parameter distribution $f(a)$ within each group. Thus, this approach leads to better prediction performance on the unobserved student responses. Ideally, a fully adaptive testing approach results in the best prediction performance. However, in some real-world educational scenarios, adaptivity is forbidden due to either fairness concerns or technology limitations. In these scenarios, applying Rank-Rasch in either a fully non-adaptive manner

(e.g., in traditional classrooms) or a batch-adaptive manner (e.g., in today's MOOCs) will not only meet these constraints but also provide comparable prediction performance to the fully adaptive approach, as demonstrated by the experimental results above.

VI. Conclusion

We have shown that the simple algorithm of sorting item level parameters of the Rasch model is theoretically optimal when ranking items for a large number of students. We have validated our results using both theory and experiments on real-world educational datasets with over 3000 students. Moreover, we show how the sorting algorithm can be applied in a batch-adaptive manner, connecting the algorithm to many existing fully adaptive item selection algorithms [3], [5].

The results in this paper motivate simple algorithms for ranking items in a Massive open online course (MOOC). Our future work will involve working closely with the instructor of a MOOC and design an experiment to showcase the effectiveness of using the sorting algorithm vs using standard policy by an instructor for selecting test items.

Appendix

Recall from (A2) that $a_1, \ldots, a_N$ are i.i.d. samples of a random variable. Thus, for a fixed i, the entropy $E(a_i, \mu_j) = H(Y_{i,j}|a_i, \mu_j)$, for $j = 1, \ldots, N$, are also i.i.d. samples of a random variable. Furthermore, it is easy to see that $0 \leq E(a_i, \mu_j) \leq \log(2)$. Using Hoeffding's inequality [12], we have

$$\mathbb{P}\left(\left|\overline{E}(\mu_i) - \mathbb{E}_a[E(a, \mu_i)]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2N\epsilon^2}{\log 2}\right),$$

where $\overline{E}(\mu_i) = \frac{1}{N}\sum_{j=1}^{N} E(a_i, \mu_j)$. According to Lemma 1, we have

$$\mathbb{E}_a[E(a, \mu_1)] < \cdots < \mathbb{E}_a[E(a, \mu_Q)], \tag{10}$$

if the questions are sorted as in (7) and Assumption (A2) holds. Thus, we see that the ranking σ obtained using Rank-Rasch can be interpreted as sorting $\mathbb{E}_a[E(a, \mu_i)]$ for $i = 1, \ldots, Q$.

Define the event $\mathcal{E}_i = \left|\overline{E}(\mu_i) - \mathbb{E}_a[E(a, \mu_i)]\right| < E_{\min}/2$, where $E_{\min}$ is defined in (8). A sufficient condition for $\sigma = \sigma^{\text{opt}}$ is that that the events $\mathcal{E}_1, \ldots, \mathcal{E}_Q$ all hold. Therefore, we have

$$\begin{aligned}\mathbb{P}(\sigma = \sigma^{\text{opt}}) &\geq \mathbb{P}\left(\bigcap_{i=1}^{Q} \mathcal{E}_i\right) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{Q} \mathcal{E}_i^c\right) \\ &\geq 1 - 2Q\exp(-NE_{\min}^2/(\log 2)), \end{aligned} \tag{11}$$

where we used the union bound and (10) with $\epsilon = E_{\min}/2$ to get (11).

Define $E(a, \mu) = \log(1 + e^{(a-\mu)}) - \dfrac{(a-\mu)}{1 + e^{-(a-\mu)}}$. We have the following lemma.

Lemma 1. *Under the assumptions in Theorem 1, $g(\mu) = \mathbb{E}_a[E(a, \mu)]$ is symmetric and non-increasing for $\mu \geq 0$.*

Proof. We first prove symmetry. Using the definition of $g(\mu)$, we have

$$\begin{aligned} g(-\mu) &= \int_{-\infty}^{\infty} f(a)E(a, -\mu)\mathrm{d}a \\ &= \int_{-\infty}^{\infty} f(-a)E(-a, -\mu)\mathrm{d}a \\ &= \int_{-\infty}^{\infty} f(a)\left(\log(1 + e^{(-a+\mu)}) - \frac{(-a+\mu)}{1 + e^{(a-\mu)}}\right)\mathrm{d}a \\ &= \int_{-\infty}^{\infty} f(a)\left(\log\frac{(1 + e^{(-a+\mu)})e^{-a+\mu}}{e^{-a+\mu}} - \frac{(-a+\mu)}{1 + e^{(a-\mu)}}\right)\mathrm{d}a \\ &= g(\mu), \end{aligned}$$

where the last step follows by simple algebra. Next, we prove that $g(\mu)$ is non-increasing for $\mu \geq 0$. It is sufficient to show that $g'(\mu) \leq 0$ whenever $\mu \geq 0$. If $\mu \geq 0$, then

$$\begin{aligned} g'(\mu) &= \int_{-\infty}^{\infty} f(a)E'(a, \mu)\mathrm{d}a \\ &= \int_{-\infty}^{\infty} f(a)\frac{(a-\mu)}{(1 + e^{-(a-\mu)})(1 + e^{(a-\mu)})}\mathrm{d}a. \end{aligned}$$

Let $b = a - \mu$. We have,

$$\begin{aligned} g'(\mu) &= \int_{-\infty}^{\infty} f(b+\mu)\frac{b}{(1 + e^{-b})(1 + e^{b})}\mathrm{d}b \\ &= \int_{-\infty}^{0} f(b+\mu)\frac{b}{(1 + e^{-b})(1 + e^{b})}\mathrm{d}b \\ &\quad + \int_{0}^{\infty} f(b+\mu)\frac{b}{(1 + e^{-b})(1 + e^{b})}\mathrm{d}b \\ &= \int_{\infty}^{0} f(-b+\mu)\frac{-b}{(1 + e^{-b})(1 + e^{b})}\mathrm{d}(-b) \\ &\quad + \int_{0}^{\infty} f(b+\mu)\frac{b}{(1 + e^{-b})(1 + e^{b})}\mathrm{d}b \\ &= \int_{0}^{\infty} f(-b+\mu)\frac{-b}{(1 + e^{-b})(1 + e^{b})}\mathrm{d}b \\ &\quad + \int_{0}^{\infty} f(b+\mu)\frac{b}{(1 + e^{-b})(1 + e^{b})}\mathrm{d}b \\ &= \int_{0}^{\infty} \frac{b}{(1 + e^{-b})(1 + e^{b})}(f(b+\mu) - f(-b+\mu))\mathrm{d}b \\ &= \int_{0}^{\infty} \frac{b}{(1 + e^{-b})(1 + e^{b})}(f(b+\mu) - f(|b-\mu|))\mathrm{d}b \\ &\leq 0, \end{aligned}$$

where we have used the assumption that $f(a)$ is symmetric and non-increasing for $a \geq 0$ and the fact that $b + \mu \geq |b - \mu| \geq 0$ when $b > 0$ and $\mu \geq 0$. □

References

[1] H. Gulliksen, *Theory of Mental Tests*. Wiley New York, 1950.

[2] H. Chang and Z. Ying, "A global information approach to computerized adaptive testing," *Applied Psychological Measurement*, vol. 20, no. 3, pp. 213–229, Sep. 1996.

[3] ——, "Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests," *The Annals of Statistics*, vol. 37, no. 3, pp. 1466–1488, June 2009.

[4] W. J. van der Linden and P. J. Pashley, "Item selection and ability estimation in adaptive testing," *Elements of Adaptive Testing*, pp. 3–30, Jan. 2010.
[5] W. J. van der Linden and C. A. W. Glas, *Computerized Adaptive Testing: Theory and Practice*. Springer Netherlands, 2000.
[6] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*. MESA Press, 1993.
[7] J. Champaign, K. F. Colvin, A. Liu, C. Fredericks, D. Seaton, and D. E. Pritchard, "Correlating skill and improvement in 2 MOOCs with a student's time on tasks," in *Proc. 1st ACM Conf. on Learning at Scale*, Mar. 2014, pp. 11–20.
[8] F. G. Martin, "Will massive open online courses change how we teach?" *Communications of the ACM*, vol. 55, no. 8, pp. 26–28, Aug. 2012.
[9] R. Agrawal, B. Golshan, and E. Terzi, "Grouping students in educational settings," in *Proc. 20th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, Aug. 2014, pp. 1017–1026.
[10] F. B. Baker and S. Kim, *Item Response Theory: Parameter Estimation Techniques*. CRC Press, 2004.
[11] R. P. Chalmers, "MIRT: A multidimensional item response theory package for the R environment," *Journal of Statistical Software*, vol. 48, no. 6, pp. 1–29, May 2012.
[12] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.