Peer-Mediated Testing

Igor Labutov
Cornell University
Ithaca, NY
iil4@cornell.edu

Kelvin Luu
Cornell University
Ithaca, NY
kl583@cornell.edu

Thorsten Joachims
Cornell University
Ithaca, NY
tj@cs.cornell.edu

Hod Lipson
Cornell University
Ithaca, NY
hod.lipson@cornell.edu

ABSTRACT

With the growing interest in large scale on-line education, fueled in part by the recent emergence of MOOCs (Massively Open Online Courses), comes an important problem of assessing competency of (typically many) learners. We propose a novel methodology of *peer-mediated testing*, where students participate in the roles of both a student and a teacher, under complete autonomy. To scale assessment-generation to the web, we propose a joint framework for crowdsourcing both the *assessment material* (in the form of a quiz), and the *assessment* (in the form of ranking) of the participants. We evaluate the effectiveness of the proposed model via a user-study in which participants are both — question "generators" and question "answerers", and compare the quality of generated ranking with the rankings obtained through assessment with expert(teacher)-curated questions. The central question that we address in our work is: *"Can students generate accurate assessment of their knowledge under complete autonomy?"*

1. INTRODUCTION

While the transmission of teaching material has benefited significantly from the digital medium, assessment methodology has changed little from an age-old tradition of instructor-generated and instructor-graded tests. Recently, the issue of scaling the grading of such tests to MOOC-scale has received attention from a computational domain, in the form of optimized peer-assisted grading [19, 17, 13, 29] and methods that assist teachers in grading effectively [5, 22]. While grading plays an integral role in any form of assessment, the generation of assessment material itself, i.e. tests, presents an equally important challenge for addressing the scaling of assessment methods. Moreover, while curated courses, such as those offered by MOOCs like Coursera, offer high-quality course-ware, their throughput is fundamentally limited by the number of interested and willing instructors who face a significant commitment in producing even a single course. Meanwhile, technical documentation in the form of heterogeneous on-line tutorials, e-books, lecture notes, video lectures are growing on the web, and play an increasing role as both supplemental and primary sources in personalized, individual learning. Unfortunately few of these sources come with assessment material. If available, assessment quizzes, would allow the learner to self-reflect on the areas in which he or she is lacking, and help provide feedback to guide the learner towards additional material. An assessment mechanism would also facilitate ranking of the learners on their depth of understanding of the material, similar to the "top-scorer" list in an arcade game.

Given an online setting of incoming learners (participants), where each learner is presented with a focused selection of reading material, our task is 1) present the learner with an optimal set of questions from a question-bank generated by other participants, 2) request the learner to generate one or more questions and 3) update the model of the questions and learners. The model of the learner refers to the parameters that capture the intrinsic "ability" of the learner. The model of the question refers to the parameters that capture the "quality" and "difficulty" of the question. Under such parametrization, we seek to generate quizzes for each participant, such as to most effectively inform their ranking, i.e. to minimize the number of trials required for generating an accurate ranking. Details of the proposed approach are outlined in the next section. The central questions that our work aims to answer include:

- **Can students generate an accurate assessment of their knowledge under complete autonomy?** Fidelity of the ranking is measured with with respect to gold set of rankings, obtained from a set of teacher-curated questions for the presented material.

- **What defines a good question?** Specifically, what does the ranking of the question "generators" and question "answerers" say about the quality of a question, i.e the extent to which a question predicts a consistent ranking.

- **What is the most effective way to facilitate learners to generate "good questions"?** Is it better to let the learners "play the teacher" and generate questions explicitly for other learners? Or is it more effective to request learners to ask genuine questions to

.

which they do not know the answers, and thus contribute questions implicitly?

2. PEER-MEDIATED TESTING

We formalize the proposed problem of *peer-mediated testing* with three subtasks:

- Learning correct answers for student-generated questions
- Learning ranking of students
- Optimal test generation

The subtasks are illustrated with the following scenario. Consider a user arriving to a Wikipedia page augmented with the testing model proposed in this work. After reading an article on the page, the user is then presented with a test consisting of 10 multiple choice questions. Each question, in addition to a set of answers, contains the "None of the above" (NOA) option. If the user chooses the NOA option, she is then offered an opportunity to provide an additional answer through the means of typing it in directly. Following the test, the user is then asked if she can contribute an additional question that could be used to make the test better [1]. A score and rank (amongst all other users) is then provided to the user as feedback (either immediately or with some delay).

The following elements of the scenario correspond to the subtasks outlined above: 1) from the potentially large set of user-provided "free-response" answers for any given question, find the "most correct" and "least correct" answers, 2) from the user's selections and free-response contributions, find an optimal rank of the user among other participating users (who may not have seen an identical test), and 3) discover an optimal subset of questions (constrained by the total number of questions), and for each question an optimal set of 4 answers that will be most informative in inferring an updated ranking of the users.

3. RELATED WORK

3.1 One-dimensional assessment

Psychologists have found that ability on various cognitive tasks is correlated, and can be reduced to a single dimension [2, 3]. Psychometric tools for adapting tests to the single "aptitude" parameter have been developed, based on the "item-response theory" [12, 3], and applied to adaptive tests such as the GRE. We apply this notion in our work; we will focus exclusively on a form of assessment that is one-dimensional, i.e. a single parameter summarizes the general ability of the user.

3.2 Multi-dimensional assessment

While our model is intended for the purposes of ranking students on a single ability scale, for the purpose of assigning a grade, for example, the task of general assessment of intelligence is more complicated. Robert Sternberg's theory of successful intelligence [25] posits multiple dimensions of intelligence, each of which is integral to the success in every-day life. Assessment methodology relying on Sternberg's theory have been discussed in [24, 23]. A computational approach to assessment (and generation of assessment material) from a multidimensional perspective has recently been proposed by [27], and offers an adaptive approach for assessing the learner's knowledge of set of concepts (concepts corresponding to dimensions). Recently, techniques for providing feedback in large-scale classrooms [5, 22] have been proposed, and constitute an important part of the assessment pipeline.

3.3 Probabilistic models for ranking

A significant amount of work had been done in developing principled methods for learning and updating distributions over rankings (for applications in and outside the scope of grading), i.e. [10, 18, 19, 30], primarily with an application in the domains of search engines and online multiplayer competitions [14], [3, 17] do not explicitly concern with rankings, but apply a principled method, based on graphical models, for estimating a student's performance and skill level on multiple choice tests (where the tests are all fixed, and curated by a teacher). Our approach employs a probabilistic model, but seeks to incorporate the dynamic process of question generation and allocation in a principled manner. Additionally, while [3, 17] focuses primarily on grading the learners with respect to a fixed test, we seek to directly obtain a global ranking of the learners. While a ranking can be obtained on a fixed test, a large pool of learner-generated questions means that no two learners are likely to take the same exact test (same set of questions) — giving no meaningful interpretation to individual test scores, yet still provide a valid global ranking of learners. Additionally, for space reasons, we do no review here the extensive and relevant literature on optimal experiment design [11].

3.4 Peer-led learning

Assessment, while an integral part of learning [21], is not unique in its potential to be mediated entirely by the students. While student-generated questions have been shown [7, 15, 6] to be a valuable pedagogical resource for enhancing learning, peer-led learning in general has seen wide success across a number of disciplines in the "wild", specifically in the form of the Peer-led Team Learning (PLTL) program adopted in a set of universities nation-wide (U.S) [26, 8], where student peers participate as both students, and "teachers". Our work on peer-mediated assessment forms, to the best of our knowledge, the first computational approach to harnessing the *autonomy* of students for enhancing a part of the learning experience.

4. APPROACH

We approach the task of ranking answers and ranking users jointly. Given an indefinite and possibly growing set of answers for each question, we seek to maintain a consistent ranking of the answers as new users both click on the existing answers and provide new answers to old questions. From a stream of user clicks and original "free-response" contributions, we seek to update the ranking of the answers and the ranking of the users simultaneously.

[1] note that in our experiments, a question was requested from the user before presenting existing questions, for reasons described later

5. MODELS

We present two models that perform joint user/answer ranking. Both models are fully Bayesian, maintaining priors on both rankings, and updating them according to the clicks that each answer receives.

At any given point, the state of the system consists of the questions $q_i \in Q$, students s_i, and a set of answers $\{a_{ij}\}_{q_i}$ for each question. Samples from the ranking distributions over rankings of question and answers are obtained implicitly, by sorting on the hidden "ability" α_i parameter of each user, and the "correctness" parameter β_{ij} of each answer a_{ij}. Additionally, we maintain a "difficulty" parameter γ_{q_i} for each question $q_i \in Q$. The observed variables in the models are the clicks $z_{a_{ij}}^{s_k} \in \{clicked, \neg clicked\}$ that an answer a_{ij} receives from a user s_k. We aim to infer the posteriors over α_i, $\{\beta_{ij}\}_{a_{ij}}$ and γ_{q_i}.

5.1 Model 1

We denote the set of answers $\{a_{ij}\} \in O(s_k)$ as answers originally contributed to by user s_k ("free-response" answers), and $\{a_{ij}\} \in C(s_k)$ as answers observed by the user s_k, i.e. answers that appeared as choices in the multiple choice selection of questions that the user s_k was presented during her session. The generative process of Model 1 is as follows:

- **For each** question $q_j \in Q$:
  - Draw $\gamma_j \sim \mathcal{N}(0, 1)$
- **For each** student $s_k \in S$:
  - Draw $\alpha_k \sim \mathcal{N}(0, 1)$
  - **For each** answer $a_{ij} \in O(s_k)$:
    - Draw $\beta_{ij} \sim \mathcal{N}(\alpha_k, 1)$
  - **For each** answer $a_{ij} \in C(s_k)$:
    - Draw $z_{ij}^k \sim$ Bernoulli(θ) where $\theta = P(a_{ij} > \gamma_{q_i})$

Intuitively, the generative process can be summarized as follows: user's abilities and question difficulties are normally distributed about some mean that is irrelevant. Then 1) when contributing original answers to questions, each user contributes an answer proportional to her ability, and 2) when deciding whether to click on an answer, the choice is made proportional to the probability that the correctness of the answer exceeds the difficulty level of the question (i.e. the correctness threshold). This captures the intuition that easier questions are likely to receive more correct answers. This intuition is also followed by [3].

The likelihood in this model is based on pairwise comparison, akin to ELO [9], Glicko [10] and TrueSkill [14] models. In this model, the pairwise comparisons are between the user-contributed answers and the question's correctness threshold. These comparisons may also be between answer pairs (i.e. clicking on user A's answer, and not clicking on user B's answer can be interpreted as A's answer is ranked higher than B's answer), although this model is not investigated in this paper. Note that, unlike in [3], we allow each user to select any number of correct answers, including NOA, which is more natural in the case of user-generated answers, where there may be multiple contributed answers that are correct.

An important missing component in the model is the absence of any dependence of question-specific parameters on the users that generated them. These dependencies are being explored as part of ongoing work.

5.2 Model 2

Model 2 incorporates the user's ability α_i as variance into the likelihood of the click, with the intuition that preferences expressed by more able users are likely to be more certain. For the convenience in developing the inference algorithm, we employ the logistic approximation to the cumulative normal function that parametrizes θ in the model above:

$$P(z_{ij}^k = click) = \frac{\exp(f(\alpha_k)(\beta_{ij} - \gamma_i))}{1 + \exp(f(\alpha_k)(\beta_{ij} - \gamma_i))} \quad (1)$$

where we choose $f(\alpha_k) = \exp(\alpha_k)$, as a monotonic transformation of the user's ability that forces positivity on the variance parameter. Under this parametrization, users with very low ability will click with probability .5, regardless of the question and answer parameters. This form of parametrization is similar to [28]. We note that in our experiments, Model 2 performs at least as well as (and in most cases outperforms) Model 1. Henceforth in the paper, we will refer to Model 2 as The Model.

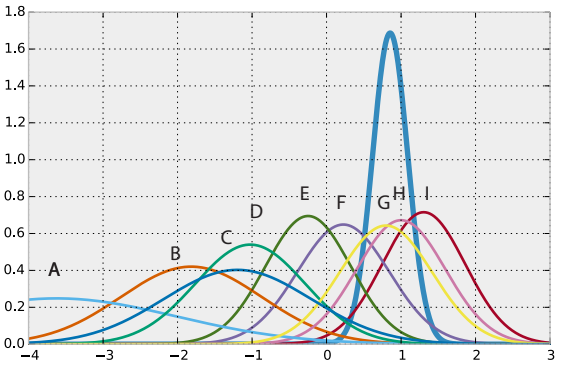
5.3 Inference

We employ the Variational Message Passing (VMP) [31] algorithm for performing inference in these models. VMP provides a general method for performing variational inference in conjugate-exponential models by passing sufficient statistics of the variables to the neighbors, which are used in turn to update their natural parameters. For the only non-conjugate term (likelihood), we perform a Metropolis-Hastings sampling subroutine that computes the moments to be used in the remainder of the message passing routine. For Model 1 inference, Infer.NET [16] was used as well, as it also provided the logistic factor, but with a variational approximation that proved to be much more efficient. Figure 1 illustrates the result of inference for one question, and a set of answers for that question, from the dataset collected using Mechanical Turk (described in the next section). In addition to each a_{ij}, which may be interpreted as the "obviousness of correctness" (a larger negative value corresponds to "more obviously wrong", and a more positive value corresponds to "more obviously correct"), the difficulty of the question q_i is embedded on the same scale. Recall that in the generative definition of our models, the "correctness" of an answer a_{ij} is judged with respect to the difficulty of the question q_i, which can be interpreted as the "threshold of correctness". Answers that lie to the right of the "threshold" can be interpreted as correct, and answers to the left as incorrect.

6. OPTIMAL TEST DESIGN

Test generation, i.e. selection of a subset of questions and answers (task 3), as outlined in the introduction, is an integral component of a system that relies on a constant stream of new questions and answers. Selecting a subset of the generated questions and answers is important for two reasons: 1) to filter "spam" questions and answers for the purpose of presenting users with only meaningful questions, and 2) reducing the size of the test to a manageable number

Question: What are the two mechanisms for genetic diversity in a population and how do they differ?



Answers

A crossing over and independent assortment are the two mechanisms for genetic diversity in a population

B mutation and sexual reproduction,mutation is a change in DNA,

C mutation is a change in DNA and sexual reproduction which takes DNA from the parents to form a new type

D Natural selection and sexual selection

E phenotypic and genotypic

F evolution

G Genetic diversity in a population comes from two main mechanisms: mutation and sexual reproduction. Mutation, a change in DNA, is the ultimate source

H Genetic diversity in a population comes from two main mechanisms: mutation and sexual reproduction.

I Mutation and sexual reproduction

Figure 1: Embedding of answers after inference. Thick blue curve indicates the q_i variable, which can be interpreted as a "correctness threshold".

of questions, while preserving the "discriminative" ability for the purpose of ranking students. In this section we pose a criterion for selecting an optimal subset of answers for each question that addresses both of these goals. Furthermore, an additional constraint on the algorithm for selecting the subset of questions and answers to be included in the test, is an effective exploration of the streaming online submissions.

6.1 Approach

For the task of selecting the most discriminating (for the purpose of user ranking) set of answers, we employ the information gain on the user's s_k ability parameter α_k as the objective for answer selection. In the framework of optimal Bayesian experiment design [20], the task is to select a subset of answers that maximize the expected information gain on α_k for a typical user. Expected information gain $\mathbb{E}[IG_{ij}(\alpha_k)]$ for the user s_k, for the answer a_{ij}, can be expressed as follows:

$$\sum_{v\in\{click,\neg click\}} P(z_{ij}^k = v|\alpha_k) D_{KL}\left(P(\alpha_k)||P(\alpha_k|z_{ij}^k = v)\right) \tag{2}$$

where the expectation is taken over the choices of the user when presented with the answer a_{ij}, i.e. whether to click or skip the answer, and D_{KL} is the Kullback-Leibler divergence, between the prior and the posterior over the ability of the user s_k in the above expression. For the purpose of designing a True/False question, for example, the optimal answer a_{ij}^* is simply the answer with the greatest expected information gain. For presenting a two-choice multiple choice question, we follow the following heuristic: select one "correct" answer (the answer whose posterior mean over β_{ij} is greater than the "correctness threshold" γ_i of the corresponding question q_i) with the greatest expected information gain among other "correct" answers, and select one "wrong" answer (using the reverse of the strategy used for the correct answer) with the greatest information gain among "wrong" answers. This heuristic is needed, as our answer selection process is greedy, i.e. adds new answers to the selection set independently of the selections already made. The expected information gain can be computed for larger sets jointly, however, we do not consider this extension in the current work.

6.2 Example: Which answers are uninformative?

We now illustrate the expected information gain computation for the answers that lie on the two extremes of the "correctness" scale. Consider, for example, an 'ambiguous" answer a_{ij} that lies almost on the "correctness threshold" q_i, i.e. $\mathbb{E}[\beta_{ij}] \approx \mathbb{E}[\gamma_i]$, where expectations are with respect to the posteriors. It is easy to see that the likelihood term in Equation 1 reduces to a constant, yielding a posterior over α_k that is very close to the prior, with a vanishing KL-divergence for both terms in Equation 2. As expected, presenting ambiguous answers is uninformative with respect to the ability of the student who is going to see the answer as a choice in a multiple choice (or a true/false) question.

At the other extreme, consider an "obvious answer", i.e. $|\mathbb{E}[\beta_{ij}]| \gg |\mathbb{E}[\gamma_i]|$. It's easy to see that the term in the sum of Equation 2 corresponding to the $P(z_{ij}^k = click|\alpha_k)$ likelihood will vanish, as this likelihood approaches 1 for a sufficiently large distance between γ_i and β_i (Equation 1), yielding a prior that is close the posterior, and consequently a vanishingly small KL-divergence. The second term in the sum of Equation 2 will also vanish, as in this case, the likelihood term $P(z_{ij}^k = \neg click|\alpha_k)$ vanishes for a sufficiently large distance between γ_i and β_i (Equation 1). As expected, presenting an answer that is obvious, is also unlikely to be informative about the ability of the user to whom this answer is presented as a choice in a multiple choice question, as both poorly- and well- abled users are likely to answer it correctly. Agreeing with the intuition, optimal answers for less able users are those that are more obvious either way (right or wrong), and vice-versa.

6.3 Exploration vs. Exploitation of answers

The consequence of presenting the "most informative" an-

swers only will be a neglect of new answers which arrive continuously in an interactive setting. As such, in addition to exploiting the existing informative answers, we seek to explore the "informativeness" of new answers. One consequence of an unexplored answer is its broad distribution over the support of "correctness", which naturally correspond to broad distributions over the expected information gain . The selection of an answer can now be viewed in the context of utility maximization in a stochastic setting, where the utility of an answer (its expected information gain) is random. Analogous to the multi-armed bandit setting, each answer (arm) returns a an expected information gain (reward) governed by some distribution, unknown to the algorithm (player). In this setting, the reward (expected information gain) is observed after performing posterior inference following the choice of the user (i.e. clicking or skipping the presented answer(s)), following which the reward (expected information gain) distributions for each arm (answer) are updated. This form of probability matching exploration/exploitation strategy is analogous to Bayesian Bandits [20] and Thompson sampling [1]. In this setting, an arm to play (answer to present) is selected by sampling the reward distributions, and picking the arm (answer) corresponding to the greatest reward (expected information gain) in the sample.

It is worthy to note, that this form of exploration/exploitation can likely be avoided, if the expected information gain is computed jointly for all latent variables in the model, which requires higher dimensional integration, but is nevertheless tractable. The approach presented here, however, is a much simpler alternative.

7. BASELINES

We implemented two baselines that capture the intuition of the joint user/answer ranking process:

7.1 Baseline 1

As in two models discussed above, the baselines also compute the α_k "ability" parameter for each user and the $\beta_{a_{ij}}$ "correctness" parameter for each answer. We define the correctness of an answer as follows:

$$\beta_{a_{ij}} = \frac{n_{ij}^{clicked}}{n_{ij}^{clicked} + n_{ij}^{\neg clicked}}$$

where $n_{ij}^{clicked}$ is the number of times answer a_{ij} appeared as an answer to a question q_j and was clicked on, and $n_{ij}^{\neg clicked}$ is the number of times answer a_{ij} appeared as an answer to a question q_j and was not clicked on, capturing the naive intuition that the consensus answer is the correct answer.

The ability of the user is computed as follows:

$$\alpha_k = \sum_{a_{ij} \in O(s_k)} \beta_{a_{ij}} + \sum_{a_{ij} \in C(s_k)} \beta_{a_{ij}}$$

which is the "total correctness" score accumulated by a given user. The right term sums the scores of the answers on which the user had clicked, and the left term sums the scores of the answers which the user contributed. The quantity is normalized if the number of questions per user differs.

7.2 Baseline 2

Baseline 2 aims to improve on the naive consensus approach to evaluating the answer's degree of correctness. We replace the count $n_{ij}^{clicked}$ with the weighted count $\tilde{n}_{ij}^{clicked}$ defined as follows:

$$\tilde{n}_{ij}^{clicked} = \sum_{s_k \in S} \sum_{q_j \in Q} \alpha_k 1[a_{ij} \in choices(s_k, q_j)]$$

where $choices(s_k, q_j)$ is the set of answers shown to user s_k for question q_j. The algorithm then iterates between computing α_k for each user and β_{ij} for each answer.

7.3 Bachrach et al. Baseline

The Bachrach et al. model (titled D.A.R.E. for Difficulty-Ability-REsponse estimation) [3] is a graphical model designed for the task for assessing user ability, and inferring the correct answers to a an existing test (they employ an IQ test for their study). The model provides a framework for adaptive testing, also based on the expected information gain criterion. Their model, however, does not consider a scale of correctness for the answers, and does not consider a problem of selecting a subset of answers for the purpose of test design, as all of the questions and answers are considered to be fixed, as provided by the instructor. Additionally, our model incorporates the conditional dependencies between users and the answers they generate, and more naturally postulates the likelihood explicitly capturing users' pairwise preferences between answers. As the Bachrach et al. cannot naturally deal with questions with a changing set of answers, we constrain the presentation of the questions to contain only one answer, i.e. a True/False question. As such, if a given question contains n contributed answers, the Bachrach et al. model would be instantiated with n distinct questions.

8. EXPERIMENTS

We evaluate the models and baselines on 1) simulated data and 2) real-world user study performed on the Amazon Mechanical Turk platform.

8.1 MTurk Data collection

We develop a custom framework for the user studies on MTurk. We employ three reading passages (two SAT texts from the reading comprehension sections (41 and 50 users respectively), and a chapter on evolution from the OpenStax Biology textbook [4] (avg. 35 users)). Each user is asked to 1) contribute 1 question and 2) answer a set of 10 questions in that order during the task. Questions with less than 2 answers are presented as "free-response" questions, i.e. users are asked to type in the answer. Questions with more than 2 answers are presented as multiple choice, with at most 4 existing answers as choices, chosen randomly. Additionally, each question provides a "None of the above" option, which when clicked, allows the user to type in the answer directly, contributing it to the answer set of the question.

Gold Questions. Half (5) of the questions shown to each user as actual SAT/Biology questions for that passage, mixed in random order with the user-generated questions, and are used to assess the correlation between the rankings obtained from the two sets of questions.

8.2 Evaluation Methodology

8.2.1 Absolute rank agreement with expert

. We evaluate the agreement between the rankings predicted by our model and the rankings obtained by sorting on the students' performance on the gold questions. This metric evaluates the ability of the model to generate an accurate assessment of the students, where the accuracy is defined by the degree of agreement with the expert (i.e. ranking generated via expert-generated questions).

8.2.2 Relative rank agreement with pair of experts

. Absolute agreement metric alone, however, may not be a fair evaluation strategy, as it disregards the baseline agreement that exists across experts. A more fair evaluation would compare the agreement of the rankings between the model and an expert, with the agreement between a pair of experts. We simulate a pair of experts by presenting 10 gold standard questions per student, and attributing half of the gold questions to each "expert".

9. RESULTS

9.1 Metrics

We employ the *Precision@K* metric. The *Precision@K* only considers the fraction of users predicted to be in the top-K users who were also in the top-K users in the gold ranking. *Precision@K* may be a useful metric in certain scenarios, such as finding the top performers in a MOOC, for awarding bonus points, or granting job interviews

9.2 MTurk

9.2.1 Absolute rank agreement with expert

Gold rankings for each user are obtained by sorting the users on the number of correctly answered gold questions. *Prec@K*, thus, in this case is a measure of correlation between the student- and teacher- induced rankings. Note that in this experiment, students generated a question prior to seeing other questions (including gold questions).

9.2.2 Relative rank agreement with pair of experts

We simulate a pair of experts by partitioning the gold-standard questions into 2 sets, and computing the gold-rankings induced by the two sets. We compute the average *Prec@K* between 1) every possible partition of the 10 gold questions into two sets, and 2) between each of the resulting partitions and the model-predicted ranking.

9.3 Analysis

9.3.1 Absolute rank agreement

We observe that for two sets of experiments (corresponding to different reading passages) (Figure 3 a, b), the performance of the *Prec@K* metric of the Model 2 exceeds both baselines, random, and either performs comparably to the Bachrach et al. baseline (Microsoft baseline in plots). Additionally, we compute an average *Prec@K* over three independent experiments (with the same questions) using one of the SAT passages. From the statistical analysis of the results, we cannot yet conclude that the difference between our model and the baselines is significant with 95% confidence ($p = 0.19$ at $K = 0.5$ and *Prec@K*(Model 2)=0.94, *Prec@K*(Baseline 2)=0.60), primarily due to only averaging over three runs (at this point limited by cost, and time).

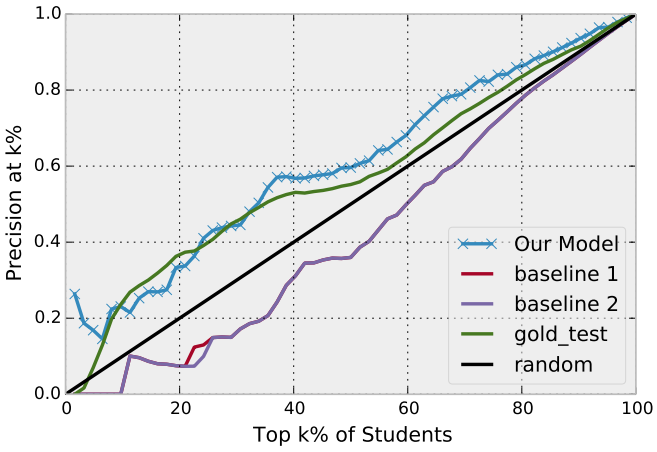


Figure 2: Average *Prec@K* for *gold* (computed between every possible partition of gold questions into two sets) and *our model* (computed between each of the resulting partitions of gold questions and model-predicted ranking). Diagonal line corresponds to the theoretical random guess baseline, i.e. if our algorithm randomly shuffled the gold rankings, we would expect to do no better or worse than the black diagonal.

From the analysis of the data, however, it is likely that the difference is significant, pending additional runs of the experiment. One important conclusion from our results is the sensitivity of the results to the type of content material. For example, the results in Figure 3 a,b correspond to the reading passages taken from two different SAT passages, and the difference in performance of our model (Model 2) can be observed between the two passages.

9.3.2 Relative rank agreement

We observe from Figure 2 that the average *Prec@K* for the between-expert and the expert-model conditions is on-par, indicating that any two independent experts are as likely to agree on the ranking of the students as is our model based on student-generated questions with any of the two experts. A single test, however, may not in all cases justify the independence assumption (needed for simulating two experts) when segregated into two sets. Averaging across all possible partitions as we do is likely to mitigate the issue partially, but a controlled study with questions generated by a pair of independent experts is called for in further work.

9.4 Active Test Design

Our evaluation of the active test design consists of 1) qualitative analysis of the generated questions and 2) the number of original answers users are likely to contribute to the question. The second point requires some clarification. Recall that if a user is not satisfied with any one answer, he or she may opt for the None-of-the-above option, at which point they may choose to contribute a new, original answer. The goal of active test design, as described above, is that the presented choices contain both the right and (one or more) wrong answers. Intuitively, we would expect questions generated in this way would be likely to receive far less new contributions (original answers), than the questions whose

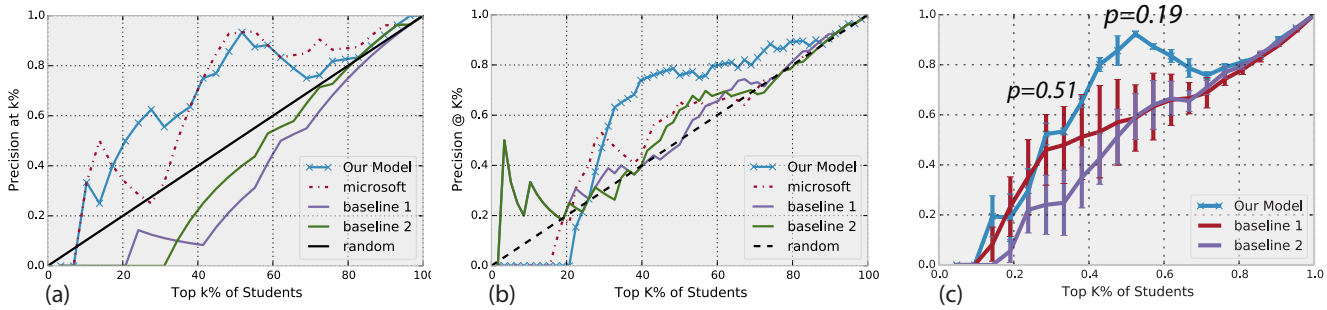


Figure 3: Precision@k% for three sets of experiments . a) SAT Dataset 1 (1 experiment) b) SAT Dataset 2 (1 experiment). c) SAT Dataset 1 (3 experiments averaged), p-values from pairwise t-test reported for a select set of k%. Diagonal line corresponds to the theoretical random guess baseline, i.e. if our algorithm randomly shuffled the gold rankings, we would expect to do no better or worse than the black diagonal. Note that Our Model refers to Model2

answer choices were selected at random, or with some other less optimal strategy. We compare the total number of newly generated answers for one experiment for three conditions: *Optimal*, *Random* and *Reverse*. The *Optimal* condition generates new answers according to the Bayesian Bandit strategy described earlier, *Random* presents answers by sampling a subset of answers uniformly at random, and the *Reverse* condition presents the answers in the least optimal way (i.e. opposite of what the *Optimal* condition presents, with some caveats of implementation that we will not discuss here). We observe (Table 1) that the largest number of newly generated answers corresponds to the *Reverse* condition, while the least to the *Random* condition. These results, however, are inconclusive, as they have not been obtained from a sufficient number of experiments at this point.

Tables 2 and 3 present two questions and their corresponding answers from the dataset based on the Evolution reading passage, together with the probabilities that correspond to the likelihood of that answer appearing in the choice set of a multiple choice question featuring these answers. These probabilities can be viewed as parameters of a multinomial distribution computed (in a deterministic way) from the distributions over the expected information gain, as described earlier (sampling from this multinomial distribution is a heuristic surrogate to performing direct sampling from the original distributions, that was easier to implement and debug). We observe, for the question in Table 2, that the incorrect answers "I don't know" receives the lowest likelihood of appearance as a choice (0.0813), whereas "mutation and sexual reproduction" is the likeliest wrong answer to be presented. The latter choice is a far less obvious wrong answer, and is likely to be more informative about the skill of the user presented with this answer as a choice, than the the former. Observe from the example question in Table 3, that the multiple lexical variations of the correct answer (finches) all receive nearly equal probability of appearance, as expected. The wrong answer case is particularly interesting in this example. Note that while the "Floreana mockingbird" is a historically correct answer, there is no mention of the "Floreana mockingbird" in the presented passage. Consequently it had received far fewer clicks from other users, leading it becoming an "obviously incorrect" answer. We can infer from this example, that this behavior will be typical of a system that relies on user-generated questions and answers. Users with extraneous sources of information, who present correct answers that are not found in the text, will unfortunately suffer a penalty for providing such answer. On the other hand, we can interpret that the scale of "correctness" to which we referred numerously, should be recast as the "correctness with respect to the presented text only", an important outcome of student-generated testing that we discovered through this work.

Optimal	Random	Reverse
6	0	10

Table 1: Number of new answers contributed for the three conditions, corresponding to the algorithm used for selecting answer subset for each question: Optimal (described above), Random, Reverse (corresponding to the inverse of Optimal)

Question:	What are the three possible outcomes on the phenotype, from genetic changes caused by mutation?
Correct Answers	
Answer	**Probability**
Regression, no change, progression	0.4585
reduced fitness, beneficial effect on fitness, no effect on fitness	0.5415
Incorrect Answers	
Answer	**Probability**
I dont know	0.0813
mutation and sexual reproduction	.9187

Table 2: Sample Question 1. Probabilities correspond to the likelihood of the answer to be chosen as a choice in a multiple choice question, and is computed from distribution over the information gain for the respective answer.

Question: What birds helped to inspire Darwin's theories?	
Correct Answers	
Answer	**Probability**
finches	0.2599
Ground finches	0.2494
finches	0.2394
finches	0.2512
Incorrect Answers	
Answer	**Probability**
Floreana mockingbird helped inspire Darwin's theory of evolution	0.0872
galopogos birds	0.9128

Table 3: Sample Question 2. Probabilities correspond to the likelihood of the answer to be chosen as a choice in a multiple choice question, and is computed from distribution over the information gain for the respective answer.

10. EXTENSIONS AND FUTURE WORK

The generality of the proposed model allows for its application in natural settings, such as academic and technical forums, e.g. StackExchange and Piazza, where interactions between users also take the form of question generation and question answering. StackExchange matches up quite well with our model; the forums offer The former, in particular, would be an interesting case. StackExchange forums offer a rich source of user interaction data, in the form of questions, proposed answers, with corresponding up-vote "scores" from users. Patterns of user activity, such as who answered whose question, and who received more up-votes in answering a question, couldn be utilized for the purpose of ranking users. The latter, on the other hand, has the advantage of having expert rankings obtained from the instructor-curated exams.

11. DISCUSSION AND CONCLUSION

We show, through custom-designed experiments on Mechanical Turk, and natural experiments in a technical and an academic forum, that fully autonomous assessment (assessment without any interventions from the teacher) is likely to be possible. Experiments on Mturk show that the agreements between expert and implicit assessment are moderate, but on-par with agreement between any two experts. Finally, we learned about the interpretation of the "correctness of an answer" in the context of student-generated testing. Specifically, the model is unlikely to infer the correct answers that are outside the scope of the presented material, which can be seen as both a limitation and a strength of the model. For the purpose of learning a correct answer to some question, the model will be limited primarily to the answers that can be found in the presented text. For the purpose of test generation, however, the model is likely to generate only the choices that are likely to be found in the reading material — an important, and a non-trivial challenge faced by human test makers.

12. REFERENCES

[1] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. *arXiv preprint arXiv:1111.1797*, 2011.

[2] A. Anastasi. Psychological testing . 1961.

[3] Y. Bachrach, T. Graepel, T. Minka, and J. Guiver. How to grade a test without knowing the answers—a bayesian graphical model for adaptive crowdsourcing and aptitude testing. *arXiv preprint arXiv:1206.6386*, 2012.

[4] R. G. Baraniuk. Opening education. *The Bridge: Linking Engineering and Society*, 43(2):41–47, 2013.

[5] S. Basu, C. Jacobs, and L. Vanderwende. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the ACL (TACL)*, 2013.

[6] C. Chin. Student-generated questions: Encouraging inquisitive minds in learning science. *Teaching and Learning*, 23(1):59–67, 2002.

[7] C. Chin and D. E. Brown. Student-generated questions: A meaningful aspect of learning in science. *International Journal of Science Education*, 24(5):521–549, 2002.

[8] T. Eberlein, J. Kampmeier, V. Minderhout, R. S. Moog, T. Platt, P. Varma-Nelson, and H. B. White. Pedagogies of engagement in science. *Biochemistry and Molecular Biology Education*, 36(4):262–273, 2008.

[9] A. E. Elo. *The rating of chessplayers, past and present*, volume 3. Batsford London, 1978.

[10] M. E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394, 1999.

[11] M. E. Glickman and S. T. Jensen. Adaptive paired comparison design. *Journal of statistical planning and inference*, 127(1):279–293, 2005.

[12] R. K. Hambleton, H. Swaminathan, and H. J. Rogers. *Fundamentals of item response theory*. Sage, 1991.

[13] J. Hamer, K. T. Ma, and H. H. Kwong. A method of automatic grade calibration in peer assessment. In *Proceedings of the 7th Australasian conference on Computing education-Volume 42*, pages 67–72. Australian Computer Society, Inc., 2005.

[14] R. Herbrich, T. Minka, and T. Graepel. TrueskillâĎć: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, pages 569–576, 2006.

[15] A. King. Facilitating elaborative learning through guided student-generated questioning. *Educational Psychologist*, 27(1):111–126, 1992.

[16] T. Minka, J. Winn, J. Guiver, and D. Knowles. Infer .net 2.4, 2010. microsoft research cambridge, 2010.

[17] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in moocs. *arXiv preprint arXiv:1307.2579*, 2013.

[18] F. Radlinski and T. Joachims. Active exploration for learning rankings from clickthrough data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 570–579. ACM, 2007.

[19] K. Raman and T. Joachims. Methods for ordinal peer grading. *arXiv preprint arXiv:1404.3656*, 2014.

[20] S. L. Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

[21] L. A. Shepard. The role of assessment in a learning

culture. *Educational researcher*, pages 4–14, 2000.

[22] R. Singh, S. Gulwani, and A. Solar-Lezama. Automated feedback generation for introductory programming assignments. In *Proceedings of the 34th ACM SIGPLAN conference on Programming language design and implementation*, pages 15–26. ACM, 2013.

[23] S. E. Stemler and R. J. Sternberg. Using situational judgment tests to measure practical intelligence. *Situational judgment tests: Theory, measurement, and application*, pages 107–131, 2006.

[24] S. E. Stemler, R. J. Sternberg, E. L. Grigorenko, L. Jarvin, and K. Sharpes. Using the theory of successful intelligence as a framework for developing assessments in ap physics. *Contemporary Educational Psychology*, 34(3):195–209, 2009.

[25] R. J. Sternberg. *The triarchic mind: A new theory of human intelligence*. Penguin Books New York, 1989.

[26] P. Varma-Nelson, M. S. Cracolice, and D. Gosser. Peer-led team learning: A student–faculty partnership for transforming the learning environment. *Invention and Impact: Building Excellence in Undergraduate Science, Technology, Engineering, and Mathematics (STEM) Education*, pages 16–18, 2004.

[27] D. Vats, C. Studer, A. S. Lan, L. Carin, and R. G. Baraniuk. Test size reduction for concept estimation. In *International Conference on Educational Data Mining (EDM)*, 2013.

[28] M. N. Volkovs and R. S. Zemel. A flexible generative model for preference aggregation. In *Proceedings of the 21st international conference on World Wide Web*, pages 479–488. ACM, 2012.

[29] A. Vozniuk, A. Holzer, and D. Gillet. Peer assessment based on ratings in a social media course. In *The 4th International Conference on Learning Analytics and Knowledge LAK14*, number EPFL-CONF-189957, 2014.

[30] R. C. Weng and C.-J. Lin. A bayesian approximation method for online ranking. *The Journal of Machine Learning Research*, 12:267–300, 2011.

[31] J. Winn, C. M. Bishop, and T. Jaakkola. Variational message passing. *Journal of Machine Learning Research*, 6(4), 2005.