ABSTRACT

Difficult test questions can be made easy by providing a set of possible answer options of which most are obviously wrong. In the education literature, a plethora of instructional guides exist for crafting a suitable set of wrong choices (distractors) in order to probe the students’ understanding of the tested concept. The art of multiple-choice question design thus hinges on the question-maker’s experience and knowledge of the potential misconceptions. In contrast, we advocate a data-driven approach, where correct and incorrect options are assembled directly from the students’ own past submissions. Large-scale online classroom settings, such as massively open online courses (MOOCs), provide an opportunity to design optimal and adaptive multiple-choice questions that are maximally informative about the students’ level of understanding of the material. We deploy a multinomial-logit discrete choice model for the setting of multiple choice (MC) testing, derive an optimization objective for selecting optimally discriminative option sets, and demonstrate the effectiveness of our approach via synthetic experiments and a user study. We furthermore showcase an application of our approach to crowd-sourcing tests from technical online forums.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous

Keywords
Adaptive Learning, Assessment, Crowdsourcing, Optimal Testing

1. INTRODUCTION

A finite set of alternatives for the student to pick from—the key feature of multiple choice questions (MCQs) that makes them attractive in grading, is the very thing that makes good MCQs notoriously difficult to design [14]. Incorrect options, also known as distractors, should ideally be picked from a representative set of misconceptions that students commonly share. But even if this set is representative, the question might still fail to distinguish between students who were “close” to the correct answer, and those who were clueless. In the adaptive testing literature [10, 19], the questions themselves are selected to be at a level that is appropriate for the student, such that their responses result in the most accurate estimate of their knowledge. In this work, we pursue the same goal, but at the level of designing a single question via selecting a set of options to present as potential answers. This problem is not a straightforward extension of the classic adaptive testing problem for two reasons: (i) from an application perspective, only recently with the advent of web-scale learning platforms are we able to leverage the massive number of student submissions and answer click-through logs to generate rich, adaptive and data-driven questions that exploit actual student misconceptions; (ii) from a technical level, selecting choices is inherently a batch optimization problem, i.e., all options must be considered jointly during optimization; this is in stark contrast to question selection, which typically assumes independence between questions and finds the optimal set in a greedy fashion.

The contributions of our work are summarized as follows:
• We derive an objective set function for selecting an optimal set of choices in a nominal choice model, given the estimated user ability, and we investigate the solutions across different regimes of student ability.
• We collect and release a dataset used in our experiments: a “U.S. States Quiz” dataset, where users were given an MCQ quiz testing their knowledge of the names of states.
• We propose a new paradigm of data-driven test design by leveraging data from technical online forums, and showcase the applicability of this model to the task of MCQ design from StackExchange posts.

2. RELATED WORK

Education In the education literature, multiple choice testing has received significant attention, studying a broad range of aspects of MCQ design, e.g., to ensure validity [7, 6], to decide on the optimal number of choices [14, 8], and to design good distractors [9, 6]. In an empirical study [7], Haladyna and Downing concluded that the key in multiple choice item design was “not the number of distractors but the quality of distractors,” where almost unanimously high-quality distractors are considered to be those that represent common student misconceptions [8]. Thissen et al. [16] develop a graphical analysis method of distractors based on the response statistics in the context of a nominal item response model, with the goal of facilitating post-hoc analysis of multiple choice items. Computational methods have been proposed for the task of multiple choice item design, but are restricted to specific domains, such as vocabulary [3], grammar testing [2], and topic-specific comprehension [12]. In all of these methods, however, distractors are generated automatically based on the structure of the problem domain, e.g., types of common errors that students make. We are not aware of any work that directly optimizes for a distractor choice set based on the data on past student submissions.

Active Learning and Adaptive Testing The field of adaptive testing borrows techniques from the areas of active learning and optimal experiment design. Adaptive testing is classically posed as a task of item set optimization (classically in an online setting, see Chapter 7 of [5]) for an optimal experiment design. Adaptive testing is classically posed as a task of item set optimization (classically in an online setting, see Chapter 7 of [5]) for an optimal experiment design. Adaptive testing borrows techniques from the areas of active learning and Adaptive Testing

3. MODEL

In motivating our model, we require it to have the following three features:
• The model specifies a probability of a student choosing a particular option as a function of that student’s ability and that option’s correctness, such that students with higher ability are more likely to pick the most correct option (we will discuss this aspect in detail below).
• An “ideal” student (with the greatest attainable ability) chooses the correct option with probability 1.
• A student with the least attainable ability makes their choice uniformly at random.

For simplicity, we require that there is one correct option, leaving the remaining options as distractors that lie on a continuum of apparent correctness, i.e., options that vary in how difficult they are to discern from the answer, and such that a more able student is more likely to discern the correct option.

A multinomial logit model with a partial order constraint on the choices and a non-negativity constraint on the student’s ability, exhibits all of the desired properties above. Specifically, we use the following model:

\[
P(s_i \text{ picks option } j \mid \theta_i, \{\beta_j\}_{j \in Q}) = \frac{\exp(\theta_i \beta_j)}{\sum_{\beta_j \in Q} \exp(\theta_i \beta_j)}
\]

such that \(\beta_j > \beta_j^*, \forall \beta_j \in Q \setminus \beta_j^*, \text{ and } \theta_i \geq 0, \forall i.
\]

Here, \(s_i\) is student \(i\) with ability \(\theta_i\), \(\{\beta_j\}_{j \in Q}\) is the set of option parameters of question \(Q\) encoding the apparent correctness of each option and \(\beta_j^*\) is the correct option. Without the explicit partial order constraint on choices and the non-negativity constraint on subjects, the model would capture the relative preference of subjects towards choices, and in psychometrics is known as the nominal response model [5], also related to the more general multidimensional unfolding models [4, 15] often used to investigate the relationship between subjects and preferences. In our setting, the non-negativity constraints on the ability \(\theta_i\), combined with the partial order constraints on the option parameters are critical to obtain the desired interpretation of the \(\theta_i\) parameters, namely as capturing the ability of the students. More importantly, performing optimal option subset selection under this model and these constraints will result in subsets that are most informative about the students’ abilities.

It is important to also realize the limitations and additional assumptions underlying this model. The most significant limitation that we will address later is what is known as the independence of irrelevant alternatives (IIA) assumption [13]. The IIA assumption is violated whenever the two options are not inherently different—in the setting of reusing student responses as potential options in the test this would happen if the two options are either completely identical or paraphrases of each other. We leave dealing with the problem of non-independence of alternatives to future work, and we discuss the potential approaches to this problem in our context in the future work section.

The additional consequence of this model (and one of our requirements) is that in a limit of a student with very low ability, that student will guess the correct answer uniformly at random. Note that this does not imply that the model does not capture common misconceptions, in fact it does so naturally by placing probability mass on the more difficult distractors, proportional to the ability of the student.

We now compare our model with respect to two related models: the classical Rasch model [5] and the recent model by Bachrach et al. [1].

3.1 Relationship to the Rasch model

The classical dichotomous Rasch model defines the likelihood of a student answering a question correctly as a function of the question’s difficulty and the student’s ability, i.e., it is agnostic to the actual choice made by the student in the MCQ setting. The likelihood of a student with ability \(\theta_i\) getting the question with difficulty \(q_j\) correct is given by:

\[
P(s_i \text{ correctly answers } j \mid \theta_i, q_j) = \frac{1}{1 + \exp([-\theta_i - q_j])}.
\]
To gain intuition about how our model encodes question difficulty, consider the case of only two options: the correct option with parameter $\beta^*_j$ and the incorrect option with parameter $\beta_j$. We can then similarly write the likelihood of the student getting such a question correct as:

$$P(s_i \text{ correctly answers } j \mid \theta, \Delta_{j-i}) = \frac{1}{1 + \exp(-\theta \Delta_{j-i})},$$

where $\Delta_{j-i} = \beta^*_j - \beta_j$, which is always positive by our definition. By analogy with the Rasch likelihood, $\Delta_{j-i}^{-1}$ captures a similar notion of question difficulty: the farther apart the two options in their parameter values, the “easier” the question is.

Observe that in the case of more than two options, the likelihood becomes:

$$P(s_i \text{ right on } j \mid \theta, \{\Delta_{j-i}\}) = \frac{1}{1 + \sum_{k \in \mathcal{Q}} \exp(-\theta \Delta_{j-k})},$$

where now we have an exponential with a distance $\Delta_{j-k}$ between the correct option and all remaining options. We can see that the probability of getting the question right approaches one when the all of the distances between the correct answer and all the remaining choices, scaled by the ability parameter, are sufficiently large to drive all of the exponents to zero. It’s worthy of noting that by relying on the information about the relative correctness of the choices, our model will be able to estimate the relative ability of the students even if they all answered the question incorrectly, i.e. some answers are more incorrect than others.

### 3.2 Relationship to the model by Bachrach et al.

Recently Bachrach et al. [1] extended the dichotomous Rasch to account for the observation of the actual choice, with the goal of inferring the correct answers from choice click-through alone (i.e., in an unsupervised way). The model can be used in place of the model by Bachrach et al., in a fully unsupervised setting, i.e., without the partial order constraints on choices, in which case it is equivalent to the traditional discrete choice model.

The fundamental difference, however, is in the capability of our formulation to be used for the task of optimal choice set design. While it is relatively straightforward in the proposed model, the same task is not possible in the model by Bachrach et al. The underlying reason is that the question difficulty in our model is implicitly defined through the relative distances between option parameters, and thus allows us to directly optimize difficulty via an explicit objective that relates option sets and question difficulty, allowing to tune the informativeness of the question for a specific user. In the Bachrach et al model, however, the difficulty and choice parameters are decoupled, making it impossible to derive an objective that relates the expected informativeness of a question about a student and a set of presented choices.

### 4. OPTIMAL CHOICE SETS

We pose the problem of optimal choice set design in the traditional active-learning setting: maximize information on the student ability parameter $\theta$ as a function of a selected observation (choice set). Given a set of choices $C$, user with ability $\theta$ is presented with a subset $C' \subseteq C, C' = \{\beta_k\}_{k \in K}$, where $\beta_k$ is a “correctness” of choice $k$. This user then chooses one item from $C'$ with probability:

$$P(z = k \mid \{\beta_k\}_k, \theta) = f(\theta; C') = \prod_{k \in K} \frac{\exp(\theta \beta_k)}{\sum_{k' \in K} \exp(\theta \beta_{k'})}.$$

We are interested in finding a subset $C^* \subseteq C$ that is optimal in some sense for the user with a given ability. In the setting of optimal experiment design, we are interested in choosing $C^*$ that results in the smallest variance of the maximum likelihood estimator of $\theta$ which can be found by maximizing the Fisher information for a subset $C^*$:

$$\mathcal{I}(\theta; C^*) = \mathbb{E} \left[ \frac{\partial^2}{\partial \theta^2} \log f(\theta; C') \right],$$

which—after some extensive, but straightforward algebraic manipulations (see the Appendix)—can be shown to be maximized by the solution to the following non-linear optimization problem:

$$\begin{align*}
\text{maximize} & \quad \sum_{i \in \mathcal{J}} \sum_{x_i} \left( \beta_i - \bar{\beta}_i \right)^2 \exp(\theta \beta_i + \bar{\beta}_i) \\
\text{subject to} & \quad x_i \in \{0, 1\}, \forall x_i \\
& \quad \sum x_i \leq K
\end{align*}$$

To build intuition about the types of choice sets this objective prefers, it is instructive to consider a simple case with only two choices $\{\beta_1, \beta_2\}$. In this case the objective reduces to:

$$\left( \beta_i - \bar{\beta}_i \right)^2 \frac{\exp(\theta \beta_1 + \bar{\beta}_1)}{\exp(\theta \beta_1 - \bar{\beta}_1) + \exp(\theta \beta_2 - \bar{\beta}_2)}.$$ 

where we define $\Delta_{ij} = \beta_i - \beta_j$.

$$\Delta_{ij}^2 \frac{\exp(-\theta \Delta_{ij}) + \exp(\theta \Delta_{ij}) + 2}{\exp(-\theta \Delta_{ij}) + \exp(\theta \Delta_{ij}) + 2}.$$ 

From these expressions, we can see that the Fisher information grows approximately as $\Delta_{ij}^2$ for small values of $\Delta_{ij}$, and decays exponentially for larger values of $\Delta_{ij}$, with an optimal $\Delta_{ij}$ depending on the value of $\theta$. This behavior resembles the traditional adaptive setting of maximizing information in a Rasch model by picking a
question of the optimal difficulty. We can gain additional insight into the optimal spacing of the two choices $\Delta_j$ for a student of a specific ability $\theta_i$. The maximum of the above expression is a solution to:

$$\frac{\theta_i \Delta_{ij} + 2}{\theta_i \Delta_{ij} - 2} = \exp(\theta_i \Delta_{ij}).$$

Consider the case where the student’s ability $\theta_i$ increases. It becomes clear from the above that $\Delta_j$ has to consequently decrease to maintain equality. Intuitively, we gain most about the student’s ability by showing more difficult distractors to more able students and vice versa.

### 4.1 Asymptotically optimal choices

Evidently, it is more instructive to investigate the solutions to optimal choice sets beyond that of two options. Consider two limiting cases: a student with a very large ability ($\theta_i \rightarrow \infty$), and a student with a very low ability ($\theta_i \approx 0$).

**Case $\theta_i \rightarrow \infty$.**

It is straightforward to show that in the limit of “infinite ability,” the information is guaranteed to go to zero. However, the rate at which it goes to zero depends on the choice set, and allows us to gain some insight into the kinds of choice sets that will be preferred for users with a large ability parameter. The log of the information function will have a linear asymptote, with the slope determined by the largest exponential in the numerator and the denominator. This slope of the asymptote is then given by:

$$\lim_{\theta \rightarrow \infty} \log F(\theta) = 2 \log(\beta_{\text{max}} - \beta_{\text{max}-1}) + \theta (\beta_{\text{max}} + \beta_{\text{max}-1} - 2 \beta_{\text{max}}),$$

where $\beta_{\text{max}}$ indicates the largest $\beta$ in the set and $\beta_{\text{max}-1}$ as the second largest $\beta$. Maximizing with respect to $\beta_{\text{max}-1}$, we get:

$$\beta_{\text{max}-1} = \beta_{\text{max}} - \frac{2}{\theta}.$$

Clearly, the greatest Fisher information for large values of $\theta$ will be obtained when $\beta_{\text{max}-1} \approx \beta_{\text{max}}$, i.e., when the distance between the two top choices approaches zero. That means that when $\theta \rightarrow \infty$, only the two best choices play a role in defining an optimal choice set, and the remaining choices become irrelevant.

**Case $\theta_i \approx 0$.**

In the limiting case of $\theta \rightarrow 0$, the objective reduces to:

$$\max \frac{1}{K^2} \sum_{k} \sum_{k'}^{K} (\beta_k - \beta_{k'})^2,$$

which will yield solutions with maximum total inter-choice distance. If, for example, the model had an ability to pick continuous values for $\beta$, and they were constrained to be in an interval, the solution to the above will be to put half of the choices on each endpoint of the interval (the proof is given in the Appendix). The intuition behind this solution is as follows: a student who is guessing ($\theta = 0$) will have an equal probability of picking any choice regardless of that choice’s underlying parameter $\beta_k$. However, the resulting parameter update on $\theta$ after observing that student’s choice, will depend on how the $\beta_k$ were spaced out. If we imagine that $n_0$ choices were placed at the $\beta = 0$ extreme and $n_1$ were placed at the $\beta = 1$ extreme (consider that all choices were spaced on a unit interval for illustration), then the expected information gain on the user whose current estimate of ability is $\theta = 0$ is:

$$\mathbb{E}[J(\theta)] = \frac{n_0}{n_0 + n_1} J(\theta) + \frac{n_1}{n_0 + n_1} J(\theta).$$

Here, $J(\theta)$ and $J(\theta)$ is the information gain conditioned on either one of the two outcomes: student clicks one of the choices whose $\beta = 0$ or one of the choices whose $\beta = 1$, respectively. For the purpose of illustration, it is sufficient to realize that $J(\theta)$ roughly corresponds to the change in $\theta$, should that student pick the option corresponding to $\beta = 1$, which we expect to be significant since we initially assume a student with $\theta = 0$ (one not very likely to pick a correct option). This provides some intuition as to why selecting multiple choices at the “correct extreme” $\beta = 1$ (i.e., $n_1 > 1$) maximizes the expected information on $\theta$. See Figure 9 for an example.

### 5. PARAMETER INFERENCE

We maximize the log-likelihood of all observations, where the $i$th observation is a set $\{s_i, c_{i1}, \ldots, c_{iC}\}$ and $s_i$ is the student who picks choice $c_{ij}$ out of a presented set of choices $C$. The goal is to infer all choice parameters for each question $j \{\beta_k\}$ and all student parameters $\{\theta_i\}$ for each student $s_i$. This optimization problem is bi-convex, i.e., convex in $\{\theta_i\}$ and $\{\beta_k\}$ separately, but not jointly. We perform alternating maximization, fixing each set of parameters and optimizing the other in turn. We use LBFGS [20] for optimization in an unconstrained case. We impose soft constraints via modifying the objective: for each pair of answers in the partial order (where a pair consists of a correct answer and one wrong answer), we add an observed Bernoulli-distributed random variable that with large probability preserves the order. In contrast to this approach, we found that direct constrained optimization, via sequential least squares programming was significantly slower. To guarantee identifiability and to speed up parameter estimation, we use $\ell_2$-norm regularization on student and choice parameters, with the regularization factor 0.1 for all experiments.

### 6. SYNTHETIC EXPERIMENTS

We conduct simulations on data that was sampled from the likelihood defined by our model. The goal of the synthetic experiments is to (i) validate the correctness of the inference algorithms, (ii) study the effect of the optimal sampling strategy, and (ii) investigate the effects of incorporating partial order constraints on learning the student parameters.

#### 6.1 Unconstrained setting

The simulation is performed as follows: 100 student parameters are sampled from a uniform distribution over student ability parameters; 50 questions with 50 options each are generated, where each options is independently sampled from a uniform distribution over question difficulties. We first evaluate the results of the unconstrained inference (i.e., by not specifying correct answers via partial order constraints), and allowing the model to infer the automatically from data (similar to the setting of Bachrach et al. in “How to Grade a Test Without knowing the Answers”). We summarize the performance of the inference algorithm via rank correlation of the inferred and ground truth ranking for students and choices respectively. We use Precision@k as a metric of rank correlation in both cases (Figure 2), and experiment with varying the number of choices that a student sees during the simulated test (between 2 and 25). We observe that providing more choices improves the resulting rank correlation, i.e., we are better in distinguishing good students from poor students, and good choices from bad choices.
6.2 Partial order constraints

The value in partial order constraints depends on the distribution of choice parameters $\beta_{jk}$. If the choice parameters are well-separated, the resulting discrete distribution over choices will have low entropy, with most of the mass on the best choice, in which case the correct answer will likely be inferred without the additional constraints. Partial order constraints help most when the choice parameters are distributed closely together (recall that proximity of choices in our model is a proxy for the “difficulty” of the question). Figure 3 illustrates this phenomenon by showing the resulting rank correlation for a synthetic population of students with the same parameters, evaluated on the choice sets of increasing variance of the choice parameters. Adding constraints does not help when the questions are “sufficiently easy,” i.e., choices are well-separated.

6.3 Optimal choice sets

We now evaluate the choice subset selection optimization objective derived in Section 3. Similar to the previous two experiments, we generate a simulated classroom with a fixed number of students (100), but in this case perform parameter inference sequentially after each student answers a question and evaluate rank correlation across students only. For every question, we sample choice sets according to three different sampling strategies: (i) random: choices are drawn uniformly at random, (ii) optimal-individual: optimal choice set is selected for each student according to their ability, and (iii) optimal-average: optimal choice set is selected according to the mean ability of the student population. We make the following observation: performing optimization for the “average” user (average ability across all students) is not outperformed when we optimize for each user’s parameter individually (i.e adaptively). We propose the following explanation: individual ability parameters will have greater variance in comparison to the inferred mean ability of the population due to difference in sample size (by a factor of the number of students). As such, the advantage of individually optimal choice sets is offset by the poorer individual parameter estimates that are needed to generate the optimal sets.

7. USER STUDY: “US STATES QUIZ”

We performed a real-world study to evaluate the importance of data-driven choice set selection in the context of a quiz that asks users to name states of the United States. In this setting, we considered a question to be a specific state which the subject is required to identify by picking a correct choice out of a set of options. This
problem serves as an excellent platform for evaluating our model for the following reasons:

1. **Large choice set**: For each state, there are 50 alternatives that can be used as potential options as part of a smaller set of choices.

2. **Ease of evaluation**: The fact that the set of possible answers to each question is finite allows us to use the raw score on a question where all 50 options are presented as the “ground-truth” of the subject’s knowledge in this domain. Any other test based on only a subset of options (and consequently the method used to obtain the options) can be evaluated against this “ground-truth” by measuring the correlation of the two scores.

3. **Independence of alternatives**: Full independence of alternatives unlikely holds in this setting. Consider, for example, that the user is aware that the state in question is on the west coast and knows which of the options belong to which coast. In this case, adding more options from the east coast should not reduce the probability of that user choosing states on the west coast, which is what happens in the multinomial logit model. This violation, however, is not as severe as in the case of repeated or paraphrased answers (like in the classic “red bus/blue bus” example).

4. **Large range of “good” and “bad” choices**: Not all distractors in this setting are “created equal”: intuitively we should expect that some states, like those that border the correct state, to be easily mistaken for the correct answer.

5. **Common knowledge** Since state knowledge is “common knowledge” we do not burden the test participants with an additional learning stage, such as a reading comprehension.

### 7.1 Data collection

Mechanical Turk workers residing in the U.S. were solicited to a task titled “How well do you know U.S. states?”, and briefly described as a quick quiz to test one’s knowledge of the U.S. states that consists of two stages:

1. **Stage I (fullMCQ)**: Subjects are presented with a map of the U.S. with a randomly highlighted state and 50 options, one for each state that they are required to pick. This selection is made for every one of the 50 states, presented in a random order. Subjects are not revealed the correct answer, and are discouraged from looking up the answers externally.

2. **Stage II (subsetMCQ)**: The same subject then repeats the test, but now with only 4 options for each of the 50 states. Options are chosen according to two strategies: Random and Optimal described in more detail below.

Two experiments were conducted (Exp1, Exp2) under two different conditions:

1. **(Exp1) Random**: \(N = 110\) During the second stage of the task when only 4 choices are presented (subsetMCQ), the choices are selected uniformly at random from the 50 options.

2. **(Exp2) Optimal**: \(N = 67\) During the second stage of the task (subsetMCQ), the choices are selected according to the optimization objective derived in Section 2. Data collected during the Random condition is used to fit the model parameters to be used for optimizing the subsets (described in detail next).

### 7.2 Evaluation

We propose two strategies for empirically assessing the quality of an MCQ test via two correlation metrics:

1. **Within-subject correlation** The performance of the subject at the first stage of the task (FullMCQ) serves as a ground-truth score of that subject’s knowledge of the domain. The correlation of the performance score of the same subject on the same set of questions, but with only a subset of choices, provides a measure of quality of the presented choice sets.

2. **Between-subject correlation** A good test should also discriminate between subjects of different levels of ability. If, for example, student A ranks higher than student B by their raw score on the fullMCQ, we should expect this ordering to be preserved if we were to instead rank the students based on their performance on the subsetMCQ test. We use Kendall Tau—a measure of rank correlation—on subsets of students ordered according to their performance in the fullMCQ and subsetMCQ tests.

### 8. RESULTS

#### 8.1 Within-subject correlation

Figure 5 compares the subjects’ scores according to their performance on the FullMCQ and subsetMCQ tests, split by condition: Random and Optimal, where performance is defined as the fraction of states that were named correctly in each test. Both plots indicate that subjects with a high score on one test also attain a high score on the other test, which is expected. The critical difference between the two conditions, however, is that of 40% of the subjects that attained a full-score (all correct) on the subsetMCQ in the Random condition, less than 4% of them attained a full score on the fullMCQ.

The subsetMCQ test where the choices are generated according to the Optimal strategy helps remove the full-score bias in the score distribution on the subsetMCQ test. Specifically, less than 17% of the subjects attain full score on the subsetMCQ designed according to the Optimal strategy. Additionally, Pearson’s correlation in the Optimal condition is 0.89, in contrast to 0.78 in Random.

#### 8.2 Between-subject correlation

We now focus on the quality of the ranking across subjects obtained on the subsetMCQ test between the Optimal and Random strategies. Our hypothesis is that a test designed to elicit maximum information about the subject’s knowledge should result in

![Figure 5: Within-subject correlation between raw score attained on the subsetMCQ and fullMCQ tests separated by choice set design strategy.](image-url)
Figure 6: Rank correlation between subjects according to the choice-set design strategy

Figure 7: Pairwise subject score differences in FullMCQ and subsetMCQ tests by choice-set design strategy. Each point in this graph is a pair of subjects. Difference in scores between the two subjects on one test in the Optimal condition exhibits less variance in predicting the score difference for the same pair of subjects on a different test.

Figure 8: Example of choice parameters $\beta_j$ for the question “Kansas” (heavy red border), and a selection of 3 optimal distractors for the mean population ability

...a higher quality discrimination across subjects of different levels of knowledge (abilities), and thus yield a more accurate ranking of the subjects. We obtain ranking of subjects obtained by sorting the students on their raw score on the subsetMCQ, and as in the within-subject analysis, evaluate it against the “ground-truth” ranking obtained by ordering the students by their raw score on the fullMCQ test. We compute ranking correlation by sampling a random set of 50 subjects and computing the Kendall Tau correlation score for the Random and Optimal conditions, repeating the process for 1000 iterations and report the statistics in Figure 6.

We observe that rank correlation in the subjects exposed to the Optimal condition significantly outperforms rank correlation of the subjects in the Random condition ($p$-value=0 by permutation test), confirming our hypothesis: the test that optimizes information about the student’s ability implicitly optimizes the accuracy of the ranking of the students.

9. APPLICATION: CROWDSOURCING TESTS FROM STACKEXCHANGE

One application that we explore in this paper is to the task of generating multiple choice tests from technical forum data. Technical forums, like StackExchange and similar QA websites like Piazza and Quora, exhibit a typical structure: (i) a user posts a question on the forum, (ii) other users propose solutions by submitting answers, and (iii) users vote on what they consider to be the best answer to the original question. Forums that follow this structure provide an opportunity to apply our model for optimal question generation where choice subsets are selected from the user submissions. Our hypothesis is that the benefit of creating assessment content dynamically from technical forums is three-fold:

- Large technical forums like StackExchange are repositories of real-world problems and solutions, where the solutions are of varying correctness and quality. A test generated from this data is likely to consist of relevant real-world problems.

- Choices created from real user submissions are likely to capture common misconceptions that other people are likely to share and thus, are potentially good distractors.

- The scope of sites like StackExchange could potentially facilitate test-generation customized to narrow target domains and subdomains of interest in areas for which testing material does not readily exist. One, for example, might want a test in a specific area of Parenting or a test that combines together multiple specific domains in programming.
• For each user $\phi_i \in S$:
  - Draw user ability $\phi_i \sim \mathcal{N}(0, \sigma^2_{\text{prior}})$
• For each answer $j$ created by user $i$:
  + Draw $\beta_{ij} \sim \mathcal{N}(\phi_i, \sigma^2_{\text{prior}})$
• Draw $\mu_\theta \sim \text{TruncNormal}(0, \sigma^2_{\text{prior}})$
• Draw $\sigma^2_{\phi} \sim \text{Inv-Gamma}(\alpha_{\text{prior}}, \beta_{\text{prior}})$
• For each question $i$:
  + For each vote for some answer of question $i$ at time $t$
    + Draw voter ability $\theta_{ik} \sim \mathcal{N}(\mu_\theta, \sigma^2_{\phi})$
    + Draw vote $z_{ik} \sim \text{Discrete}(\pi_{ik}^{(1)}, \ldots, \pi_{ik}^{(N)})$ where $C_i$ is a set of answers available for question $i$ at time $t$ and $\pi_{ik}^{(l)} = \exp(\theta_{ik} \beta_{ij}) / \sum \exp(\theta_{ik} \beta_{ij})$

### 9.2 Generative Model and Inference

We formalize the above model with a Bayesian generative story given above. We put normal priors on the answer and user parameters, and a truncated-normal prior on the voter ability, to ensure non-negativity. The high-level description of the story is as follows: users with ability $\phi_i$ contribute answers to questions whose correctness $\beta_{ij}$ is normally distributed about the creator’s ability, i.e., more able users are able to create higher-quality answers. When a vote on a particular answer is made, the voter with ability $\theta_{ik}$ at time $t$ observes a set of answers that exist at that time and makes a selection according to a discrete distribution parameterized by a softmax, where voters with greater ability are more likely to pick the best choice. We use Variational Message Passing for inference, a deterministic approximate posterior inference algorithm, implemented in the Infer.NET package [11]. We perform inference on three StackExchange forums: Biology (620 users, 638 questions), Physics (3,487 users, 5,262 questions), Parenting (1,820 users, 1,503 questions).

### 9.3 Examples

We present qualitative analysis of the results via examples in Figure 10, which provide some insight into the advantages and issues with applying our model to real-world forum data at the task of question generation. Full end-to-end evaluation of the quality and effectiveness of the generated questions will require user-studies, which we leave for future work. Figure 10 displays posteriors over answer correctness parameters for four questions, with the highlighted and annotated answers belonging to the optimal pair of answers, where optimality is determined by the optimality criterion given by Equation 4. As in the previous experiment, we optimize for answer sets for the “average user,” i.e., whose ability is given by the average posterior mean of $\theta$. In practice, the optimization can be carried out in an online, adaptive setting. We use the MAP estimates as parameters in the optimization objective. Finally, in selecting choice pairs, we require that the “most correct” choice (one with the highest posterior mean) always appears in the set, making the selection problem essentially one of finding a good distractor.

The examples in Figure 10 are given with their respective forum name and question ID, and can be easily viewed in more detail by finding them by going to a StackExchange page for that question. For example, for the top left question in Figure 10 (147346),
We have proposed a method for optimal choice selection in the task of optimal test design. Our response model is closely related to a discrete choice model, where the variance parameter encodes the ability of the user. This formulation, unlike related models such as [5, 1], allows us to explicitly optimize for optimal choice sets, where optimality is specified in terms of estimator efficiency such as (5, 1), allowing us to explicitly optimize for optimal choice the ability of the user. This formulation, unlike related models

10. DISCUSSION

We focus on the limitations and extensions of the current model to use-cases where users interactively submit answers, which are then re-used as choices in a multiple choice test, similar to the technical forum setting described in the previous section. The application scenario may be a large MOOC-like classroom, where hundreds of students submit answers to a question, and these answers are to be re-used either for the same or the next batch of students in the form of a well-designed multiple choice question. Two important challenges arise in this setting:

- Independence of irrelevant alternatives Many students are likely to submit similar answers, either lexical, syntactic or semantic variations. In such cases, the assumption that choices are made independently is violated. If the same optimization objective were to be used to select optimal choices when some of the choices are identical, the resulting choice sets would likely feature such identical choices. A model that explicitly clusters or models the correlation structure between choices is needed. This aspect constitutes our ongoing work.

- Exploitation vs. exploration of choices In a big classroom, where new answers arrive constantly, it is important that in addition to optimizing the choice set to learn best about the user, the model learns about the choice parameters themselves. The right system must balance between exploiting the choices

Figure 10: Example StackExchange questions with posterior distributions over choice correctness parameters. Two optimal choices are highlighted and annotated. See text for more detail.

can be found at: http://physics.stackexchange.com/questions/147346. Questions 14736, 14609 and 776 are examples where the distractors are all plausible incorrect answers (the correct answer in every question is marked with “A”). Question 8996, however, is a common example of a generated choice set, where the distractor is also a correct answer, yet it appeared less popular for another reason, e.g., it was incomplete, had little supporting evidence, or was simply not a commonly-known answer (the case for question 8996) and therefore received significantly fewer votes. In our setting, we argue that having an explicit constraint that the distractor is wrong is not necessary—it is sufficient if the user can tell apart the best answer from the remaining answers. However, if the dimension of quality is orthogonal to correctness, e.g., if one of the answers is better phrased or contains additional illustrations, the question will not serve its purpose in differentiating those users that know the answer from those that do not. This limitation is potentially less severe in areas where the answer is constrained to be of a particular format, e.g., if the answer is computer code like in StackOverflow, where often multiple submitted answers are correct, but one exhibits the best performance. We leave the full study of the application of this model to test generation from technical forums for future work.
that are optimal for learning about the user, and choices whose parameters are uncertain.

Finally, we believe that the most promising application of models such as the one proposed here will be in the area of data-driven test design, where the data is in the form of existing questions and answers, found in technical forums. As such places on the web can only be expected to grow in the immediate future, so will the need to leverage real-life problems and solutions to create relevant, real-world tests in the rapidly growing and evolving technical domains.

12. APPENDIX

12.1 A. Derivation of Optimal choice set

We derive the discrete optimization objective for choosing optimal sets of answers by maximizing Fisher information of the $\theta$:

$$\mathcal{J}(\theta;C') = \mathbb{E}\left[ \frac{\partial^2}{\partial \theta^2} \log f(\theta;C') \mid \theta \right]$$

$$= \mathbb{E}\left[ \frac{\partial}{\partial \theta} \left( \delta(z = k) \theta g_k - \sum_{k} \frac{\beta_k \exp(\theta \beta_k)}{\sum_{k} \exp(\theta \beta_k)} \right) \right]$$

$$= -\sum_{k} \frac{\beta_k^2 \exp(\theta \beta_k)}{\sum_{k} \exp(\theta \beta_k)} + \sum_{k} \frac{\beta_k \exp(\theta \beta_k)}{\sum_{k} \exp(\theta \beta_k)^2} \frac{\sum_{k} \exp(\theta \beta_k)}{\sum_{k} \exp(\theta \beta_k)^2} = -\frac{\sum_{k} \beta_k^2 \exp(\theta \beta_k)}{\sum_{k} \exp(\theta \beta_k)^2} + \frac{\sum_{k} \beta_k \exp(\theta \beta_k)}{\sum_{k} \exp(\theta \beta_k)^2}$$

$$= \frac{\sum_{k} \beta_k \exp(\theta \beta_k) + (\sum_{k} \beta_k \exp(\theta \beta_k))^2}{\sum_{k} \exp(\theta \beta_k)^2}$$

$$= \frac{\sum_{k} \beta_k \exp(\theta \beta_k) + \sum_{k} \exp(\theta \beta_k)^2}{\sum_{k} \exp(\theta \beta_k)^2} = \sum_{k} \exp(\theta \beta_k) = \sum_{k} \exp(\theta \beta_k) = \sum_{k} \exp(\theta \beta_k)$$

$$= \sum_{k} \exp(\theta \beta_k) = \sum_{k} \exp(\theta \beta_k)$$

$$\sum_{k} \exp(\theta \beta_k)^2$$

Claim: The information on $\theta$ is maximized when half of the choices are placed at one end of the interval, and half of the choices are placed at the other end.

Proof: Consider an arbitrary placement of $\{\beta_k\}$ on a unit interval. Pick any $\beta_k$. Moving $\beta_k$ to either end of the interval will guarantee an improvement of the objective. To see that let $x = \beta_k$. We are then interested in finding $x$ that maximizes (ignoring terms that do not depend on $x$):

$$\sum_{k} (\beta_k - x)^2$$

This is a quadratic function with a minimum at $x_{\text{min}} = \frac{\sum_{k} \beta_k}{K+1}$. Since $0 \leq \beta_k \leq 1$, we have that $0 \leq x_{\text{min}} \leq 1$, i.e. regardless of the placement of the other choices, the objective is always improved by moving $\beta_k$ to one of the endpoints. Suppose that we place $n_0$ choices at the $\beta = 0$ endpoint and $n_1$ choices at the $\beta = 1$ endpoint. The objective value is then given by $\min(n_0, n_1)$, which is maximized when half of the choices are placed on one end and half of the choices on the other (if number of choices is odd, placement of one choice is arbitrary). □

13. REFERENCES


