Optimally Discriminative Choice Sets in Discrete Choice Models: Application to Data-Driven Test Design

Igor Labutov
Electrical and Computer Engineering
Cornell University
Ithaca, NY
iil4@cornell.edu

Frans Schalekamp
Operations Research and Information Engineering
Cornell University
Ithaca, NY
fms9@cornell.edu

Kelvin Luu
Department of Computer Science
University of Washington
Seattle, WA
kellu@cs.washington.edu

Hod Lipson
Department of Mechanical Engineering
Columbia University
New York, NY
hod.lipson@columbia.edu

Christoph Studer
Electrical and Computer Engineering
Cornell University
Ithaca, NY
studer@cornell.edu

ABSTRACT

Difficult multiple-choice (MC) questions can be made easy by providing a set of 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) that enable the assessment of the students' understanding. The art of MC 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. In this work, we (i) develop a multinomial-logit discrete choice model for the setting of MC testing, (ii) derive an optimization objective for selecting optimally discriminative option sets, (iii) propose an algorithm for finding a globally-optimal solution, and (iv) demonstrate the effectiveness of our approach via synthetic experiments and a user study. We finally showcase an application of our approach to crowd-sourcing tests from technical online forums.

Keywords

Adaptive Learning, Assessment, Crowdsourcing, Optimal Testing

KDD '16, August 13-17, 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2939672.2939879

CCS Concepts

•**Applied computing** → **Computer-assisted instruction;**

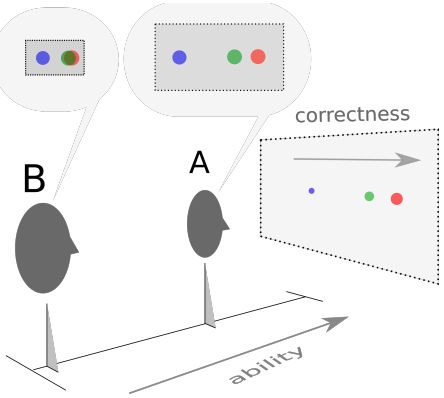

Figure 1: Geometric intuition behind the choice of an optimal set of options in a multiple choice test: we imagine the subject's distance from the wall to be inversely related to their "ability to see." If asked which colored dot painted on the wall is the right-most dot, subjects closer to the wall would be more likely to answer the question correctly, and subjects farther away would be most likely to guess. The question of optimal choice set design can then be posed as: "where on the wall do we paint the dots, such that I best learn about your distance to the wall based on your answers?"

1. INTRODUCTION

The design of a good set of options in multiple-choice questions (MCQs) is notoriously difficult [15]. 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 [11, 20], 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, i.e., to select 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 we are 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 (though test bank optimization is an exception, see Chapter 7 in [5]) . The main contributions of our work are summarized as follows:

- We propose an objective function for selecting an optimal set of choices in a discrete choice model, given the estimated user ability, and we investigate the solutions across different regimes of student ability.
- We propose an algorithm for finding a globally-optimal option choice set.
- 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 U.S. 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

2.1 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 (i.e. does the question measure learning outcomes?) [8, 7], to decide on the optimal number of choices [15, 9], and to design good distractors [10, 7]. In an empirical study [8], Haladyna and Downing concluded that the key in multiple choice item design was "not the number of distractors but the quality of distractors." They find that, almost unanimously, high-quality distractors are considered to be those that represent common student misconceptions [9]. Thissen et al. [17] developed 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 a posteriori analysis of multiple choice items. Computational methods have been proposed for the task of multiple choice item design (i.e. designing a question and its choices), but are restricted to specific domains, such as vocabulary [3], grammar testing [2], or topic-specific comprehension [13]. For all these methods, however, distractors are generated automatically based on the structure of the problem domain. We are unaware of prior results that directly optimize for a distractor choice set based on the data of past student submissions.

2.2 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 [11, 20]), where the optimization objective is related to the estimator efficiency, typically of student ability (see Chapter 7 of [5] for an overview). More recently, methods based on the principle of estimator efficiency have been applied to the task of test-set reduction [18] in the context of a multidimensional extension of the Rasch model (SPARFA) [19]. We can view choice-set optimization as a non-trivial generalization of optimal test-design that was traditionally explored in the setting of item-set optimization only. We argue that this extension will become particularly relevant in rapidly growing, data-intensive educational settings, where a subset of real student submissions can be efficiently selected into an optimal distractor set for a student of a specific ability level.

3. MODEL

In formulating our model, we require it to exhibit the following three properties:

Property 1 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 of greater ability are more likely to pick the most correct option (we will discuss this aspect in detail below).

Property 2 A "perfect" student (with the highest attainable *ability*) chooses the correct option with probability 1.

Property 3 A student with the lowest attainable *ability* makes their choice uniformly at random.

For simplicity, we require that there is exactly 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 correct 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 *apparent correctness* of each choice β_j and a non-negativity constraint on the student's *ability* θ_i, exhibits all of the properties above. Specifically, we use the following statistical model:

$$P(i \text{ picks option } j \mid \theta_i, \{\beta_j\}_{j \in C}) = \frac{\exp(\theta_i \beta_j)}{\sum_{j' \in C} \exp(\theta_i \beta_{j'})}, \quad (1)$$

where j is the option index, $\beta_{j^*} > \beta_j$, $\forall j \in C \setminus j^*$, and j^* is the correct option. Furthermore, we assume $\theta_i \geq 0$, $\forall i$, where θ_i is the ability of student i, and $\{\beta_j\}_{j \in C}$ is the set of option parameters presented to the student, encoding the *apparent correctness* of each option. Without an explicit partial order constraint on the choices and a non-negativity constraint on the students, the model would capture the relative preference of subjects towards choices. In psychometrics this model is known as the nominal response model [5] and is also related to the more general multidimensional unfolding models [4, 16] often used to investigate the relationship between subjects and preferences. In our setting, the non-negativity constraint on the ability θ_i, combined with the partial order constraints on the option parameters $\{\beta_j\}$ are critical to obtaining the desired interpretation of the ability parameter θ_i, namely as

capturing the *ability* of the student (larger values indicate greater ability). One can easily verify that Property 2 and Property 3 are both satisfied by considering the limiting behavior of (1) when $\theta_i = 0$ and $\theta_i = \infty$ respectively. Property 1 is satisfied as a result of $P(i \text{ picks option } j^* \mid \cdot)$ (i.e., the probability of student i picking a correct option) being a monotone function of θ_i. As a consequence, 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 also important to understand the limitations and additional assumptions underlying this model. The most significant limitation is what is known as the *independence of irrelevant alternatives* (IIA) assumption [14]. The IIA assumption is violated whenever the two options are not inherently different. For example, in the setting of reusing student responses as potential options in a test, this would occur if the two options are either completely identical or are paraphrases of each other. We leave dealing with the problem of IIA to future work.

To place our model in the context of existing work, we compare it with two closely related models: the classical Rasch model [5] and the recent model proposed 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' ability, i.e., it is agnostic to the actual choice made by the student in an MCQ setting. The likelihood of student i with ability θ_i getting the question j with difficulty q_j correct is given by:

$$P(i \text{ correctly answers } j \mid \theta_i, q_j) = \frac{1}{1 + \exp(-(\theta_i - q_j))}.$$

To gain intuition about how our model encodes question *difficulty*, consider the case of only two options: the correct option with parameter β_{j^*} and the incorrect option with parameter β_j. We can now express the likelihood of the student answering this question correctly using our model as follows:

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

where $\Delta_{j^*-j} = \beta^*_j - \beta_j$, which is positive by definition (since $\beta_{j^*} > \beta_j$). By analogy with the Rasch likelihood, $\Delta^{-1}_{j^*-j}$ captures a similar notion of question *difficulty*: the farther apart are the two options in the parameter space, the "easier" is the resulting question.

When the question contains more than two options, the likelihood of the student answering the question correctly can be expressed as:

$$P(i \text{ right on } j \mid \theta_i, \{\Delta_{j^*-j}\}_j) = \frac{1}{1 + \sum_{j \in Q} \exp(-\theta_i \Delta_{j^*-j})},$$

where an exponential term containing the distance Δ_{j^*-j} between the correct option and every remaining option now appears in the denominator. Observe that the probability of getting the question right approaches one only when the correct option parameter (scaled by ability θ_i) is well-separated from every other option (distractor). An important advantage offered by modeling individual choices is that the model's estimate of the students' abilities will not only depend on which questions were answered correctly and incorrectly, but also on the nature of the incorrect answers chosen. Consequently, our model could distinguish between the abilities of two students, even if both of these students answered all questions incorrectly.

3.2 Relationship to 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 (simplified) generative process of their model is defined as follows:

$$z_{ij} \sim P(i \text{ correctly answers } j \mid \theta_i, q_j)$$

$$P(i \text{ picks option } k \mid z_{ij}, \pi_k) = \begin{cases} \pi_k & \text{if } z_{ij} = 1 \\ 1/K & \text{otherwise,} \end{cases}$$

which can be interpreted as a mixture model of two components: (i) if the student answers the question correctly ($z_{ij} = 1$), the student picks option k with probability π_k and (ii) if the student answers the question incorrectly ($z_{ij} = 0$), the student picks an option uniformly at random (i.e., $\pi_k = 1/K$ where K is the number of options). The probability of the student answering correctly $P(z_{ij} = 1 \mid \theta_i, q_j)$ is parametrized by the standard Rasch model described in Section 3.1[1].

While both our model and the model by Bachrach *et al.* can be used for the task of estimating student and question parameters in the absence of annotated data (i.e., answer key), the fundamental distinction lies in the capability of our formulation to be used for the task of optimal choice set design—a task that is not feasible with the model by Bachrach *et al.* The underlying reason for this distinction is because our model implicitly couples choice parameters and question difficulty (see Section 3.1), allowing us to tune question difficulty to students of varying ability levels by optimizing over choice sets. In contrast, in the Bachrach *et al.* model, the question 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 formulate the problem of optimal choice set design as active learning—query a user (student) with an instance (choice set) such that the expected outcome (student's answer) maximizes information about the unknown parameters (student ability). Given a question's complete set of potential answer options Q, a student with ability θ is presented with a subset $C \subseteq Q$. We are interested in finding a subset $C^* \subseteq Q$ that is optimal in some sense for the user with a given ability. Specifically, we are interested in choosing C that results in the smallest variance of the maximum likelihood estimator of θ, which is equivalent to C with the maximum Fisher information w.r.t. θ:

$$C^* = \operatorname{argmax} \mathcal{I}(\theta; C) \tag{2}$$

where Fisher information of set C, $\mathcal{I}(\theta; C)$, is given by

$$\mathcal{I}(\theta; C) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(\theta; C) \middle| \theta \right]. \tag{3}$$

[1]This is a slight oversimplification of the original model (ignoring question "discriminability" parameter) proposed by Bachrach *et al.*, but captures its key aspects for our purpose.

Here, $f(\theta; C)$ is the likelihood function in (1). It can be shown that the solution to the above is the following combinatorial optimization problem[2]:

$$\begin{aligned} \underset{\{x_n\}}{\text{maximize}} \quad & \frac{\sum_i^N \sum_{j>i}^N x_i x_j (\beta_i - \beta_j)^2 \exp(\theta[\beta_i + \beta_j])}{\sum_i^N \sum_j^N x_i x_j \exp(\theta[\beta_i + \beta_j])} \\ \text{subject to} \quad & x_n \in \{0,1\}, \forall n \in Q \\ & \sum_{n=1}^{N} x_n \leq K, \end{aligned} \tag{4}$$

where $\{x_n\}_{n=1\ldots N}$ are indicator variables ($x_n \in \{0,1\}$) that select choices from Q to be included in C, $N = |Q|$ (i.e., the total number of potential options) and K is the maximum permissible size of C (e.g., four options).

4.1 Asymptotically optimal choices

We now investigate the nature of the optimal choice sets. Consider two limiting cases: a student with a large ability ($\theta_i \to \infty$), and a student with a low ability ($\theta_i \to 0$).

Case $\theta_i \to \infty$: It is straightforward to show that in the limit of "infinite ability," the information will go to zero. However, the rate at which it goes to zero depends on the choice set, allowing us to gain insight into the kinds of choice sets that will be "preferred" for users with a large ability. The logarithm of the information function will have a linear asymptote, with the slope dominated by the largest exponential in the numerator and the denominator. We can show that as $\theta \to \infty$, only the two choices with the largest values of β remain relevant (i.e. $\{\beta_{max}, \beta_{max-1}\}$), with the optimal spacing between them, $\beta_{max} - \beta_{max-1}$, given by:

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

Clearly, the greatest Fisher information for large values of θ will be obtained when $\beta_{\text{max}-1} \approx \beta_{\text{max}}$, i.e., when the distance between the two top choices approaches zero.

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

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

where K is the number of options we seek to display to the student and k indexes over those options. The solution to the above can be obtained by choosing a subset of the choices from Q with the smallest β ("left-most" or "incorrect" choices) and a subset of choices from Q with the largest β ("right-most" or "correct" choices) (proof omitted). The intuition behind this solution requires some explanation. It is instructive to consider the optimal solution in the case of only two choices. The optimal "spacing" between the correct choice and the distractor (Δ_{ij}) will lie somewhere between 0 and ∞, but where exactly depends on our prior belief about the ability of the student (θ). An intuitive interpretation of this solution can be gained by relying on a related notion of *information gain*: the expected distance (KL-divergence) between the prior and the posterior (after observing the choice) on θ (ability). Information gain exhibits the same limiting behavior: when the two choices are infinitely far apart ($\Delta_{ij} \to \infty$), the student will always pick the correct option regardless of their ability—thus, the posterior will not be updated as a consequence of their choice (hence, no information gain). In the extreme of the two choices spaced very close together, the student will always "flip a coin" between them, again giving away no information about their ability. It is this last scenario that will be fundamental to understanding the optimal choice set (with more than two choices) when $\theta = 0$.

Consider now introducing additional choices into the choice set. Appealing to the *information gain* interpretation, we again consider the prior-posterior gain of each potential choice (there are K of them now). As in the case with only two choices, the prior-posterior gain for each option will be non-zero if the student has a "more than a coin-flip" chance of choosing the better option (i.e., giving the student an opportunity to demonstrate their ability), from which it follows that the remaining options must be sufficiently far apart for a student with $\theta \approx 0$. Because under the prior of $\theta = 0$ each outcome (choice) is equally likely, the expected information gain is a sum of of such prior-posterior gains. It follows then that the spacing configuration that maximizes total inter-choice distance will also maximize the expected information gain. It also explains why there should be a large "dead-zone" (i.e., no other choices) between the choices separated at the "correct" and the "incorrect" extremes: inserting even a single choice in the middle will result in the student with $\theta \approx 0$ flipping a coin between the choices at the "correct" extreme and the choice in the middle, neutralizing all of the prior-posterior gain.

4.2 Optimization algorithm

Although problem (4) is a non-linear combinatorial optimization problem, we show that it can be transformed into a series of integer linear programs (ILPs) which can be used to find a globally optimal solution. See Appendix A for the details and analysis of our algorithm.

5. SYNTHETIC EXPERIMENTS

5.1 Parameter learning

The simulation is performed as follows: 100 student ability parameters (θ_i) are sampled from a uniform distribution; 50 questions with 20 options each are generated, where each option parameter β is independently sampled from a zero-mean normal distribution. We evaluate a range of variances for the distribution over choice parameters and study its effect on the quality of the inferred parameters.

We summarize the performance of the inference algorithm via (i) rank correlation of the inferred and ground truth rankings of students and (ii) the accuracy in identifying the correct answers in questions. We use Kendall Tau as a metric of rank correlation. Kendall Tau returns a quantity in the range $[-1, +1]$, where $+1$ indicates perfect correlation (every pair of students in both rankings is in a consistent order), -1 when the rankings are inverted, and 0 when the rankings are not correlated. In predicting the correct answer for a question, recall that in our model, the choice with the largest parameter β is interpreted as the correct answer (see Section 3). Accuracy in predicting correct answers, therefore, is defined as a fraction of questions where the predicted correct answer matches the ground-truth correct answer.

Figures 2(a) and 2(b) depict accuracy and rank correlation as a function of the number of choices (i.e., multiple choice options) presented in each question, and as a function of the

[2]derivation ommited due to space limitation

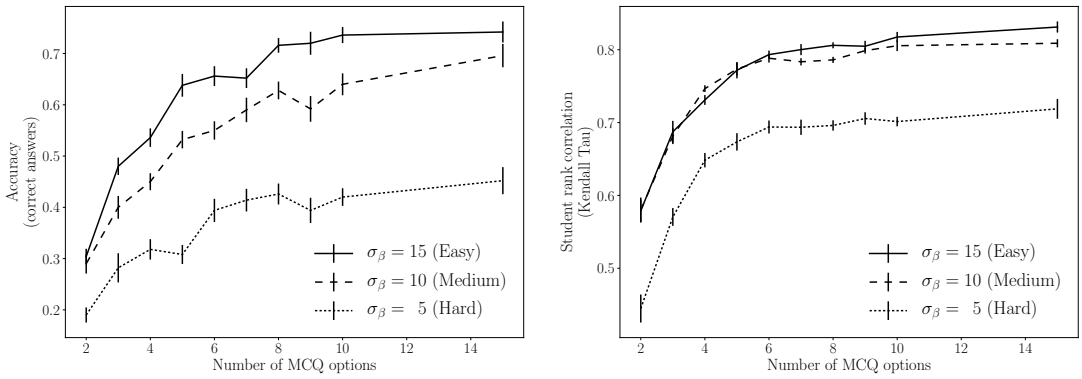

Figure 2: Performance (simulation) of the model in (a) predicting the correct answers in a set of questions and (b) ranking students by their ability, as a function of (i) number of choices shown in each question and (ii) variance of the choice parameter distribution (shown as standard deviation σ_β). "Easier" questions correspond to those with a "wider" spread between choice parameters (i.e., higher variance). We can conclude the following on the basis of these results: (i) more choices improves performance in both, predicting correct answers and ranking students, but with diminishing returns, and (ii) showing "easier" questions generally improves performance in correct answer prediction and ranking, however, the model is able to rank well even when the correct answers are more difficult to identify (see Section 5). Note that the random baseline for accuracy in (a) is 5% as there are 20 choices for each question in the simulation.

variance of the distribution over choice parameters β. Recall that the variance of the distribution from which we sample the choice parameters β is inversely related to the difficulty of the resulting question. As we discussed in Section 3.1, the question becomes "easy" (i.e., students of lower θ will have a high probability of getting it right) when the choice parameters are "spread out" (which is achieved when the choices are sampled from a high-variance distribution). Both Figure 2(a) and Figure 2(b) indicate that (i) more choices result in better performance (higher accuracy in identifying correct answers and higher rank correlation between the true and inferred student rankings), and (ii) "easier" questions (i.e., questions whose choice parameters are sampled from a high-variance distribution) generally result in better accuracy and rank correlation.

It is worthwhile to analyze the observation that student rank-correlation (Figure 2(b)) remains the same between the "Easy" and "Medium" conditions, while accuracy (Figure 2(a)) drops considerably. This can be attributed to the fact that in inferring the ability parameter of a student, the model relies jointly on the parameters of every choice in the set, i.e., not only on whether the chosen option was correct. As a result, while the ordering of the top two choices may be incorrect (resulting in an incorrect prediction of the correct answer), the remaining choices still play an important role in inferring student parameters (and thus in the quality of the ranking).

5.2 Optimal choice sets

We now evaluate the choice subset selection optimization objective introduced in Section 4. We again generate a simulated classroom with 50 students and 50 questions[3]. In contrast to the experiment in Section 4, here we perform parameter inference sequentially after each student answers

[3]Student and choice parameters were sampled from uniform distributions with support $(0, 1)$ and $(0, 100)$ respectively

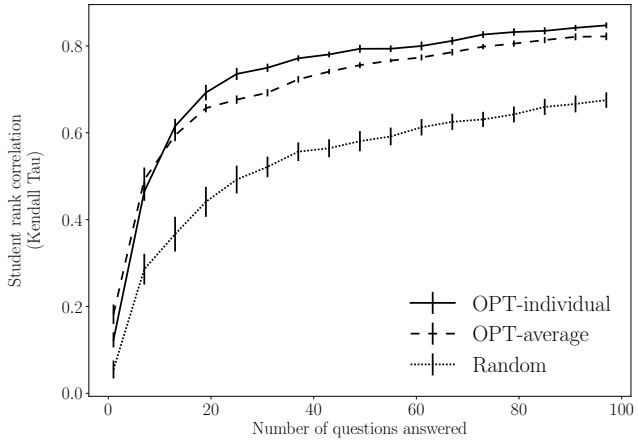

Figure 3: Rank correlation between the true and inferred student rankings as a function of the number of questions answered by each simulated student, separated by the choice sampling strategy. Optimizing choice sets according to the proposed objective (OPT-average and OPT-individual) results in better rank correlation with fewer questions compared to when the choice sets are sampled randomly. Optimizing choice sets according to the individual student abilities (OPT-individual) marginally improves performance over optimizing choice sets based on the average student ability (OPT-average).

a question, simulating an adaptive testing scenario. For every question, we sample choice sets of size 2 according to three different sampling strategies: (i) **random**: choices are drawn uniformly at random, (ii) **OPT-individual**: the optimal choice set is selected for each student according to that student's estimated ability parameter, and (iii) **OPT-average**: the optimal choice set is selected according to the average estimated ability of the student population (i.e. choice sets are identical for each student). Figure 3 compares the performance across the three conditions using the student rank correlation metric introduced in Section 5. On the basis of these results, we draw the following conclusions: (i) presenting choice sets optimized using the objective introduced in Section 4 with the inferred parameters achieves significantly better rank-correlation and with fewer questions than when the choice sets are sampled randomly; (ii) optimizing choice sets based on the individual student parameters marginally improves performance over optimizing choice-sets to the average ability of the student population. Note, however, that in practice the exact gains will vary depending on the nature of the student and choice parameter distributions.

6. 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 person is required to identify by picking a correct choice out of a set of options (other states). This problem serves as an excellent platform for evaluating our model for two reasons:

1. **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 user's knowledge in this domain. Any other test based on only a subset of the options (and consequently a method used to obtain the options) can be evaluated against this "ground-truth" by measuring the correlation of the two scores.

2. **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. This provides an opportunity for a data-driven method to excel in finding "good" choice sets for building effective questions.

6.1 Data collection

Mechanical Turk workers residing in the U.S. were solicited to a task titled "How well do you know U.S. states?", which was briefly described as a quick quiz to test one's knowledge of the U.S. states, consisting of two stages:

1. **Stage I (fullMCQ)**: Workers 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 choose from. This selection is made for every one of the 50 states, presented in random order. Workers are not revealed the correct answer, and are discouraged from looking up the answers externally.

2. **Stage II (subsetMCQ)**: The same workers then repeat 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 for how the multiple choice options were sampled:

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 introduced in Section 4. Data collected during the **Random** condition is used to fit the model parameters to be used for optimizing the subsets. The subsets are optimized for the average ability of the users in the **Random** condition (this corresponds to the **OPT-average** strategy introduced in Section 5).

6.2 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 worker in the first stage of the task (**FullMCQ**) serves as a ground-truth score of that worker's knowledge of the domain. The correlation of the performance score (fraction of correctly identified states) of the same worker on the same set of questions, but with only a subset of the choices, provides a measure of quality of the presented choice sets.

2. **Between-subject correlation** A good test should also discriminate between workers of different levels of ability. If, for example, student A ranks higher that student B according to 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 students ordered according to their performance on the **fullMCQ** and **subsetMCQ** tests.

7. RESULTS

7.1 Within-subject correlation

Figure 4 compares the workers' 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 workers 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 the 40% of the workers 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 workers 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**.

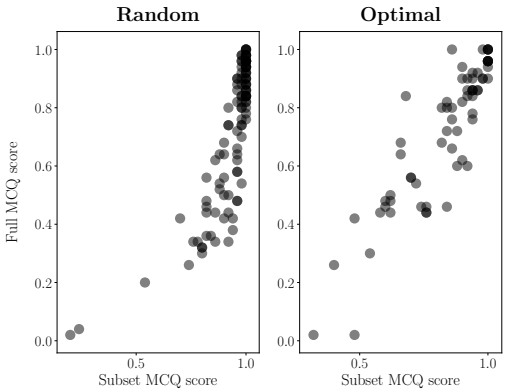

Figure 4: Within-subject correlation between raw scores attained on the subsetMCQ and fullMCQ tests separated by choice set design strategy—choice sets optimized according to the proposed objective yield better within-subject score correlation than choice sets sampled randomly.

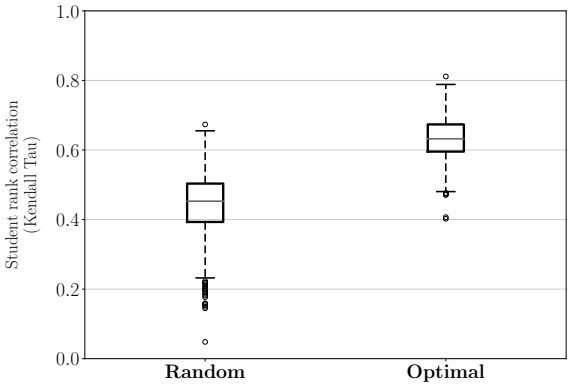

Figure 5: Rank correlation between workers ranked according to the raw scores attained on the subsetMCQ and fullMCQ tests, separated by choice set design strategy – choice sets optimized according to the proposed objective yield better rank correlation than choice sets sampled randomly.

7.2 Between-subject correlation

We now focus on the quality of the workers' ranking using the raw scores obtained on the **subsetMCQ** test between the **Optimal** and **Random** strategies. Our hypothesis is that a test designed to elicit maximum information about the worker's knowledge should result in a higher quality discrimination across workers of different levels of knowledge (abilities), and thus yield a more accurate ranking of the workers. We obtain a ranking of workers by sorting everyone according to 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 rank correlation by sampling a random set of 50 workers and computing Kendall Tau for the **Random** and **Optimal** conditions, repeating the process for 1000 iterations and report the statistics in Figure 5.

We observe that rank correlation in the workers given a **subsetMCQ** test with the **Optimal** choice set significantly outperforms rank correlation of the workers given a **subsetMCQ** test with a **Random** choice set (p-value=0 by permutation test), confirming our hypothesis: *a test that optimizes information about the student's ability implicitly optimizes the accuracy of the ranking of the students.*

8. CROWDSOURCING TESTS FROM FORUMS

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, 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. The potential benefit of creating assessment content dynamically from technical forums is:

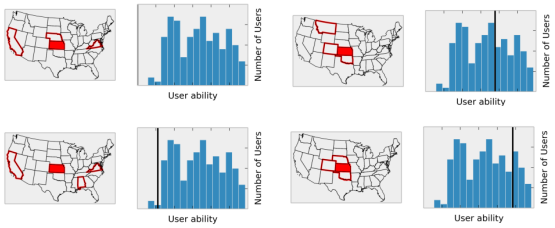

Figure 6: Visualization of optimal choice sets for the question "Kansas", optimized to students of varying prior ability parameter (black vertical bar, displayed over the empirical distribution of inferred student abilities). Observe that as ability increases, choices become clustered closer to the true answer, making the correct answer more difficult to discern.

1. 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.

2. 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.

8.1 Modeling users and questions

We describe how we adapt our model to the setting of a generic technical online forum that fits the structure described above, i.e., it contains user-submitted questions, user-submitted answers and user votes for each answer. Exactly as in the problem of "U.S. States Quiz," we endow each choice (answer post) with a real-valued parameter β_{ij}, but where in this case i is an index of the user that contributed that answer and j is an index of the question which this answer answers. For modeling convenience, we explicitly distinguish between users that contribute an answer, and users that vote

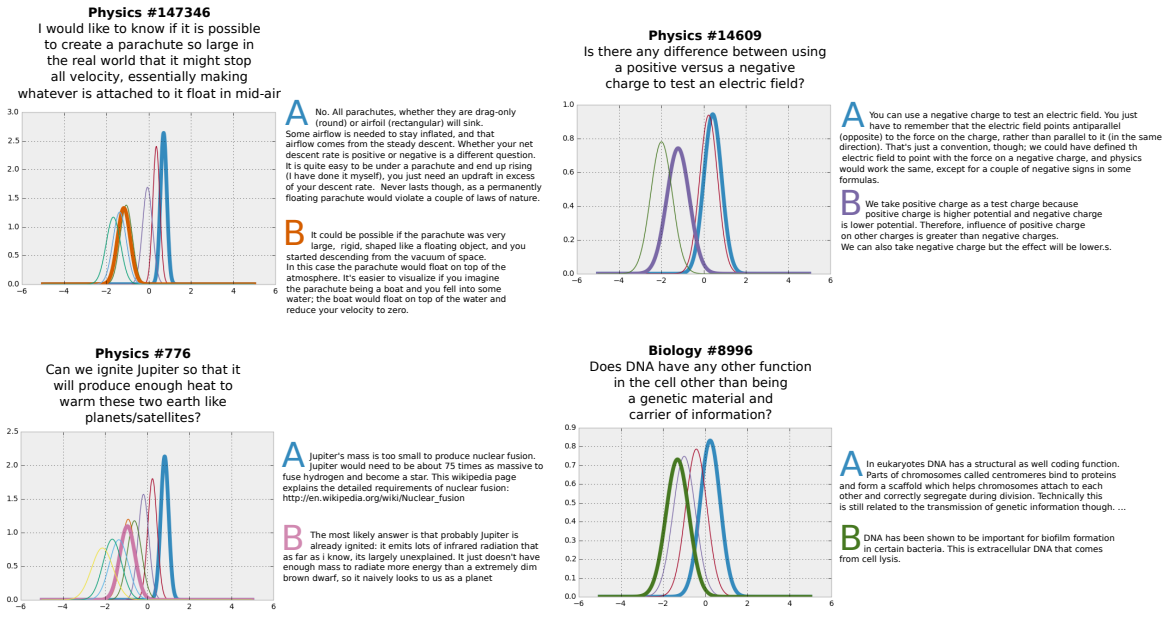

Figure 7: Example StackExchange questions with posterior distributions over choice correctness parameters. Two optimal choices are highlighted and annotated. See Section 8.3 for details.

for a particular answer. "Voting users" are modeled the same way as the users who answer multiple choice questions in our discrete choice model, i.e., their strictly positive "choosing ability" θ appears as a coefficient of the choice correctness in parametrizing the discrete distribution over choices[4]. "Contributing users" are endowed with an "answering ability" parameter ϕ, which parameterizes the distribution over answer correctness parameters for answers contributed by that user. This allows us to share statistical strength of "good" and "bad" answers that are created by the same users, e.g., users that contribute poor answers in general (answers that receive few upvotes) will be informative in inferring answer parameters in other questions they answered, where the voting information may be sparse.

8.2 Generative Model and Inference

We formalize the above model with a Bayesian generative story shown on the right. 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 ϕ_i contribute answers to questions whose correctness β_{ij} is normally distributed about the creator's ability, i.e., more able users are able to create higher-quality answers. Later at some time t, a voter with ability θ_k observes a set of answers C_q^t (to question q) that have been created up to time t and makes a selection according to the discrete distribution parametrized by (1), 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, provided in the Infer.NET package [12]. 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).

For each user $i \in S$:

- Draw user ability $\phi_i \sim \mathcal{N}(0, \sigma^2_{prior})$

  For each answer j created by user i:
  - Draw $\beta_{ij} \sim \mathcal{N}(\phi_i, \sigma^2_{prior})$

Draw $\mu_\theta \sim \text{TruncNormal}(0, \sigma^2_{prior})$

Draw $\sigma^2_\theta \sim \text{Inv-Gamma}(\alpha_{prior}, \beta_{prior})$

For each question q:

- **For each** vote in question q at time t
  - Draw voter ability $\theta_{qk} \sim \mathcal{N}(\mu_\theta, \sigma^2_\theta)$
  - Draw vote $z_{qk} \sim \text{Discrete}\left(\{\pi_{qk}^{(i,j)}\}_{(i,j)\in C_q^t}\right)$ where C_q^t is a set of answers available for question q at time t and

$$\pi_{qk}^{(i,j)} = \frac{\exp(\theta_{qk}\beta_{ij})}{\sum_{(i',j')\in C_q^t}\exp(\theta_{qk}\beta_{i'j'})}$$

8.3 Examples

We present a qualitative analysis of the results via examples in Figure 7, which provide some insight into the advantages and issues with applying our model to real-world forum data

[4]unfortunately StackExchange datasets do not reveal the identity of the "voters", thus we assume that each vote is contributed by a distinct "voter"

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 7 displays posteriors over answer correctness parameters for four questions, with the highlighted and annotated answers belonging to the optimal choice set, where optimality is determined by the optimality criterion introduced in Section 4. As done in Section 6.1, we optimize the choice sets for an "average user," i.e., whose ability is given by the posterior mean of θ. 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 7 are given with their respective forum name and a question ID, and can be viewed in more detail by finding them on the StackExchange site. For example, the top left question in Figure 7 (147346), 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 may be correct, but only 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.

9. DISCUSSION

We have proposed a method for optimal choice selection for 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 identify optimal choice sets, where optimality is specified in terms of estimator efficiency on the user ability parameter. We have demonstrated that the resulting choice sets are selected on the basis of how easily the choices are mistaken for one another, highlighting one of the principles of multiple choice question design: *good distractors must capture common misconceptions.* We also look ahead to the application of this model to data-driven crowd-sourced assessment generation from technical forums, and briefly highlight challenges and potentials of this paradigm.

Acknowledgements

The work of I. Labutov was supported in part by a grant from the John Templeton Foundation provided through the Metaknowledge Network at the University of Chicago. The work of C. Studer was supported in part by Xilinx Inc. and by the US NSF under grants ECCS-1408006 and CCF-1535897. Computational resources were sponsored in part by grants from Amazon and Microsoft.

APPENDIX

A. A PRACTICAL ALGORITHM TO FIND AN OPTIMAL CHOICE SET

The mathematical programming formulation (4) for finding the optimal choice set has binary decision variables and even when these are relaxed to take on real values between 0 and 1, the objective function is a nonlinear function. In this section we describe a practical way of finding an optimal solution, which uses an Integer Linear Programming (ILP) solver as a subprocedure. The idea is to introduce new binary variables that represent the product of two decision variables (which can be enforced using linear constraints), replace the objective function by just the numerator of the original objective function, and add a constraint that bounds the denominator of the original objective function. This problem given in (5) is an ILP, for which we invoke the subroutine. The (basic) procedure now is the following: an upper bound on the denominator is given (at first infinity), the best solution is found, given that bound (which can be found using an ILP), then the bound is lowered to slightly below the denominator given by the current solution. This is repeated until all possible denominators are considered. The best overall solution is kept.

$$
\begin{aligned}
\text{maximize} \quad & z = \sum_{i,j:i<j} y_{ij}(\beta_i - \beta_j)^2 \exp(\theta[\beta_i + \beta_j]) \\
\text{subject to} \quad & \sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j]) \leq B \\
& y_{ij} \leq x_i, \forall i, j \\
& y_{ij} \leq x_j, \forall i, j \\
& y_{ij} \geq x_i + x_j - 1, \forall i, j \\
& \sum x_i \leq K \\
& x_i \in \{0, 1\}, \forall i \\
& y_{ij} \in \{0, 1\}, \forall i, j
\end{aligned}
\tag{5}
$$

The Algorithm

1. Set $\delta = 2\exp(\min_i \beta_i^2)$.

   Set $B \leftarrow \infty$, and solve ILP (5). (The current solution is denoted by y_{ij} and z.)

2. Let $B_{\text{eff}} \leftarrow \sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j])$, let $r_{\text{best}} \leftarrow z/B_{\text{eff}}$, and let $B_{\text{best}} \leftarrow B_{\text{eff}}$.

3. Repeat while $B_{\text{eff}} > 0$:

   (a) Set $B \leftarrow \min\{z/r_{\text{best}}, B_{\text{eff}} - \delta\}$.
   Solve ILP (5). (The current solution is denoted by y_{ij} and z.)

   (b) Let $B_{\text{eff}} \leftarrow \sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j])$.
   If $B_{\text{eff}} > 0$ and $z/B_{\text{eff}} > r_{\text{best}}$ then set $r_{\text{best}} \leftarrow z/B_{\text{eff}}$ and set $B_{\text{best}} \leftarrow B_{\text{eff}}$.

Proof of Correctness

CLAIM 1. *[6] For any $x_i, x_j \in \{0,1\}$ we have $x_i x_j = y_{ij}$ for $y_{ij} \in \{0,1\}$, when the following three constraints are satisfied: $y_{ij} \leq x_i$, $y_{ij} \leq x_j$ and $y_{ij} \geq x_i + x_j - 1$.*

PROOF. If $x_i = 0$ or $x_j = 0$, y_{ij} has to equal 0 as well because of the first two constraints. If $x_i = x_j = 1$ then the third constraint forces y_{ij} to be 1. □

CLAIM 2. *After every execution of the while loop of the algorithm r_{best} is equal to the best objective value of (4) with the additional constraint $\sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j]) \geq B_{eff}^{after}$, where the superscript "after" indicates the values at the end of the loop.*

PROOF. Proof by induction on the number of executions of the while loop. The base case is when the while loop is not executed yet (0 executions of the while loop). At that moment z is the maximum objective value of (5) with $B = \infty$. So any solution where $\sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j]) \geq B_{\text{eff}}$ has objective value that does not exceed z. Therefore the objective of (4) (which is equal to the quotient of objective and the constraint) cannot exceed r_{best}, under $\sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j]) \geq B_{\text{eff}}$.

Induction step: Suppose the claim is true after $k-1$ executions of the while loop. In iteration k, z is the maximum objective value of (5) with $B = \min\{z/r_{\text{best}}^{\text{start}}, B_{\text{eff}}^{\text{start}} - \delta\}$, where the superscript "start" indicates the values at the start of the loop. By the same argument as above the objective of (4) cannot exceed $z/B_{\text{eff}}^{\text{after}}$, under $\sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j]) \geq B_{\text{eff}}^{\text{after}}$ and $\sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j]) \leq B$. By the induction hypothesis $r_{\text{best}}^{\text{start}}$ is equal to the best objective value of (4) with the additional constraint $\sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j]) \geq B_{\text{eff}}^{\text{start}}$. Because $r_{\text{best}}^{\text{after}}$ is set to the maximum of these values, we have proved the claim as long as there is no better solution when $\sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j]) \geq B$ and $\sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j]) \leq B_{\text{eff}}^{\text{start}}$. Note that the choice of δ ensures that there is no solution such that $\sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j]) > B_{\text{eff}}^{\text{start}} - \delta$ and $\sum_{i,j} y_{ij} \exp(\theta[\beta_i + \beta_j]) < B_{\text{eff}}^{\text{start}}$. Finally, because z is an upper bound on the objective value of (5) at every execution of the while loop, we know that the denominator can be constrained to be at most $z/r_{\text{best}}^{\text{before}}$ before we can find an improved solution. □

COROLLARY 1. *The algorithm above finds a (globally) optimal solution to (4).*

B. REFERENCES

[1] 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.

[2] I. Bejar. A sentence-based automated approach to the assessment of writing: A feasibility study. *Machine-Mediated Learning*, 2(4):321–332, 1987.

[3] J. C. Brown, G. A. Frishkoff, and M. Eskenazi. Automatic question generation for vocabulary assessment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 819–826. Association for Computational Linguistics, 2005.

[4] C. H. Coombs. Psychological scaling without a unit of measurement. *Psychological review*, 57(3):145, 1950.

[5] G. H. Fischer and I. W. Molenaar. *Rasch models: Foundations, recent developments, and applications*. Springer Science & Business Media, 2012.

[6] R. Fortet. L'Algèbre de Boole et ses applications en recherche opérationnelle. *Cahiers Centre Etudes Rech. Oper. no.*, 4:5–36, 1959.

[7] T. M. Haladyna. *Writing Test Items To Evaluate Higher Order Thinking*. ERIC, 1997.

[8] T. M. Haladyna and S. M. Downing. Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1):51–78, 1989.

[9] T. M. Haladyna and S. M. Downing. How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, 53(4):999–1010, 1993.

[10] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333, 2002.

[11] F. M. Lord. *Applications of item response theory to practical testing problems*. Routledge, 1980.

[12] T. Minka, J. Winn, J. Guiver, and D. Knowles. Infer .net 2.5. *Microsoft Research Cambridge*, 2012.

[13] R. Mitkov, L. An Ha, and N. Karamanis. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(02):177–194, 2006.

[14] P. Ray. Independence of irrelevant alternatives. *Econometrica: Journal of the Econometric Society*, pages 987–991, 1973.

[15] M. C. Rodriguez. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2):3–13, 2005.

[16] P. H. Schönemann. On metric multidimensional unfolding. *Psychometrika*, 35(3):349–366, 1970.

[17] D. Thissen, L. Steinberg, and A. R. Fitzpatrick. Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, 26(2):161–176, 1989.

[18] D. Vats, C. Studer, A. S. Lan, L. Carin, and R. Baraniuk. Test-size reduction for concept estimation. In *Educational Data Mining 2013*, 2013.

[19] A. E. Waters, A. Lan, C. Studer, and R. G. Baraniuk. Learning analytics via sparse factor analysis. In *Personalizing education with machine learning, nips 2012 workshop*, 2012.

[20] D. J. Weiss. Improving measurement quality and efficiency with adaptive testing. *Applied psychological measurement*, 6(4):473–492, 1982.