Iterative Methods for Ranking Students with Noisy Questions

Gerdus Benade Wolfgang Gatterbauer Nam Ho-Nguyen R. Ravi

1 June 2017

Abstract

We study the problem of ranking students by their abilities, solely based on responses to student-sourced multiple-choice questions. This addresses the crucial problem of scaling automatic assessment of students to very large class sizes. Current state-of-the-art methods (i) assume student responses obey a parameterized model, (ii) were designed for situations with trusted questions, and (iii) are not scalable (we observe empirically that the running time is quadratic in the number of students). In this paper, we define an axiomatic framework for robust ranking algorithms, as well as a new model for simulating ill-posed questions. We marry ideas from work in truth discovery with properties from item response theory to devise several new algorithms with strong axiomatic guarantees and *linear* scalability. We prove that 'translation invariant' and 'anti-symmetric' ranking methods are not affected by ill-posed questions, and that our new methods satisfy these properties. Computational experiments demonstrate the viability of these algorithms.

1 Introduction

In the last decade, technology has had a profound impact on the way students learn. Massive Open Online Courses (MOOCs) allow anyone with an internet connection access to high quality online classes, attracting thousands of students in each class. Large class sizes have created new challenges in creating *scalable assessments* that actively test student comprehension in large courses: While automatically gradable exercises such as Multiple Choice Questions (MCQs) allow scaling the task of grading, creating topical and relevant questions is a labor-intensive task. Motivated by the findings of "learning by explaining" [10], we have been building and using a student-sourced question creation and curation system.

In this work, we focus on a crucial component of this system: How can we assess students solely based on their responses to questions contributed by other students? A key challenge is that some of these questions may be misleadingly worded or simply wrong. We illustrate what is meant by an ill-posed question with an example from one of our classes:

Example 1 (Ill-posed question)**.** *A student created the following question based on a case study and specified a) as correct.*

> What is the main reason for Zara's market dominance?
> a) Zara's one-week output time versus the six-month industry average.
> b) Zara's factory in Spain is a competitive advantage.
> c) Zara's in-house design and creation teams.
> d) Zara produces fewer products to create the appearance of scarcity and increase demand.

Roughly half of the students selected answer a), while one quarter each chose answers c) or d). The idea behind the specified answer is correct, however, Zara's turnaround time is two weeks, not one. Students who had carefully read the case knew a) to be false and selected either c) or d), both of which are reasonable. The question is "ill-posed" or "misleading" as students with better understanding of the case are less likely to choose the answer specified as correct.[1]

[1]The "discrimination score" of this question (defined in Section 2.1) is -0.4, showing that better students were less likely to pick the answer defined as correct. Intuitively, answering this question "correctly" should rather deduct than add points to the student score.

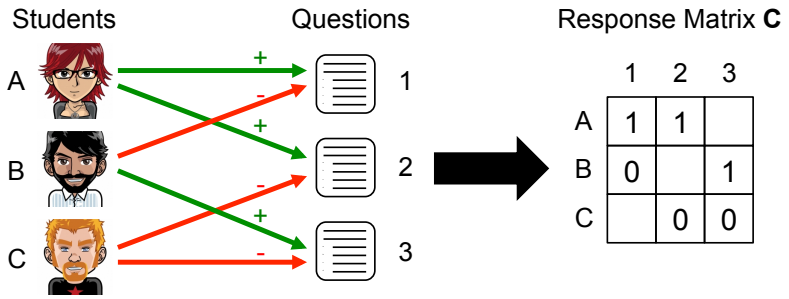


Figure 1: We extend existing work in *truth discovery* [34]. Students are assigned potentially ill-posed questions of unknown quality; we want to accurately rank students based on their *responses* to questions (+ as correct, - as incorrect).

Problem. Our goal is to develop a fast and un-supervised algorithm that ranks students by their abilities , based on the responses to student-sourced questions. Formally, assume that m students answer a set of n questions of unknown quality. Let $\mathbf{C}$ be the $m \times n$ binary response matrix where $C_{ij} = 1$ when student i is presented question j and answers in accordance with the specified "correct" solution, and $C_{ij} = 0$ otherwise. To accommodate large class sizes and potentially thousands of questions, students may be presented with only a constant number of questions independent of the size of n. In this case $\mathbf{C}$ will have missing entries. The main research question we address in this paper is the following:

> *How can we rank students by their abilities, solely based on the student responses* $\mathbf{C}$*, given that questions are created by students and may be ill-posed?*

We will build on truth discovery methods (Figure 1) while guided by widely held assumptions in Item Response Theory (IRT), although our methods do not assume any generative models.

Student ranking desiderata. Before discussing our general approach, we suggest criteria for any student ranking approach.

1. *Accuracy*: The inferred ranking should be similar to the ranking of student abilities derived from traditional approaches.
2. *Robustness to misleading questions*: Since questions are student-sourced and potentially ill-posed, ranking methods should work well even if there are many misleading questions.[2]
3. *Resilience to missing responses*: We want low drop in accuracy as the proportion of questions that students do not attempt increases. This is particularly relevant for large class sizes.
4. *Scalability* : We envisage use in an online system which provides a real-time ranking of students and updates every time a student submits a new response. Such a ranking method should be fast and scale to very large class sizes.
5. *Transparency and fairness*: The calculation of student scores needs to be transparent and explainable to the students. A student's score should either be the sum of question scores answered "correctly," or the percentage of available points scored when only presented with a subset of the questions.
6. *Identification of ill-posed questions*: This is useful both for ranking more accurately, and for identifying students who may benefit from additional attention.

Overall Approach. We propose several new ranking methods that fulfill our desiderata based on the same framework as the popular *Hyperlinked Induced Topic Search (HITS)* algorithm of Kleinberg [21]. Specifically, starting from an initial vector of student scores $\mathbf{s}_0$, estimate question quality $\mathbf{q}$ from the responses $\mathbf{C}$. Let every student's score the sum of the weights of the questions answered correctly, i.e. $s_i = \sum_j C_{ij} q_j$, or in vector form, $\mathbf{s} = \mathbf{Cq}$. If student i is only presented with a subset of question N_i, compute the percentage of the maximum available score achieved, i.e. $s_i = \sum_{j \in N_i} C_{ij} q_j / \sum_{j \in N_i} \max\{0, q_i\}$. Recalculate question and student scores iteratively until a fixed point is reached. At termination, this framework provides a numerical score for each student which may be used for ranking or assigning grades.

This framework is inherently transparent, since student scores are computed in a natural way. A

[2] In one of our classes, we found up to 35% of questions in an early homework to be misleading, i.e. with negative discrimination score.

question's weight can be interpreted as measure of its quality which may influence our assessment of both the creator, and the students who attempted it.

HITS satisfies our transparency requirement, however, we show HITS fails to identify, and is not robust to, misleading questions.

Our new methods and results. We provide a new definition of what it means for a ranking method to be robust to misleading questions. We suggest two properties or ranking functions, *translation invariance* and *antisymmetry*, and show that ranking methods exhibiting them are robust to ill-posed questions.

A simple modification of HITS, called *centered HITS (cHITS)*, satisfies these properties. cHITS can be extended to be more sensitive to question quality, we call this the *biserial update (BSRL)* method. BSRL can be interpreted as a maximum likelihood fit for a linear regression model. Since the question responses are binary, a natural improvement is to use a logistic instead of linear regression model. We term this the *logistic regression (LogR) update* method. All three methods – cHITS, BSRL and LogR – are robust, transparent, and inherently scalable.

We carry out extensive computational evaluations of our three new methods to verify that they satisfy our desiderata. We find that LogR is as accurate as MIRT (the state-of-the-art maximum likelihood inference program based on Item Response Theory, see Section 2.1) across a wide range of parameter settings while being significantly more scalable and able to identify misleading questions, *even when the maximum likelihood estimator knows the IRT model from which the data was generated.* Crucially, LogR performs better than MIRT on experiments with real-world data collected during the instruction a class.

2 Related work

Our work is loosely related to previous work in crowd-labeling (see [31, 13], for example), however, in the interest of space we only discuss the two most relevant streams of literature in detail.

2.1 Item Response Theory (IRT)

Item Response Theory (IRT) [4] is widely used to assess students, e.g. in the Scholastic Aptitude Test (SAT) [22] and Graduate Record Examinations (GRE) [20]. It models the probability of a student providing a correct response as a function of latent traits describing student ability and item factors characterizing the question. In the widely studied two parameter logistic (2PL) model, the *ability* of student i is captured by a single latent trait variable θ_i, and question j has a *difficulty* β_j and a *discrimination factor* α_j. The probability that student i answers questions j correctly is modeled as

$$\mathbf{P}(C_{ij} = 1|\theta_i, \alpha_j, \beta_j) = \left(1 + e^{-\alpha_j(\theta_i - \beta_j)}\right)^{-1} \tag{1}$$

and responses are assumed to be conditionally independent.

The discrimination factor a_i determines the maximal slope of the *item characteristic curve (ICC)* $f_j(\theta) = \mathbf{P}(C_{ij} = 1|\theta, a_j, b_j)$, which occurs at the inflection point where $\theta_i = b_j$, and the probability of a correct response is $\frac{1}{2}$. Figure 2 shows how a change in difficulty or discrimination factor affects the ICC.

When $a_j > 0$, as is the case for trusted questions, the probability of answering a question correctly increases with θ. For student-sourced questions, note that if $a_j < 0$ the probability of answering question j "correctly" decreases with increasing student ability. A completely non-discriminating question has $\alpha_j = 0$.

Example 1 (Continued)**.** *The question in Example 1 had $\alpha \approx -0.4$. Figure 3 shows its ICC, along with a representative sample of the responses received. Stronger students were less likely to answer this ill-coded question correctly according to the submitted solution.*

The basic 2PL model has been extended in various ways, e.g. to handle multi-dimensional latent traits [2, 12]. Several software packages estimate IRT parameters using maximum likelihood methods and their extensions (BILOG, MULTILOG, PARAM-3PL and LTM [29] and MIRT [9] in R). We collectively refer to the set of such methods as "MIRT" in reference to the `R`-package MIRT used in our experiments.[3] These methods face at least four challenges in our scenario: (1) Their validity relies on

[3]This is not traditional, in the Item Response Literature "MIRT" is often an acronym for "Multi-dimensional Item Response Theory".

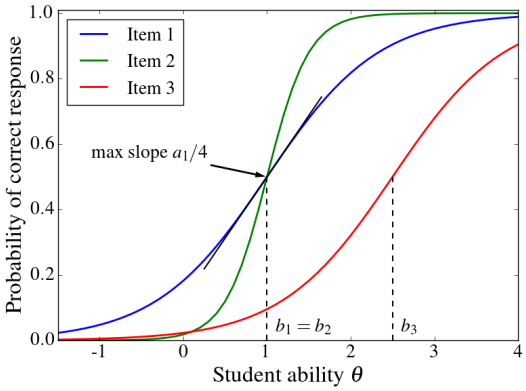


Figure 2: ICCs for the 2PL model: Item 1 and 2 have equal difficulty, but item 2 has higher discrimination ($a_2 = 4$ vs. $a_1 = 1.5$). Item 3 has higher difficulty than item 1 ($b_3 = 2.5$), and equal discrimination.

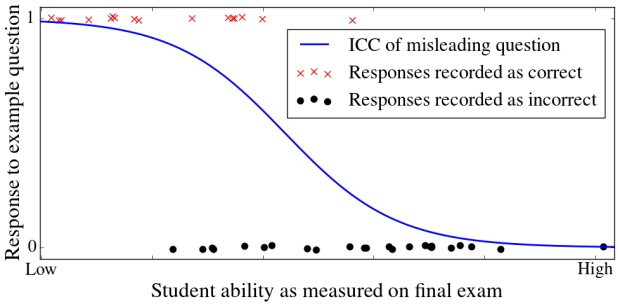


Figure 3: (Example 1 continued) Stronger students are more likely to answer a misleading question incorrectly.

assuming very specific generative models; (2) Maximum likelihood methods may lack transparency, especially for more complex latent trait models; (3) IRT models face significant challenges in convergence when different questions have appreciable differences in their discrimination factors [1] ; and (4) The computation of marginal maximum likelihood estimators involves numerical integration of intractable integrals and assumes that student abilities come from a normal distribution. These computations do not scale gracefully with the number of students.

By contrast, our methods are guaranteed to converge, scale linearly with the number of students and are transparent. Our methods are agnostic to generative models and provably robust against ill-posed questions. Surprisingly, our experiments show that our algorithms are comparable to the 2PL maximum likelihood estimators even when the instance is generated with a 2PL model. In the limited real-world data we have available, our methods are superior.

2.2 Truth inference

The *Hyperlinked Induced Topic Search (HITS)* algorithm of Kleinberg [21] is popular in the area of trust and verification of claims [16]. This iterative method updates hub scores as the sum of the authority scores it is linked to, and then an authority's score as the sum of the hub scores it links to. This is repeated (after normalizing) until convergence. The method is fast and scalable to very large datasets.

This algorithm has been extended for the problem of estimating "trustworthiness" and "believability" scores from networks where the trustworthiness of the actors are unknown or uncertain [16, 26]. Variants like PageTrust [11] and EigenTrust [19], attempt to estimate the trustworthiness or quality of sources, which is similar to our problem of identifying good students and high-quality questions. Other variants of these methods (such as AverageLog, TruthFinder, Investment, etc.) have been extensively studied, are typically extremely fast, and have been proven in practice in a wide variety of settings [3, 15, 24, 25, 32, 33]. Our setting is similar, we have bipartite graph with students in one partition and questions in the other, with an edge between student i and question j exactly when $C_{ij} = 1$. We will use HITS as an

important baseline. Neither HITS, nor any of the other methods mentioned in a survey [16, 23] satisfy the properties that are critical for iterative methods to be robust to ill-posed questions. Our methods are more accurate than HITS, provably robust to ill-posed questions and able to identify misleading questions.

3 A Framework for Ranking Students

In this section we discuss our model and solution approach in more detail. Specifically, we avoid assuming that student responses follow some generative model but rather assume that there exist a true ordering of the students defined by their probabilities of answering an arbitrary question (drawn from some distribution of questions) correctly. This model is broad enough to encompass several known models, including those of Item Response Theory. Given this true ordering of students, it is natural to assume that a stronger student has a higher probability of answering an arbitrary question correctly. However, in this paper we focus instead on robustness to misleading questions and do not require this assumption for our results.

We identify a framework for iterative algorithms inspired by the HITS algorithm and show that if an iterative ranking algorithm calibrates question weights in a *translation invariant* and *antisymmetric* way, the method is robust to the introduction of misleading questions.

3.1 Sufficient Information for Ranking

Let us recall our problem definition. Suppose that m students, each with unknown ability θ_i, $i \in [m]$, answer a set of n questions, each of unknown quality. Let $\mathbf{C}$ be the $m \times n$ binary matrix where $C_{ij} = 1$ when student i answered question j correctly. Assume throughout this discussion that every student answers every question, except when explicitly stated otherwise (the case of missing responses in investigated in Section 4.7). We wish to recover a ranking of the students which depends only on $\mathbf{C}$ and is as close as possible to the actual ranking by their unknown abilities θ_i. For this, we need to describe the relation between the abilities and the response matrix $\mathbf{C}$.

We assume that there exists some "true" ranking of the students by their abilities θ_i and that a better student is more likely to answer an arbitrary question correctly. This is referred to as the *sufficient information* assumption. More precisely, we assume that the questions are generated by some random process, and that a student of ability θ, in expectation, answers a randomly generated question correctly with probability $p(\theta)$. Our sufficient information assumption is equivalent to requiring

$$\theta > \theta' \implies p(\theta) > p(\theta').$$

Notice that while this conditions holds *on average*, it may not hold for any one specific question. In particular, we expect a better student to have a smaller probability of answering a misleading question correctly.

The implications of this assumption under a 2PL model is discussed in the appendix. More generally, if all questions are generated by the same model and subsequently designated as "*misleading*" or not, then this assumption implies that less than half the questions are misleading. In the absence of sufficient information, we do not expect any reasonable ranking algorithm to recover the true ranking of the students. We make no further assumptions about student abilities or question properties. Indeed, for our results, $p(\theta)$ need not have an explicitly defined functional form.

3.2 An Iterative Framework for Ranking

Our task is to to infer an n-dimensional vector of question weights $\mathbf{q}$, and an m-dimensional vector of student test scores $\mathbf{s}$. Our proposed ranking framework defines a solution by a system of multiple constraints. We use iterative updates to find a fixed point $(\mathbf{s}, \mathbf{q})$. The solution is defined by three parts:

1. *Initialization*: Start with an initialization of the student scores $\mathbf{s}_0$. For some methods, the choice of $\mathbf{s}_0$ does not affect the resulting fixed point. In the absence of additional information, one may initialize the student scores with the number of questions a student answered correctly: $\mathbf{s}_0 = \mathbf{C}\mathbf{1}_n$.
2. *Question weights*: Calibration functions $f : \{0,1\}^{m \times n} \times \mathbb{R}^m \to \mathbb{R}^n$ calculate question weights $\mathbf{q}^{(t)}$ based on student scores and the response matrix. Every method is defined by a different calibration

Input: Response matrix $\mathbf{C}$, initial student scores $\mathbf{s}_0$
Output: Student scores $\mathbf{s}$, question weights $\mathbf{q}$

$\mathbf{s} \leftarrow \mathbf{s}_0$ // initialize student scores
repeat
  $\mathbf{q} \leftarrow f(\mathbf{C}, \mathbf{s})$ // update question weights
  $\mathbf{s} \leftarrow \mathbf{C}\mathbf{q}$ // update student scores
until *convergence or iteration limit*

Algorithm 1 Framework for transparent student rankings

function and our goal is to find functions with appropriate mathematical properties. The functions we discuss all decompose into identical functions $g : \{0,1\}^m \times \mathbb{R}^m \to \mathbb{R}$, so that the update rule for $q_j^{(t)}$ uses only the jth column $\mathbf{C}_j$,

$$q_j^{(t)} \leftarrow g(\mathbf{C}_j, \mathbf{s}^{(t-1)}). \tag{2}$$

3. *Student scores*: Transparency demands that student test scores are the weighed sum of the questions correctly answered by the students, so

$$\mathbf{s}^{(t)} \leftarrow \mathbf{C}\,\mathbf{q}^{(t)}. \tag{3}$$

Iterative updates are performed until the student and question scores converge (see Algorithm 1).

While we require student scores to be assigned transparently, (2) allows the question weights to be determined in arbitrarily complex ways. Still, we argue that a reasonable calibration function g should possess the following two properties:

1. Let $g(\mathbf{c}, \mathbf{s})$ be the computed question weight based on the student scores $\mathbf{s}$ and student responses $\mathbf{c}$ (corresponding to a column in $\mathbf{C}$). Then the question weight should not change if all student scores are increased by the same number. More formally, we say that g is *translation invariant* if

$$\forall \mathbf{c} \in \{0,1\}^m, \mathbf{s} \in \mathbb{R}^m, \alpha \in \mathbb{R}.\ g(\mathbf{c}, \mathbf{s} + \alpha\mathbf{1}) = g(\mathbf{c}, \mathbf{s}) \tag{4}$$

2. If the question response vector $\mathbf{c} \in \{0,1\}^m$ is flipped to $\mathbf{1} - \mathbf{c}$, then the question score $g(\mathbf{1} - \mathbf{c}, \mathbf{s})$ should reflect this. We say that g is *antisymmetric* if

$$\forall \mathbf{c} \in \{0,1\}^m, \mathbf{s} \in \mathbb{R}^m.\ g(\mathbf{1} - \mathbf{c}, \mathbf{s}) = -g(\mathbf{c}, \mathbf{s}). \tag{5}$$

Intuitively, translation invariance states that adding a constant to each student score should not change the calibration score, since we are chiefly concerned with rankings and the underlying ranking stays the same. Antisymmetry ensures the calibration function g is sensitive to a reversal of the student responses to a particular question (as may occur in the case of an ill-coded question). Reversing the sign of the calibration function for inverted responses is motivated by the use of negative scores to flag ill-coded questions.

In Section 4, we discuss two baseline and three novel ranking methods derived by modifying the calibration functions g in order to satisfy properties (4) and (5). Before that, we show that these properties help satisfy one of our desiderata: robustness to misleading questions.

3.3 A Model for Misleading Questions

Let $\mathbf{C}$ be a response matrix and let $\mathbf{C}_j$ refer to the j-th column of $\mathbf{C}$. Let $J \subseteq [n]$ be a subset of misleading questions, *i.e.* questions whose responses have been inverted $\mathbf{C}'_j = \mathbf{1}_m - \mathbf{C}_j$.

This is a simple form of generating misleading questions from good questions that is well justified: Consider a good MCQ j with two potential answers and response vector $\mathbf{C}_j$. If the correct answer was erroneously specified as the incorrect answer, the response vector for this ill-posed question would be exactly $\mathbf{C}'_j = \mathbf{1} - \mathbf{C}_j$. This response-flipping model is also consistent with IRT: In the 2PL generative model, the probability of a particular response $\mathbf{c}$ being generated from a question with difficulty b and discrimination factor a is identical to the probability of the flipped response $\mathbf{1}_m - \mathbf{c}$ being generated from

a question with identical difficulty but negated discrimination factor $-a$. This follows directly from the functional form of the probability of any student i getting a question j correct in the 2PL model:

$$\frac{1}{1+e^{-a_j(\theta_i-b_j)}} = 1 - \frac{1}{1+e^{-(-a_j)(\theta_i-b_j)}}.$$

Formally, let $\mathbf{C}'$ be the matrix with these inverted responses replacing the corresponding columns in $\mathbf{C}$. If g is antisymmetric, then the score vectors calculated from the current $\mathbf{q}$ using $\mathbf{C}$ and $\mathbf{C}'$ will imply the same student ranking.

Lemma 1 (Rank invariance to response-flipping)**.** *Suppose we have a function $g : \{0,1\}^m \times \mathbb{R}^m \to \mathbb{R}$ that satisfies antisymmetry* (5)*. Let $\boldsymbol{\sigma} \in \mathbb{R}^m$ be a student score vector, and $\mathbf{C}_j \in \{0,1\}^m$, $j \in [n]$, be a set of binary vectors. Fix an arbitrary subset $J \subseteq [n]$ and define two vectors $\mathbf{s}, \mathbf{s}' \in \mathbb{R}^m$ from using the two update rules* (2) *and* (3) *with $\mathbf{C}$ and $\mathbf{C}'$ respectively:*

$$s_i = \sum_{j \notin J} C_{ij} g(\mathbf{C}_j, \boldsymbol{\sigma}) + \sum_{j \in J} C_{ij} g(\mathbf{C}_j, \boldsymbol{\sigma})$$
$$s'_i = \sum_{j \notin J} C_{ij} g(\mathbf{C}_j, \boldsymbol{\sigma}) + \sum_{j \in J} (1 - C_{ij}) g(\mathbf{1} - \mathbf{C}_j, \boldsymbol{\sigma}).$$

Then the elements of $\mathbf{s}$ and $\mathbf{s}'$ differ by the same constant and, hence, imply the same student ordering.

Proof. We rewrite s'_i using antisymmetry of g:

$$\begin{aligned} s'_i &= \sum_{j \notin J} C_{ij} g(\mathbf{C}_j, \boldsymbol{\sigma}) + \sum_{j \in J} (1 - C_{ij}) g(\mathbf{1} - \mathbf{C}_j, \boldsymbol{\sigma}) \\ &= \sum_{j \notin J} C_{ij} g(\mathbf{C}_j, \boldsymbol{\sigma}) + \sum_{j \in J} (C_{ij} - 1) g(\mathbf{C}_j, \boldsymbol{\sigma}) = s_i - \sum_{j \in J} g(\mathbf{C}_j, \boldsymbol{\sigma}). \end{aligned}$$

The result now follows since each s'_i is simply s_i offset by the same constant $\sum_{j \in J} g(\mathbf{C}_j, \boldsymbol{\sigma})$. □

We next show that if g is translation invariant (4) in addition to being antisymmetric (5), then the rankings resulting from using $\mathbf{C}$ and $\mathbf{C}'$, respectively, will remain the same across multiple iterations.

Theorem 2 (Iterative rank invariance)**.** *Let g be a calibration rule satisfying* (4) *and* (5)*. Let $\mathbf{q}^{(t)}, \mathbf{s}^{(t)}$ and $\boldsymbol{\gamma}^{(t)}, \boldsymbol{\sigma}^{(t)}$ be question and student scores at iteration t generated from $\mathbf{C}$ and $\mathbf{C}'$ respectively, both starting from the same $\mathbf{s}^{(0)} \in \mathbb{R}^m$. Then, at every iteration t, the elements of $\mathbf{s}^{(t)}$ and $\boldsymbol{\sigma}^{(t)}$ differ by the same constant $\alpha^{(t)}$ and, hence, imply the same student ordering.*

Proof. We prove this by induction on t. The base case $t = 1$ follows by applying Lemma 1 with $\mathbf{s}^{(0)} = \mathbf{s}_0$. Suppose now that it holds for $t - 1 \geq 1$, that is, $\boldsymbol{\sigma}^{(t-1)} = \mathbf{s}^{(t-1)} + \alpha^{(t-1)} \mathbf{1}_m$. By translation invariance (4), the question scores $\gamma_j^{(t-1)} = g(\mathbf{C}'_j, \boldsymbol{\sigma}^{(t-1)}) = g(\mathbf{C}'_j, \mathbf{s}^{(t-1)})$. It follows from Lemma 1 with $\mathbf{s} = \mathbf{s}^{(t-1)}$ that $\boldsymbol{\sigma}^{(t)}$ differs from $\mathbf{s}^{(t)}$ by a constant. This completes the induction hypothesis, and the proof. □

The implication of this result is that a *translation invariant* and *antisymmetric* calibration leads to a transparent method for ranking students that is *robust to misleading questions*. This holds irrespective of the generative model underlying the creation of $\mathbf{C}$.

Interestingly, our results suggest that an appropriate calibration function will give a robust ranking method for *any* fraction of misleading questions, *if given the same starting s_0*. Loosely, this states that the response vectors $\mathbf{C}_j$ and $\mathbf{C}'_j$ provide the same amount of information about students' relative abilities.

In practice, however, too many ill-posed questions will affect the initial estimate of student abilities and result in a different ranking. For example, if $\mathbf{s}_0 = \mathbf{C}\mathbf{1}_n$ and $|J| > n/2$, we expect $\mathbf{s}'_0$ to invert the ranking given by $\mathbf{s}_0$.

4 Five Methods for Ranking

We first review two simple baseline methods: using the number of correct responses, and HITS [21], and show they neither satisfy translation invariance (4), nor asymmetry (5). We derive three novel update rules satisfying properties (4) and (5). Finally, we discuss the computational complexity of these methods and how to modify the algorithms when students only answer $k < n$ questions.

4.1 Number of correct responses (AvgSc)

The simplest way to rank students is by the *number of recorded correct responses* (AvgSc). All questions have equal weight, so we set $g^{\text{AvgSc}}(\mathbf{c}, \mathbf{s}) := 1$. No iteration is required. Clearly, g^{AvgSc} does not satisfy (5).

Even though ranking students by the number of correct responses submitted is very simple, we can show under the sufficient information assumption that this method of ranking is *consistent*. This means that for a fixed number of students, the probability of recovering the correct ranking tends to 1 as the number of questions tends to infinity.

Proposition 1 (Consistency of AvgSc)**.** *Ranking students by the number of correct responses they submit is consistent.*

Proof. Denote by $p_i := p(\theta_i)$ the probability that student $i \in [m]$ answers a randomly generated question correctly. We define $\epsilon := \min_{i,i'} \{|p_i - p_{i'}| : p_i \neq p_{i'}\}$ to be the smallest nonzero difference between the p_i. Let Y_i be the number of correct responses of student i, and observe that $Y_i \sim \text{Bin}(n, p_i)$. We prove that with high probability, the number of correct responses follows the ranking of the p_i.

$$\begin{aligned}\mathbb{P}[Y_i \text{ ranking differs from } p_i \text{ ranking}] &\leq \mathbb{P}[\exists\, i \text{ s.t. } |Y_i/n - p_i| > \epsilon/2] \\ &\leq \sum_{i=1}^{m} \mathbb{P}[|Y_i/n - p_i| > \epsilon/2] \\ &\leq 2m \exp(-n\epsilon^2/2).\end{aligned}$$

Here, the second inequality is the union bound, and the third inequality is from applying Hoeffding's inequality. Consistency then follows by the sufficient information assumption. □

Consistency acts as a convenient sanity check that ranking by the number of correct responses is, in fact, a reasonable baseline which will find the correct ranking if given enough information about the students. However, in the presence of questions of different qualities it may be beneficial to weigh better questions more heavily and to identify misleading questions. Clearly, g^{AvgSc} does not satisfy (5).

4.2 Hubs and Authorities (HITS)

The popular HITS algorithm [21] attempts to overcome the problems arising from assigning equal weight to every source of information by determining "quality scores" for each question. The idea is to model the information structure as a directed graph, where nodes represent sources of information, and a directed edge represents one source pointing to another. The algorithm then assigns each source a hub score based on which nodes that source points to, and an authority score based on which nodes are pointed to that source. The intuition is that good hubs are those which point to many good authorities, and good authorities are pointed to by many good hubs.

Concretely, the hub score of a source is computed as the sum of authority scores of nodes pointed to by that source, and the authority score is computed as the sum of hub scores of nodes pointing to that source. The algorithm computes these scores iteratively, starting from initial values

This algorithm forms the basis from which various variants have been applied and studied in truth discovery (Section 2.2).

We formulate a bipartite graph with $m+n$ nodes, with student nodes on one side, and question nodes on the other. The response matrix $\mathbf{C}$ gives the edge structure, with $C_{ij} = 1$ indicating the existence of an edge from question j pointing to student i. The HITS algorithm computes hub scores *only* for questions, and authority scores *only* for students (since edges only point from questions to students). We interpret HITS as saying that good questions are those answered correctly by strong students, and strong students are answering the good questions correctly.

In our formalism, the inner product calibration rule

$$g^{\text{HITS}}(\mathbf{c}, \mathbf{s}) := \mathbf{c}^\intercal \mathbf{s} \tag{HITS}$$

together with (2), (3) captures the HITS algorithm.

We can write the update rule for the student scores $\mathbf{s}^{(t)}$ as

$$\mathbf{s}^{(t)} = \mathbf{C}\mathbf{C}^\intercal \mathbf{s}^{(t-1)} = (\mathbf{C}\mathbf{C}^\intercal)^t \mathbf{s}_0,$$

where the second equality is from recursively applying the first. This corresponds to an un-normalised version of the well-studied power method for computing leading eigenvalues of the matrix $\mathbf{C}\mathbf{C}^\intercal$ [30]. In fact, $\mathbf{s}^{(t)}/\|\mathbf{s}^{(t)}\|$ directly corresponds to iterations of the power method. As long as the initial student score vector $\mathbf{s}_0$ is not orthogonal to the leading eigenspace, $\mathbf{s}^{(t)}/\|\mathbf{s}^{(t)}\|$ converges to a leading eigenvector of $\mathbf{C}\mathbf{C}^\intercal$. Since $\mathbf{s}^{(t)}/\|\mathbf{s}^{(t)}\|$ and $\mathbf{s}^{(t)}$ give the same ranking of the students, the final ranking will be given by the leading eigenvector of $\mathbf{C}\mathbf{C}^\intercal$.

It is possible to show that as the number of questions increase to infinity (while m stays fixed), the probability ranking students correctly converges to 1.

Proposition 2 (Consistency of HITS)**.** *Ranking students according to their HITS weights is consistent.*

Proof. Without loss of generality, we assume for this proof that the indices for $p_i := p(\theta_i)$ are sorted in decreasing order, i.e., $p_i > p_k$ if $i < k$. Furthermore, we define

$$\epsilon = \min_{i,i',k} \{p_{i'} |p_i - p_k| : i' \neq i, k,\, p_i \neq p_k\}.$$

Recall that the score vector $\mathbf{s}$ is the leading eigenvector of the matrix $\mathbf{M} := \mathbf{C}\mathbf{C}^\intercal$. Suppose that $\mathbf{M}$ satisfies the following order property: $M_{ii} > M_{kk}$ for $i < k$ and $M_{ii'} > M_{ki'}$ for $i < k$, $i' \neq i, k$. Then the score vector is also in decreasing order, i.e., $s_1 \geq \ldots \geq s_m$, and thus will recover the ranking of the p_i. To see this, observe that $\mathbf{s} \in \arg\max_{\|\mathbf{x}\|=1} \mathbf{x}^\intercal \mathbf{M}\mathbf{x}$, and if there exists $s_i < s_k$ when $i < k$, then through some simple algebra, exploiting the ordering property and the fact that $\mathbf{M} \geq 0$, we can show that switching the entries s_i and s_k will increase the quadratic form $\mathbf{s}^\intercal \mathbf{M}\mathbf{s}$, which is a contradiction.

The proof now reduces to showing that $\mathbf{M}$ satisfies the ordering property with high probability. Observe that diagonal entries $M_{ii} \sim \mathrm{Bin}(n, p_i)$ and off-diagonal entries $M_{ik} \sim \mathrm{Bin}(n, p_i p_k)$. By our assumption on the ordering of the p_i, we can observe that $\mathbb{E}[\mathbf{M}]$ satisfies the ordering property. Thus, $\mathbf{M}$ will also satisfy the ordering property if it is not too far away from $\mathbb{E}[\mathbf{M}]$. More precisely, $\mathbf{M}$ satisfies the ordering property if each entry M_{ik} is within $n\epsilon_2$ of its expected value np_ip_k (or np_i if $i = k$). Thus, the probability that $\mathbf{s}$ returns an incorrect ranking is bounded by the probability of this event not happening, which we can estimate as follows:

$$\begin{aligned}
&\mathbb{P}[\mathbf{s} \text{ ranking differs from } p_i \text{ ranking}] \\
&\leq \mathbb{P}\left[\exists\, i, k \text{ s.t. } |M_{ik} - \mathbb{E}[M_{ik}]| > n\epsilon/2\right] \\
&\leq \sum_{i \leq k} \mathbb{P}\left[|M_{ik} - \mathbb{E}[M_{ik}| > n\epsilon/2\right] \\
&\leq \sum_{i \leq k} 2\exp(-n\epsilon^2/2) \\
&= m(m+1)\exp(-n\epsilon^2/2),
\end{aligned}$$

where the third inequality is from Hoeffding's inequality. This shows that, with high probability, ranking students by HITS recovers their ranking by p_i, and consistency then holds by sufficient information. □

The update rule g^{HITS} does not satisfy translation invariance (4) or antisymmetry (5), so we do not expect it to perform well in the presence of misleading questions. Additionally, if $\mathbf{s}_0 \geq 0$ (component-wise), then $\mathbf{q}^{(t)} \geq 0$. We can therefore not identify ill-posed questions by their negative scores, which is one of our desiderata discussed in Section 1.

4.3 Centered HITS (cHITS)

A simple modification to HITS gives us translation invariance. Specifically, we first "center" the student score vector, then compute the question weights as before. This results in the update

$$g^{\mathrm{cHITS}}(\mathbf{c}, \mathbf{s}) := g^{\mathrm{HITS}}(\mathbf{c}, \mathbf{s} - \bar{s}\mathbf{1}_m) = \mathbf{c}^\intercal (\mathbf{s} - \bar{s}\mathbf{1}_m), \tag{cHITS}$$

where $\bar{s} := \mathbf{1}_m^\intercal \mathbf{s}/m$ is the mean student score. By construction, this new update rule satisfies translation invariance (4). cHITS allows negative question scores, as opposed to (HITS), and has the potential to identify misleading questions. Remarkably, this simple fix to HITS also gives us antisymmetry (5). The next result follows immediately after substituting the above definition of $g^{\text{cHITS}}(\mathbf{c}, \mathbf{s})$ and observing that $\mathbf{1}_m^\intercal(\mathbf{s} - \bar{s}\mathbf{1}_m) = 0$.

Lemma 2 (Antisymmetry of cHITS)**.** *Let $\mathbf{s} \in \mathbb{R}^m$ and $\mathbf{c} \in \{0,1\}^m$ be arbitrary. Then $g^{\text{cHITS}}(\mathbf{1}_m - \mathbf{c}, \mathbf{s}) = -g^{\text{cHITS}}(\mathbf{c}, \mathbf{s})$.*

It now follows trivially from Theorem 2 and lemma 2 that the cHITS update rule is robust to misleading questions.

Corollary 1. *cHITS is robust to misleading questions.*

Finally, we show that cHITS is equivalent to an eigenvector problem, which also shows it is convergent. Denoting $\mathbf{W} = \mathbf{I}_m - \mathbf{1}_m\mathbf{1}_m^\intercal/m$, we can write the updates in matrix form as

$$\begin{aligned}\mathbf{q}^{(t-1)} &= \mathbf{C}^\intercal \mathbf{W} \mathbf{s}^{(t-1)} \\ \mathbf{s}^{(t)} &= \mathbf{C}\mathbf{q}^{(t-1)}.\end{aligned}$$

For the student scores $\mathbf{s}^{(t)}$ this becomes

$$\mathbf{s}^{(t)} = \mathbf{C}\mathbf{C}^\intercal \mathbf{W} \mathbf{s}^{(t-1)} = (\mathbf{C}\mathbf{C}^\intercal \mathbf{W})^t \mathbf{s}_0, \tag{6}$$

where the second equality follows from recursively applying the first. Like the updates for HITS, (6) is simply an un-normalised power method update, except that $\mathbf{C}\mathbf{C}^\intercal\mathbf{W}$ is not symmetric. However, we can still show that the iterations converge to a transformed eigenvector of a symmetric matrix.

Theorem 3. *Suppose $\mathbf{s}^{(t)}$ are computed according to* (6)*. Then $\mathbf{s}^{(t)}/\|\mathbf{s}^{(t)}\| \to \mathbf{C}\mathbf{C}^\intercal\mathbf{e}$, where $\mathbf{e}$ is the eigenvector of the symmetric positive definite matrix $\mathbf{W}^\intercal\mathbf{C}\mathbf{C}^\intercal\mathbf{W}$ with largest eigenvalue that is not orthogonal to $\mathbf{s}_0$.*

Proof. First observe $\mathbf{W}\mathbf{W}^\intercal = \mathbf{W} = \mathbf{W}^\intercal$, thus (6) is equivalent to

$$\mathbf{s}^{(t)} = (\mathbf{C}\mathbf{C}^\intercal\mathbf{W}\mathbf{W}^\intercal)^t \mathbf{s}_0.$$

Now observe that by simply reordering the parentheses, we have

$$\begin{aligned}(\mathbf{C}\mathbf{C}^\intercal\mathbf{W}\mathbf{W}^\intercal)^t &= \mathbf{C}\mathbf{C}^\intercal\mathbf{W}(\mathbf{W}^\intercal\mathbf{C}\mathbf{C}^\intercal\mathbf{W})^{t-1}\mathbf{W}^\intercal \\ &= \mathbf{C}\mathbf{C}^\intercal(\mathbf{W}^\intercal\mathbf{C}\mathbf{C}^\intercal\mathbf{W})^{t-1}.\end{aligned}$$

Let $\mathbf{e}_i$, $i \in [m]$ be an orthogonal basis of eigenvectors of $\mathbf{W}^\intercal\mathbf{C}\mathbf{C}^\intercal\mathbf{W}$, with corresponding eigenvalues λ_i in descending order. Then $\mathbf{s}_0 = \sum_{i=1}^m \mathbf{e}_i^\intercal \mathbf{s}_0 \mathbf{e}_i$, hence

$$\mathbf{s}^{(t)} = \mathbf{C}\mathbf{C}^\intercal(\mathbf{W}^\intercal\mathbf{C}\mathbf{C}^\intercal\mathbf{W})^{t-1}\mathbf{s}_0 = \sum_{i=1}^m \lambda_i^{t-1}\mathbf{e}_i^\intercal\mathbf{s}_0\mathbf{C}\mathbf{C}^\intercal\mathbf{e}_i.$$

Therefore, letting λ be the eigenvalue corresponding to the vector $\mathbf{e}$ from the statement of the theorem,

$$\begin{aligned}\frac{\mathbf{s}^{(t)}}{\|\mathbf{s}^{(t)}\|} &= \frac{\sum_{i=1}^m \lambda_i^{t-1}\mathbf{e}_i^\intercal\mathbf{s}_0\mathbf{C}\mathbf{C}^\intercal\mathbf{e}_i}{\left\|\sum_{i=1}^m \lambda_i^{t-1}\mathbf{e}_i^\intercal\mathbf{s}_0\mathbf{C}\mathbf{C}^\intercal\mathbf{e}_i\right\|} \\ &= \frac{\sum_{i=1}^m (\lambda_i/\lambda)^{t-1}\mathbf{e}_i^\intercal\mathbf{s}_0\mathbf{C}\mathbf{C}^\intercal\mathbf{e}_i}{\left\|\sum_{i=1}^m (\lambda_i/\lambda)^{t-1}\mathbf{e}_i^\intercal\mathbf{s}_0\mathbf{C}\mathbf{C}^\intercal\mathbf{e}_i\right\|}.\end{aligned}$$

Considering first the numerator, if $\lambda_i > \lambda$, by assumption we have $\mathbf{e}_i^\intercal\mathbf{s}_0 = 0$, so these terms disappear. For $\lambda_i < \lambda$, $(\lambda_i/\lambda)^{t-1} \to 0$ as $t \to \infty$. Thus the numerator of $\mathbf{s}^{(t)}$ converges to a vector in the eigenspace of λ, multiplied by $\mathbf{C}\mathbf{C}^\top$, and by continuity the denominator simply normalises the resulting vector. This gives us our result. □

Since $\mathbf{s}^{(t)}/\|\mathbf{s}^{(t)}\|$ and $\mathbf{s}^{(t)}$ differ only by a positive multiplicative factor, they give the same ranking. Thus, Theorem 3 states that the final ranking is the same as $\mathbf{C}\mathbf{C}^\intercal\mathbf{e}$, i.e., the ranking problem reduces to computing an eigenvalue.

4.4 Biserial update (BSRL)

We now explore a possible improvement to cHITS. Consider two questions, one is answered correctly by half of the students and the other by only a small fraction of students. It seems reasonable that a correct answer to the latter question should be worth more points than a correct answer to the former. A similar argument suggests weighing a question answered correctly by the vast majority of students more heavily places a larger relative penalty on the few students answered it incorrectly. A natural way to do this is to scale a question's response vector by its standard deviation. Let $f_j := \mathbf{1}_m^\intercal \mathbf{C}_j/m$ denote the fraction of students who answered question j correctly. The resulting update rule is

$$q_j^{(t+1)} = \frac{\mathbf{C}_j^\intercal\big(\mathbf{s}^{(t)} - \bar{s}^{(t)}\mathbf{1}_m\big)}{\|\mathbf{C}_j - f_j\mathbf{1}_m\|} = \frac{(\mathbf{C}_j - f_j\mathbf{1}_m)^\intercal\big(\mathbf{s}^{(t)} - \bar{s}^{(t)}\mathbf{1}_m\big)}{\|\mathbf{C}_j - f_j\mathbf{1}_m\|}. \tag{7}$$

The second equality holds because $f_j\mathbf{1}_m^\top\left(\mathbf{s}^{(t)} - \bar{s}^{(t)}\mathbf{1}_m\right) = 0$.

This update rule has close ties to the *cosine similarity* between $\mathbf{C}_j$ and $\mathbf{s}^{(t)}$ and the least-squares solution when fitting a linear curve to the relationship between $\mathbf{C}_j$ and $\mathbf{s}^{(t)}$, however, we prefer to interpret it as a correlation.

Our third update rule called BSRL ("biserial") is thus defined by

$$g^{\mathrm{BSRL}}(\mathbf{c},\mathbf{s}) := \mathrm{Corr}(\mathbf{c},\mathbf{s}) = \frac{(\mathbf{c} - \bar{c}\mathbf{1}_m)^\intercal\,(\mathbf{s} - \bar{s}\mathbf{1}_m)}{\|\mathbf{c} - \bar{c}\mathbf{1}_m\|\,\|\mathbf{s} - \bar{s}\mathbf{1}_m\|}, \tag{BSRL}$$

where $\bar{x} = \mathbf{1}^\intercal\mathbf{x}/m$ is the mean of $\mathbf{x} \in \mathbb{R}^m$. Compared to (7), we have an extra factor in the denominator which affects all question scores similarly, and does not change the resulting ranking of students.

Like cHITS, BSRL is robust to misleading questions. We first show that g^{BSRL} is antisymmetric (5).

Lemma 3 (Antisymmetry of BSRL)**.** *Let $\mathbf{s} \in \mathbb{R}^m$ and $\mathbf{c} \in \{0,1\}^m$ be arbitrary. Then $g^{\mathrm{BSRL}}(1-\mathbf{c},\mathbf{s}) = -g^{\mathrm{BSRL}}(\mathbf{c},\mathbf{s})$.*

Proof. Let $\bar{c}$ be the mean of $\mathbf{c}$. Then $\bar{c}' = 1 - \bar{c}$ is the mean of the complement vector $\mathbf{1} - \mathbf{c}$. It follows that

$$\begin{aligned} g^{\mathrm{BSRL}}(\mathbf{1}-\mathbf{c},\mathbf{s}) &= \mathrm{Corr}(\mathbf{s},\mathbf{1}-\mathbf{c}) \\ &= \frac{\langle \mathbf{s} - \bar{s}\mathbf{1}, \mathbf{1} - \mathbf{c} - (1-\bar{c})\mathbf{1}\rangle}{\|\mathbf{s} - \bar{s}\mathbf{1}\|\|\mathbf{1} - \mathbf{c} - (1-\bar{c})\mathbf{1}\|} \\ &= \frac{\langle \mathbf{s} - \bar{s}\mathbf{1}, -\mathbf{c} + \bar{c}\mathbf{1}\rangle}{\|\mathbf{s} - \bar{s}\mathbf{1}\|\|-\mathbf{c} + \bar{c}\mathbf{1}\|} \\ &= -g^{\mathrm{BSRL}}(\mathbf{c},\mathbf{s}). \end{aligned}$$

□

Since BSRL is a correlation, it trivially satisfies translation invariance (4). It then follows directly from Theorem 2 that, if the biserial update rule is initialized with a student vector $\mathbf{s}^{(0)}$, then it is robust to misleading questions.

Corollary 2. BSRL *is robust to misleading questions.*

Finally, we show that BSRL is equivalent to an eigenvector problem, which also shows it is convergent. Denote $\mathbf{W} = \mathbf{I}_m - \mathbf{1}_m\mathbf{1}_m^\intercal/m$ and let $\mathbf{D}_C$ to be the $n \times n$ diagonal matrix with entries $1/\|\mathbf{C}_j - \mathbf{1}_m f_j\|$. In matrix form, the updates are

$$\begin{aligned} \mathbf{q}^{(t-1)} &= \mathbf{D}_C\mathbf{C}^\intercal\mathbf{W}\mathbf{s}^{(t-1)} \\ \mathbf{s}^{(t)} &= \mathbf{C}\mathbf{q}^{(t-1)}. \end{aligned}$$

The update rule written just in terms of the student scores $\mathbf{s}^{(t)}$ is

$$\mathbf{s}^{(t)} = \mathbf{C}\mathbf{D}_C\mathbf{C}^\intercal\mathbf{W}\mathbf{s}^{(t-1)} = (\mathbf{C}\mathbf{D}_C\mathbf{C}^\intercal\mathbf{W})^t\mathbf{s}_0, \tag{8}$$

where the second equality follows by recursively applying the first. As with HITS and cHITS, these are un-normalised iterations of the power method, except that $\mathbf{C}\mathbf{D}_C\mathbf{C}^\intercal\mathbf{W}$ is not symmetric. Hence, as before, we can show that the final ranking of BSRL is equivalent to computing an eigenvector.

Theorem 4. *Suppose* $\mathbf{s}^{(t)}$ *are computed according to* (8)*. Then* $\mathbf{s}^{(t)}/\|\mathbf{s}^{(t)}\| \to \mathbf{C}\mathbf{D}_C\mathbf{C}^\intercal\mathbf{e}$*, where* $\mathbf{e}$ *is the eigenvector of the symmetric positive definite matrix* $\mathbf{W}^\intercal\mathbf{C}\mathbf{D}_C\mathbf{C}^\intercal\mathbf{W}$ *with largest eigenvalue that is not orthogonal to* $\mathbf{s}_0$*.*

Proof. The proof of this is similar to that of Theorem 3, by observing that

$$\mathbf{s}^{(t)} = (\mathbf{C}\mathbf{D}_C\mathbf{C}^\intercal\mathbf{W}\mathbf{W}^\intercal)^t\mathbf{s}_0$$

and

$$(\mathbf{C}\mathbf{D}_C\mathbf{C}^\intercal\mathbf{W}\mathbf{W}^\intercal)^t = \mathbf{C}\mathbf{D}_C\mathbf{C}^\intercal(\mathbf{W}^\intercal\mathbf{C}\mathbf{D}_C\mathbf{C}^\intercal\mathbf{W})^{t-1}.$$

The proof now follows in the exact same fashion as in that of Theorem 3. □

4.5 Logistic update (LogR)

BSRL measures the correlation between $\mathbf{s}$ and the question response vectors $\mathbf{C}_j$. Since question responses are binary, *logistic regression* may be a more appropriate way to model this relationship.

In logistic regression, we aim to predict each response C_{ij} with the estimated student score s_i by assuming $\mathbf{P}(C_{ij} = 1 \mid s_i) = (1 + \exp(-a_j s_i - b_j))^{-1}$, where a_j, b_j are parameters to be learned. The responses $\mathbf{C}_j$ are positively correlated with $\mathbf{s}$ when $a_j > 0$, and vice versa for negative a_j.

The logistic regression based update rule is therefore

$$g^{\text{LogR}}(\mathbf{c}, \mathbf{s}) := \arg_a \max_{a,b} \sum_{i=1}^{m} \left[c_i(x_i) - \ln\left(1 + \exp(x_i)\right)\right], \tag{LogR}$$

with $x_i = as_i + b$, which is simply the Maximum Likelihood Estimation (MLE) for the logistic regression model specified above.

By repeating the arguments used for BSRL, we can show that LogR is robust.

Corollary 3. *LogR is robust to misleading questions.*

The proof of this is again a direct application of Theorem 2, given Lemmas 4 and 5 below which show that the LogR update satisfies translation invariance (4) and antisymmetry (5).

Lemma 4 (Translation invariance of LogR)**.** *Let* $\mathbf{c} \in \{0,1\}^m$*,* $\mathbf{s} \in \mathbb{R}^m$ *and* $\alpha \in \mathbb{R}$ *be arbitrary. Then*

$$g^{LogR}(\mathbf{c}, \mathbf{s} + \alpha\mathbf{1}_m) = g^{LogR}(\mathbf{c}, \mathbf{s}).$$

Proof. Let

$$a^+, b^+ = \arg\max_{a,b} \sum_{i=1}^{m} [c_i(a(s_i + \alpha) - b) - \log\left(1 + \exp(a(s_i + \alpha) - b)\right)], \text{and}$$

$$a^*, b^* = \arg\max_{a,b} \sum_{i=1}^{m} \left[c_i(as_i - b) - \log\left(1 + \exp(as_i - b)\right)\right].$$

We will show that $a^+ = a^*$ and $b^+ = b^* + \alpha a^+$.

Substituting $a' = a, b' = b + \alpha a$ into the former objective gives us latter objective, so the former problem upper bounds the latter problem. Substituting $a' = a, b' = b - \alpha a$ into the latter objective gives us the former objective, thus the latter problem upper bounds the former problem. Therefore the two problems are equivalent, and any optimal solution to the former problem can be constructed into a solution for the latter problem with the same a-term. □

Lemma 5 (Antisymmetry of LogR)**.** *Let* $\mathbf{c} \in \{0,1\}^m$ *be a fixed binary vector, and* $\mathbf{s} \in \mathbb{R}^m$ *be a fixed real vector. Then*

$$g^{LogR}(\mathbf{c}, \mathbf{s}) = -g^{LogR}(\mathbf{1} - \mathbf{c}, \mathbf{s}).$$

Proof. Observe that

$$\begin{aligned}
&\arg\max_{a,b} \sum_{i=1}^{m} \left[(1-c_i)(as_i-b) - \log\left(1+\exp(as_i-b)\right)\right] \\
&= \arg\max_{a,b} \sum_{i=1}^{m} \left[c_i(-as_i+b) + \log\left(\frac{\exp(as_i-b)}{1+\exp(as_i-b)}\right)\right] \\
&= \arg\max_{a,b} \sum_{i=1}^{m} \left[c_i(-as_i+b) - \log\left(1+\exp(-as_i+b)\right)\right] \\
&= -\arg\max_{a,b} \sum_{i=1}^{m} \left[c_i(as_i-b) - \log\left(1+\exp(as_i-b)\right)\right].
\end{aligned}$$

□

4.6 Runtime Discussion

Computing the student score vector $\mathbf{s}$ for AvgSc takes time $O(mn)$. HITS essentially computes two matrix-vector multiplications with $\mathbf{C}$, at cost $O(mn)$ each iteration. Centered HITS requires a additional pass through the vector of student scores, however, it retains the $O(mn)$ asymptotic complexity. BSRL consists of one matrix-vector product, which is $O(mn)$, and n correlation computations, at $O(m)$ each. The total cost per iteration is $O(mn)$. For LogR, each iteration we solve n different two-dimensional convex programs, each of which takes $O(\log(1/\epsilon))$ iterations to solve up to accuracy ϵ when using interior point solvers. Since the objective is a sum of m terms, computing gradients for this sum takes $O(m)$ time. The total cost per iteration is $O(mn\log(1/\epsilon))$.

4.7 Methods for missing responses

Suppose each student is assigned only a subset $k < n$ of questions. Since students may have different point totals available to them, we need to ensure that the grading is both *transparent* and *fair*.

Transparent aggregation. For any fixed question weights, it is natural to let a student's score be the fraction of the available points which they received. Let student i attempt the set of questions indexed by N_i. The best score achievable on the subset of questions N_i is $\sum_{j\in N_i} \max\{0, q_j\}$. Observe that on the ill-coded questions we expect a perfect student to answer "incorrectly" and score 0 instead of $q_j < 0$. Student scores are the fraction of this maximum achievable score attained, we replace (3) with

$$s_i = \frac{\sum_{j\in N_i} C_{ij} q_j}{\sum_{j\in N_i} \max\{0, q_j\}}. \tag{9}$$

It is also necessary to modify the calibration functions our methods use to take missing responses into consideration:

1) AvgSc: No modification beyond dividing by the number of questions attempted is required.

2) HITS: We modify g^{HITS} analogously to (10). Let M_j be the set of students who attempted question j. Now

$$g^{\text{HITS-M}}(\mathbf{C}_j, \mathbf{s}) = \frac{\sum_{i\in M_j} C_{ij} s_i}{\sum_{i\in M_j} \max\{0, s_i\}}.$$

3) Centered HITS: As for HITS, using the centered vector of student scores, so $g^{\text{cHITS-M}}(\mathbf{C}_j, \mathbf{s}) = g^{\text{HITS-M}}(\mathbf{C}_j, \mathbf{s} - \bar{s}\mathbf{1}_m)$.

4) Biserial: $\text{Corr}(\mathbf{C}_j, \mathbf{s}^{(t)})$ can be expressed as a sum over students – we limit the expression to the terms for students who attempted question j. Let $\mathbf{C}_j^j$ and $\mathbf{s}^j$ denote vectors $(C_{ij})_{i\in M_j}$ and $(s_i)_{i\in M_j}$, respectively. Now $g^{\text{BSRL-M}}(\mathbf{C}_j, \mathbf{s}) = \text{Corr}(\mathbf{C}_j^j, \mathbf{s}^j) = g^{\text{BSRL}}(\mathbf{C}_j^j, \mathbf{s}^j)$

5) LogR: We fit a logistic curve through the points $(C_{ij}, s_i)_{i\in[m]}$ to determine the weight of question j. If only a subset of the students attempt a question, we fit the curve through $(C_{ij}, s_i)_{i\in M_j}$ instead with the calibration rule $g^{\text{LogR-M}}(\mathbf{C}_j, \mathbf{s}) = g^{\text{LogR}}(\mathbf{C}_j^j, \mathbf{s}^j)$.

Table 1: Percentage of correct responses when $\nu = 0$.

$\omega =$	-1	-0.5	0	0.5	1
$\mathbb{E}[\mathbb{P}(\text{correct})]$	0.94	0.81	0.49	0.21	0.04

5 Experimental evaluation

Experimental setup. We evaluate the methods on both synthetic data and real-world data. Our algorithms are implemented in Python, and we compare to the widely used R package MIRT [9]. Experiments are run on a machine with a 2.3GHz CPU and 64GB RAM.

Summary of findings. Our key findings are:

(1) *Accuracy & robustness*: When students attempt all questions, both LogR and BSRL are robust to the presence of at least 40% misleading questions, and are at least as accurate as MIRT across all parameter regimes (Figure 4).

(2) *Resilience*: When students answer only a subset of questions, LogR is consistently the most accurate method when questions are easy, and the robust methods are comparable to MIRT when questions are difficult (Figure 5).

(3) *Identification* When questions are easy, LogR and BSRL identify ill-coded questions much better than MIRT; when questions are harder all methods perform well (Figure 6).

(4) *Scalability* We observe that our methods scale linearly in n and are orders of magnitude faster than MIRT (Figure 7).

(5) When sampling scenarios from a real-world dataset, LogR leads to a statistically significant increase in ranking accuracy over MIRT.

We conclude that LogR satisfies the desiderata set out in Section 1.

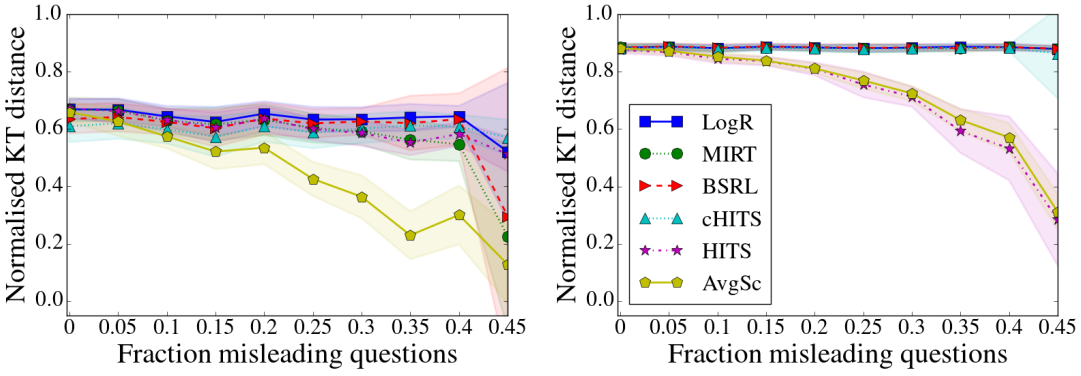
5.1 Experiments on synthetic data

2PL model. We simulate data with the 2PL model from Section 2.1. We have three vectors of parameters $\mathbf{alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}$ for the question discrimination factors and difficulties, and the student abilities. The ground truth (GT) rank of the students is according to their abilities θ_i. All parameters are uniformly and independently distributed: $\theta_i \sim U(\theta_{\min}, \theta_{\max})$, $\alpha_j \sim U(0, \alpha_{\max})$, $\beta_j \sim U(\beta_{\min}, \beta_{\max})$. We normalize the parameter space without loss of generality by drawing $\theta_i \in U(0, 1)$, and set $\alpha_{\max} = 4$. We assume the range of β is similar to that of θ, so $\beta_{\max} = \beta_{\min} + 1$, and characterize the relationship between student abilities and question difficulties by $\omega := \beta_{\min}$.

Table 1 illustrates the expected fraction of correct responses for different values of ω. On average, we expect questions to have smaller difficulties than the students' abilities (it is challenging to write difficult and relevant questions), implying $\omega < 0$. This regime also corresponds to what we have encountered in our classrooms ($\omega \leq 0$) and we distinguish between the two extremes: $\omega = 0$, when every student submits a question with difficulty roughly equal to his ability; and $\omega = -1$, when students submit *easy* questions, with difficulty significantly lower than their ability. Results for intermediate values of ω may be interpolated from these extremes and are omitted. In order to analyze the impact of *ill-posed questions*, we parameterize the fraction of misleading questions by flipping the sign $\alpha_j \leftarrow -\alpha_j$ for a fraction of ν questions.

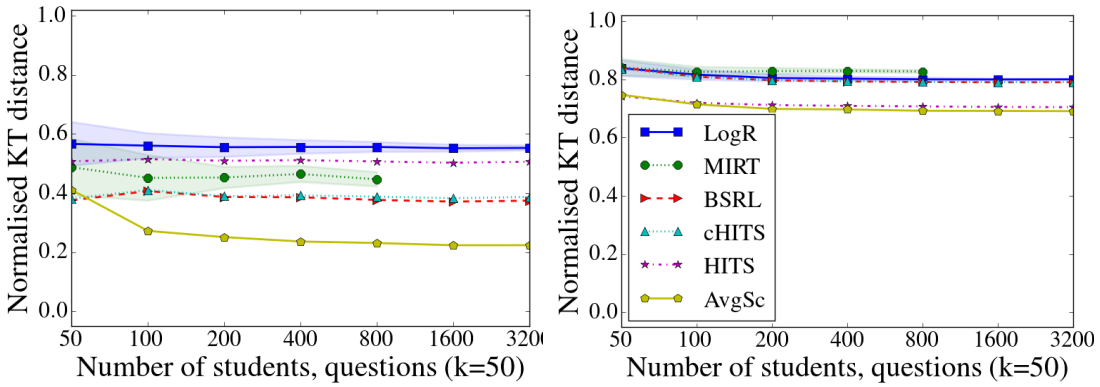
Methods. For all methods described in Section 3 we set $\mathbf{s}_0 = \mathbf{C1}_n$ and add a normalisation step at every iteration so that $||q^{(t)}|| = 1$. This scaling prevents numerical difficulties and does not affect the rankings. We compare these methods against a maximum likelihood estimator for the 2PL model computed by the `mirt`-package [9] (MIRT). For further details on the maximum likelihood estimation for IRT models and the `mirt` package, see [6, 8, 9, 18].

Accuracy. We measure the accuracy of an output ranking using two metrics: the *normalised Kendall Tau* (KT) score between two rankings, and the *average normalised displacement* (AD) of a student as compared to the true ranking. The normalised KT score is the average number of concordant pairs between the two rankings minus the average number of discordant pairs, divided by the number of pairs, this value ranges in $[-1, 1]$. A randomly generated ranking has expected score 0 from the ground truth, since we expect 50% of the pairwise orderings in a random ranking to be correct. The normalised displacement of a student is the distance from his true position, divided by the number of students in



(a) KT: $\omega = -1$ (b) KT: $\omega = 0$

Figure 4: Section 5.1.1: Robustness as ν changes.



(a) KT: $\nu = 0.2, \omega = -1$ (b) KT: $\nu = 0.2, \omega = 0$

Figure 5: Section 5.1.2: Students answer a constant $k = 50$ questions while $n = m$ increases from 50 to 3 200.

the class. Note that higher values are better for KT-scores, while the opposite is true for AD. These two metrics generally lead to similar conclusions, so we report AD only when particularly interesting.

5.1.1 Accuracy & Robustness to misleading questions

We compare the accuracy of our methods while varying the fraction of ill-coded questions $\nu \in \{0, 0.05, 0.1, \ldots, 0.45\}$. Figures 4a and 4b represent the normalized KT score of our experiments with 100 students and 100 questions, averaged over 500 runs.

Results. cHITS, BSRL and LogR match MIRT in terms of performance and robustness across both metrics. Interestingly, the three methods show no decrease in performance when the fraction of misleading question increases up to $\nu = 0.4$. This is a stronger property than implied by theory: Theorems 1, 2 and 3 imply the quality of the ranking remain unchanged for increasing ν as long as $\mathbf{s}_0$ is unchanged. However, in our experiments $s_0 = \mathbf{C1}$ which changes with ν, yet we still witness remarkable robustness to misleading questions. The baseline methods AvgSc and HITS degrade significantly as misleading questions are introduced. When questions are easy ($\omega = -1$) and there are many ill-coded questions ($\nu \geq 0.35$) our novel, provably robust methods outperform MIRT.

5.1.2 Resilience to missing responses

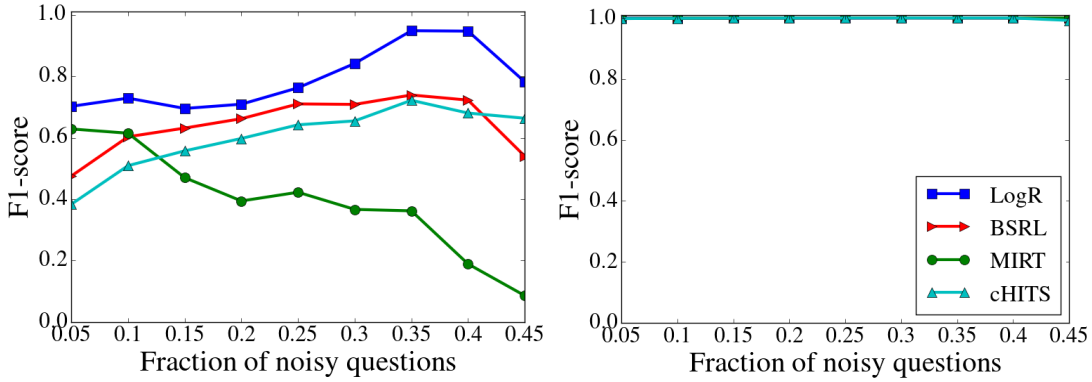
We fix the number of questions answered by each student at $k = 50$ and increase $m = n$ up to 3 200. The noise parameter ν is varied in $\{0, 0.2, 0.4\}$. We report results averaged over 100 repetitions for $\nu = 0.2, \omega \in \{-1, 0\}$ in Figure 5. We were unable to scale MIRT to $n \geq 800$ (also see fig. 7).

Results. When $\omega = 0$, MIRT is marginally more accurate than LogR. HITS and AvgSc perform

worst. When $\omega = -1$, LogR is significantly better than MIRT. We also observe clear separation between LogR and the other provably robust iterative methods.

It is extremely encouraging that accuracy does not degrade as class size increases and $\mathbf{C}$ becomes more sparse.

5.1.3 Identification of misleading questions



(a) F1-score: $\omega = -1$

(b) F1-score: $\omega = -0$

Figure 6: Section 5.1.3: F1-score for identifying misleading questions as ν changes.

Figure 6 reports the F1-score (harmonic mean of precision and recall) for the classification task of identifying ill-posed questions for the experiment in Section 5.1.1. The ground truth is taken to be those questions with $a_j < 0$: a method classifies a question correctly if $\text{sgn}(a_j) = \text{sgn}(q_j)$. Recall neither AvgSc nor HITS can flag misleading questions.

Results. When $\omega = 0$, all methods identify misleading questions almost perfectly. When $\omega = -1$ the iterative methods, and in particular LogR, have significantly higher f1 scores than MIRT. Closer investigation reveals that while MIRT's precision remains comparable LogR's, its recall drops precipitously for $\nu \in [0.1, 0.3]$.

5.1.4 Scalability

We fix $n = 100$, $\omega = 0$, $\nu = 0.2$, and increase m between 100 and 50 000. Figure 7 summarizes the runtimes.

Results. Our novel methods scale linearly in the number of students and are orders of magnitude faster than MIRT. The difference in speed between BSRL and LogR is likely due to the fact that in BSRL there is a simple closed form expression for computing question weights, while LogR requires solving an optimization program. As discussed in [17], computing the marginal maximum likelihood for the IRT generative models involves intractable integrals. When these integrals are evaluated using Gaussian quadrature [7], the number of evaluation points increase dramatically in the number of latent trait parameters to be estimated. Adaptive quadrature methods [27, 28] reduce the number of points to evaluate, however, we do not expect these methods to scale gracefully with the number of students and questions. Our experiments suggest a growth rate quadratic in the number of students even when students are described using a single latent variable (their one-dimensional ability).

5.2 Real data

In our experiments with synthetic data, LogR performed as well, and even better than, MIRT across a variety of metrics and parameter regimes. We now compare these methods on a real-world data set collected during the teaching of a course.

Setup. We collected a dataset of student-sourced questions and responses.As part of regular homeworks, students were asked to submit multiple-choice questions that relate to that week's class material, and also answer such questions created by fellow students (recall Example 1). The dataset contains

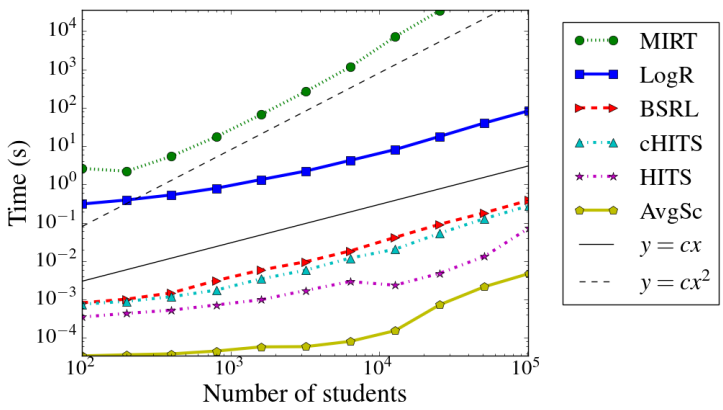


Figure 7: Runtime as the number of students increase while the number of questions $n = 100$ remains constant: The iterative methods scale linearly in the number of students.

Table 2: Results on real data (32 students and 223 question)

	KT	AD
LogR	**0.535**	**0.167**
MIRT	0.424	0.202

32 students and 223 contributed questions. The average student answered approximately 40% of the questions and the quality of the questions, and whether they are ill-coded or not, is not known. We use the final class scores of the students to compute their ground truth ranking.

In addition to finding a single ranking based on the given response matrices, we repeatedly sample 16 students and 110 questions to test for statistically significant differences in performance.

Results. Table 2 compares the performance of LogR and MIRT on the real-world dataset. LogR is more accurate than MIRT in terms of both KT-score and average displacement. The difference of 3.5% in the average displacement between LogR and MIRT means that, on a class size of 1 000, we expect LogR to rank students 35 positions closer to their true positions than MIRT.

When sampling subsets of students and questions 100 times, the improvement of LogR over MIRT is statistically significant at the $p = 0.01$ confidence level for both metrics.

6 Conclusions

We investigated the problem of *ranking students in the presence of ill-posed questions*. We identified properties that make ranking methods robust to the presence of such questions, suggest several algorithms that fulfill these properties, and experimentally confirm their usefulness. In particular, our new iterative update methods based on biserial correlation and on logistic regression scale linearly in the number of students while matching the performance of current state-of-the art parameter estimation methods for IRT models (which are *not transparent* to students and *do not scale* to large class sizes).

Our theoretical analysis and experiments suggest that these methods have great potential for grading students in a way that is (1) accurate, (2) robust to misleading questions, (3) resilient to missing responses, (4) linearly scalable, (5) transparent and fair.

References

[1] E. B. Andersen. A goodness of fit test for the rasch model. *Psychometrika*, 38(1):123–140, 1973.

[2] E. B. Andersen. Estimating latent correlations between repeated testings. *Psychometrika*, 50(1):3–16, 1985.

[3] R. Andersen, C. Borgs, J. Chayes, U. Feige, A. Flaxman, A. Kalai, V. Mirrokni, and M. Tennenholtz. Trust-based recommendation systems: an axiomatic approach. In *WWW*, pages 199–208, 2008.

[4] F. K. Baker and S. Kim. *Item Response Theory: Parameter Estimation techniques (2nd Ed)*. Marcel Dekker Inc., 2004.

[5] D. P. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[6] R. D. Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51, 1972.

[7] R. D. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459, 1981.

[8] L. Cai and D. Thissen. Modern approaches to parameter estimation in item response theory. *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*, pages 41–59, 2014.

[9] R. P. Chalmers et al. mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48(6):1–29, 2012.

[10] M. T. H. Chi, M. W. Lewis, P. Reimann, and R. Glaser. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13:145–182, 1989.

[11] C. de Kerchove and P. Van Dooren. The pagetrust algorithm: How to rank web pages when negative links are allowed? In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 346–352. SIAM, 2008.

[12] S. E. Embretson. A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56(3):495–515, 1991.

[13] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In *SIGMOD*, pages 61–72, 2011.

[14] M. R. N. w. c. b. A. B. Frederic M. Lord. *Statistical Theories of Mental Test Scores*. Addison-Wesley, 1968.

[15] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *WSDM*, pages 131–140, 2010.

[16] M. Gupta and J. Han. Heterogeneous network-based trust analysis: a survey. *ACM SIGKDD Explorations Newsletter*, 13(1):54–71, 2011.

[17] M. Jeon and F. Rijmen. Recent developments in maximum likelihood estimation of mtmm models for categorical data. *Frontiers in psychology*, 5, 2014.

[18] M. S. Johnson et al. Marginal maximum likelihood estimation of item response models in r. *Journal of Statistical Software*, 20(10):1–24, 2007.

[19] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*, pages 640–651. ACM, 2003.

[20] N. M. Kingston and N. J. Dorans. The feasibility of using item response theory as a psychometric model for the gre aptitude test. *ETS Research Report Series*, 1982(1), 1982.

[21] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *JACM*, 46(5):604–632, 1999.

[22] F. M. Lord, M. R. Novick, and A. Birnbaum. Statistical theories of mental test scores. 1968.

[23] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *COLING*, pages 877–885, 2010.

[24] J. Pasternack and D. Roth. Generalized fact-finding. In *WWW*, pages 99–101, 2011.

[25] J. Pasternack and D. Roth. Latent credibility analysis. In *WWW*, pages 1009–1021, 2013.

[26] J. Pasternack, D. Roth, and V. G. V. Vydiswaran. Information trustworthiness. AAAI Tutorial, 2013.

[27] J. C. Pinheiro and D. M. Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, 4(1):12–35, 1995.

[28] S. Rabe-Hesketh, A. Skrondal, A. Pickles, et al. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1):1–21, 2002.

[29] D. Rizopoulos. ltm: An r package for latent variable modeling and item response theory analyses. *Journal of statistical software*, 17(5):1–25, 2006.

[30] Y. Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

[31] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *arXiv preprint arXiv:1606.09632*, 2016.

[32] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *TKDE*, 20(6):796–808, 2008.

[33] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *PVLDB*, 5(6):550–561, 2012.

[34] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng. Truth inference in crowdsourcing: Is the problem solved? *PVLDB*, 10(5):541–552, 2017.

A Additional Material

A.1 Sufficient information in the IRT model

Denote by $p(\theta) \in [0,1]$ the probability that a student of ability $\theta \in \mathbb{R}$ will answer a randomly generated question correctly. One of the key assumptions in Section 3 is sufficient information: a student with higher ability is more likely to answer a randomly chosen question correctly than a student with lower ability. Formally, this is ensured by assuming that $p(\theta)$ is an increasing function in θ. We now explore conditions needed for this to be satisfied under the 2PL IRT model specified in (1). We note that our results can be extended easily to the more general 3PL model.

Under (1), we can write

$$\begin{aligned} p(\theta) &= \mathbf{E}_{a,b}\left(\frac{1}{1+\exp(-a(\theta-b))}\right) \\ &= \int_{a,b} \frac{1}{1+\exp(-a(\theta-b))} f(a,b)\, da\, db, \end{aligned}$$

where the expectation is taken over the randomness in the question parameters a, b resulting from the question generation method, and $f(a,b)$ is a density function. To ensure that $p(\theta)$ is non-decreasing in θ, we can equivalently ensure that the derivative $dp/d\theta \geq 0$. This is given by

$$\frac{dp}{d\theta} = \int_{a,b} f(a,b) g(a;\theta,b)\, da\, db,$$

where $g(a;\theta,b) := \frac{a\exp(-a(\theta-b))}{(1+\exp(-a(\theta-b)))^2}$. Now observe that for *any* θ and b, the function $g(a;\theta,b)$ is an odd function of a, i.e., $g(-a;\theta,b) = -g(a;\theta,b)$, thus we have

$$\begin{aligned} \frac{dp}{d\theta} &= \int_{a,b} f(a,b) g(a;\theta,b)\, da\, db \\ &= \int_b \int_{a=0}^{\infty} \left(f(a,b)g(a;\theta,b) - f(-a,b)g(a;\theta,b)\right)\, da\, db \\ &= \int_b \int_{a=0}^{\infty} g(a;\theta,b)\left(f(a,b) - f(-a,b)\right)\, da\, db. \end{aligned}$$

A sufficient condition for $dp/d\theta \geq 0$ is

$$\int_{a=0}^{\infty} g(a;\theta,b)\left(f(a,b) - f(-a,b)\right) da \geq 0,$$

and since $g(a;\theta,b) \geq 0$ for $a \geq 0$, a sufficient condition for this is $f(a,b) \geq f(-a,b)$ for any given $a \geq 0$ and b. In terms of question generation, this condition states that for every $|a|$, the likelihood of the question being reliable ($a \geq 0$) is at least as much as it being bad ($a \leq 0$).

In summary, to ensure that the sufficient information assumption holds, we require that the question generation process is at least as likely to generate good questions than bad questions.

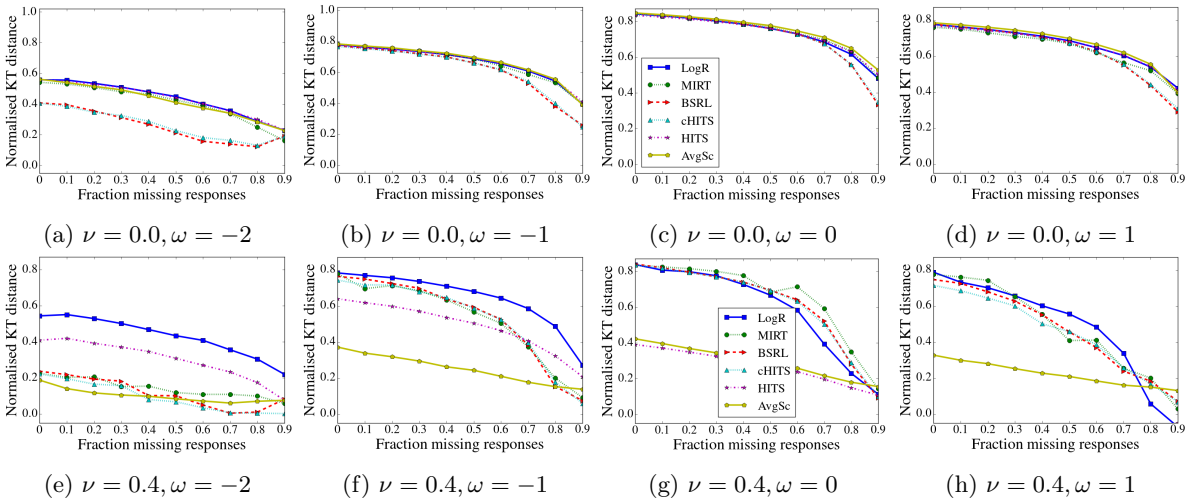
A.2 Additional computational results

We report the results for additional parameter regimes which were omitted from the main body of the paper, mainly the results testing the resiliency of our methods to missing responses under different noise levels.

A.2.1 Resilience to missing responses

In the main body we report results for $\nu = 0.2$, here we include the results for $\nu = 0$ and $\nu = 0.4$ in Figure 8. We again omit $\omega = 2$ since this parameter regime is highly unlikely to reflect reality.

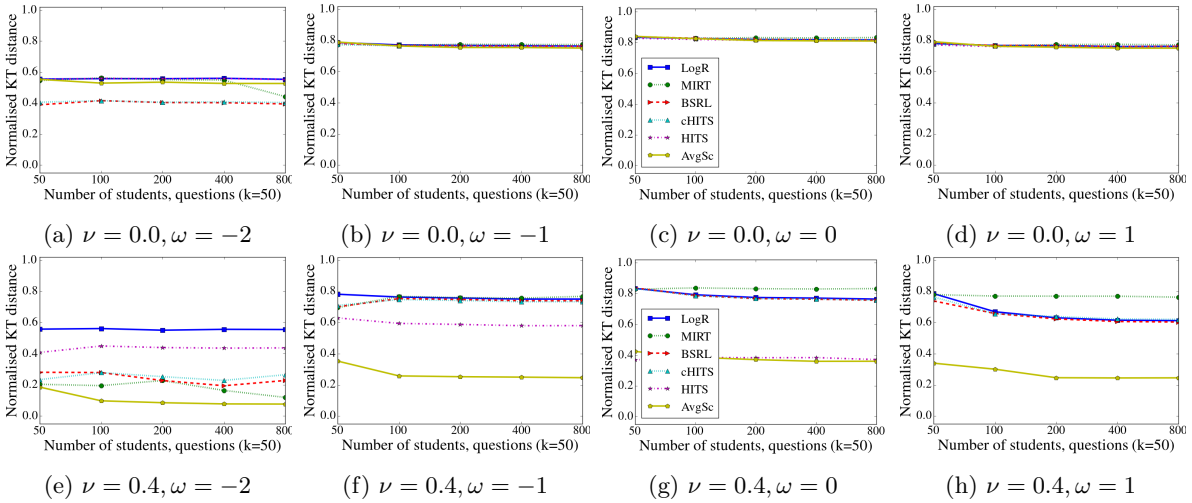
We similarly report additional results for the experiments where we fix $k = 50$ and increase the size of the class in Figure 9.



(a) $\nu = 0.0, \omega = -2$ (b) $\nu = 0.0, \omega = -1$ (c) $\nu = 0.0, \omega = 0$ (d) $\nu = 0.0, \omega = 1$

(e) $\nu = 0.4, \omega = -2$ (f) $\nu = 0.4, \omega = -1$ (g) $\nu = 0.4, \omega = 0$ (h) $\nu = 0.4, \omega = 1$

Figure 8: Resilience of different methods an increasing fraction of missing responses under different parameter regimes.



(a) $\nu = 0.0, \omega = -2$ (b) $\nu = 0.0, \omega = -1$ (c) $\nu = 0.0, \omega = 0$ (d) $\nu = 0.0, \omega = 1$

(e) $\nu = 0.4, \omega = -2$ (f) $\nu = 0.4, \omega = -1$ (g) $\nu = 0.4, \omega = 0$ (h) $\nu = 0.4, \omega = 1$

Figure 9: Resilience of different methods to missing responses when students answer a constant $k = 50$ questions and the class size $m = n$ increases.

It is interesting to see in both these experiments that when there are no misleading questions, AvgSc which considers only a student's fraction of correct responses performs very well. We observe as before that in the presence of misleading questions, LogR is comfortably the most accurate method when $\nu < 0$, and that MIRT performs well when $\nu = 0$.

A.3 Handling missing responses: Imputation

The preceding discussions of missing responses assumed that we exclude them from our computations in a principled way. Alternatively, we can estimate the value of a missing response based on the parts of $\mathbf{C}$ which are known. This ties in to a large literature on collaborative filtering.

For example, can estimate a missing response of student i on question j by looking at the responses of 'similar' students to this question. Let $N_{ik} = N_i \cap N_k$, in other words, the questions attempted by both student i and student k. We propose using the biserial correlation coefficient between student i and k's vector of responses to this set of jointly attempted questions as a measure of similarity. Formally, let χ_{ik} be the correlation between $(C_{i\ell})_{\ell \in N_{ik}}$ and $(C_{k\ell})_{\ell \in N_{ik}}$ then the average similarity weighted score of

student i is

$$S_{ij} = \frac{\sum_{k \in M_j} \chi_{ik} \times 2(C_{kj} - 1)}{|M_j|}$$

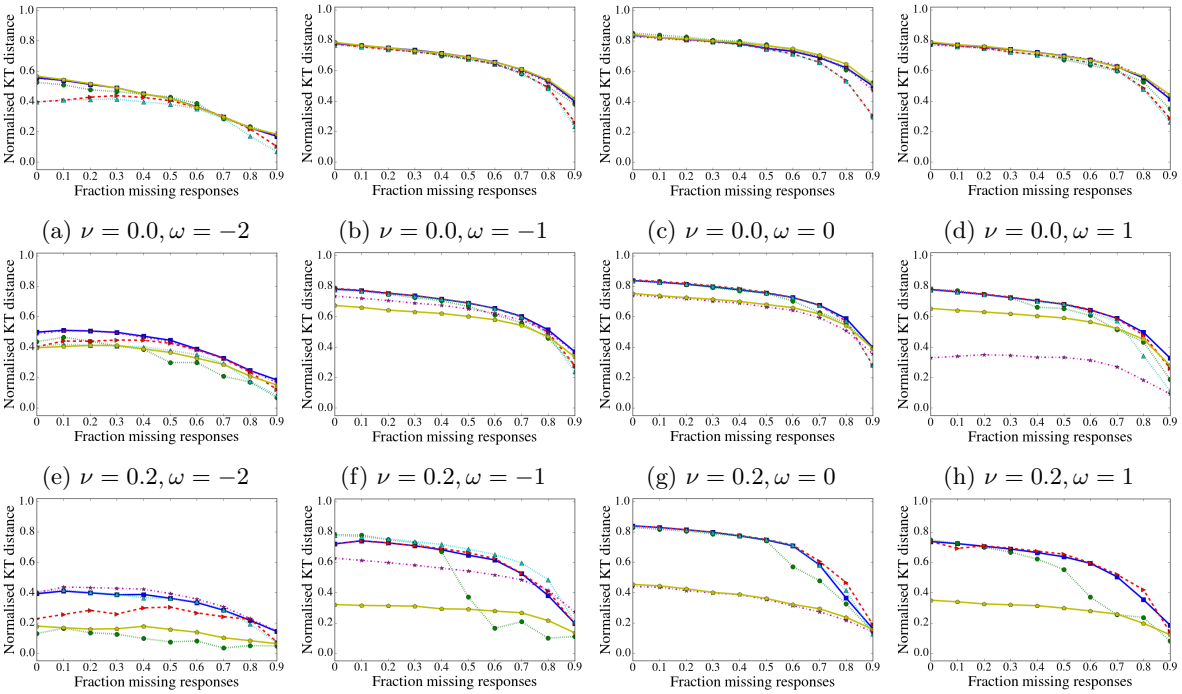
and we estimate $\tilde{C}_{ij} = \frac{1}{2}(1 + S_{ij})$. Observe for example that if $\chi_{ik} = 1$, in other words the response of i to the questions N_{ik} is identical to the responses of student k, then student k contributes exactly $2C_{kj}$ to S_{ij}. The imputed matrix $\tilde{\mathbf{C}}$ (or a rounded version of it) can be used in all our update methods, however, this comes at the cost of transparency since a student's score will now be the result of his submitted answers and our estimates of his answers for questions he did not attempt.

A similar imputation method based on the similarity between questions can be used, indeed, these two methods are closely related to techniques in item- and user-based collaborative filtering.

AvgSc, HITS, cHITS and BSRL can be run without modification on the imputed matrix $\tilde{\mathbf{C}}$. MIRT and LogR require a binary matrix of responses, we form this $\hat{\mathbf{C}}$ by randomly rounding each entry of $\tilde{\mathbf{C}}$ so that $P[\hat{C}_i j = 1] = \tilde{C}_{ij}$ for all $i \in [m], j \in [n]$.

Figure 10 shows the performance of the algorithms on such an imputed matrix for $n = m = 50$. A comparison to Figure 5 and **??** shows that although imputation does not necessarily improve the performance of the best method in every parameter regime significantly, it does seem to assist some of the methods, for example BSRL, in remaining competitive in a larger number of parameter regimes. These results do not change significantly when using question-similarity instead of student-similarity for imputation, in part because $n = m = 50$.

The usefulness of imputation in practice will depend on the ranking method used (for example it may be more useful when using BSRL compared to LogR) and the premium placed on transparency.



(a) $\nu = 0.0, \omega = -2$ (b) $\nu = 0.0, \omega = -1$ (c) $\nu = 0.0, \omega = 0$ (d) $\nu = 0.0, \omega = 1$

(e) $\nu = 0.2, \omega = -2$ (f) $\nu = 0.2, \omega = -1$ (g) $\nu = 0.2, \omega = 0$ (h) $\nu = 0.2, \omega = 1$

(i) $\nu = 0.4, \omega = -2$ (j) $\nu = 0.4, \omega = -1$ (k) $\nu = 0.4, \omega = 0$ (l) $\nu = 0.4, \omega = 1$

Figure 10: Resilience of different methods an increasing fraction of missing responses under different parameter regimes after estimating the values of the missing entries in the response matrix.