# JAG: A Crowdsourcing Framework for Joint Assessment and Peer Grading

**Igor Labutov**
Carnegie Mellon University
Pittsburgh, PA 15213

**Christoph Studer**
Cornell University
Ithaca, NY 14853

## Abstract

Generation and evaluation of crowdsourced content is commonly treated as two separate processes, performed at different times and by two distinct groups of people: content creators and content assessors. As a result, most crowdsourcing tasks follow this template: one group of workers generates content and another group of workers evaluates it. In an educational setting, for example, content creators are traditionally students that submit open-response answers to assignments (e.g., a short answer, a circuit diagram, or a formula) and content assessors are instructors that grade these submissions. Despite the considerable success of peer-grading in massive open online courses (MOOCs), the process of test-taking and grading are still treated as two distinct tasks which typically occur at different times, and require an additional overhead of grader training and incentivization. Inspired by this problem in the context of education, we propose a general crowdsourcing framework that fuses open-response test-taking (content generation) and assessment into a single, streamlined process that appears to students in the form of an explicit test, but where everyone also acts as an implicit grader. The advantages offered by our framework include: a common incentive mechanism for both the creation and evaluation of content, and a probabilistic model that jointly models the processes of contribution and evaluation, facilitating efficient estimation of the quality of the contributions and the competency of the contributors. We demonstrate the effectiveness and limits of our framework via simulations and a real-world user study.

## 1 Introduction

Crowdsourcing open-ended content—for example, seeking an answer to “What is the best way to brew a cup of coffee?”—requires both: finding people *willing* to contribute an original answer and finding *competent* people to identify the correct answer(s) among these contributions. Traditional crowdsourcing mechanisms separate the two tasks into two stages: one group of workers contribute the content, and another group evaluates it. However, it is natural to assume that the same workers who contribute “good” original content will also be competent at evaluating the contributions of other workers on the same topic, and vice versa. Consequently, if our ultimate goal is to identify “good” content and competent contributors, it is natural to seek a statistical model that captures this intuition formally. In this work, we focus on the specific instance of this problem in the context of education, although our approach applies to general crowdsourcing tasks.

Recently, massive open online courses (MOOCs) brought crowdsourcing into the realm of education, via a task known as *peer-grading*. In its traditional form, peer-grading assigns to each student a secondary role of a grader. Students are responsible for validating the correctness of other students’ solutions in order to assign a score, typically in accordance with a rubric provided by an instructor. The advantage of peer-grading is its flexibility to students’ submissions, which may range from short answers, to essays, diagrams, code, or entire projects (Kulkarni et al. 2015). In a practical deployment, however, peer-grading, faces the same challenges as crowdsourcing. First, each student differs in their ability to grade, and the grades assigned by different students must be reconciled in a reasonable way. Second, grading becomes an additional burden on the students, and mechanisms must be put in place that not only incentivize participation and effort, but prevent students from “gaming” the process. Only recently, research in peer-grading has started to address all these challenges (Piech et al. 2013; Raman and Joachims 2014; Wu et al. 2015).

The remainder of this paper will focus on the problem of open-response assessment and grading in the context of education. Nevertheless, the presented framework, model, and algorithms are not limited to this context, and can be applied to many crowdsourcing tasks where the goal is to identify high-quality contributions made by competent contributors.

### JAG: An alternative to Peer Grading

We present a novel, alternative approach to peer-grading that naturally resolves the challenges of grade aggregation and incentivization. We propose *joint assessment and grading (JAG)*, which fuses peer grading and assessment into a single, streamlined process by re-framing grading as additional testing. Our approach is motivated by the fact that a “grader” that has no answer key, when presented with the listing of other students’ answers, is no different than a test-taker facing a multiple-choice question (with multiple possible correct or incorrect answers). In other words: a student selecting what they believe to be the correct answer in a multiple choice question (MCQ) constructed from the open-response submissions of other students is in effect simultaneously (i) grading the other students and (ii) being assessed by their ability to select the correct answer. In peer-grading, we already face the challenge of noisy inputs from the (potentially unmotivated) graders. By re-framing the act of grading as that of MCQ testing, the source of the apparent noise in grading becomes distributed according to the ability of the students in the class.

The proposed mechanism of JAG combines the advantages of both worlds: the structure of multiple choice questions and the flexibility to general response types offered by *peer-grading*. First, by constructing the MCQs directly from students' open-response submissions, the questions naturally capture the distribution of misconceptions present in the population of students being tested, requiring little to no instructor input. Second, our framework offers a mechanism for automatically grading open-response submissions, thus facilitating greater student engagement and higher-order thinking characteristic to open-response questions (Haladyna 1997). Third, by re-framing the task of grading as that of testing, the students are incentivized in the context of a familiar task: namely by expending their effort towards correctly answering an MCQ, they are implicitly directing that effort towards grading other students' submissions. At the same time, the students are not burdened with (what they may perceive as) a "thankless" job of grading, but instead in the process of answering the additional MCQs, they are provided with an additional opportunity to demonstrate their knowledge.

In this paper, we formalize the process of JAG as a statistical estimation problem. At the heart of our approach is the traditional Rasch model (Rasch 1993) that captures the interaction between student abilities and question difficulties in determining the likelihood of a student answering a question correctly. We develop an expectation maximization (EM) algorithm for estimating the parameters of the proposed model in an unsupervised setting (i.e., in absence of an answer key), and demonstrate the effectiveness of our framework through a real-world user-study conducted on Amazon's Mechanical Turk. Additionally, we investigate the key properties and limitations of our approach via simulations.

## 2 Related Work

Our work builds on the recent progress in two distinct areas: *crowd-sourcing* and *peer-grading*, that we unite and extend within our JAG framework.

### Crowdsourcing

An important task in *crowdsourcing* is known as *label-aggregation*, and is concerned with the problem of optimally recovering some underlying ground truth (e.g., image class label) from a number of (unreliable) human judgements. See (Hung et al. 2013) for a detailed review. In the context of education, the task of automatically identifying the correct answers from open-response submissions is closely related to the task of label aggregation. Within the field of crowd-sourcing, the work of (Dawid and Skene 1979; Whitehill et al. 2009; Bachrach et al. 2012) are the most related to our approach. (Dawid and Skene 1979) was the first to suggest an expectation maximization (EM) algorithm for label aggregation, motivated by a clinical setting of making a diagnosis. More recently, (Whitehill et al. 2009) extended this approach to model the variation in task difficulty in the context of image labeling. In the context of education, (Bachrach et al. 2012) has proposed a statistical model for aggregating answers from "noisy" students, with the goal of automatically identifying the correct answers to MCQs. They deploy an expectation propagation (EP) algorithm for Bayesian inference, and demonstrate the ability to infer correct answers accurately in a setting of an IQ test. Our work can be seen as a generalization of the method proposed in (Whitehill et al. 2009; Bachrach et al. 2012), where we explicitly model the dependence among question choices and students that generate those choices in the context of answering open-response questions.

### Peer-grading

Much of the recent research in *peer-grading* addresses a related problem of aggregating a number of "noisy" grades submitted by students in a statistically principled manner. Models such as the ones in (Piech et al. 2013; Raman and Joachims 2014) pose the problem of peer-grading as that of statistical estimation. Since traditional grading assumes that graders are in possession of a grading rubric, statistical models of peer-grading are concerned primarily with accounting for the reliability and bias of graders in evaluating assignments against a gold-standard. In contrast to such "explicit" models of grading, we view grading as an implicit process that results as a by-product of students' genuine attempt to answer MCQs constructed from the open-response submissions of other students. As such, we do not require additional "grader-specific" parameters, as grading in our framework is subsumed by the response model (model of how students answer questions as a function of their ability and question difficulty). We note, however, that one of the proposed models in (Piech et al. 2013) explicitly couples grading and ability parameters in an attempt to capture the intuition that better students may also be better graders. This intuition can be viewed as being taken to its extreme in our setting: removing the boundary between grading and test-tasking ensures that better students are more reliable graders by construction.

## 3 Model

### Fully observed setting

We start by reviewing the classic IRT Rasch model that will serve as the foundation of our approach. Consider a set of students $S$ and a set of questions $Q$, where a student $i \in S$ is endowed with an ability parameter $s_i \in \mathbb{R}$, and each question $j \in Q$ is endowed with a difficulty parameter $q_j \in \mathbb{R}$ (note that we capitalize all sets in our notation). To simplify the notation, we will often overload $s_i$ to refer to both, the student index $i$ and their ability, depending on the context; the same applies to $q_j$, which we use to refer to the question itself as well as its difficulty. The well-established 1-PL IRT Rasch model (Rasch 1993) expresses the probability that the student $s_i$ answers question $q_j$ correctly via the following likelihood function:

$$P(z_{i,j} \mid s_i, q_j) = \frac{1}{1 + \exp\left(-z_{ij}(s_i - q_j)\right)}, \qquad (1)$$

where $z_{i,j} \in \{+1, -1\}$ is the binary outcome of student $s_i$'s attempt of question $q_j$; we use $+1$ and $-1$ to designate correct and incorrect responses, respectively. If we are in the possession of an answer key for each question, then we also know $\{z_{i,j}\}$, $\forall i, j$ (we will refer to this as the *fully observed* setting). This allows us to estimate the ability of each student and the difficulty of each question by maximizing the likelihood of all outcomes under our model:

$$\{s_i, \forall i, q_j, \forall j\} = \operatorname*{argmax}_{s_i, q_i} \prod_{z_{i,j} \in D} P(z_{i,j} \mid s_i, q_j), \qquad (2)$$

where $D = \{z_{ij}\}$ is the set of outcomes (e.g., of a test).

## Partially observed setting

Consider now the setting where some (or all) of the outcomes $z_{i,j} \in D$ are not observed. In practice, this is the case, for example, when the answer key to some of (or all) the questions is not available. In our setting, where the choices in the multiple choice question are in fact other students' submissions, the correctness of these submissions are not known a priori. Let $A_j$ be the set of open-response answers submitted by a subset of students in $S_{\text{open}} \subseteq S$ in response to the question $q_j$ . At some later time, a student $s_i \in S_{\text{mcq}} \subseteq S$ is presented with the same question $q_j$, but in the form of a multiple-choice question, with the options being exactly the answers in $A_j$ (note that $S_{\text{mcq}}$ need not be disjoint with $S_{\text{open}}$). The student $s_i$ is informed that there may be zero or more correct answers in the set of options in $A_j$ and they are instructed to select "all that apply." The student $s_i$ goes through each option in $A_j$ and submits a response to that option. Let $y_{i,j} = \{y_{i,j}^k\}$ be the set of such responses made by student $s_i$ on the set of answers $A_j$, where $y_{i,j}^k \in \{+1, -1\}$ is the student $s_i^{\text{th}}$ selection on the $k^{\text{th}}$ answer (option) in $A_j$. In other words the variables $y_{i,j}^k$ are the observations of whether the student $s_i$ selected answer $k$ to question $j$ (i.e., that student judged that particular answer to be correct). In what follows, we describe the statistical model that relates the student and question parameters which we are interested in estimating, to the set of response observations. Our model consists of two components: (i) the *open-response* component that models the students (and their responses) that generate open-response answers, and (ii) the *multiple choice model* component that models the students (and their responses) that are presented with the multiple choice version of each question.

**Open-response model:** Because we do not know whether the submitted open-response answers are correct, we treat the correctness of each submission as a hidden variable $z_{i,j} \in \{+1, -1\}$; this allows us to express the component of the overall likelihood of our data, responsible for the open-response answers only, as follows:

$$P(\{z_{i,j}\} \mid S_{\text{open}}, Q) = \prod_{z_{i,j}} P(z_{i,j} \mid s_i, q_j),$$

where $P(z_{i,j} \mid s_i, q_j)$ is the Rasch likelihood given in (1). Note that we drop the $k$-superscript notation for the $z_{i,j}$ variables because each student is assumed to provide at most one open-response submission to each question (since $k$ indexes the answers to a specific question). The observed responses to the multiple-choice version of each question (described next) will provide the necessary data to estimate the parameters in the model, including the hidden variables $z_{i,j}$, i.e., the correctness of each open-response submission.

**Multiple choice model:** Now consider the setting where each question is presented in the form of an MCQ. Recall that a student answering a multiple choice question is presented with multiple options, each generated by some (other) student in the set $S_{\text{open}}$, and where several options (or even no options) may be correct. The intuition that we want to capture in our model is that a student of great relative ability (i.e., $s_i \gg q_j$) will select ($y_{i,j}^k = +1$) the option (i.e., judge it as being correct) *if* that option is actually correct ($z_j^k = +1$). The same student will *not* select that option ($y_{ij}^k = -1$) if that option is incorrect ($z_j^k = -1$). At the same time, a student of poor relative ability (i.e., $s_i \ll q_j$) will not not be able to identify the correct answer, regardless of whether the option is correct, i.e., they will guess. This intuition can be captured by the following function that parametrizes the likelihood of student $s_i$ selecting the option $k$ to question $q_j$:

$$\begin{aligned} &P\left(y_{i,j}^k \mid s_i, q_j, z_j^k\right) \\ &\qquad = \frac{1}{2}\left(\frac{1}{1 + \exp\left(-y_{ij}^k z_j^k (s_i - q_j)\right)} + 1\right). \qquad (3) \end{aligned}$$

One can easily verify that this likelihood satisfies the requirements outlined above by considering every combination of the assignment to $y_{i,j}^k$ and $z_j^k$, and taking the limits of $s_i - q_j \to \infty$ (high relative ability) and $q_i - s_j \to \infty$ (poor relative ability). Note that this time we drop the index $i$ (index of the student who generated the option $k$ in question $q_j$) in $z_j^k$, as in the above, we use $s_i$ to refer to the student answering the multiple choice version of the question. Note that the above likelihood follows the same intuition as proposed by (Bachrach et al. 2012), but in a setting with an arbitrary number of choices and one correct answer.

If we make a leap of assuming conditional independence between the student $s_i^{\text{th}}$ responses to each option in a multiple choice question (conditional on $s_i$, $q_j$ and $z_j^k$), then we can express the likelihood of observing every response to every multiple choice question as follows:

$$\begin{aligned} &P(\{Y_j\} \mid S_{\text{mcq}}, Q, \{Z_j\}) \\ &\qquad = \prod_{s_i \in S_{\text{mcq}}} \prod_{q_j \in Q} \prod_{\substack{y_{i,j}^k \in Y_j \\ z_j^k \in Z_j}} P(y_{i,j}^k \mid s_i, q_j, z_j^k). \end{aligned}$$

The assumption of conditional independence requires some additional justification in our setting. Intuitively, we are justified in claiming conditional independence when we believe that the set of conditioning variables accounts for everything that may be shared across observations, such that the only remaining source of the variance is noise. For example, observations of different students answering the same question on the test are conditionally independent given the difficulty of that question. In modeling the likelihood of a student selecting each option in a multiple choice question, however, we overlook the potential for the options to be related. In an extreme example, two options may be identical or paraphrases of each other, which we expect to be common-place when these options are generated by students in a large classroom. In this case, conditional independence no longer holds without an introduction of additional conditioning variables that group the related options in some way. This problem can, to some extent, be mitigated by pre-processing and clustering similar answers before displaying them as options in a MCQ. We consider this strategy in our work.

To complete our model, we combine the *open-response* and the *multiple-choice* components:

$$P(\mathbf{y}, \mathbf{z} \mid \boldsymbol{s}, \boldsymbol{q}) = \underbrace{P(\mathbf{y} \mid \mathbf{z}, \boldsymbol{s}, \boldsymbol{q})}_{\text{multiple choice}} \underbrace{P(\mathbf{z} \mid \boldsymbol{s}, \boldsymbol{q})}_{\text{open response}}, \tag{4}$$

where we adopt vector notation for the variables and parameters in our model to facilitate the development of the learning algorithm in Section 4. In order to give the dimensions for each of the variables in (4), assume that each student in $S_{\text{open}}$ provides an open-response answer to each of the questions in $Q$ and that each student in $S_{\text{mcq}}$ also answers each question in $Q$ (which entails providing a response to each option contained in a given question). Under these assumptions then, $\mathbf{z} \in \{+1, -1\}^{|S_{\text{open}}||Q|}$, $\mathbf{y} \in \{+1, -1\}^{|S_{\text{open}}||S_{\text{mcq}}||Q|}$, $\boldsymbol{s} \in \mathbb{R}^{|S_{\text{mcq}} \cup S_{\text{open}}|}$, and $\boldsymbol{q} \in \mathbb{R}^{|Q|}$.

## 4 Parameter Learning

We now derive the expectation maximization (EM) algorithm for obtaining an approximate maximum likelihood estimate (MLE) of the parameters $\boldsymbol{s}$ and $\boldsymbol{q}$ of the model in (4). We briefly outline the key steps in obtaining the algorithm.

**E-step:** We compute the expectation of the log-likelihood (the logarithm of Equation 4) with respect to the unobserved variables $\mathbf{z}$ which yields a function $f(\boldsymbol{s}, \boldsymbol{q})$ of the parameters $\boldsymbol{s}$ and $\boldsymbol{q}$ only. The expectation is performed with respect to the posterior distribution of $\mathbf{z}$ given a previous estimate of $\boldsymbol{s}$ and $\boldsymbol{q}$ (or an initial guess).

**M-step:** We obtain an updated estimate of parameters $\boldsymbol{s}$ and $\boldsymbol{q}$ by maximizing $f(\boldsymbol{s}, \boldsymbol{q})$ obtained in the E-step.

The above procedure iterates until convergence. Below we give both steps explicitly in the context of the *joint assessment and grading (JAG)* framework.

**E-step:** Let $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{q}}$ be an intermediate estimate of the parameters. Conditioning on these estimates, the posterior of $z_j^k$ (correctness of answer (option) $k$ to question $q_j$) is a Bernoulli random variable with the probability of being correct given by (up to a normalizing constant):

$$P(z_j^k = 1 \mid \hat{\boldsymbol{s}}, \hat{q}_j) \propto$$
$$\underbrace{P(z_j^k = 1 \mid \hat{s}_{i'}, \hat{q}_j)}_{\text{open response}} \underbrace{\prod_{s_i \in S_{\text{mcq}}} P\left(y_{i,j}^k \mid \hat{s}_i, \hat{q}_j, z_j^k = 1\right)}_{\text{multiple choice responses}}. \tag{5}$$

The posterior over the answer correctness $z_j^k$ naturally integrates two sources of information: (i) the likelihood that the student who generated the answer was correct, and (ii) the likelihood that the students answering the multiple choice version of the question selected this answer as correct (note that $s_{i'} \in S_{\text{open}}$ and $s_i \in S_{\text{mcq}}$). Each likelihood is parametrized by the model's current estimate of the students' abilities and question difficulties, and as a consequence gives more weight to the signal coming from the more able students.

**M-step:** The expectation of the log-likelihood with respect to $\mathbf{z}$ yields an expression that is a weighted linear combination of (log-) Rasch-likelihoods (given in (3) and (1) respectively), and can be easily maximized with a small modification to an existing Rasch solver to account for the constants. We use the L-BFGS algorithm (Zhu et al. 1997) in order to perform this optimization step.

**Initialization:** Note that while the M-step is convex, the joint optimization problem in $\mathbf{z}$, $\boldsymbol{s}$, and $\boldsymbol{q}$ is not convex, and in general the EM algorithm will only yield an approximate solution and may get trapped in local optima. The problem becomes more pronounced in datasets with few interactions, e.g., small classrooms. As such, initialization plays an important role in determining the quality of the obtained solution. A natural heuristic for initializing the posteriors over $\mathbf{z}$ is with the fraction of "votes" given to the answer (i.e., fraction of students that identified the answer as correct). This heuristic was also suggested in (Dawid and Skene 1979).

## 5 Experiments with Synthetic Data

In order to understand the behavior of our framework in a hypothetical classroom, we evaluate the model on a series of synthetically generated datasets. As our model attempts to infer the correctness of each answer entirely from the choices made by students in answering multiple choice questions, an important concern is the limitation of inference on difficult questions. Difficult questions are questions where we can expect the majority of students to be unable to identify the correct answers, and present a challenge to any model that relies on aggregating judgements. The model's ability to recover the correct answer despite the majority being incorrect, fundamentally requires the model to leverage its estimates of students' abilities so as to weigh the judgements of better students proportionally higher. Also note that we are concerned with questions of great *relative difficulty* (with respect to the ability of the students in the class), not absolute difficulty.

We can simulate an entire spectrum of regimes that present a varying degree of difficulty to inference, and evaluate the model's performance in correctly inferring the correct answers in each regime. We accomplish this by generating a synthetic population of students and questions with a fixed expected *relative competency* (i.e., $\mathbb{E}[s - q] = k$, where $s \sim p(s)$ and $q \sim p(q)$), performing inference with our model on the generated observations, and computing the fraction of correctly inferred correct answers (accuracy) for different $\mathbb{E}[s - q]$. Note that $\mathbb{E}[s - q]$ is a quantity that conveniently summarizes the classroom in terms of its "competency" relative to the testing material. Large values of $\mathbb{E}[s - q]$ indicate that the students are well-prepared, and most will answer the questions correctly.

### Simulation procedure

We let $p(s) = \mathcal{N}(\mu_s, \sigma = 2)$ and $p(q) = \mathcal{N}(\mu_q, \sigma = 2)$. We generate a synthetic classroom with the following parameters $|S_{\text{open}}| = 10$, $|S_{\text{mcq}}| = 10$, and $|Q| = 15$, where

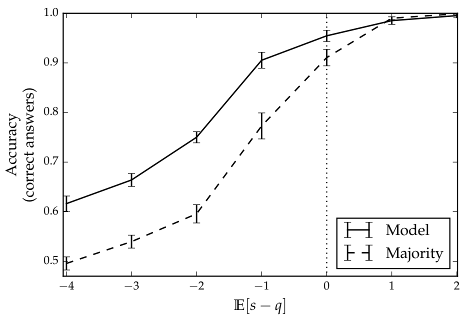

Figure 1: Accuracy in predicting the correct answers on synthetic data, as a function of the average relative competency in the classroom (measured in the multiples of standard deviations of the distributions). The simple majority-vote baseline performs comparably with our model for class distributions with large relative competency (since the majority of the students answer most questions correctly). The model significantly outperforms the baseline in the regime of lower relative competency (i.e., when most questions are too difficult for the majority of the students).

every student in $S_{\text{open}}$ submits an open-response answer to every question in $Q$, every student in $S_{\text{mcq}}$ responds to every question (which entails providing a response to every option) and $|S_{\text{mcq}} \cap S_{\text{open}}| = \emptyset$. We then sample hidden ($\mathbf{z}$) and observed ($\mathbf{y}$) variables from Bernoulli distributions parametrized by (2) and (3) respectively. Note that in these simulations and in the real-world experiments described in the next section, we assume that $S_{\text{mcq}}$ and $S_{\text{open}}$ are disjoint (i.e., $|S_{\text{mcq}} \cap S_{\text{open}}| = \emptyset$). This assumption drastically simplifies the design of real-world experiments, as the process of administering open-response and multiple-choice questions can be done in separate stages. Note, however, that this does not affect the generality of the model; in fact, this regime is more challenging from the perspective of parameter estimation, since each student generates only one type of observation (i.e., open-response or multiple choice answer, but never both).

Figure 1 illustrates the performance of the model as a function of the expected relative competency of the students ($\mathbb{E}[s - q]$). We compare the performance of our model to a simple *majority* baseline (i.e., label the answer as correct if the majority of the students select it). As expected, the majority baseline works best when the relative competency of the class is high (since most students will correctly identify the correct answers). The performance degrades significantly in the regime where the relative competency is negative (i.e., most students are expected to answer the questions incorrectly). Observe that the model is able to maintain a significant performance margin (>10%) over the baseline even in the regime of low relative competency.

## 6 Real-World Experiments

We emulate a classroom setting on the Amazon Mechanical Turk platform by soliciting Mechanical Turk workers to participate in a reading comprehension task. The study was conducted in two separate phases with a different set of workers in each: (i) the *open-response task* and (ii) the *multiple choice task*. In each task, a worker was presented with an article[1], followed by a set of 15 questions. In the *open-response task*, the questions were displayed in an open-response format, and the workers were asked to type in their response. In the *multiple choice task*, the same 15 questions were presented in a multiple-choice format, with the choices aggregated from the open-response submissions obtained in the *open response task*. The answers collected in the *open response task* were clustered semi-automatically before being displayed as choices in the *multiple choice task*. The clustering step aggregated identical answers or answers within a few characters in difference (for example due to spelling errors), and semantically identical answers were then grouped manually (e.g., paraphrases). Clustering answers is a critical pre-processing step as it ensures that a reasonable number of choices is shown as part of the multiple choice question, as well as that the conditional independence assumption discussed in Section 3 holds.

In total, 15 workers participated in the *open-response task* and 82 workers participated in the *multiple choice task*. A total of 225 open-response submissions were generated in response to the total of 15 comprehension questions, resulting in 101 distinct choices after clustering.

### Results

We evaluate the effectiveness of our JAG framework on the data collected via Amazon's Mechanical Turk using two performance metrics: (i) accuracy in predicting the correctness of each answer and (ii) quality of the predicted ranking of the students. We evaluate our algorithm in a *semi-supervised* setting where we provide a set of partially labeled items, i.e., we label correctness for a subset of the answers. This represents a practical use-case of our framework—instead of being entirely hands-off, an instructor may choose to manually grade a subset of the students' answers to improve the performance of automatic inference. We evaluate two versions of our model: **EM +open** and **EM -open** in addition to the majority baseline described in Section 5:

- **EM +open:** The full model as described in Section 3 and Section 4.
- **EM -open:** A subset of the **EM +open** model lacking the *open-response* component described in Section 3. In other words, during inference the model does not leverage any information about the ability of the answer generator, and relies entirely on the multiple choice responses to infer the correctness of the answers.

**Predicting answer correctness** Figure 2 depicts accuracy as function of the amount of labeled data (accuracy was computed with respect to a gold-standard annotation of correctness for each answer, performed by one of the authors of

[1] Unit 7.2 (Language) from the OpenStax Psychology textbook.

the paper). From it we conclude that (i) the full model (**EM +open**) significantly outperforms both the majority baseline and **EM -open**, (ii) the **EM +open** performs very well without any labeled data ($\approx 86\%$ accuracy), (iii) adding labeled data improves performance, and (iv) the *open-response* component of the model (one that is lacking in the **EM -open** model) is critical in significantly boosting performance, i.e., incorporating information about the answer creator is valuable in inferring the correctness of each answer.

**Predicting student ranking** Although predicting the correctness of each answer is itself a valuable intermediate output, a motivating use-case of our framework is to assess the students' competency. A ranking of the students by their expertise is one example of summative assessment, and may be valuable in identifying students that excel or are in need of additional help. We evaluate the quality of the rankings produced by our model in the following way: (i) use the gold-standard annotation for the correctness of each answer to fit a standard Rasch model, identifying the abilities $\boldsymbol{s}_{\text{gold}}$ of each student (both in $S_{\text{open}}$ and $S_{\text{mcq}}$), (ii) obtain the ability parameters using our model (**EM +open** and **EM -open**) (trained with a varying amount of labeled data) and (iii) rank the students according to each set of parameters and compute rank correlation. 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 are in a consistent order), $-1$ when the rankings are inverted, and $0$ when the rankings are not correlated.

Figure 3 and Figure 4 depict rank correlation as a function of the amount of labeled answers for the students in sets $S_{\text{open}}$ (workers in the **open response task**) and $S_{\text{mcq}}$ (workers in the **multiple choice task**) respectively. We observe that (i) incorporating partially labeled set of answers improves rank correlation, (ii) the **EM +open** model performs superior to or on par with the majority baseline (note that **EM -open** is not relevant when ranking the students in the $S_{open}$ set).

## 7 Conclusion

In this work, we have developed a novel framework for crowdsourced content generation and evaluation, referred to as joint assessment and grading (JAG). In the context of education, JAG offers a powerful alternative to classical peer-grading, as it naturally fuses test-taking and grading into a unified, streamlined process with a common incentive mechanism. Furthermore, our framework is general enough to be applied to many different crowdsourcing tasks where the goal is to generate and identify high-quality contributions.

## 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 National Science Foundation (NSF) under grants ECCS-1408006 and CCF-1535897. Computational resources were sponsored in part by grants from Amazon and Microsoft.

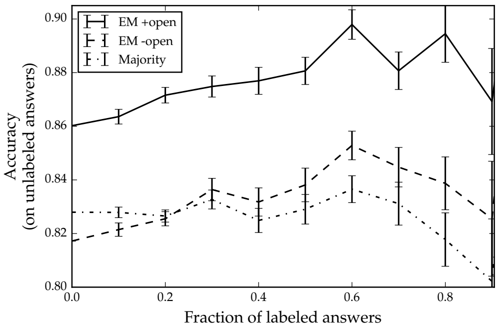

Figure 2: Accuracy in predicting the correct answers in the dataset collected on Mechanical Turk. The model that incorporates both the *open-response* and *multiple choice* components (**EM +open**) significantly outperforms the model that only incorporates the multiple choice component (**EM -open**) and a simple majority-vote baseline.

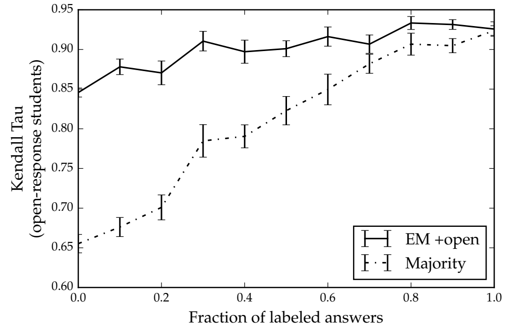

Figure 3: Rank correlation (kendall-tau) for students submitting open-response answers ($S_{open}$) between the model-inferred ranking (**EM +open**) and the ranking obtained using the gold-standard correctness labels for each answer.

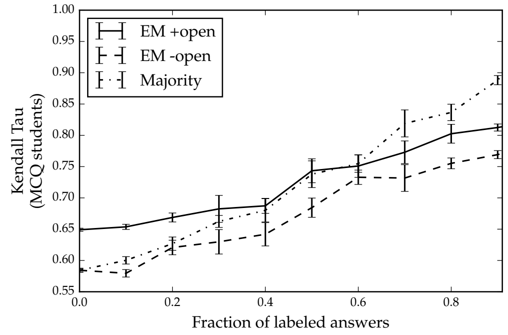

Figure 4: Rank correlation (kendall-tau) for students submitting multiple choice answers ($S_{mcq}$) between the model-inferred ranking (**EM +open** and **EM -open**) and the ranking obtained using the gold-standard labels for each answer.

## References

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

Dawid, A. P., and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics* 20–28.

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

Hung, N. Q. V.; Tam, N. T.; Tran, L. N.; and Aberer, K. 2013. An evaluation of aggregation techniques in crowdsourcing. In *Web Information Systems Engineering–WISE 2013*. Springer. 1–15.

Kulkarni, C.; Wei, K. P.; Le, H.; Chia, D.; Papadopoulos, K.; Cheng, J.; Koller, D.; and Klemmer, S. R. 2015. Peer and self assessment in massive online classes. In *Design Thinking Research*. Springer. 131–168.

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

Raman, K., and Joachims, T. 2014. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1037–1046. ACM.

Rasch, G. 1993. *Probabilistic models for some intelligence and attainment tests.* ERIC.

Whitehill, J.; Wu, T.-f.; Bergsma, J.; Movellan, J. R.; and Ruvolo, P. L. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, 2035–2043.

Wu, W.; Daskalakis, C.; Kaashoek, N.; Tzamos, C.; and Weinberg, M. 2015. Game theory based peer grading mechanisms for moocs.

Zhu, C.; Byrd, R. H.; Lu, P.; and Nocedal, J. 1997. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)* 23(4):550–560.