Methods for Ordinal Peer Grading

Karthik Raman
Department of Computer Science
Cornell University, Ithaca NY 14853
karthik@cs.cornell.edu

Thorsten Joachims
Department of Computer Science
Cornell University, Ithaca NY 14853
tj@cs.cornell.edu

ABSTRACT

Massive Online Open Courses have the potential to revolutionize higher education with their wide outreach and accessibility, but they require instructors to come up with scalable alternates to traditional *student evaluation*. Peer grading – having students assess each other – is a promising approach to tackling the problem of evaluation at scale, since the number of "graders" naturally scales with the number of students. However, students are not trained in grading, which means that one cannot expect the same level of grading skills as in traditional settings. Drawing on broad evidence that *ordinal* feedback is easier to provide and more reliable than *cardinal* feedback [5, 38, 29, 9], it is therefore desirable to allow peer graders to make ordinal statements (e.g. "project X is better than project Y") and not require them to make cardinal statements (e.g. "project X is a B-"). Thus, in this paper we study the problem of automatically inferring student grades from ordinal peer feedback, as opposed to existing methods that require cardinal peer feedback. We formulate the ordinal peer grading problem as a type of rank aggregation problem, and explore several probabilistic models under which to estimate student grades and grader reliability. We study the applicability of these methods using peer grading data collected from a real class — with instructor and TA grades as a baseline — and demonstrate the efficacy of ordinal feedback techniques in comparison to existing cardinal peer grading methods. Finally, we compare these peer-grading techniques to traditional evaluation techniques.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation, Theory

Keywords

Peer Grading, Ordinal Feedback, Rank Aggregation

KDD'14, August 24–27, 2014, New York, NY, USA.

ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623654.

1. INTRODUCTION

The advent of *MOOCs* (Massive Online Open Courses) promises unprecedented access to education given their relatively low costs and broad reach, empowering learning across a diverse range of subjects for anyone with access to the Internet. Classes frequently have upwards of 20000 students, which is orders of magnitude larger than a conventional university class. Thus, instructors are forced to rethink classroom logistics and practices so as to scale to MOOCs.

One of the key open challenges is *student evaluation* for such large classes. Traditional assessment practices, such as instructors or teaching assistants (TAs) grading individual student assignments, are simply infeasible at this scale. Consequently, assignments in most current MOOCs take the form of simple multiple-choice questions and other schemes that can be graded automatically. However, relying on such rigid testing schemes does not necessarily serve as a good indicator of learning and falls short of conventional test-design standards [16, 17]. Furthermore, such a restrictive testing methodology limits the learning outcomes that can be tested, or even limits the kinds of courses that can be offered. For example, liberal-arts courses and research-oriented classes require more open-ended assignments and responses (e.g., essays, project proposals, project reports).

Peer grading has the potential to overcome the limitations outlined above while scaling to the size of even the largest MOOCs. In peer grading, students — not instructors or TAs — provide feedback on the work of other students in their class [14, 22], meaning that the number of "graders" naturally grows with the number of students. While the scaling properties of peer grading are attractive, there are several challenges in making peer grading work.

One key challenge lies in the fact that students are not trained graders, which argues for making the feedback process as simple as possible. Given broad evidence that for many tasks *ordinal* feedback is easier to provide and more reliable than *cardinal* feedback [5, 38, 29, 9], it is therefore desirable to base peer grading on ordinal feedback (e.g. "project X is better than project Y"). Unfortunately, all existing methods for aggregating peer feedback into an overall assessment require that students provide *cardinal* feedback (e.g. "project X is a B-"). Furthermore, the efficacy of simple techniques for aggregating cardinal feedback, such as averaging, has been questioned [7, 10, 30]. While probabilistic machine learning methods have recently been proposed to address these challenges [32], they still face the problem that students may be grading on different scales. For example, students may have a preconception of what constitutes a B+

based on the university they come from. These scales may also be non-linear as the difference between an A+ and an A may not be the same as the difference between a C+ and a C.

To overcome the problems of cardinal feedback, we introduce the task of *ordinal peer grading* in this paper. By having students give ordinal statements and not cardinal statements as feedback, we offload the problem of developing a scale from the student onto the peer grading algorithm. The key technical contributions of this paper lie in the development of methods for ordinal peer grading, where the goal is to automatically infer an overall assessment of a set of assignments from ordinal peer feedback. Furthermore, a secondary goal of our methods is to infer how accurately each student provides feedback, so that reliable grading can be incentivized (e.g., as a component of the overall grade). To this effect, we propose several machine learning methods for *ordinal peer grading*, which differ by how probability distributions over rankings are modeled. For these models, we provide efficient algorithms for estimating assignment grades and grader reliabilities.

To study the applicability of our methods in real-world settings, we collected peer-assessment data as part of a university-level course. Using this data, we demonstrate the efficacy of the proposed ordinal feedback techniques in comparison to the existing cardinal feedback techniques. Furthermore, we compare our ordinal peer grading methods with traditional evaluation techniques that were used in the course in parallel. Using this classroom data we also investigate other properties of these techniques, such as their robustness, data dependence and self-consistency. Finally, we analyze the responses to a survey completed by students in the classroom experiment, indicating that most students found the peer grading experience (receiving and providing feedback) helpful and valuable.

2. THE PEER GRADING PROBLEM

We begin by formally defining the peer grading problem, as it presents itself from a machine learning perspective. We are given a set of $|D|$ *assignments* $D = \{d_1, ..., d_{|D|}\}$ (*e.g.,* essays, reports) which need to be graded. Grading is done by a set of $|G|$ graders $G = \{g_1, ..., g_{|G|}\}$ (*e.g.,* student peer grader, reviewers), where each grader receives a subset $D_g \subset D$ to assess. The choice of assignments for each grader can be uniformly random, or can follow a deterministic or sequential design. In either case, the number of assignments that any grader assesses $|D_g|$ is much smaller than the total number of assignments $|D|$ (*e.g.,* $|D_g| \approx 10$).

Each grader provides feedback for his or her set of assignments D_g. Ordinal and cardinal peer grading differ in the type of feedback a grader is expected to give:

Cardinal Peer Grading (CPG): In cardinal peer grading, each grader g provides cardinal-valued feedback for each item $d \in D_g$. Typically, this is a numeric or categorical response which we denote as $y_d^{(g)}$ (*e.g.,* Likert scale, letter grade).

Ordinal Peer Grading (OPG): In ordinal peer grading, each grader g returns an ordering $\sigma^{(g)}$ (possibly with ties) of his or her assignments D_g, indicating relative but not absolute quality. More generally, ordinal feedback could also consist of multiple pairwise preferences, but we focus on the case of a single ordering in this paper.

$D_g (\subset D)$	Set of items graded by grader g
$s_d (\in \Re)$	Predicted grade for item d (larger is better)
$\eta_g (\in \Re^+)$	Predicted reliability of grader g
σ_g	Ranking feedback (with possible ties) from g
$r_d^{(\sigma)}$	Rank of item d in ranking σ (rank 1 is best)
ρ_g	Set of pairwise preference feedback from g
$d_2 \succ_\sigma d_1$	d_2 is preferred/ranked higher than d_1 (in σ)
$\pi(A)$	Set of all rankings over $A \subseteq D$
$\sigma_1 \sim \sigma_2$	$\exists$ way of resolving ties in σ_2 to obtain σ_1

Table 1: Notation overview and reference.

Independent of the type of feedback that graders provide, the goal in peer grading is twofold.

We call the first goal *grade estimation*, which is the task of estimating the true quality of the assignments in D from the grader feedback. We distinguish between two types of grade estimation, which differ by how they express assignment quality. In *ordinal grade estimation*, the goal is to infer a ranking $\hat{\sigma}$ of all assignments in D that most accurately reflects some true ordering (by quality) σ^*. In *cardinal grade estimation*, the goal is to infer a cardinal grade $\hat{s}_d$ for each $d \in D$ that most accurately reflects each true grade s_d^*. Note that the type of feedback does not necessarily determine whether the output of grade estimation is ordinal or cardinal. In particular, we will see that some of our methods can infer cardinal grades even if only given ordinal feedback.

The second goal is *grader reliability estimation*, which is the task of estimating how accurate the feedback of a grader is. Estimating grader reliability is important for at least two reasons. First, identifying unreliable grades allows us to downweight their feedback for grade estimation. Second, and more importantly, it allows us to incentivize good and thorough grading by making peer grading itself part of the overall grade. In the following, we will typically represent the reliability of a grader as a single number $\eta_g \in \Re^+$.

In the following sections, we derive and evaluate methods for grade estimation and grader reliability estimation in the Ordinal Peer Grading setting.

2.1 Related Work in Rank Aggregation

The grade estimation problem in Ordinal Peer Grading can be viewed as a specific type of rank aggregation problem. Rank aggregation describes a class of problem related to combining the information contained in rankings from multiple sources. Many popular methods used today [15, 25, 11] build on classical models and techniques such as the seminal work by Thurstone [39], Mallows [28], Bradley & Terry [8], Luce [27] and Plackett [33]. These techniques have been used in different domains, each of which have branched off their own set of methods.

Search Result Aggregation (also known as **Rank Fusion** or **Metasearch**) has the goal of merging search result rankings from different sources to produce a single output ranking. Such aggregation has been widely used to improve over the performance of any single ranker in both supervised and unsupervised settings [3, 34, 40, 31]. Rank aggregation for search differs from Ordinal Peer Grading in several aspects. First, grader reliability estimation is not a goal in itself. Second, the success of search result aggregation depends mostly on correctly identifying the top items, while grade estimation aims to accurately estimate the full ranking. Third, ties and data sparsity are not an issue in search result aggregation, since (at least in principle) input rankings are total orders over all results.

Algorithm 1 Normal Cardinal-Score (NCS) Algorithm (called $\mathbf{PG}_1$ in [32]) is used as a baseline in our experiments

$s_d \sim \mathcal{N}(\mu_0, \frac{1}{\gamma_0})$	▷ True Scores
$\eta_g \sim Gamma(\alpha_0, \beta_0)$	▷ Grader Reliability
$b_g \sim \mathcal{N}(0, \frac{1}{\gamma_1})$	▷ Grader Bias (Only for NCS+G)
$y_d^{(g)} \sim \mathcal{N}(s_d + b_g, \frac{1}{\eta_g})$	▷ Observed Cardinal Peer Grade
Estimate $\hat{s}_d, \hat{\eta}_g$ and $\hat{b}_g$	▷ Using MLE

Social Choice and **Voting Systems** perform rank aggregation on preferences that a set of individuals stated over competing items/interests/candidates. The goal is to identify the most preferred alternatives given conflicting preferences [2]. Commonly used aggregation techniques are the *Borda count* and other Condorcet voting schemes [3, 13, 26]. These methods are ill-suited for the OPG problem, as they do not model voter reliability, typically assume rankings of all alternatives (or at least leave the choice of alternatives up to the voter), and usually focus on the top of the rankings.

Crowdsourcing is probably the most closely related application domain, where the goal is to merge the feedback from multiple *crowdworkers* [19, 6]. Due to the differing quality of these workers, modeling the worker reliability is essential [35, 11]. The key difference in our setting is that the number of items is large and we would like to correctly order all of them, not just identify the top-few.

Rank-aggregation has also been used for other settings such as multilabel/multiclass classification (by combining different classifiers) [23] or for learning player skills in a gaming environment [18]. Is is impossible to survey the vast literature on this topic and thus we refer the interested reader to a comprehensive survey on the topic [24]. These techniques have also been adapted for educational assessment [4], via a graphical model based approach, for modeling the difficulty of questions and estimating the correct answers in a crowdsourced setting. However these techniques are neither applicable for a peer grading setting nor can they handle open-ended answers (like essays).

2.2 Related Work in Peer Grading

With the advent of online courses, peer grading has been increasingly used for large classes with mixed results [7, 10, 30]. While most previous uses of peer grading have relied on simple estimation techniques like averaging cardinal feedback scores, recently a probabilistic learning algorithms has been proposed for peer grade estimation [32]. However, this method requires that students provide *cardinal* scores as grades. A second limitation of the method in [32] is that they incentivize grader reliability by relating it to the grader's own assignment score. However, such a setup is inappropriate when there are groups (such as our setting) or where external graders/reviewers are used (*e.g.,* conference reviewing). In addition, such an indirect incentive is harder to communicate and justify compared to the direct grader reliability estimates used in our case. Lastly their approach requires that each student grades some assignments that were previously graded by the instructor in order to estimate grader reliability. This seems wasteful, given that students are only able to grade a small number of assignments in total. We empirically compare their cardinal peer grading technique (Algorithm 1, using MLE instead of Gibbs sampling) with the ordinal peer grading techniques proposed in this paper.

Overall, given the limited amount of attention that the peer grading problem has received in the machine learning literature so far, we believe there is ample opportunity to improve on the state-of-the-art and address shortcomings that currently exist [36], which is reinforced by concurrent work on the topic by others [12, 37].

3. ORDINAL PEER GRADING METHODS

In this section, we develop ordinal peer grading methods for grade estimation and then extend these methods to the problem of grader reliability estimation. **Our methods are publicly available as software at** `www.peergrading.org`, where we also provide a **web service for peer grade estimation**. These methods require as data an i.i.d. sample of orderings

$$S = (\sigma^{(g_1)}, ..., \sigma^{(g_{|G|})}), \tag{1}$$

where each ordering sorts a subset of assignments according to the judgment of grader g_i.

3.1 Grade Estimation

Our grade estimation methods are based on models that represent probability distributions over rankings. In particular, we extend Mallow's Model (Sec 3.1.1), the Bradley-Terry model (Sec 3.1.3), Thurstone's model (Sec 3.1.4), and the Plackett-Luce model (Sec 3.1.5) as appropriate for the ordinal peer grading problem.

3.1.1 Mallows Model (MAL and MALBC)

Mallow's model [28] describes a distribution over rankings σ in terms of the distance $\delta(\bar{\sigma}, \sigma)$ from a central ranking $\bar{\sigma}$, which in our setting is the true ranking σ^* of assignments by quality.

$$P(\sigma|\bar{\sigma}) = \frac{e^{-\delta(\bar{\sigma},\sigma)}}{\sum_{\sigma'} e^{-\delta(\bar{\sigma},\sigma')}} \tag{2}$$

While maximum likelihood estimation of σ^* given observed rankings is NP-hard for many distance functions [13, 34], tractable approximations are known for special cases. In this work, we use the following tractable **Kendall-τ distance** [20], which assumes that both rankings are total orderings over all assignments.

DEFINITION 1. *We define the Kendall-τ Distance δ_K between ranking σ_1 and ranking σ_2 as*

$$\delta_K(\sigma_1, \sigma_2) = \sum_{d_1 \succ_{\sigma_1} d_2} \mathbb{I}[[d_2 \succ_{\sigma_2} d_1]] \tag{3}$$

It measures the number of incorrectly ordered pairs between the two rankings. In our case, the rankings that students provide can have ties. We interpret these ties as *indifference* (*i.e.,* agnostic to either ranking), which leads to the following model, where the summation in the numerator is over all total orderings σ' consistent with the weak ordering σ.

$$P(\sigma|\bar{\sigma}) = \frac{\sum_{\sigma' \sim \sigma} e^{-\delta(\bar{\sigma},\sigma')}}{\sum_{\sigma'} e^{-\delta(\bar{\sigma},\sigma')}} \tag{4}$$

Note also that the input ranking σ may only sort a subset of assignments. In such cases, we appropriately restrict the normalization constant in (4). For Kendall-τ distance, this normalization constant can be computed efficiently, and it only depends on the number of elements in the ranking.

$$Z_M(k) = \prod_{i=1}^{k} \left(1 + e^{-1} + \cdots + e^{-(i-1)}\right) = \prod_{i=1}^{k} \frac{1 - e^{-i}}{1 - e^{-1}}$$

Algorithm 2 Computing MLE ranking for Mallows Model

1: $C \leftarrow D$ ▷ C contains unranked items
2: **for** $i = 1 \ldots |D|$ **do**
3: **for** $d \in C$ **do**
4: $x_d \leftarrow \sum_{g \in G} \eta_g |d' \in C : d' \succ_{\sigma_g} d| - |d' \in C : d \succ_{\sigma_g} d'|$
5: $d^* \leftarrow \min_{d \in C} x_d$ ▷ Select highest scoring item
6: $r_{d^*}^{(\hat{\sigma})} \leftarrow i$ ▷ Rank as next item
7: $C \leftarrow C/d^*$ ▷ Remove d^* from candidate set
8: **return** $\hat{\sigma}$

The numerator can likewise be computed efficiently. Note that ties in the grader rankings $\sigma^{(g)}$ do not affect the normalization constant under the interpretation of indifference.

Under this modified Mallow's model, the maximum likelihood estimator of the central ranking $\hat{\sigma}$ is

$$\hat{\sigma} = \texttt{argmax}_{\sigma} \left\{ \prod_{g \in G} \frac{\sum_{\sigma' \sim \sigma^{(g)}} e^{-\delta_K(\sigma, \sigma')}}{Z_M(|D_g|)} \right\}. \quad (5)$$

Computing the maximum likelihood estimate $\hat{\sigma}$ as an estimate of the true ranking by quality σ^* requires finding the *Kemeny-optimal aggregate*, which is known to be NP-hard [13]. However numerous approximations have been studied in the rank aggregation literature [13, 21, 1]. In this work we use a simple greedy algorithm as shown in Algorithm 2.

As an alternative algorithm for computing the estimated ranking, we utilize a Borda count-like approximation for the Mallows model (which we denote as MAL_{BC}), where Line 2 of Algorithm 2 is replaced with

$$x_d \leftarrow \sum_{g \in G} r_d^{(\sigma^{(g)})}.$$

3.1.2 Score-Weighted Mallows (MALS)

Mallow's model presented above has two shortcomings. First, it does not output a meaningful cardinal grade for the assignments, which makes it applicable only to ordinal grade estimation. Second, the distance δ_K does not distinguish between misordering assignments that are similar in quality from those that have a large quality difference.

To address these two shortcomings, we propose an extension which estimates cardinal grades $\hat{s}_d$ for all assignments. To this effect, we introduce the following score-weighted ranking distance, which scales the distance induced by each misranked pairs by its estimated grade difference.

Definition 2. *The* **score-weighted Kendall-τ distance** δ_{SK} *over rankings* σ_1, σ_2 *given cardinal scores* s_d *is*

$$\delta_{SK}(\sigma_1, \sigma_2 | s) = \sum_{d_1 \succ_{\sigma_1} d_2} (s_{d_1} - s_{d_2}) \mathbb{I}[[d_2 \succ_{\sigma_2} d_1]]. \quad (6)$$

Treating ties in the grader rankings as described above results in a score-weighted version of the Mallows model (MALS). We use the following maximum a posteriori estimator to estimate the scores $\hat{s}$.

$$\hat{s} = \texttt{argmax}_s \left\{ Pr(\mathbf{s}) \prod_{g \in G} \frac{\sum_{\sigma' \sim \sigma^{(g)}} \exp\left(-\delta_{SK}(\hat{\sigma}, \sigma' | s)\right)}{\sum_{\sigma' \in \pi(D_g)} \exp\left(-\delta_{SK}(\hat{\sigma}, \sigma' | s)\right)} \right\} \quad (7)$$

Note that $\hat{\sigma}$ can be obtained by sorting items as per $\hat{s}_d$. $Pr(\hat{\mathbf{s}}) = \prod_{d \in D} Pr(\hat{s}_d)$ is the prior on the latent item scores. In our experiments we model $Pr(\hat{s}_d) \sim \mathcal{N}(0, 9)$, and use the same prior in all of our methods. While the resulting objective is not necessarily convex, we use Stochastic Gradient Descent (SGD) for grade estimation and initialize the grades using a scaled-down Mallows solution.

3.1.3 Bradley-Terry Model (BT)

The above models define distributions over rankings as a function of a ranking distance, and they require approximate methods for solving the maximum likelihood problem. As an alternative, we can utilize rank aggregation models based on distributions over pairwise preferences, since a ranking of n items can also be viewed as a set of preferences over the $\binom{n}{2}$ item pairs. The Bradley-Terry model [8] is one model for pairwise preferences, and it derives a distribution based on the differences of underlying item scores s_d through a logistic link function.

$$P(d_i \succ_{\rho^{(g)}} d_j | s) = \frac{1}{1 + e^{-(s_{d_i} - s_{d_j})}} \quad (8)$$

Since each preference decision is modeled individually, the feedback from the grader could be a (possibly inconsistent) set of preferences that does not necessarily have to form a consistent ordering. The following is the maximum a posteriori estimator used in this paper.

$$\hat{s} = \texttt{argmax}_s \left\{ Pr(\mathbf{s}) \prod_{g \in G} \prod_{d_i \succ_{\rho^{(g)}} d_j} \frac{1}{1 + e^{-(s_{d_i} - s_{d_j})}} \right\} \quad (9)$$

The resulting objective is (jointly) log-convex in all of the estimated grades $\hat{s}_d$, with the gradients taking a simple form. Hence SGD can be used to estimate the global optimal grades efficiently. We treat ties as the absence of a preference. One can also extend this model to incorporate ties more explicitly, but we do not discuss this for brevity.

3.1.4 Thurstone Model (THUR)

An alternate to the logistic link function of the Bradley-Terry model is to utilize a normal distribution for the pairwise preferences. Like the Bradley-Terry model, the resulting Thurstone model [39] model can be understood as a random utility model using the following process: For each pair of items d_i, d_j, the grader samples (latent) values $x_{d_i}^{(g)} \sim \mathcal{N}(s_{d_i}, \frac{1}{2})$ and $x_{d_j}^{(g)} \sim \mathcal{N}(s_{d_j}, \frac{1}{2})$, and then orders the pair based on the two values. The mean of the normal distribution of d_i is the quality s_{d_i}. Maximum a posteriori estimation of the scores s requires maximization of the following function:

$$\hat{s} = \texttt{argmax}_s \left\{ Pr(\mathbf{s}) \prod_{g \in G} \prod_{d_i \succ_{\rho^{(g)}} d_j} \mathcal{F}(s_{d_i} - s_{d_j}) \right\} \quad (10)$$

$\mathcal{F}$ is the CDF of the standard normal distribution. This objective function is log-convex and we use SGD to optimize it.

3.1.5 Plackett-Luce Model (PL)

A drawback of the pairwise preference models is that they can be less expressive than models built on distributions over rankings. An extension to the Bradley-Terry model (the Plackett-Luce model [33]) allows us to use distributions over rankings, while still retaining convexity and simplicity

Algorithm 3 Alternating SGD-based Minimization

Require: $N \geq 0$ (Number of iterations), Likelihood L

1: $Obj \leftarrow -\log L$
2: $\hat{\mathbf{s}} \leftarrow SGD_S(Obj, \eta = \mathbf{1})$ ▷ Est. scores w/o reliabilities
3: **for** $i = 1 \ldots N$ **do**
4: $\quad \eta \leftarrow SGD_G(Obj, \hat{\mathbf{s}})$ ▷ Estimate reliabilities
5: $\quad \hat{\mathbf{s}} \leftarrow SGD_S(Obj, \eta)$ ▷ Est. scores with reliabilities
6: **return** $\hat{\mathbf{s}}, \eta$

of gradient computation. This model can be best understood as a multi-stage experiment where at each stage, an item d_i is drawn (w/o replacement) with probability $\propto e^{s_{d_i}}$. The probability of observing ranking $\sigma^{(g)}$ under this process is:

$$P(\sigma^{(g)}|s) = \prod_{d_i \in D_g} e^{s_{d_i}} / \left(e^{s_{d_i}} + \sum_{d_i \succ_{\sigma^{(g)}} d_j} e^{s_{d_j}} \right)$$

The resulting maximum a posteriori estimator is

$$\hat{s} = \texttt{argmax}_s \left\{ Pr(\mathbf{s}) \prod_{g \in G} \prod_{d_i \in D_g} \frac{e^{s_{d_i}}}{e^{s_{d_i}} + \sum_{d_i \succ_{\sigma^{(g)}} d_j} e^{s_{d_j}}} \right\}. \quad (11)$$

3.2 Grader Reliability Estimation

While the methods discussed in Section 3.1 allow us to estimate assignment grades from ordinal feedback, they still do not give us means to directly estimate grader reliabilities $\hat{\eta}_g$. However, there is a generic way of extending all methods presented above to incorporate grader reliabilities. Using Mallow's model as an example, we can introduce $\hat{\eta}_g$ as a variability parameter as follows:

$$Pr(\sigma|\bar{\sigma}, \eta_g) = \frac{\sum_{\sigma' \sim \sigma^{(g)}} \exp\left(-\eta_g \delta_K(\bar{\sigma}, \sigma')\right)}{Z_M(\eta_g, |D_g|)} \quad (12)$$

The resulting estimator of both $\hat{\sigma}$ and $\hat{\eta}$ is

$$\hat{\sigma}, \hat{\eta} = \texttt{argmax}_{\sigma,\eta} \left\{ \prod_{g \in G} Pr(\eta_g) \frac{\sum_{\sigma' \sim \sigma^{(g)}} \exp\left(-\eta_g \delta_K(\sigma, \sigma')\right)}{Z_M(\eta_g, |D_g|)} \right\}, \quad (13)$$

where $Pr(\hat{\eta}_g)$ is the prior on the grader reliability. In this work we use a *Gamma* prior $\hat{\eta}_g \sim Gamma(10, 0.1)$.

Similarly, the other objectives can also be extended in this manner as seen in Table 2. While many of the extended objectives, such as the one above in Eq. (13), are convex in the grader reliabilities $\hat{\eta}_g$ (for given $\hat{\sigma}$), they unfortunately are not jointly convex in the reliabilities *and* the estimated grades. We thus use an iterative alternating-minimization technique, which alternates between minimizing the log-objective to estimate the assignment grades and minimizing the log-objective to estimate the grader reliabilities. This iterative alternating approach using stochastic gradient descent is used for all joint estimation tasks in this paper. Note that methods which estimate the reliabilities using Algorithm 3 are denoted by a **+G** suffix to the method, while those that simply estimate the assignment grades are represented by the method name alone.

4. EXPERIMENTS

In the following we present experiments that compare ordinal and cardinal peer grading methods. We evaluate their ability to predict instructor grades, their variability, their robustness to bad peer grading, and their ability to identify bad graders. We also present the results from a qualitative student survey to evaluate how students perceived the peer grading process.

Method	Score	Cnvx	Estimator
MAL+G	No	No	$Pr(\eta) \prod_{g \in G} \sum_{\sigma' \sim \sigma^{(g)}} \exp(-\hat{\eta}_g \delta_K(\hat{\sigma}, \sigma')) / Z_M(\hat{\eta}_g, \lvert D_g \rvert)$
MALS+G	Yes	No	$Pr(\hat{\mathbf{s}}, \eta) \prod_{g \in G} \sum_{\sigma' \sim \sigma^{(g)}} \exp(-\hat{\eta}_g \delta_{SK}(\sigma^{(g)}, \hat{\sigma}, F)) / Z(\cdot)$
BT+G	Yes	Yes	$Pr(\hat{\mathbf{s}}, \eta) \prod_{g \in G} \prod_{d_i \succ_{\rho(g)} d_j} 1/(1+e^{-\hat{\eta}_g (s_{d_i} - s_{d_j})})$
THUR+G	Yes	Yes	$Pr(\hat{\mathbf{s}}, \eta) \prod_{g \in G} \prod_{d_i \succ_{\rho(g)} d_j} \mathcal{F}(\sqrt{\hat{\eta}_g}(s_{d_i} - s_{d_j}))$
PL+G	Yes	Yes	$Pr(\hat{\mathbf{s}}, \eta) \prod_{g \in G} \prod_{d_i \in D_g} 1/(1 + \sum_{d_i \succ_{\rho(g)} d_j} e^{-\hat{\eta}_g (s_{d_i} - s_{d_j})})$

Table 2: Summary of the ordinal methods studied which model the grader's reliabilities, including the ability to output cardinal scores and if the resulting objective is convex in these scores.

4.1 Data Collection in Classroom Experiment

We use a real dataset consisting of peer feedback, TA grades, and instructor grades for evaluating the peer grading methods proposed in this paper. This data was collected as part of a senior-undergraduate and masters-level class with an enrollment of about 170 students. The class was staffed with 9 Teaching Assistants (TAs) that participated in grading, and a single Instructor. This size of class is attractive, since it is large enough for collecting a substantial number of peer grades, while at the same time allowing traditional instructor and TA grading to serve as a baseline. The availability of instructor grades makes our data different from other peer-grading evaluations used in the past (e.g., [32]). We are happy to provide the data to other researchers subject to IRB approval.

The dataset consists of two parts that were graded independently, namely the *poster presentation* and the *final report* of an 8-week long course project. Students worked in groups of 3-4 students for the duration of the project, and there were a total of 44 project groups. While student worked in groups, peer grading was performed individually via the Microsoft Conference Management Toolkit (CMT) system. The peer grading process was performed single-blind for the posters and double-blind for the reports, and the reviewer assignments were made uniformly at random. Students were given clear directives and asked to focus on aspects such as *novelty* and *clarity* (among others) while determining their grade. They were also asked to justify their grade by providing feedback comments. Students were told that a part of their grade depends on the quality of their peer feedback.

All grading was done on a 10-point (cardinal) Likert scale, where 10 was labeled "perfect", 8 "good", 5 "borderline", 3 "deficient" and 1 "unsatisfactory". This will allow us to compare cardinal and ordinal peer grading methods, where ordinal methods merely use the ordering (possibly with ties) implied by the cardinal scores. Note that in a true application of ordinal peer grading accuracy could improve, since it would allow simplifying the grading instructions and reduce cognitive overhead if students did not have to worry about the precise meaning of specific cardinal grades.

The following describes the grading processes used at each stage, and Table 3 summarizes some of the key statistics.

Data Statistic	PO	FR
Number of Assignments	42	44
Number of Peer Reviewers	148	153
Total Peer Reviews	996	586
Total TA Reviews	78	88
Participating TAs	7	9
Per-Item Peer Grade Devn.	1.16	1.03

Set	Who?	Mean	Devn.
PO	Peers	8.16	1.31
	TAs	7.46	1.41
	Meta	7.55	1.53
FR	Peers	8.20	1.35
	TAs	7.59	1.30
	Instructor	7.43	1.16

Table 3: Statistics for the two datasets (PO=Poster, FR=Report) from the classroom experiment along with the staff (TAs/Meta/Instructor) and student grade distributions.

4.1.1 *Grading Process for Poster Presentations*

The poster presentations took place in a two-hour poster session. Two groups did not present their poster. Students were encouraged to rotate presenting their poster. This likely increased variability of grades, since different reviewers often saw different presenters. Students and TAs took notes and entered their reviews via CMT afterwards.

The *TA Grades* were independent, meaning that the TAs did not see the peer reviews before entering their review. There were on average 1.85 TA reviews for each poster.

The *Peer Grades* totaled on average 23.71 reviews for each poster, with each peer reviewer reviewing 6.73 posters on average.

The final *Meta Grade* for each poster was determined as follows. One of the TAs that already provided an independent review was selected as a meta-reviewer. This TA was asked to aggregate all the arguments brought forward in the reviews and make a final grade on the same 10-point scale. The instructor oversaw this process, but intervened only on very few grades.

4.1.2 *Grading Process for Final Projects*

At the end of the project, groups submitted a report of about 10 pages in length. The reviewing process was similar to that of the poster presentations, but with one important difference — namely that all project reports were graded by the TAs and the instructor without any knowledge of the peer reviews, as detailed below.

On average each report received 13.32 *Peer Grades* as the overall score on each of the peer reviews (students were also asked for component scores like "clarity", etc.).

Each report also received two *TA Grades*, which the TAs submitted without knowledge of the peer reviews.

Finally, each report received an *Instructor Grade*, following the traditional process of project grading in this class. The instructor and head TA each graded half the projects and determined the grade based on their own reading of the paper, taking the TA reviews as input. These grades were provided without viewing the peer reviews. We can therefore view the instructor grades as an assessment that is entirely independent of the peer grades (in contrast to the Meta Grades for the posters, which have some dependency).

4.2 Evaluation Metrics

A commonly used measure for reporting student performance (among many standardized tests) is the percentile rank relative to all students in the class. Following this practice, we use percentile rank as the grade itself (a letter grade can easily be derived via curving), and report ranking metrics as our main indicators of performance. In particular, we use the following variant of Kendall-τ that accounts for ties.

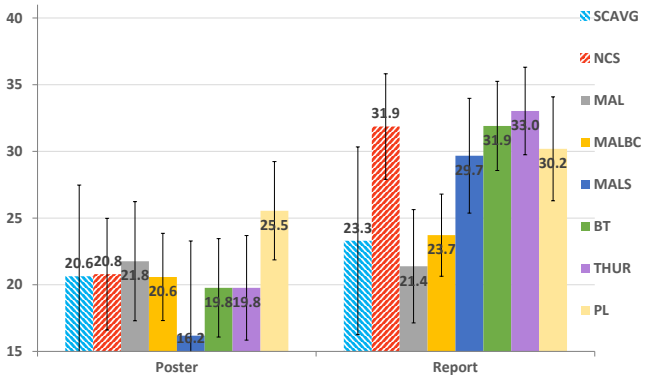

Figure 1: Comparing peer grading methods (w/o grader reliability estimation) against Meta and Instructor Grades in terms of $\mathcal{E}_K$ (lower is better).

$$\tau_{KT}(\sigma_1,\sigma_2)=\sum_{d_1\succ_{\sigma_1} d_2}\mathbb{I}[[d_2\succ_{\sigma_2} d_1]]+\frac{1}{2}\mathbb{I}[[d_1\approx_{\sigma_2} d_2]] \quad (14)$$

Note that this measure is not symmetric, assuming that the first argument is a target ranking and the second argument is a predicted ranking. It treats ties in the target ranking as *indifference*. Ties in the predicted ranking are treating as a lack of information, incurring a $\frac{1}{2}$ error (*i.e.,* equivalent to breaking ties randomly). Such a correction is necessary for evaluation purposes, since otherwise predicted rankings with all ties (which convey no information) would incur no error. Normalizing $\tau_{KT}(\sigma_1,\sigma_2)$ and accounting for the fact that we may have more than one target ranking leads to the following error measure.

Definition 3. *Given a set of target rankings S_g, we define the* **Kendall-τ error** $\mathcal{E}_K$ *of predicted ranking σ_I as:*

$$\mathcal{E}_K(\sigma_I)=\frac{100}{|S_g|}\sum_{\sigma_t\in S_g}\frac{\tau_{KT}(\sigma_t,\sigma_I)}{\max_{\sigma\in\pi(D)}\tau_{KT}(\sigma_t,\sigma)} \quad (15)$$

This error *macro-averages* the (normalized) τ_{KT} errors for each target ranking. Due to the normalization, they lie between 0 (indicating perfect agreement) and 100% (indicating reversal with target rankings). A random ranking has expected $\mathcal{E}_K$ error of 50%.

4.3 How well does Ordinal vs. Cardinal Peer Grading Predict Final Grade?

The first question we address is in how far peer grading resembles the grades given by an instructor. Specifically, we investigate whether ordinal peer grading methods achieve similar performance as cardinal peer grading methods, even though ordinal methods receive strictly less information.

For all methods considered in this paper, Figure 1 shows the Kendall-τ error $\mathcal{E}_K$ compared to the Meta Grades for the Posters, and compared to the Instructor Grades for the Reports. The errorbars show estimated standard deviation using bootstrap-type resampling.

On the posters, none of the methods show significantly worse performance than another method. In particular, there is no evidence that the cardinal methods are performing better than the ordinal methods. A similar conclusion also holds for the reports. However, here the ordinal methods based on Mallow's model perform better than the car-

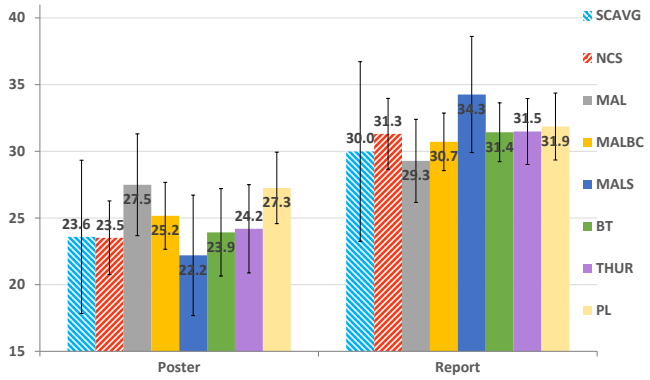

Figure 2: Comparing peer grading methods (w/o grader reliability estimation) against TA Grades in terms of $\mathcal{E}_K$, using TA grades as the target ranking.

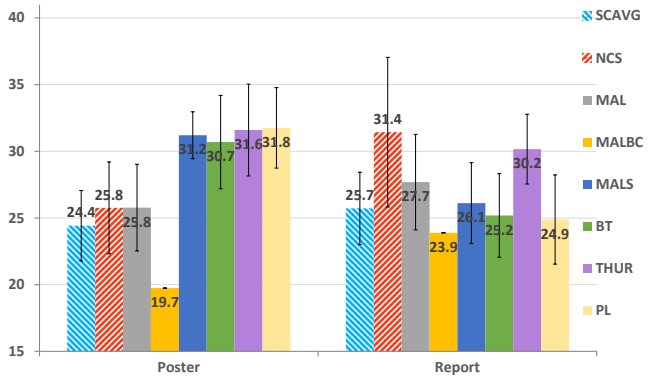

Figure 3: Self-consistency of peer-grading methods (w/o grader reliability estimation) in terms of $\mathcal{E}_K$.

dinal NCS[1] method [32] (see Algorithm 1), as well as some of the other ordinal methods. Simply averaging the cardinal scores of the peer graders, which we call Score Averaging (SCAVG), performs surprisingly well.

In summary, most methods achieve an $\mathcal{E}_K$ between 20% and 30% on both problems, but all have large standard deviations. The $\mathcal{E}_K$ appears lower for the posters than for the projects, which can be explained by the fact that the Meta Grade was influenced by the peer grades. But how good is an $\mathcal{E}_K$ between 20% and 30%?

4.4 How does Peer Grading Compare to TA Grading?

We now consider how Peer Grading compares to having each assignment graded by a TA. For medium sized classes, TA grading may still be feasible. It is therefore interesting to know if TA grading is clearly preferable to Peer Grading when it is feasible. But more importantly, the inter-judge agreement between multiple TAs can give us reference points for the accuracy of Peer Grading.

As a first reference point, we estimate how well the TA Grades reflect the Meta Grades for the posters and the Instructor Grades for the reports. In particular, we consider a grading process where each assignment is graded by a single TA that assigns a cardinal grade. Each TA grades a fraction of the assignments, and a final ranking of the assignments is then computed by sorting all cardinal grades. We call this grading process *TA Grading*.

We can estimate the $\mathcal{E}_K$ of TA grading with the Meta Grades and the Instructor Grades, since we have multiple TA grades for most assignments. We randomly resample a TA grade from the available grades for each assignment, compute the ranking, and then estimate mean and standard deviation of the $\mathcal{E}_K$ over 5000 samples. This leads to a mean $\mathcal{E}_K$ of 22.0 ± 16.0 for the posters and 22.2 ± 6.8 for the reports. Comparing these to the $\mathcal{E}_K$ of the peer grading methods in Figure 1, we see that they are comparable to the performance of many peer grading methods — *even though the $\mathcal{E}_K$ of TA grading is favorably biased. Note that Meta Grades and the Instructor Grades were assigned based on the same TA grades we are evaluating against.*

[1]We tuned the hyperparameters of the NCS model to maximize performance. We also used a fixed grader reliability parameter in the NCS model, since it provided better performance than with reliability estimation (NCS+G).

To avoid this bias and provide a fairer comparison with TA grading, we also investigated how consistent peer grades are with the TA grades, and how consistent TA grades are between different TAs. Figure 2 shows the $\mathcal{E}_K$ of the peer grading methods when using TA Grades as the target ranking for both the Posters and the Reports. Variances were again estimated via bootstrap resampling. Note that TA Grades were submitted without knowledge of the Peer Grades. Overall, the peer grades have an $\mathcal{E}_K$ with the TA Grades that is similar to the $\mathcal{E}_K$ with the respective Final grades considered in the previous subsection. Again, there is no evidence that the ordinal peer grading methods are less predictive of the TA Grades than the cardinal peer grading methods.

To estimate $\mathcal{E}_K$ between different TAs, we use the following resampling procedure. In a leave-one-out fashion, we treat the grades of a randomly selected TA as the target ranking and compute the predicted ranking by sampling from the other TAs grades as described above. Averaging over 5000 repetitions reveals that the $\mathcal{E}_K$ between the TAs is 47.5 ± 21.0 for the posters and 34.0 ± 13.8 for the reports.

These numbers can be compared to the $\mathcal{E}_K$ of peer grading methods in Figure 2. For the Reports, peer grades are roughly as consistent with the TA grades as other TA grades are. For the posters the peer grading methods are substantially more predictive of TA grades than other TA grades. The reason for this is at least twofold. First, the peer grading methods have access to much more data, which reduces variability (especially since presentations were not always given by the same student). Second, the peer grading methods have enough data to correct for different grading scales, while offsets in grading scales can have disastrous consequences in TA grading.

Finally, we also consider the self-consistency of the peer grading methods. Analogous to the self-consistency of TA grading, we ask how similar are the grades we get if we repeat the grading procedure with a different sample of assessments. We randomly partition peer reviewers into two equally sized datasets. For each peer grading method, we perform grade estimation on both datasets, which generates two rankings of the assignments. Ties in these rankings are broken randomly to get total orderings. Figure 3 shows the $\mathcal{E}_K$ between the two rankings (over 20 sampled partitions). For the posters, peer grading is substantially more self consistent than TA grading, and for the reports all peer grading methods have lower $\mathcal{E}_K$ estimates than TA grading as well.

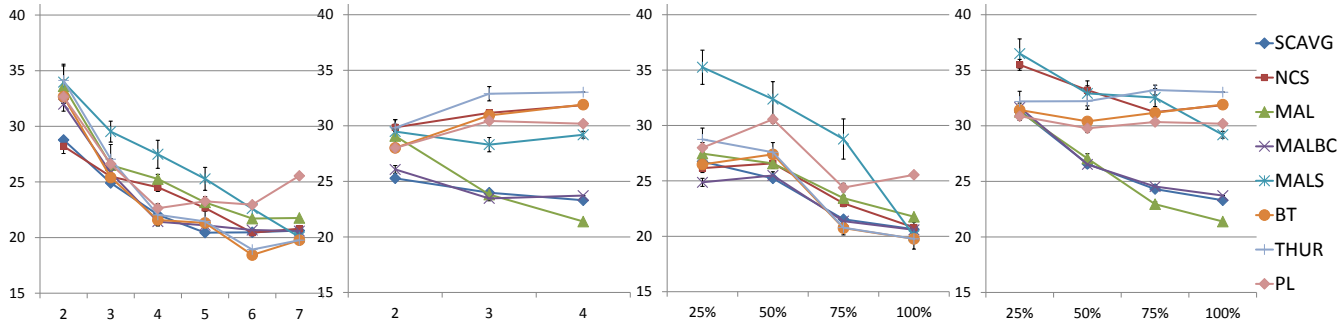

Figure 4: Change in $\mathcal{E}_K$ performance of peer grading methods (using Meta and Instructor Grades as target ranking) when varying the number of assignments assigned to each reviewer for Posters (first from left) & Reports (second), and when varying the number of peer reviewers for Posters (third) Reports (last).

Overall, we conclude that there is no evidence that TA grading would have led to more accurate grading outcomes than peer grading.

4.5 How does Grading Accuracy Scale with the Number of Peer Reviews?

How many reviewers are necessary for accurate peer grading, and how many reviews does each peer grader need to do? To gauge how performance changes with the number of peer reviews, we performed two sets of experiments. First, we created 20 smaller datasets by downsampling the number of peer reviewers. The results are shown in the two rightmost graphs of Figure 4. Overall, the methods degrade gracefully when the number of reviewers is reduced. Overall, we find that most ordinal methods scale as well as cardinal methods for both datasets.

A second way of increasing or reducing the amount of available data lies in the number of assignments that each student grades. Thus we repeated the experiment, but instead downsampled the number of assignments per reviewer (corresponding to a lower workload for each grader). The leftmost two plots of Figure 4 show the results. Again, we find that performance degrades gracefully.

4.6 Can the Peer Grading Methods Identify Unreliable Graders?

Peer grading can only work in practice, if graders are sufficiently incentivised to report an accurate assessment. This can be achieved by giving a grade also for the quality of the grading. In the following, we investigate whether the grader reliability estimators proposed in Section 3.2 can identify graders that are not diligent.

For both the posters and the projects, we add 10 "lazy" peer graders that report random grades drawn from a normal distribution whose mean and variance matches that of the rest of the graders[2]. For the ordinal methods, this results in a random ordering. We then apply the peer grading methods, estimating the respective reliability parameters η_g for each grader using 10 iterations of the alternating optimization algorithm. We then rank graders by their estimated η_g.

Figure 5 (top) shows the percentage of lazy graders that rank among the 20 graders with the lowest η_g. The error bars show standard error over 50 repeated runs with different lazy graders sampled. Most ordinal methods significantly

[2]Otherwise it would be easy to identify these graders.

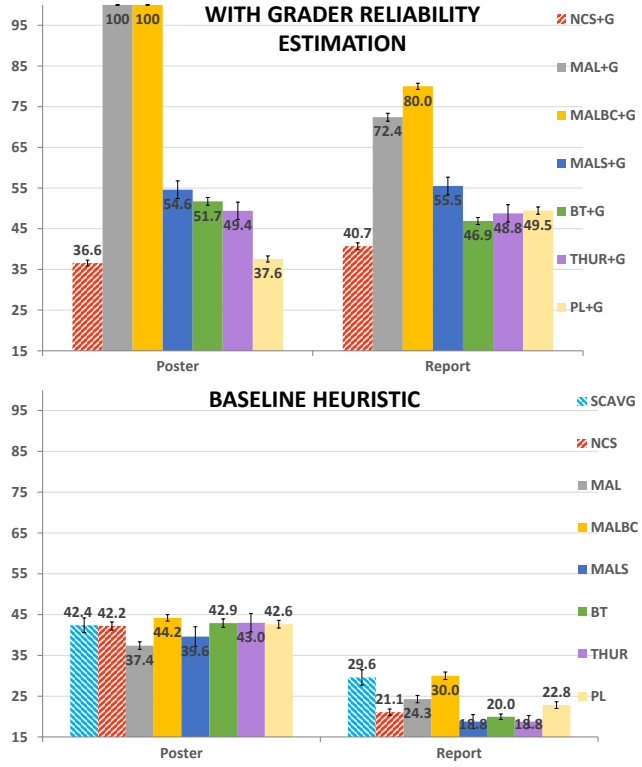

Figure 5: Percentage of times a grader who randomly scores and orders assignments is among the 20 least reliable graders (*i.e.*, bottom 12.5%).

outperform the cardinal NCS method for both the posters and the reports. The variants of Mallow's model perform very well, identifying around 70-80% of the lazy graders for the reports and all 10 lazy graders for the posters. The better performance for the posters than for the reports was to be expected, since students provide 7 instead of 4 grades.

Figure 5 (bottom) shows the results of a heuristic baseline. Here, grade estimation without reliability estimation is performed, and then graders are ranked by their $\mathcal{E}_K$ with the estimated ranking $\hat{\sigma}$. For almost all methods, this performs worse, clearly indicating that reliability estimation is superior in identifying lazy graders. We find similar results even when there are 100+ lazy graders, and we investigate robustness in the following section.

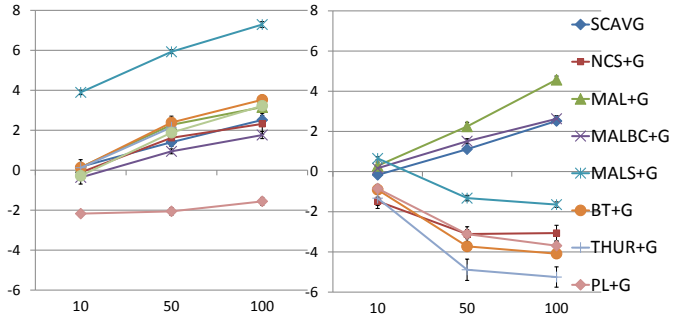

Figure 6: Change in $\mathcal{E}_K$ (using Instructor and Meta Grades as target ranking) for (Left) Posters and (Right) Final Reports with the addition of an increasing number of *lazy* graders *i.e.*, $\mathcal{E}_K$(With Lazy)$-$ $\mathcal{E}_K$(Without Lazy). A negative value indicates that performance improves on adding this noise.

Method	Posters		Reports	
	Runtime	Runtime (+G)	Runtime	Runtime (+G)
NCS	0.32 ±0.03	7.0 ±0.55	0.20 ±0.03	4.6 ±0.25
MAL	0.01 ±0.00	6.1 ±0.11	0.01 ±0.00	2.5 ±0.03
MAL_{BC}	0.01 ±0.00	5.1 ±0.08	0.01 ±0.00	2.5 ±0.03
MALS	151.4 ±12.39	418.7 ±9.10	2.0 ±0.13	4.2 ±0.16
BT	0.46 ±0.06	5.6 ±0.38	0.21 ±0.02	2.2 ±0.10
THUR	57.9 ±0.76	490.1 ±7.45	12.2 ±0.86	120.8 ±1.03
PL	0.36 ±0.03	4.2 ±0.08	0.18 ±0.01	2.0 ±0.10

Table 4: Average runtime (with and without grader reliability estimation) and their standard deviation of different methods in CPU seconds.

4.7 How Robust are the Peer Grading Methods to Lazy Graders?

While Section 4.6 showed that reliability estimation in ordinal peer grading is well-suited for identifying lazy graders, we would also like to know what effect these lazy graders have on grade estimation performance. We study the robustness of the peer grading methods by adding an increasing number of lazy graders. Figure 6 shows the change in $\mathcal{E}_K$ (w.r.t. Instructor/Meta grades) after adding 10/50/100 lazy graders (compared to the $\mathcal{E}_K$ with no lazy graders). We find that in most cases performance does not change much relative to the variability of the methods. Interestingly, in some cases performance also improves on adding this noise. A deeper inspection reveals that noise is most beneficial for methods whose original $\mathcal{E}_K$ performance was weaker than that of the other methods. For example, the Thurstone model showed the weakest performance on the Reports and improves the most.

4.8 How Computationally Efficient are the Peer Grading Methods?

While prediction accuracy is the prime concern of grade inference, computational efficiency needs to be sufficient as well. Table 4 show the average runtimes and their standard deviations for the posters and the reports. All methods are tractable and most finish within seconds. The Score-Weighted Mallows model is less efficient for problems where a each grader assesses many assignments, since the gradient computations involves computing the normalization constant (which involves summing over all rankings). However, training scales linearly with the number of graders. Another method that requires more time is the Thurstone model. The main bottleneck here is the computation of the gradient as it involves a lookup of a CDF value from the normal distribution table.

Question A) Was getting peer feedback helpful?		**Question B**) Was providing peer feedback valuable?	
A_1	Yes, it was helpful.	B_1	Yes it was a valuable experience
A_2	Helpful, but not as much as instructor feedback.	B_2	Yes, it was valuable, but with caveats (e.g. took lot of time).
A_3	Somewhat helpful (e.g. only few comments were helpful).	B_3	Only little value (e.g. was too difficult / lacked the grading skills)
A_4	No / Not really / Did not help much.	B_4	Not valuable / Not really valuable.
A_5	Other / Missing	B_5	Other / Missing

Table 5: Response categories for survey questions.

4.9 Do Students Value Peer Grading?

A final point that we would like to explore is that peer grading is not only about grade estimation, but also about generating useful feedback. In particular, the cardinal or ordinal assessments were only a small part of the peer feedback. Peer graders had to write a justification for their assessment and comment on the work more generally.

To assess this aspect of peer grading, a survey was conducted at the end of class as part of the course feedback process. This survey included two questions about the student's peer grading experience in the class; more specifically, about how *helpful* the feedback they received was, and how *valuable* the experience of providing feedback was to them. Both questions were to be answered in free-form text. Of the 161 students that participated in the project, 120 students responded to at least one of the questions, with 119 answering the question about receiving feedback (mean response length in characters: 62.93; stdev: 77.22) and 118 the question about providing feedback (mean: 100.36; stdev: 105.74). Following standard practice from survey analysis, we created five categories for coding these open-ended responses as show in Table 5. While the first four categories (roughly) follow a decreasing scale of approval, the last serves as a catch-all (including missing responses).

All free-text responses were manually assigned to these categories by four external annotators (who were not involved with the class and had not seen the comments before). For all the 237 student comments (*i.e.,* responses), the annotators were asked to choose the category that was *most appropriate/best describes the comment.* To check inter-annotator agreement we used the Fleiss Kappa measure. κ values of 0.8389 and 0.6493 for the two questions indicate high annotator agreement. The final assignment of response to category was done by majority vote among the four annotators (score of 0.5 each if tied between categories).

Table 6 summarizes the results of the survey after coding. Overall, around 68% found it at least somewhat helpful to receive peer feedback, and around 74% found substantial value in providing the peer feedback. Interestingly, of the 26% of the students who expressed that receiving peer feedback was not (really) helpful to them, 17% still found it valuable to provide peer feedback. Overall, we conclude that the vast majority of students found some value in the peer grading process.

5. CONCLUSIONS

In this work we study the problem of student evaluation at scale via peer grading using ordinal feedback. We cast this as a rank aggregation problem and study different probabilistic

%	A_1	A_2	A_3	A_4	A_5	Total
B_1	**34.58**	2.08	5.83	10.00	1.67	**54.17**
B_2	5.42	0.00	5.83	7.08	1.67	20.00
B_3	0.42	2.92	2.08	2.50	0.42	8.33
B_4	2.92	0.83	5.00	5.42	0.00	14.17
B_5	0.00	0.00	0.42	1.67	1.25	3.33
Total	**43.33**	5.83	19.17	26.67	5.00	

Table 6: Results of the student survey, coded according to the categories in Table 5.

models for obtaining student grades, as well as estimating the reliability of the peer graders. Using data collected from a real course, we find that the performance of ordinal peer grading methods is at least competitive with cardinal methods for grade estimation, even though they require strictly less information from the graders. For grader reliability estimation, Mallow's model outperforms all other methods, and it shows consistently good and robust performance for grade estimation as well. In general, we find that ordinal peer grading is robust and scalable, offering a grading accuracy that is comparable to TA grading in our course.

This research was funded in part by NSF Awards IIS-1217686 and IIS-1247696, and the JTCII Cornell-Technion Research Fund.

References

[1] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: Ranking and clustering. *J. ACM*, 55(5):23:1–23:27, Nov. 2008.

[2] K. J. Arrow. *Social Choice and Individual Values*. Yale University Press, 2nd edition, Sept. 1970.

[3] J. A. Aslam and M. Montague. Models for metasearch. In *SIGIR*, pages 276–284, 2001.

[4] Y. Bachrach, T. Graepel, T. Minka, and J. Guiver. How to grade a test without knowing the answers - a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *ICML*, 2012.

[5] W. Barnett. The modern theory of consumer behavior: Ordinal or cardinal? *The Quarterly Journal of Austrian Economics*, 6(1):41–65, 2003.

[6] M. Bashir, J. Anderton, J. Wu, P. B. Golbus, V. Pavlu, and J. A. Aslam. A document rating system for preference judgements. In *SIGIR*, pages 909–912, 2013.

[7] L. Bouzidi and A. Jaillet. Can online peer assessment be trusted? *Educational Technology & Society*, 12(4):257–268, 2009.

[8] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):pp. 324–345, 1952.

[9] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: Preference judgments for relevance. In *ECIR*, pages 16–27, 2008.

[10] C.-C. Chang, K.-H. Tseng, P.-N. Chou, and Y.-H. Chen. Reliability and validity of web-based portfolio peer assessment: A case study for a senior high school's students taking computer course. *Comput. Educ.*, 57(1):1306–1316, Aug. 2011.

[11] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *WSDM*, pages 193–202, 2013.

[12] J. Diez, O. Luaces, A. Alonso-Betanzos, A. Troncoso, and A. Bahamonde. Peer assessment in moocs using preference learning via matrix factorization, 2013.

[13] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW*, pages 613–622, 2001.

[14] S. Freeman and J. W. Parks. How accurate is peer grading? *CBE-Life Sciences Education*, 9(4):482–488, 2010.

[15] J. Guiver and E. Snelson. Bayesian inference for plackett-luce ranking models. In *ICML*, pages 377–384, 2009.

[16] J. Haber. `http://degreeoffreedom.org/between-two-worlds-moocs-and-assessment`.

[17] J. Haber. `http://degreeoffreedom.org/mooc-assignments-screwing/`, Oct. 2013.

[18] R. Herbrich, T. Minka, and T. Graepel. Trueskill$^{\text{tm}}$: A bayesian skill rating system. In *NIPS*, pages 569–576, 2007.

[19] P. G. Ipeirotis and P. K. Paritosh. Managing crowdsourced human computation: a tutorial. In *WWW*, pages 287–288, 2011.

[20] M. Kendall. *Rank correlation methods*. Griffin, London, 1948.

[21] C. Kenyon-Mathieu and W. Schudy. How to rank with few errors. In *STOC*, pages 95–103, 2007.

[22] C. Kulkarni, K. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. Klemmer. Peer and self assessment in massive online classes. *ACM Trans. CHI*, 20(6):33:1–33:31, Dec. 2013.

[23] G. Lebanon and J. D. Lafferty. Cranking: Combining rankings using conditional probability models on permutations. In *ICML*, pages 363–370, 2002.

[24] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, Mar. 2009.

[25] T. Lu and C. Boutilier. Learning mallows models with pairwise preferences. In *ICML*, pages 145–152, June 2011.

[26] T. Lu and C. E. Boutilier. The unavailable candidate model: A decision-theoretic view of social choice. In *EC*, pages 263–274, 2010.

[27] R. D. Luce. *Individual Choice Behavior: A theoretical analysis*. Wiley, 1959.

[28] C. L. Mallows. Non-null ranking models. *Biometrika*, 44(1/2):pp. 114–130, 1957.

[29] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97, March 1956.

[30] M. Mostert and J. D. Snowball. Where angels fear to tread: online peer-assessment in a large first-year class. *Assessment & Evaluation in Higher Education*, 38(6):674–686, 2013.

[31] S. Niu, Y. Lan, J. Guo, and X. Cheng. Stochastic rank aggregation. *CoRR*, abs/1309.6852, 2013.

[32] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in MOOCs. In *EDM*, 2013.

[33] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975.

[34] T. Qin, X. Geng, and T.-Y. Liu. A new probabilistic model for rank aggregation. In *NIPS*, pages 1948–1956, 2010.

[35] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *JMLR*, 11:1297–1322, Aug. 2010.

[36] J. Rees. `http://www.insidehighered.com/views/2013/03/05/essays-flaws-peer-grading-moocs`.

[37] N. Shah, J. Bradley, A. Parekh, M. Wainwright, and K. Ramchandran. A case for ordinal peer-evaluation in MOOCs, 2013.

[38] N. Stewart, G. D. A. Brown, and N. Chater. Absolute identification by relative judgment. *Psychological Review*, 112:881–911, 2005.

[39] L. L. Thurstone. The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology*, 27:384–400, 1927.

[40] M. N. Volkovs and R. S. Zemel. A flexible generative model for preference aggregation. In *WWW*, pages 479–488, 2012.