Some Scaling Laws for MOOC Assessments

Nihar B. Shah
UC Berkeley
nihar@eecs.berkeley.edu

Joseph Bradley
UC Berkeley
joseph.kurata.bradley@gmail.com

Sivaraman Balakrishnan
UC Berkeley
sbalakri@berkeley.edu

Abhay Parekh
UC Berkeley
parekh@berkeley.edu

Kannan Ramchandran
UC Berkeley
kannanr@eecs.berkeley.edu

Martin J. Wainwright
UC Berkeley
wainwrig@stat.berkeley.edu

ABSTRACT

One problem that arises with the increasing numbers of students in Massive Open Online Courses (MOOCs) is that of student evaluation. The large number of students makes it infeasible for instructors or teaching assistants to grade all assignments, while current auto-grading technology is not feasible for many topics. As a result, there has recently been an increase in the use of peer-grading, where students grade each other; in this way the number of graders automatically scale with the number of students. However, in practice, peer-grading has been seen to have high error rates and has come under serious criticism.

In this paper, we take a statistical approach to assessing the feasibility of peer-grading for MOOCs. Using simple yet general models, we show that peer-grading as a standalone option will not scale, i.e., as the number of students increases, the expected number of students misgraded will grow proportionately. We then consider a hybrid approach that combines peer-grading with auto-grading. In this setting, an automated approach is used for 'dimensionality reduction', a classical technique in statistics and machine learning, and peer-grading is used to evaluate this lower dimensional set of answers. We show that this alternative approach has the potential to scale.

While most current research on assessment in MOOCs is empirical, our more theoretical approach provides a fundamental understanding of why there is such a high level of errors in current grading systems, and provides a direction for future research to solve this problem.

General Terms

Peer-grading, Educational Assessment, Scaling Laws, Massive Open Online Courses, Clustering, Dimensionality Reduction

ACM KDD 2014 Workshop on Data Mining for Educational Assessment and Feedback

1. INTRODUCTION

We discuss the scalability of grading, with special attention given to peer grading in Massive Open Online Courses (MOOCs). For MOOCs under budget constraints, instructor grading becomes infeasible as course sizes grow. Large courses require alternative approaches, such as auto-grading or peer-grading.

We focus our discussion on MOOCs for which auto-grading using pre-trained models is difficult. For subjective topics and complex problems, it is often difficult to design machine grading systems which are accurate [1]. In these cases, it is important that humans perform the grading.

Peer-grading is a system of grading where students taking a course are graded by other students in the same course. Peer-grading is a natural choice for MOOCs since the total number of graders in a peer-grading system automatically grows in proportion to the number of students enrolled. For instance, Coursera employs peer-grading in its human-computer interaction (HCI) course. Since the students are not expert graders, in this peer-grading system, the answer provided by each student is graded by 3 to 5 students. The final grade of a student is computed as the median of these individual grades [8]. Algorithms for aggregating peer-grades using probabilistic models are proposed in [14].

Research has shown (e.g., [8, 9]) that current auto-grading and peer grading systems make a large number of mistakes. Qualitative observations about the inaccuracy of MOOC assessment have led much criticism of auto-grading and peer-grading [1, 15]. For MOOC course credits to gain increased acceptance, these errors must be reduced.

In this paper, we view the problem of assessments in MOOCs through the lens of statistical analysis. Our approach is orthogonal to the largely empirical nature of the works in the literature in this field. We study the *scaling* behavior of peer grading, i.e., the behavior when the number of students gets large. Our analysis reveals that under reasonable assumptions, these systems will incorrectly grade a constant fraction of the students in expectation. This constant fraction is not a problem for small courses, where an instructor can handle complaints from students who feel they were misgraded. This is not a scalable solution for large courses since student

complaints will overwhelm instructors. This gives some insight into why current peer grading systems misgrade many students.

Current efforts to improve grading have examined many aspects: improving auto-grading models [10, 18, 7], improving models for aggregating peer-grades [14], combining auto and peer grading [9, 2], and dimensionality reduction [3, 4, 13, 16, 6, 11]. These efforts have shown useful improvements in practice. However, we argue that these methods do not change the scaling behavior of grading; at most, they decrease the constant fraction of misgraded students.

We then show that, on the upside, this scaling behavior may be improved via a *combination* of current methods used in MOOC assessments. Specifically, a combination of auto-and-peer-grading with 'dimensionality reduction' has the potential to create vanishing error rates, i.e., error rates which approach zero as MOOC course sizes grow. We outline what such a system would look like.

Importantly, our proposed framework is flexible: it allows for different grade collection systems, is independent of the algorithm employed for grade aggregation, and can also accommodate improvements such as adaptive grader assignments.

2. SCALING ANALYSIS OF PEER-GRADING

We are interested MOOCs which are *massive* and *open*. In other words, the courses are very large and offered for free or at low costs. This cost constraint prevents the number of instructors from scaling proportionally to the number of students. Thus, we may assume that instructors cannot hand-grade every student's answer.[1] Instead, grading is done via peer grading, where students evaluate their peers' answers, and/or auto-grading, where a computer software evaluates the students' answers.

Since grading requires considerable effort, we expect the average student to grade at most a few peers. We assume that most student graders perform better than random guessing but are imperfect, i.e., frequently make errors in the grading; empirical studies have shown this to be true (e.g., [8]). For our analysis, we assume that there is a "true" grade for each students' answer. The grades provided by expert graders can be assumed to match the true grades, but students (i.e., peer-graders) provide only noisy measurements of the true grade.

Theorem 1 below analyses the scaling behavior of typical peer-grading systems described above. Under weak assumptions regarding the setting, we show that in expectation, the grades of a constant fraction of students will be in error. We then extend our results to scenarios where experts such as the instructor or teaching assistants also assess a subset of the answers, and further when these assessments may be used to choose which student (peer) grades which answer. In an attempt to keep our claims independent of the model and inference algorithm, we consider a generic grading scenario and then show that under this scenario, the grades obtained from the peers for a constant fraction of students having a specific true grade are *indistinguishable* from those for students having a different true grade.

THEOREM 1. *Consider the following four settings:*

(a) There are d students enrolled in the course. Each student provides one answer, and these d answers must be graded. There are b possible values of the grade (in other words, the answers must be binned into b ordered bins). Each grader assigns the answer to one of these b bins. Each (peer-) grader grades k answers, and each answer is graded by k (peer-) graders. The choice of who-grades-whom is fixed but arbitrary. We assume that a constant fraction $\gamma \in (0,1)$ of the students are imperfect graders while the remaining fraction $(1-\gamma)$ of the students are perfect graders (the value of γ may be known to the inference algorithm). Consider any fixed $p_0, p_1, p_{2u}, p_{2d} \in (0,1]$ with $p_0 + p_1 + p_{2u} + p_{2d} = 1$. An imperfect grader, with probability p_{2u}, gives a grade that is two bins higher than the true grade; with probability p_{2d}, gives a grade that is two bins lower than the true grade; with probability p_1, gives a grade that is one bin away from the true grade; and with probability p_0, gives a grade equal to the true grade.[2] A perfect grader's evaluation always exactly matches the true grade of the answer being graded. The constants b, k, γ, p_0, p_1, p_{2u}, p_{2d} are all independent of the number of students d.

(b) In addition to the grading process of (a), a constant fraction $\zeta \in [0,1)$ of answers are also graded by experts who grade perfectly.

(c) The grading process is adaptive: The first stage is a calibration stage where experts grade a constant fraction ζd of answers, and each peer-grader grades a constant number of answers. Next, based on the results of the first stage, the choice of which student grades which answer for the remaining part is made. Finally, the grading process of (a) is applied with this choice of who-grades-whom.

(d) There are only two possible grades ('pass' and 'fail'), and an imperfect grader grades an answer correctly with probability p_0 and makes a grading error with a probability $(1-p_0)$.

Under each of these settings, in expectation, the final grades of a constant fraction of students will be in error.

The scaling results derived here imply that if the distribution of the quality of peer-graders is independent of the number of enrolled students, then as the number of students in the course increases, the number of misgraded students will increase proportionately. The student experience will suffer in such a scenario, and instructors may be faced with an overwhelming number of student complaints.

[1]The term 'answer' will be used generically to refer to exam-solutions, homeworks, or any other material submitted for evaluation by the students.

[2]In order to simplify the setting, we ignore the possible non-existence of a bin that is 'two bins higher' or 'two bins lower'; this assumption is inconsequential to the scaling law.

The remainder of this section discusses the assumptions and some extension of Theorem 1. The assumption of having the parameter γ independent of the total number of enrolled students d means that the proportion of good students does not (significantly) increase when the number of total students scales up. We assume b to be independent of d since it is typical of courses to keep the number of grade-levels independent of the number of students enrolled. The independence of parameters p_0, p_1, p_{2u}, p_{2d} from d means that the average abilities of the students do not vary (significantly) when the number of students enrolled in the course increases. Finally, and very importantly, we assume k to be a constant (and not increasing in d) since grading is a task that requires significant time and effort on the part of the grader, and we must impose a limit on the number of papers the students need to peer-grade.

Given the formulation of Theorem 1, the math behind the proof is fairly simple. We show that under the four settings considered in the theorem, the probability that any individual student will be in error is lower bounded by a constant, which implies that in expectation, a constant fraction of students will be in error. Details of the proofs are provided in the appendix. A similar argument holds for the case when the grades collected from the graders are on a finer scale (with more than b levels), or when the peer-grading is ordinal, i.e., where students are asked to compare two or more answers instead of assigning numeric scores to the answers [17]. Peer- grading may be generalized to also include grading by people not taking the course, e.g., by people who have previously taken the course. If the number of such people remains linear in the number of students, then arguments similar to those of Theorem 1 continue to apply.

The setting of Theorem 1 considers adaptivity of only upto two stages, and this stems from the requirement of providing timely feedback to students [8]. In order to ensure fairness, the use of extraneous information about the students' skills is generally avoided in the grading process [14], as assumed in the theorem. One could potentially obtain some information about the students' grading abilities from their performance in previous homeworks or tests: we conjecture that using this information will not change the scaling laws as long as the number of homeworks/tests remains independent of the number of students; attempting to prove this remains a part of future work. Finally, the model considered in Theorem 1 also subsumes 'score-based' models, as described in Corollary 2 below.

COROLLARY 2. *Consider a setting where each answer has a true real-valued 'score', and a quantization of the score forms a grade. Assume without loss of generality that the score of any answer can take any value in the interval $[0, 1]$, and that the set of k bins (grades) is a partition of this interval into k non-empty intervals. The true 'scores' of the d answers are distributed on this interval such that each interval contains some constant fraction of answers. The evaluation of an answer by a peer-grader is modeled as adding a noise to the true score of that answer, where the noise is distributed i.i.d. with some cdf F on $[-1, 1]$ (e.g., a truncated Gaussian distribution). The grade given by the peer to an answer is a quantization of this evaluation according to the grade-intervals. Assume that the distribution F assigns a non-zero probability to every non-empty interval in $[-1, 1]$.*

If the grading process from any part (a)-(d) of Theorem 1 is executed under this setting, then in expectation, the final grads of a constant fraction of students will be in error.

3. BREAKING THE BARRIER VIA DIMENSIONALITY REDUCTION

One possible means of breaking the barrier of a constant fraction of answers being graded erroneously is *dimensionality reduction*. We discuss two methods for dimensionality reduction: clustering and featurization. Clustering is the more intuitive method, but featurization is more general. These methods combine auto-grading with peer-grading in a manner to be discussed in the sequel. The peer-grading interface can remain the same as before, where students grade answers submitted by their peers.

3.1 Clustering

Suppose we use a computer program to cluster the collection of all d answers provided by the d students, with respect to the similarity of their content. The clustering algorithm is such that ideally, within each cluster, all answers have the same true grade. Note that multiple clusters could have the same true grade. Given this assumption, grading any one answer in the cluster effectively grades all answers in the cluster. The total number of answers to grade is effectively reduced.

A cluster can be graded by collecting peer-grades for any answer within the cluster. These grades may be aggregated via any reasonable algorithm, for instance, taking a median of the received grades. Finally, the answer of any student is assigned the grade that its cluster receives.

The following theorem shows that a good performance of the clustering algorithm can reduce the number of erroneously graded students to a vanishing fraction. Here, $D(p_1||p_2)$ denotes the Kullback-Liebler divergence between two Bernoulli distributions having parameters p_1 and p_2 respectively.

THEOREM 3. *Consider the following grading process: There are d students. A clustering algorithm partitions the answers of these students into Δ clusters, and one (representative) answer is selected from each cluster. The set of grades have only two possible values: pass or fail. Each student grades a constant number k of (representative) answers, and each (representative) answer receives an equal number of grades. The final grade given to all answers in a cluster is decided via a majority vote on the (peer-) grades obtained by the answers in the cluster.*

First assume that the clustering algorithm is perfect, so that all answers in a cluster share the same true grade. There are no perfect graders. Each grader grades correctly with probability $p_0 \in (0.5, 1]$ and makes an error with probability $(1 - p_0)$. Then, if the number of clusters satisfies $\Delta \leq c\frac{d}{\log d}$ and the number of answers graded by each student satisfies $k \geq \frac{c}{D(0.5||p_0)}$ for some universal constant c, then the number of students who are misgraded (in expectation) is a constant independent of d.

Now, if the clustering algorithm is imperfect, but erroneously clusters some $\beta(d)$ (non-representative) answers where $\beta(d)$ is sub-linear in d, then the number of students who are misgraded in expectation is upper bounded by the sum of $\beta(d)$ and the constant obtained above. As a result, the fraction of answers that are misgraded goes to zero as the total number of answers d grows.

The theorem says that if the answers can be clustered into clusters of average size of order $\log d$ or higher, then even when each student can grade only a constant number of answers, the number of answers graded incorrectly will be independent of the number of students d. The system can now scale to accommodate arbitrarily large numbers of students without the worry of the number of grading errors blowing up.

The theorem assumes a good performance of the clustering algorithm. This may seem contradictory to our earlier discussion on the performance of auto-grading algorithms suggesting that they do not perform very well for topics that are subjective in nature. To this end, we note that while auto-grading requires the algorithm to understand the semantics of each answer in order to grade, clustering only requires understanding the features on which similarity testing must be performed. The job of clustering may thus be viewed as a subset of the superior task of auto-grading. Indeed, designing algorithms for clustering answers for educational assessment is an area of active research in the community, e.g., [3, 4, 13, 16, 6, 11].

Let us now discuss the assumption on the performance of the clustering algorithm from a statistical perspective. We have assumed that the number of answers clustered erroneously grows sub-linearly in d. In other words, we assume that the fraction of answers clustered incorrectly reduces with an increase in the total number of answers. An intuitive justification of this assumption is that clustering is often performed by comparing answers with each other, and as the total number of answers d grows, the number of answers (in each cluster) available for comparison also grows. This intuition is supported by the literature on statistical guarantees for clustering problems (e.g., [5]) which we employ below to formalize the intuition stated above. The setting considered in the literature [5] does not exactly match our requirements, nevertheless, the results are highly encouraging due to the considerable similarity with our setting and the strong guarantees available.

PROPOSITION 4. *(Adaption of [5, Theorem 2.2]) Consider a clustering algorithm that operates in the following manner. The algorithm has a black-box comparator that, given any pair of answers, correctly identifies whether they belong to the same cluster or not with a probability at least $\frac{1}{2} + \epsilon$ (independent of all other comparisons). Here, $\epsilon \in \left(0, \frac{1}{2}\right]$ is a fixed value, unknown to the algorithm. Suppose there are $\Delta \leq c_1 \epsilon^2 \frac{d}{\log d}$ clusters, of equal size, and the value of Δ is known to the algorithm. Here, c_1 is a specific universal constant. Then there exists a clustering algorithm such that the expected number of answers clustered incorrectly is upper bounded by a universal constant (independent of d).*

REMARK 1. *The setting of [5, Theorem 2.2] differs from our setup in Theorem 3 in the following respects: We assume the average cluster-size to be at least $c \log d$ (for some constant $c > 0$) and further require that the clustering algorithm does not know the sizes of these clusters; [5, Theorem 2.2] assumes that the sizes of all clusters are identical, lower bounded by $c \log d$, and the size of each cluster is known to the algorithm. Moreover, for many parameter regimes of interest, no polynomial-time clustering algorithms are known. On the other hand, the results of [5, Theorem 2.2] provide a very strong guarantee in that only $O(1)$ answers are in wrongly clustered (in expectation), whereas for the purpose of Theorem 3, a guarantee of $o(d)$ errors would suffice.*

The scaling analysis presented in this section suggests that reducing the dimension of the answers by a logarithmic factor may suffice for designing a scalable grading system. Previous works on clustering which we reference used clustering to aid instructors in grading. This approach can lessen the burden on instructors, who can assign grades to groups of answers. Our analysis motivates and theoretically justifies the use of clustering tools [3, 4, 13, 16, 6, 11] for massive open online courses, where in conjunction with peer-grading, we show that they can help achieve scalability in the grading process.

3.2 Featurization

We briefly discuss a more general type of dimensionality reduction. Our clustering method assumes that many answers are similar enough to be declared equivalent by a clustering algorithm. One could generalize to assume that *parts* or *aspects* of many answers are similar and can be compared or clustered; we describe this as *featurization*, where the content of an answer is summarized by a set of features.

Suppose that answers may be described by Δ features. Assume that the grade of each answer may be computed as a function of these features; e.g., a simple such function for pass/fail grades would be thresholding a weighted sum of the features. Then we have reduced the problem of grading to a very traditional regression setting: answers are examples, features are computed algorithmically for each example, and peer-grades provide noisy labels for the examples.

The regression model is simply another aggregation method for peer-grades. Assume for now that we use a generalized linear model $y^{(i,t)} \sim f(w^T x^{(i)}) + \epsilon^{(i,t)}$, where $x^{(i)}$ is the feature vector for answer i, $y^{(i,t)}$ is the t^{th} peer-grade for answer i, $f(\cdot)$ is the inverse link function, and $\epsilon^{(i,t)}$ is the noise added by the peer-grader. This model generalizes simple peer-grading systems which treat all answers independently: d boolean features are indicators corresponding to the d answers, so the feature vector $x^{(i)}$ for answer i has a "1" for feature i and "0" elsewhere. The model also generalizes clustering: we have Δ boolean features corresponding to the Δ clusters.

This setup could allow grading systems to draw on extensive research on feature engineering and modeling. The success of previous work on feature-based clustering of answers in MOOCs indicates that useful features can be found. The fact that current aggregation methods for peer-grades can

be generalized by simple regression models indicates that such models are reasonable.

Depending on the choice of the regression model, one can achieve scaling results similar to those for clustering. E.g., consider a logistic regression model (with a logit link function) to classify answers as pass/fail. Then previous work [12] on logistic regression has shown that, with Δ features and $\Theta(d)$ samples (peer-grades), it suffices to have $\Delta = o(d)$.

Previous work has considered combining auto-grading with peer-grading [9]. However, that work separated auto-grading and peer-grading into two stages, rather than combining them into a joint model. The first stage used auto-grading with expert-labeled examples to compute initial grades, and the second stage used peer-grading to improve grades based on confidences in the first stage. This separation makes it unclear whether scaling laws are changed: either auto-grading makes a constant fraction of grading errors (in which case a constant fraction of errors remain after peer-grading as well), or auto-grading makes a vanishing fraction of errors (in which case peer-grading becomes unnecessary as class sizes grow).

4. CONCLUSIONS

In this paper, we gave a rudimentary analysis of the scaling properties of various grading mechanisms in MOOCs. We saw that under very simple and general models, the kinds of peer-grading systems employed today will not scale. We then showed that combining (auto-) dimensionality reduction and peer-grading has the potential to scale. Dimensionality reduction is already an active topic of research [3, 4, 13, 16, 6, 11], and the proposal of combining it with peer grading falls under the more general paradigm of combining machine and human intelligence [9, 2].

While most current research on assessment in MOOCs is empirical, this paper provided a more theoretical approach, helping understand the fundamental sources of the errors observed in current grading systems, and a path for future research to overcome those errors. Accurate, reliable, and scalable assessment will help to pave the way for MOOCs to democratize education.

5. REFERENCES

[1] Professionals against machine scoring of student essays in high-stakes assessment. `http://humanreaders.org/petition/index.php`. Retrieved June 1, 2013.

[2] V. Aggarwal, A. Minds, S. Srikant, and V. Shashidhar. Principles for using machine learning in the assessment of open response items: Programming assessment as a case study. In *NIPS Workshop on Data Driven Education*, Dec. 2013.

[3] S. Basu, C. Jacobs, and L. Vanderwende. Powergrading: a clustering approach to amplify human effort for short answer grading. *TACL*, 1:391–402, 2013.

[4] M. Brooks, S. Basu, C. Jacobs, and L. Vanderwende. Divide and correct : Using clusters to grade short answers at scale. In *Learning at Scale*, 2014.

[5] Y. Chen and J. Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*, 2014.

[6] E. Glassman, J. Scott, R. Singh, and R. Miller. Overcode: Visualizing variation in student solutions to programming problems at scale.

[7] T. Kakkonen, N. Myller, J. Timonen, and E. Sutinen. Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 29–36. Association for Computational Linguistics, 2005.

[8] C. Kulkarni, K. Pang-Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer. Peer and self assessment in massive open online classes. *ACM Transactions on Computer-Human Interaction*, 9(4), 2013.

[9] C. Kulkarni, R. Socher, M. S. Bernstein, and S. R. Klemmer. Scaling short-answer grading by combining peer assessment with algorithmic scoring. In *Learning at Scale*, 2014.

[10] L. S. Larkey. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 90–95, 1998.

[11] A. Luxton-Reilly, P. Denny, D. Kirk, E. Tempero, and S.-Y. Yu. On the differences between correct student solutions. In *Proceedings of the 18th ACM conference on Innovation and technology in computer science education*, pages 177–182. ACM, 2013.

[12] A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *NIPS*, 2002.

[13] A. Nguyen, C. Piech, J. Huang, and L. Guibas. Codewebs: scalable homework search for massive open online programming courses. In *Proceedings of the 23rd international conference on World wide web*, pages 491–502. International World Wide Web Conferences Steering Committee, 2014.

[14] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in MOOCs. In *International Conference on Educational Data Mining*, 2013.

[15] J. Rees. Peer grading can't work. `http://www.insidehighered.com/views/2013/03/05/essays-flaws-peer-grading-moocs`. March 5, 2013.

[16] S. Rogers, D. Garcia, J. F. Canny, S. Tang, and D. Kang. ACES: Automatic evaluation of coding style. Master's thesis, EECS Department, University of California, Berkeley, May 2014.

[17] N. B. Shah, J. K. Bradley, A. Parekh, M. Wainwright, and K. Ramchandran. A case for ordinal peer-evaluation in MOOCs. In *NIPS Workshop on Data Driven Education*, Dec. 2013.

[18] Y. Zhenming, Z. Liang, and Z. Guohua. A novel web-based online examination system for computer science education. In *IEEE Frontiers in Education Conference*, volume 3, 2003.

APPENDIX: PROOFS

Proof of Theorem 1.

(a) Consider any one student. When the assignment of who-grades-whom is made, the grading capabilities of the students are unknown. Hence for this particular student, the set of k graders are distributed uniformly at random among the remaining $(d-1)$ students. The probability that the k graders for this student are all imperfect is $\frac{\binom{(d-1)\gamma}{k}}{\binom{d-1}{k}}$. Applying standard bounds on binomial coefficients, we get that this quantity is lower bounded by $\frac{\gamma^k}{e^k}$. Now, the probability that the student is graded two bins lower by each of these graders is p_{2d}^k. When this happens, this student is indistinguishable from a student whose true grade is two bins lower and is graded correctly by all of its (peer-) graders. This happens with a probability at least

$$\frac{p_{2d}^k \gamma^k (1-\gamma+p_0\gamma)^k}{2e^k},$$

which is a constant independent of d. Using the linearity of expectation, we get that the expected number of students whose grades are in error is lower bounded by

$$d\frac{p_{2d}^k \gamma^k (1-\gamma+p_0\gamma)^k}{4e^k}$$

which is linear in d.

The proof above considered inferring the grade of a student from only the quality of her answer. In some situations, one may try to evaluate the student based on her performance in the peer-grading process as well. We argue that even in this case, the scaling laws will remain the same since the probability that a specific student's peer-grading performance is identical to that of another student whose true grade is two bins lower is non-zero and independent of d. The probability of these two students being indistinguishable therefore remains non-zero and independent of d, thereby ensuring that the expected number of students graded erroneously is lower bounded by a constant fraction of d.

(b) Consider any imperfect grader. Consider the event that this imperfect grader perfectly grades all the answers that are also graded by experts. The probability of this event is lower bounded by p_0^k (of course the grader does not know which answers are graded by experts, and the event represents the random chance of grading these answers perfectly). Now such an event would make an imperfect grader indistinguishable from a perfect grader. As a result, for any answer that is not graded by an expert, a lower bound on the probability that this answer is indistinguishable from an answer whose true grade is two bins lower is

$$\frac{p_{2d}^k \gamma^k {p_0^k}^2 (1-\gamma+p_0\gamma)^k}{2e^k}.$$

Thus the expected number of students whose grades are in error is lower bounded by

$$d\frac{p_{2d}^k \gamma^k {p_0^k}^2 (1-\gamma+p_0\gamma)^k (1-\zeta)}{4e^k},$$

which is linear in d.

(c) The argument in the proof of part (b) continues to hold in this setting. In (b), the performance of an imperfect grader on the answers that are graded by experts is indistinguishable from the performance of a perfect grader. When this happens, there is no information about these imperfect graders available for designing the second stage.

(d) Arguments analogous to the previous parts continue to hold, with p_{2d} replaced by $(1-p_0)$. □

Proof of Corollary 2.

First observe that our assumptions imply that a constant fraction of students have a grade that is the maximum possible, and another constant fraction of students have a grade that is two bins below the maximum possible. The final grade received by an answer depends only on these (quantized) grades given by the peer-graders, and this brings us to precisely the setting of Theorem 1. Since the distribution F assigns a non-zero probability to any non-empty interval, the probabilities of an imperfect (peer-) grader giving a grade that is two bins lower or a grade identical to the true grade are both non-zero, and i.i.d. for all evaluations. This ensures $p_0 > 0$ and $p_{2d} > 0$ in the setting of Theorem 1. Applying the proof of Theorem 1 gives the desired result. □

Proof of Theorem 3.

First consider the setting of a perfect clustering algorithm. The total number of grades obtained from the students is kd. Since there are Δ clusters, each cluster receives $\frac{kd}{\Delta}$ grades (we shall ignore any rounding effects for now since these are inconsequential to the scaling behavior). Each grade is correct with a probability $p_0 \in (0.5, 1]$, and hence, applying the Chernoff bound, we get that the probability of misgrading any individual cluster is upper bounded by

$$\exp\left(-D(0.5||p_0)\frac{kd}{\Delta}\right).$$

Thus, in expectation, the total number of misgraded students is

$$d\exp\left(-D(0.5||p_0)\frac{kd}{\Delta}\right).$$

Substituting $\Delta \leq c\frac{d}{\log d}$ and $k \geq \frac{c}{D(0.5||p_0)}$ as assumed in the statement of the theorem, we get that the total number of misgraded students is $O(1)$ in expectation.

If the clustering algorithm incorrectly clusters $\beta(d)$ (non-representative) answers, with $\beta(d) = o(d)$, then the expected total number of errors is $o(d) + O(1)$ which is still $o(d)$. □

Proof of Proposition 4.

The parameters p, q, n and K in [5, Theorem 2.2] are respectively lower bounded by $\frac{1}{2}+\epsilon$, upper bounded by $\frac{1}{2}-\epsilon$, equal to d and equal to $\frac{d}{\Delta}$ in our setting. The assumption $\Delta \leq c_1\epsilon^2\frac{d}{\log d}$, with c_1 being large enough, ensures that the condition [5, Equation (7)] required by the theorem is satisfied. While the statement of [5, Theorem 2.2] guarantees correct clustering of all answers with a probability at least $(1-c_2d^{-c_3})$ for some constants c_2 and c_3, the proof of the theorem establishes the value of the constant c_3 as 1. It follows that the number of answers clustered incorrectly is upper bounded by $d \times \left(c_2 d^{-c_3}\right) = c_2$. □