On the Validity of Peer Grading and a Cloud Teaching Assistant System

Tim Vogelsang
iversity GmbH
Berliner Straße 33
Bernau bei Berlin, Germany
t.vogelsang@iversity.org

Lara Ruppertz
iversity GmbH
Berliner Straße 33
Bernau bei Berlin, Germany
l.ruppertz@iversity.org

ABSTRACT

We introduce a new grading system, the Cloud Teaching Assistant System (CTAS), as an additional element to instructor grading, peer grading and automated validation in massive open online courses (MOOCs). The grading distributions of the different approaches are compared in an experiment consisting of 476 exam participants. 25 submissions were graded by all four methods. 451 submissions were graded only by peer grading and automated validation. The results of the experiment suggest that both CTAS and peer grading do not simulate instructor grading (Pearson's correlations: 0.36, 0.39). If the CTAS and not the instructor is assumed to deliver accurate grading, peer grading is concluded to be a valid grading method (Pearson's correlation: 0.76).

Categories and Subject Descriptors

Human-centered computing [**Human computer interaction (HCI)**]: HCI design and evaluation methods—*Laboratory experiments*; Human-centered computing [**Human computer interaction (HCI)**]: Empirical studies in HCI; Human-centered computing [**Collaborative and social computing**]: Empirical studies in collaborative and social computing

General Terms

Measurement

Keywords

MOOCs, peer grading, iversity, CTAS, validity

1. INTRODUCTION

Top-down MOOCs easily scale to hundreds of thousands of students, e.g. by means of video lectures and quizzes[19]. Instructors' feedback and grading, however, do not easily scale to these numbers of students, although this would be necessary for meaningful certification of MOOC students. When completely automated grading is not possible, as in the case of essays, more complex review models are required. In order to solve this problem and grade huge amounts of exam papers, online course providers such as Coursera, EdX, NovoEd and iversity use peer grading systems[13, 18]. As opposed to one instructor grading the exams of all students, such systems facilitate mutual grading by students.

The concept of peer grading in education is not new (cf. Section 2.2), and the designers of peer grading systems for MOOCs were already aware that students and instructors worry about reliability and validity of peer grading systems[6, 16].

The real new development is the possibility of massive scale peer grading, emerging in the context of MOOCs. When trying to scale the whole area of open educational resources (OER)[17], educators and computer scientists now face the challenge of designing peer grading systems that collect data sets rich enough to allow the computation of accurate grades by means of appropriate statistical models and algorithms. Consequently, the existing digital peer grading systems differ from one another in both design and underlying statistical models: Common features of a peer grading system are self grading[13, 22, 12] and calibration[2]. The latter means that peers' grading proficiency is adjusted according to training assignments that they grade, normally with the help of certain grading rubrics. Finally, an important property of a peer grading system is whether multiple assessments can be integrated. This can be realized with regular peer graded homework. If data from such multiple assessments can be taken into account, more advanced statistical models can be applied than in a single assessment situation[20].

This paper investigates the application of a new grading approach: The Cloud Teaching Assistant System (CTAS). Qualified teaching assistants, not belonging to the instructor's staff, are hired only for the purpose of reviewing and grading the submissions of a single MOOC assignment.

In order to overcome the potential contradiction of grading accuracy and scalability for problems like single essay grading in MOOCs, the CTAS should finally be integrated into a complex process of peer grading, professor grading, self grading and calibration, together with an appropriate statistical model. In order to achieve this goal, the role of the CTAS as a hybrid element between the students and the instructor in a MOOC has to be understood. Hence, the objective of this paper is to gain observational insights into the CTAS by looking at a concrete example, and to investigate its validity as well as its relation to instructor grading, peer

LAK'15 Mar 16-20, 2015, Poughkeepsie, NY, USA
Copyright 2015 ACM 978-1-4503-3417-4/15/03.
http://dx.doi.org/10.1145/2723576.2723633.

grading and automated validation.

2. RELATED WORK

2.1 Reliability and validity of peer grading

In order to judge the accuracy of given grades, literature distinguishes between reliability and validity of peer grading. Cho et al. remark that reliability is ignored in most studies[5] and that there are even cases where validity is misreported as reliability[24]. The measure of reliability most commonly studied is inter-rater reliability, which refers to the consistency of grades that different peers would assign to the same assessment. Validity concepts assume that the instructor provides a trustworthy grade, and hence validity of peer grading is calculated as a correlation coefficient between peer grades and instructor grades[5, 15, 3].

Examples of reported reasons to doubt both the reliability and validity of peer grading include lack of experience with the required content and methods, as well as bias due to uniformity, race, friendship[6] or even manipulation[16]. However, the hypothesis that the validity of peer grading is higher in advanced rather than beginner courses or in science and engineering rather than in other disciplines was not supported by a meta-study by Falchikov and Goldfinch, which analyzed forty-eight peer assessment studies[7]. Despite the possible problems connected to peer grading, peers are reported to be more likely to understand other student's problems than instructors [11].

The CTAS affords a new hybrid grading element which is allocated between instructor and peer grading. If we were to count the teaching assistants as among the instructor's staff, it would be reasonable to investigate the inter-rater reliability among the staff. However, since these teaching assistants are only hired for grading and not for course development, we interpret the CTAS as a separate unit. Hence, this paper investigates the validity of the CTAS, particularly in comparison to peer and instructor grading.

2.2 Digital peer grading systems

Pinkwart and Loll date the technical foundation of digital peer grading systems back to collaborative filtering techniques[14], originally describing algorithms that create and distribute system knowledge from associating users to system artifacts such as products or user profiles[9]. This concept is the basis of today's Web 2.0 technologies and applications include recommendation (amazon), tagging (flickr) or peer grading.

Before online education became massive with MOOCs, peer grading tools were already used in eLearning. One of those tools is PeerGrader (PG)[14, 8], which has special communication functionality and the ability to modify and re-submit assessments. Further, the Scaffolded Writing and Rewriting in the Discipline (SWoRD)[4] was designed for written assessments in large content courses and inspired by the reviewing and publication process of journals. Another example is Calibrated Peer Review (CPR)[21], which features both grading of training assessments for calibration and self-assessment.

CPR was used for two consecutive Coursera offerings of the course Human Computer Interaction, taught by Scott Klemmer[13] and analyzed afterwards[20] with statistical models close to the Bayesian approach used by Goldin et. al in smaller classrooms[1, 10]. EdX developed its own framework for peer grading, integrating scoring and feedback. This framework uses the three techniques self assessment, peer assessment, and artificial intelligence (AI)[18].

These peer grading systems follow the popular idea of crowdsourcing[23] in a way that shifts the burden of grading from those who create the assessment (instructors) to those who complete it. In such systems, teaching assistants naturally belong to the staff of the instructors. To the best of our knowledge none of the existing massive scale peer grading systems are explicitly designed for having teaching assistants as a hybrid third grading element in between, neither belonging to the staff, nor to the students. The CTAS is meant to fill this gap.

3. METHODS

3.1 Research Questions

As a result of the literature review in Section 2, we are mainly interested in the following questions:

- Q1: How well do MOOC students perform on automatically validated questions compared to the human graded essay question?
- Q2: How are grades given by peers, the CTAS and the instructor statistically distributed and do they reveal a bias?
- Q3: How do students' performances and the grading types differ among different scoring criteria?
- Q4: How valid are CTAS and peer grading?

The experiment is designed with the goal of shedding light on these questions and Section 4 is structured accordingly. Q4 is the main question of our investigation.

3.2 Experiment Setup

We carried out an experiment within the iversity MOOC 'Public Privacy: Cyber Security and Human Rights' with 4620 active students, using iversity's proprietary implementation of assessment and grading. At the end of this course, 476 students participated in an online exam consisting of 16 multiple-choice questions with automated validation and one essay question without automated validation. During a subsequent grading phase, 377 essays were successfully peer evaluated and graded by a total of 395 fellow course participants (*peers*). For various reasons not all students qualified for evaluating or being evaluated. Disqualification reasons include, for example, empty submissions and incomplete grading. Additionally, 25 of those essays were corrected and graded within a *Cloud Teaching Assistants System (CTAS)* as well as by the course instructor. We were able to gather four sets of grades corresponding to the three groups of graders, namely a) peers, b) the CTAS, and c) the instructor as well as d) the automated validation of multiple-choice questions. Using this data we were able to measure the correlation between peer grading, CTAS, instructor grading and automated validation. In order to calculate the correlations we use Pearson's correlation coefficient throughout this paper. This is the coefficient used in similar validity investigations [3, 5]. Where necessary, we aggregate multiple peer, CTAS or multiple-choice results by mean. We remark that multiple-choice results measure the general knowledge of a student but cannot be included into the validity analysis, which has the essay question as a basis. However, the data allows us to judge upon the validity of peer grading and CTAS, by comparing it with instructor grading.

3.3 The Peer Grading Assignment

This paper analyzes the final exam for the MOOC *Public Privacy: Cyber Security and Human Rights*, which was offered on the iversity platform from November 2013 to February 2014. It is an introductory course offered by Associate Prof. Dr. Anja Mihr (Utrecht University, Netherlands). The course systematically examines the compliance between international human rights norms, standards and mechanism within legal and political frameworks and the growing cyber security regime. A total of 33,782 students registered for this course, 4620 of which were still actively engaged in the last week of the course.

The instructor of the course, Prof. Anja Mihr, developed most of the course's content and material, including the final exam, in which 467 students participated. The final exam consisted of 16 multiple-choice questions, which aimed at assessing students' knowledge about theories and concepts introduced throughout the course. 10 of these 16 multiple-choice questions had already been answered by students throughout the course. In addition, the final exam included a peer grading assignment, where students were asked to write a short essay elaborating on the idea of the Rule of Law in Cyber Space as a means of cyber security (cf. Appendix B).

After submitting the essays, students had to grade their peers' essays according to a grading rubric consisting of five criteria: (1) thesis and central idea, (2) essay organization, (3) content, (4) balanced argumentation, and (5) citation and sources. Each criterion was rated using a five point rating scale ranging from zero to four. For a detailed description of the grading rubric confer to Appendix A.

In order to facilitate the peer grading, students were required to evaluate seven of their peers' essays, using the peer grading tool provided on the platform. In addition to scoring the essays according to the grading rubric, students also had to provide a brief feedback comment. However, they were not asked to grade their own essays based on the criteria and rating scales provided. Though only seven peer evaluations were required, a large number of students graded even more exams. Since students participating in the final exam had been required by the course syllabus to take the mid-term exam, made up of a similarly designed peer grading assignment, they were used to the process, the grading tool and the grading rubric.

3.4 The Cloud Teaching Assistant System

In addition to peer grading, the iversity platform offers the *Cloud Teaching Assistant System (CTAS)*. This is another method to process assignments that cannot be evaluated automatically. To this end, a group of qualified examiners is assembled by iversity in order to realize grading. Throughout this paper, such an examiner is referred to as *Cloud Teaching Assistant (CTA)* and the respective grading as *CTA grading*. In order to ensure their quality, CTAs are selected according to their expertise in the relevant subject as well as with the help of a criteria sheet defined in advance by instructors. The selection takes place in two steps. First of all, iversity draws up a pre-selection of candidates and presents them to the instructor. The final selection of the CTAs is made by the instructor.

	student	result	premium
1	12963	2.25	no
2	14042	3.50	no
3	14508	3.00	no
4	14532	3.75	no
5	14769	3.50	no
6	15597	3.50	no

Table 1: Results of automated multiple-choice validation for all 476 students. 16 multiple-choice questions led to 16 possible points. The column *result* contains the final multiple-choice result divided by four.

In the fall of 2013 the Cloud Teaching Assistant System was applied in the MOOC Public Privacy: Cyber Security and Human Rights to the 25 final assignments of the course participants who had registered for CTA grading. In this paper we refer to these students as *premium students*. Thus, alongside the standard peer grading, exams of these premium students were additionally graded by a group of four CTAs, such that each exam was evaluated by two CTAs. Prior to starting the grading, CTAs were required to take the MOOC themselves, to carefully examine the final assignment and the rubrics, according to which the grading had to be done. However, their grading capacities were not particularly tested, the assumption having been made that as trained professionals they knew how to grade essays, i.e. provide valid grading similar to that of the course instructor. For grading the exams the four CTAs used the same grading rubric as given to the peers (cf. Appendix A). Likewise, they also used the proprietary grading tool provided on the iversity platform to evaluate the exams.

3.5 The data sets

Throughout the analysis we work with four data sets. The third and the fourth dataset are obtained by aggregating operations on the second dataset. The first dataset (snapshot in Table 1) contains for every student his achieved result from the automatically validated multiple-choice test and whether he is a premium student or not. It contains 476 observations, one for each student. The final result is divided by four, in order to simplify the comparison to the essay grades and underlying criterion scores.

The second dataset (snapshot in Table 2) contains the essay scores for each criterion given to premium students by peers, CTAs and the instructor. It consists of 1160 observations, i.e. every premium student received an average of 46.4 scores, having been scored by one instructor, two CTAs and on average 6.28 peers in each of the five criteria.

The third dataset (snapshot in Table 3) is also composed of scoring information about premium students. For every premium student and each of the five criteria, we aggregate by grading type, i.e. the dataset contains the mean received peer score, the mean received CTA score and the received professor's score. Hence, we obtain 125 observations.

The multiple-choice evaluation is apparently not based on multiple criteria. In order to compare students' essay grades to their multiple-choice performance, we aggregated dataset 3 by taking the mean over all five criteria scores.

	criterion	scorer	scoree	score	type
106	thesis	1081825	12783	3	peer_scored
107	organization	1081825	12783	3	peer_scored
108	content	1081825	12783	4	peer_scored
109	argumentation	1081825	12783	3	peer_scored
110	sources	1081825	12783	4	peer_scored
111	thesis	353963	12783	3	peer_scored

Table 2: Results of human evaluation for all 25 premium students - not aggregated. Every row represents one exam, one grading type (peer, CTA or instructor) and one criterion.

	scoree	criterion	peer_scored	cta_scored	prof_scored
1	14781	thesis	3.00	2.50	3.00
2	12783	organization	3.00	3.00	2.00
3	12783	content	3.50	2.00	3.00
4	12783	argumentation	2.50	2.50	2.00
5	12783	sources	3.75	3.00	2.00
6	62999	thesis	3.50	3.00	3.00

Table 3: Results of human evaluation for all 25 premium students - aggregated by grading type. Every row represents one exam and one criterion. The mean criterion score per grading type can be compared via the columns peer_scored, cta_scored and prof_scored.

As a result, the fourth dataset (snapshot in Table 4) shows an overall peer, CTA and professor's grading, compared to the multiple-choice performance - again only for premium students. This dataset is the smallest with 25 observations.

4. RESULTS

4.1 Automated Validation

Using dataset 1 we identify the distribution of multiple-choice results comparing premium and non-premium students (Figure 1).

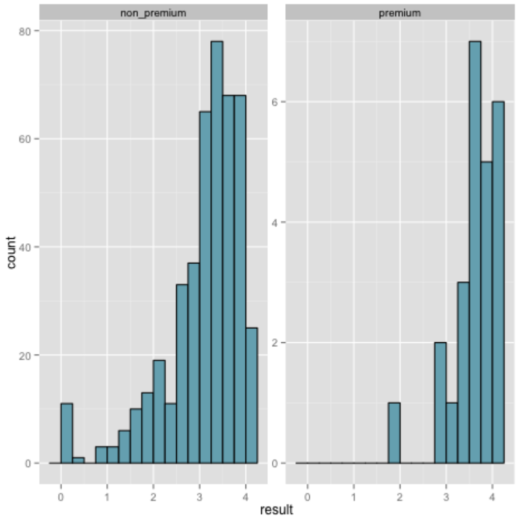


Figure 1: Result histogram from automated validation, comparing the performance of premium and non-premium students in multiple-choice questions. The x-axis represents the final multiple-choice result, originally ranging from 0 to 16, divided by four.

Both distributions are right skewed and show a high mean (premium students: 2.97, non-premium students: 3.49), which indicates a general high performance of both premium and non-premium students in the multiple-choice questions. The premium students achieved the highest results relatively more frequently than the non-premium students. However, we have to remark that there are significantly fewer premium students than non-premium students included in this dataset.

4.2 Grading distributions

Using dataset 2 we identify the distribution of scores given to premium students by peers, CTAs and the instructor (Figure 2). Note that this dataset contains a score for every student in each of the five criteria, each given by one instructor, two CTAs and multiple peers. As a result of the different sizes of the three grading groups, the scope of the three distributions is different.

The distribution of the CTA scores and instructor's scores both show a normal shape with three as the most frequently awarded and four as an unpopular score. The peers also preferred giving the score three, but with a much higher frequency than CTAs and the instructor. As another difference for peer grading, the second most common score given by peers was the score four. This heavily skews the peers' scoring distribution towards the scores three and four.
Using dataset 4 we are able to include the multiple-choice results into a density plot (Figure 3, Appendix A). The drawback of this representation is that only 25 observations are contained. However, we observe that the automatically evaluated multiple choice test shows the highest overall results among the premium students.

4.3 Criteria

With dataset 2 we compare the scoring distribution from Section 4.2 across the five criteria (Figure 4).

This comparison shows that the grading distributions observed for peers, CTAS and instructors in Section 4.2 apply to all five criteria and not only to the accumulated grades. As an exception the instructor deviates in criterion *thesis* by rarely giving the central score two. Furthermore, the peers

	gradee	peer_graded	cta_graded	prof_graded	multiple_choice
1	12783	3.15	2.60	2.40	3.50
2	62999	3.50	3.10	2.40	3.25
3	68424	3.50	3.00	1.80	3.50
4	192994	3.20	2.10	0.80	3.50
5	193615	2.57	1.50	0.40	3.25
6	239590	2.72	2.00	1.20	2.75

Table 4: Results of human and automated evaluation for all 25 premium students - aggregated by grader type and criterion. Every row represents one exam. The mean grade per grading type can be compared via the column peer_graded, prof_graded, cta_graded and multiple_choice.

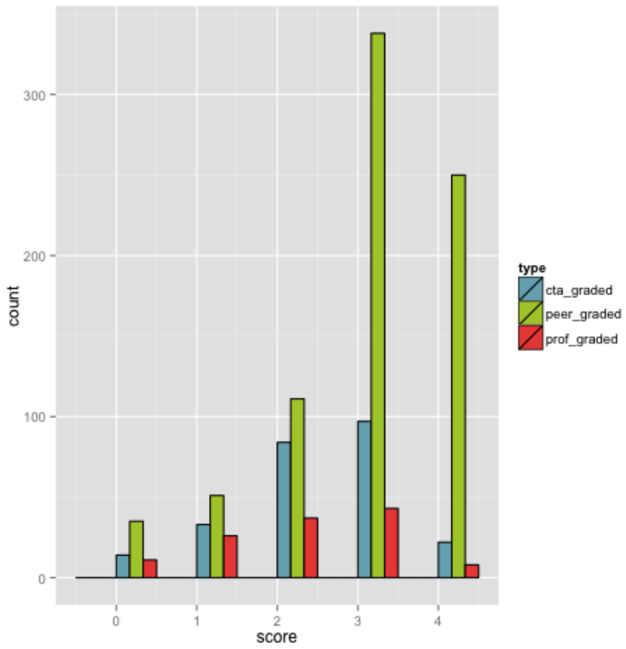


Figure 2: Scoring distributions for all 25 premium students - not aggregated by criteria, i.e every premium student is represented with 5 scores, one for each criterion. Each color represents a different grading type.

slightly deviate in criterion *sources*, by preferring to give the score four over giving score three.
Comparing the means of the three distributions for each of the five criteria (Table 5) reveals another insight: Overall, the criterion *organization* was assigned the best relative score by CTAs, but the worst relative score by the instructor.

4.4 Validity

Table 6 is a correlation matrix of all given scores by peers (peer), CTAs (cta) and the instructor (prof), extracted from dataset 3. For one exam and per criterion, peer and CTA scores are respectively aggregated by taking the mean. This means that potential reliability effects among peers or CTAs are ignored in order to establish comparison to instructor scoring. Table 7 is a correlation matrix of all given grades (aggregated by criteria). It is extracted from dataset 4 and ignores criteria effects in order to compare grading to automated validation (auto). The two correlation tables are our main instrument to judge the validity of peer grading and CTAS. Analyzing the correlations on both levels of aggregation has the advantage that we do not miss potential validity effects that cancel out when aggregating by criteria. However, since we do not observe big unexpected jumps in correlation from Table 6 to Table 7, we do not include a specific validity analysis for every single criterion.

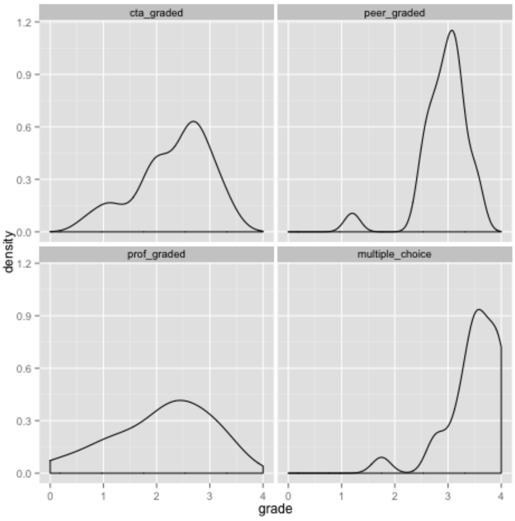


Figure 3: Grade distribution densities for all 25 premium students - aggregated by criteria. Each plot shows the grading density for one grading type: Peer grading, CTA grading, professor grading as well as automated validation (multiple choice).

First of all, the correlation between peer grading and automated validation is only almost zero (0.05). Since this result is very different for the correlation of automated validation with CTA grading and instructor grading, we cannot make a general conclusion on the relationship between student's knowledge and their performance on essay questions.

	peer	cta	prof
peer	1.00	0.66	0.35
cta	0.66	1.00	0.35
prof	0.35	0.35	1.00

Table 6: Correlation matrix of the columns in Dataset 3. Mean scores (aggregated by grading type) for all 25 premium students in each of the five criteria are included.

	criterion	peer_scored_mean	cta_scored_mean	prof_scored_mean
1	thesis	3.00	2.26	2.28
2	organization	2.87	2.42	1.80
3	content	2.99	2.34	2.32
4	argumentation	2.83	2.34	1.96
5	sources	2.95	2.24	2.08

Table 5: Average score given to the 25 premium students. Comparing the five scoring criteria across peers, CTAs and instructor.

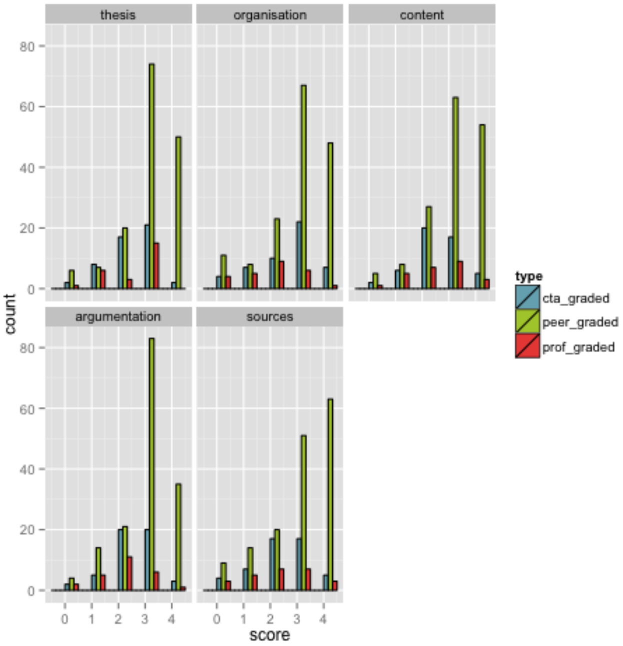


Figure 4: Scoring distributions for all 25 premium students - faceted by criteria. Each color represents a different grading type.

We notice that CTAS grading has the strongest correlation to multiple-choice performance (0.53). But we remark that this does not justify judgement on the validity of peer grading, CTAS grading or instructor grading.
Next, the instructor's grading only shows a weak correlation to the peers' grading and a similarly weak correlation to CTAs' (Table 7: 0.36, 0.39). Assuming that the instructor is an accurate grader, this finding entails that neither peers nor CTAs provide valid grading. Since the correlation coefficients only summarize the relation between two grading entities, a scatterplot is provided for CTA grading against professor grading (Figure 5), derived from dataset 6.

In contrast to this, peer and CTA grading strongly correlate (Table 7: 0.76). That means that assuming CTAs provide accurate grading, peer grading turns out to be much more valid than under the assumption that the instructor provides accurate grading.

	peer	cta	prof	auto
peer	1.00	0.76	0.36	0.05
cta	0.76	1.00	0.39	0.53
prof	0.36	0.39	1.00	0.25
auto	0.05	0.53	0.25	1.00

Table 7: Correlation matrix of the columns in Dataset 4 . Mean grades (aggregated by critera and grading type) for all 25 premium students are included.

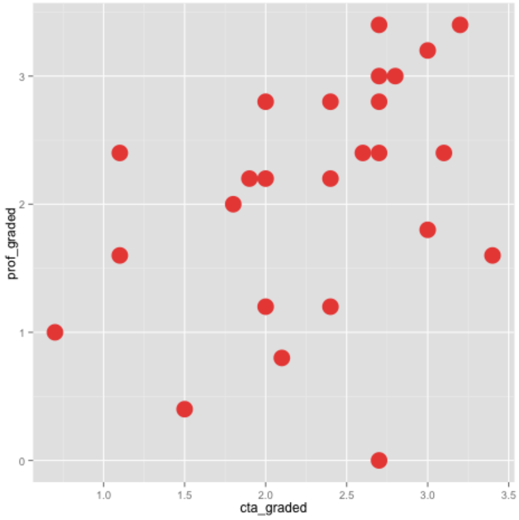


Figure 5: Scatterplot of CTA grades against professor grades - aggregated by criteria. Each dot represents the mean CTA grade compared to the instructor grade.

5. DISCUSSION

Our analysis allowed us to gather data on automated validation, peer grading and instructor grading as well as CTAS grading as a new hybrid element. The experiment was designed in a way that the four types of grading could be compared on the same exam and the three latter grading types on the same essay question within that exam. The number of exam participants was high (467), whereas the number of exams to which all types of grading were applied was rather low (25 premium students). The latter is actually a reasonable size for a classical classroom, and hardly suited to produce robust statistical results. However, peers and CTAs were acquired and recruited in the context of a massive open online course, which still makes our findings unique, compared to similar investigations in other settings.

Our grading distribution analysis reveals the skew of the distribution as a difference between peer grading and automated validation (with skew), as opposed to CTAS grading and instructor grading (without skew). The overall high multiple-choice performance is a result of the student's knowledge in combination with the test design. Making the questions more difficult or replacing the 10 familiar questions with unknown ones would probably reduce the skew in the distribution of automated validation. More interestingly, the skew of peer scoring cannot be explained by student's abilities or the test design, since the CTAs and the instructor did not produce this skew when scoring the same students on the same essay. Hence, we conclude that peer grading was biased in our setup. The precise reasons for this bias remains an open question.

Concerning scoring criteria, we found that instructor and CTAs do not agree on *organization*. A possible explanation for the professor's relatively harsh scoring of this criterion is the fact that students of the online course did not learn essay writing methodologies within the online course. This remains for further investigation with larger sample sizes, together with the question of why CTAs assign the highest scores to the criterion *argumentation*.

The most ambiguous finding from this research are the validity results. Assuming the professor to be an accurate grader, peer grading and CTAS grading were invalid (weak correlation). Ignoring the instructor and assuming that pro CTAs grade accurately, we conclude that peer grading was valid (strong correlation) in this setting. We were expecting to observe weak correlation between peer grading and instructor grading in our setting without self grading and calibration. But between CTAs and instructor we expected a strong correlation, since CTAs also hold an academic degree in the particular subject area and work as teaching assistants at their universities. It remains for further research on a larger set of courses, whether a strong correlation between CTA grading and instructor grading can be observed after modifications to the CTAS.

In order to summarize our objectives, we list new hypotheses that remain for further investigation on a larger sample size of online courses, as well as of participants.

1. Under similar experiment conditions in other MOOCs, CTAs and the instructor will most strongly disagree on the criterion *argumentation*.

2. An advanced CTA selection process, including test submissions and calibration, will increase the validity of the CTAS grading, as compared with instructor grading.

3. Under similar experiment conditions in other MOOCs, i.e. especially without calibration, peer grading will be biased as opposed to CTAS and instructor grading. Further, it will only be weakly correlated with instructor grading.

4. Implementation of a calibration process within the peer grading feature will reduce the peers' bias.

5. A calibration process combined with method and evaluation training within the online course will increase the validity of peer grading in our setting and enable the peers as a collective to outperform CTAs as measured in validity.

Testing the above hypotheses on a larger sample of MOOCs will help to understand the possibilities and limits of peer, CTA and instructor grading as separate elements. It remains a future challenge to design an accurate calibration, grading and statistical evaluation process for MOOCs, in which all three elements are combined in a scalable way.

6. ACKNOWLEDGMENTS

We thank Prof. Anja Mihr for collaborating with us and hence enabling us to conduct this research. Further, we thank Prof. Niels Pinkwart, Zhilin Zheng, Jo Corrall, André Rieck, René Kizilcec, Johannes Diermeier and Janusz Ginster for valuable feedback and contributions as well as Hannes Klöpper for having the initial idea to implement the CTAS.

7. REFERENCES

[1] K. Ashley and I. Goldin. Toward ai-enhanced computer-supported peer review in legal education. *Legal Knowledge and Information Systems - JURIX 2011: The Twenty-Fourth Annual Conference*, December 2011.

[2] S. P. Balfour. Assessing writing in moocs: Automated essay scoring and calibrated peer review. *Research and Practice in Assessment*, 8(1):40–48, June 2013.

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

[4] K. Cho and C. Schunn. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education*, 48(3):409–426, April 2007.

[5] K. Cho, C. Schunn, and R. Wilson. Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, 98(4):891–901, November 2006.

[6] W. Dancer and J. Dancer. Peer rating in higher education. *Journal of Education for Business*, 67(5):306–309, July 1992.

[7] N. Falchikov and J. Goldfinch. Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. review of educational research. *Review of Educational Research*, 70(3):287–322, September 2000.

[8] E. F. Gehringer. Electronic peer review and peer grading in computer-science courses. In *Proceedings of the 32nd SIGCSE Technical Symposium on Computer Science Education*, pages 139–143, Charlotte, North Carolina, USA, February 2001.

[9] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an informationtapestry. *Communications of the ACM*, 35(12):61–70, December 1992.

[10] I. Goldin and K. D. Ashley. Peering inside peer review with bayesian models. In *Proceedings of the 15th international conference on artificial intelligence in education*, pages 90–97, Auckland, New Zealand, June 2011.

[11] P. J. Hinds. The curse of expertise: The effects of expertise and debiasing methods on predictions of novice performance. *Journal of Experimental Psychology: Applied*, 5(2):205–221, June 1999.

[12] C. Kulkarni and S. Klemmer. Learning design wisdom by augmenting physical studio critique with online

self-assessments. *Stanford University Technical Report*, July 2010.

[13] C. Kulkarni, K. P. Wei, H. Le, D. Chia, K.Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction*, 20(6):33:1–33:31, December 2013.

[14] F. Loll and N. Pinkwart. Using collaborative filtering algorithms as elearning tools. In *Proceedings of the 42nd Hawaii International Conference on System Sciences*, Waikoloa, Big Island, Hawaii, USA, January 2009.

[15] H. Luo and A. C. Robinson. Peer grading in a mooc: Reliability, validity, and perceived effects. *Journal of Asynchronous Learning Networks*, 18(2):1–14, July 2014.

[16] B. Matthews. Assessing individual contributions: Experience of peer evaluation in major group projects. *British Journal of Educational Technology*, 25(1):19–28, January 1994.

[17] P. Mitros and F. Sun. Creating educational resources at scale. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 16–18, Athens, Greece, July 2014.

[18] P. F. Mitros and V. Paruchuri. An integrated framework for the grading of freeform responses. In *Proceedings of the Sixth International Conference of MIT Learning International Networks Consortium*, Cambridge, Massachusetts, USA, June 2013.

[19] L. Pappano. The year of the mooc. In *New York Times*, November 2012. Retreived from: `http://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html` [October 2014].

[20] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in moocs. 2013. Retrieved from: `http://arxiv.org/pdf/1307.2579.pdf` [October 2014].

[21] A. A. Russell. Calibrated peer review - a writing and critical-thinking instructional tool. In *Teaching Tips: Innovations in Undergraduate Science Instruction*, pages 67–71. NSTA Press, 2004.

[22] P. M. Sadler and E. Good. The impact of self-and peer-grading on student learning. *Educational Assessment*, 11(1):1–31, June 2006.

[23] J. Surowiecki. *The wisdom of the crowds*. Anchor Books, New York, USA, 2004.

[24] K. Topping. Peer assessment between students in college and universities. *Review of Educational Research*, 68(3):249–276, September 1998.

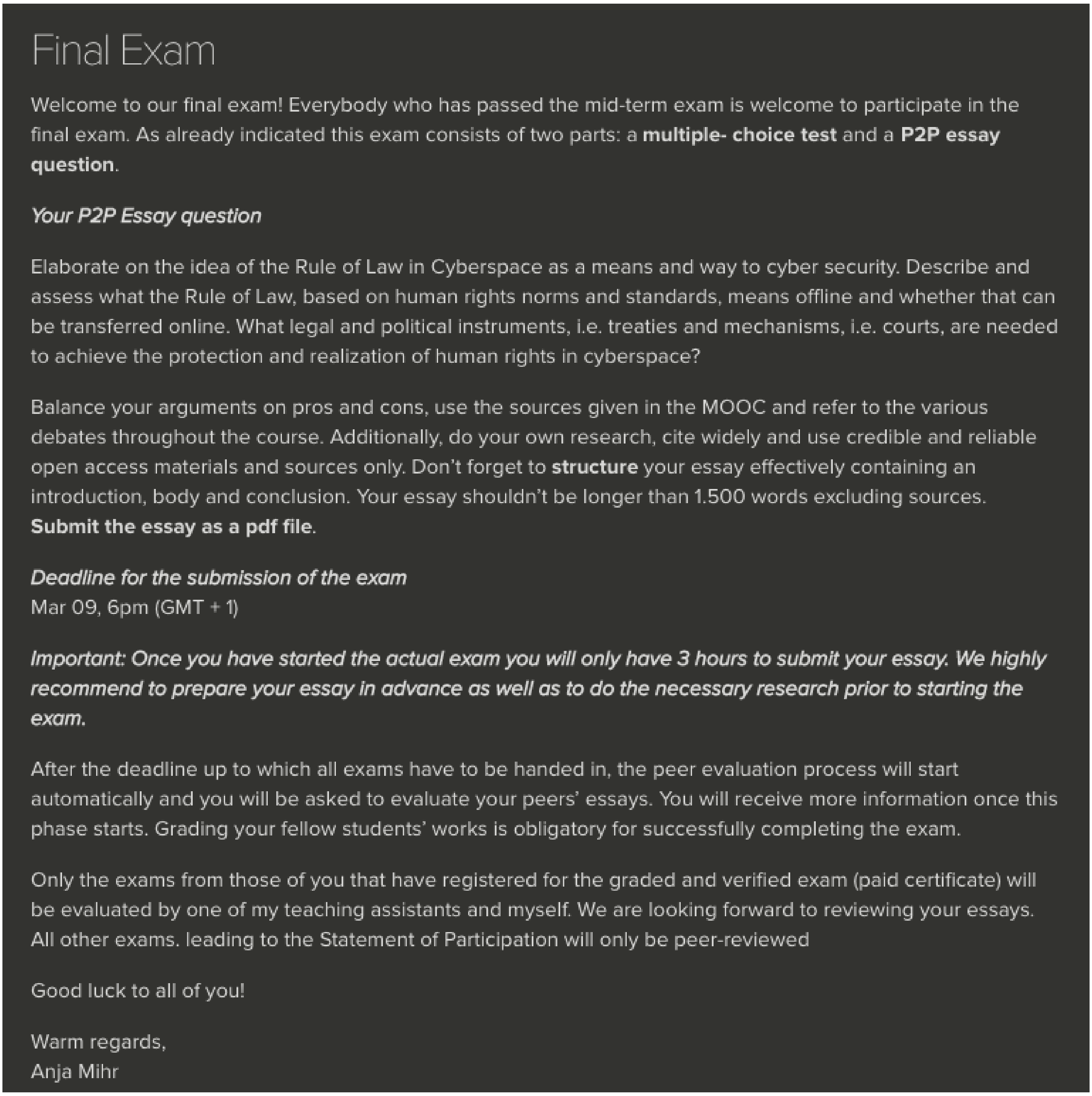
APPENDIX

A. GRADING RUBRIC

B. EXAM TEXT

	0	1	2	3	4
Thesis/Central Idea	No thesis, no main point.	Vague thesis with no idea development	Thesis exists, but unelaborated idea development.	Good thesis that is clearly stated and main idea is developed throughout the essay.	Strong/original thesis that is clearly stated and the central idea is exceptionally well-developed throughout the essay.
Organization	Essay is fewer than 3 pages or 3-5 pages but lacking an organised structure; there is no order to the essay and no paragraphs.	Essay is 3-5 pages with slightly organised structure, but missing part or all of the introduction, body, conclusion.	Essay is 3-5 pages with an introduction, body and conclusion. Paragraphs are defined, but not in a logical order, transitions are missing.	Essay is 3-5 pages with an introduction, body and conclusion. Paragraphs are well-defined, transitions exist.	Essay is 3-5 pages with an excellent organised structure. Clearly defined introduction, body and conclusion. Clear and effective transitions from one idea to the next.
Content	Does not meet assignment requirements.	Relates to the thesis, but has too little or irrelevant, unimportant, inaccurate details.	Relates to the thesis and is relevant, important and accurate, but not thoroughly developed - too broad or narrow.	Relates to and supports thesis with relevant, important and accurate details thoroughly.	Relates to and supports thesis; detail goes above and beyond the simply adequate; unique and interesting ideas; complex understanding and exploration of the topic.
Balanced Argumentation	Does not follow or support the main idea of the essay.	Poor argumentation, rather polemic.	Provides good argumentation but is only one-sided.	Well-balanced argumentation that supports the main idea of the essay and provides specific details.	Excellent balance of arguments that support the thesis in an effective way; arguments go above and beyond what was mentioned in the course.
Sources	No sources have been used, documented sources are not in standard style or many citations are inadequate.	Sources do not support essay structure and thesis; citations contain major errors or not enough sources have been used.	Sources support essay structure and thesis, but occasionally take over the thesis or citations contain some errors.	Sources support thesis and do not take over the thesis; citations are sufficient and follow general standards.	Sources supports thesis exceptionally well; citations are sufficient and follow general standards.

Table 8: Grading rubrics of the final examination of the course ”Public Privacy: Cyber Security and Human Rights”, held by Prof. Anja Mihr 2013/2014 on the iversity.org platform. Image courtesy Anja Mihr.



Final Exam

Welcome to our final exam! Everybody who has passed the mid-term exam is welcome to participate in the final exam. As already indicated this exam consists of two parts: a **multiple- choice test** and a **P2P essay question**.

Your P2P Essay question

Elaborate on the idea of the Rule of Law in Cyberspace as a means and way to cyber security. Describe and assess what the Rule of Law, based on human rights norms and standards, means offline and whether that can be transferred online. What legal and political instruments, i.e. treaties and mechanisms, i.e. courts, are needed to achieve the protection and realization of human rights in cyberspace?

Balance your arguments on pros and cons, use the sources given in the MOOC and refer to the various debates throughout the course. Additionally, do your own research, cite widely and use credible and reliable open access materials and sources only. Don't forget to **structure** your essay effectively containing an introduction, body and conclusion. Your essay shouldn't be longer than 1.500 words excluding sources. **Submit the essay as a pdf file**.

Deadline for the submission of the exam
Mar 09, 6pm (GMT + 1)

Important: Once you have started the actual exam you will only have 3 hours to submit your essay. We highly recommend to prepare your essay in advance as well as to do the necessary research prior to starting the exam.

After the deadline up to which all exams have to be handed in, the peer evaluation process will start automatically and you will be asked to evaluate your peers' essays. You will receive more information once this phase starts. Grading your fellow students' works is obligatory for successfully completing the exam.

Only the exams from those of you that have registered for the graded and verified exam (paid certificate) will be evaluated by one of my teaching assistants and myself. We are looking forward to reviewing your essays. All other exams. leading to the Statement of Participation will only be peer-reviewed

Good luck to all of you!

Warm regards,
Anja Mihr

Figure 6: Final exam of the course "Public Privacy: Cyber Security and Human Rights", held by Prof. Anja Mihr 2013/2014 on the iversity.org platform. Screenshot taken from iversity.org. Image courtesy Anja Mihr.