Peer Assessment Based on Ratings in a Social Media Course

Andrii Vozniuk
EPFL
Lausanne, Switzerland
andrii.vozniuk@epfl.ch

Adrian Holzer
EPFL
Lausanne, Switzerland
adrian.holzer@epfl.ch

Denis Gillet
EPFL
Lausanne, Switzerland
denis.gillet@epfl.ch

ABSTRACT

Peer assessment is seen as a powerful supporting tool to achieve scalability in the evaluation of complex assignments in large courses, possibly virtual ones, as in the context of massive open online courses (MOOCs). However, the adoption of peer assessment is slow due in part to the lack of ready-to-use systems. Furthermore, the validity of peer assessment is still under discussion. In this paper, in order to tackle some of these issues, we present as a proof-of-concept of a novel extension of GRAASP, a social media platform, to setup a peer assessment activity. We then report a case study of peer assessment using GRAASP in a Social Media course with 60 master's level university students and analyze the level of agreement between students and instructors in the evaluation of short individual reports. Finally, to see if both instructor and student evaluations were based on appearance of project reports rather than on content, we conducted a study with 40 kids who rated reports solely on their look. Our results convey the fact that unlike the kid evaluation, which shows a low level of agreement with instructors, student assessment is reliable since the level of agreement between instructors and students was high.

Categories and Subject Descriptors

K.3.1 [**Computers and Education**]: Computer Uses in Education; L.0.0 [**Assessment / Evaluation / Measurement**]: Assessment and Evaluation

General Terms

Design, Experimentation

Keywords

Learning activities, social media, peer assessment, ratings, MOOCs

1. INTRODUCTION

Peer assessment can be defined as *an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal- status learners* [23]. Peer assessment has been used as an evaluation tool in courses for several centuries [23] and it has recently gained more attention in the context of massive open online courses (MOOCs) [6], where it can serve as a powerful tool to achieve scalability in evaluation of complex assignments with large student cohorts.

Peer assessment can also have benefits related to the quality of learning. Several studies have shown that students can additionally learn when evaluating work of other students [22], but these findings are not universal [19]. It can also be argued that peer assessment reduces the workload for instructors, saving time in the process. However, depending on the setup and implementation time this might not be the case. Also the fact that teachers have to formalize the assessment process can benefit both students and teachers as they have to clarify the educational objectives, goals and the grading scale [23]. To be used as a replacement for instructor grading, the grades based on peer assessment should match to the grades of instructors. Most studies evaluate the accuracy of peer assessment as adequate compared to instructor evaluation, however this comparison is only useful if one assumes that instructor assessments are valid, which he suggests is *a doubtful assumption in some contexts* [23].

Even though peer review and assessment have advantages, they have not become a mainstream evaluation feature in universities. Pearce et al. [16] argue that this is in part because of difficulty of usage and price of dedicated peer review and assessment systems. They point to online tools as promising solutions to this problem but they deplore the fact that so few of them exist.

In this paper, in order to encourage the adoption of peer assessment, we present a novel extension of a social media platform (GRAASP), which can be used to setup a peer assessment activity. Then we report a case study of peer assessment using GRAASP in a Social Media course taught at the Swiss Federal Institute of Technology in Lausanne (EPFL) with 60 master students and analyze the level of agreement between the assessment of students and instructors in the evaluation of short individual reports. Our findings suggest that the level of agreement is significant. Finally, to see whether the assessment of students and instructors was solely based on appearance and could be replicated without reading the content of the reports, we conducted a boundary case experiment with 40 twelve-year old kids who assessed the reports solely based on their look. Our results show that there is no agreement between kids and students and thus suggest that students and instructors are, at least to some extent, influenced by the content of the report.

It should be highlighted that in the context of this social media course, the objective of the peer assessment is twofold. First, it aims at assessing the work of the students. Second, it aims at introducing students to the challenges and the opportunities related to ratings in social media platforms.

LAK '14, March 24 - 28 2014, Indianapolis, IN, USA
Copyright 2014 ACM 978-1-4503-2664-3/14/03 $15.00.

This paper is structured as follows. Section 2 discusses related work. Section 3 presents how GRAASP was used for peer assessment in a Social Media course. Section 4 discusses the evaluation methodology before Section 5 presents the results. Finally Section 6 wraps up with a conclusion and discussion of future work.

2. RELATED WORK

Hereafter we make a brief overview of previous findings on peer assessment and present existing peer assessment systems.

2.1 Peer assessment findings

Peer assessment is generally found to be formative [16], but students may dislike this work as they feel that this is the "teacher's responsibility" [1]. Cho et al. [3] found that if students were guided and were provided scaffolds they took their job seriously and their results where as valid as instructor ratings. This even though students themselves do not believe their ratings are reliable.

However, these positive results are not universal [24]. For example, Sadler and Good [19] find that even though student and instructor assessment are strongly correlated, they do not appear to be linked to improved understanding and higher performing students may suffer when graded by others.

Furthermore, both evaluations by students and instructors can be prone to biases. For students, these biases include inexperience in the field, inexperience in grading, and basic biases, such as rating friends favorably [7], as well as more advanced biases, such as making pacts with others [13]. Instructors can be affected by similar biases. For example the number of papers to grade can lead to rushed evaluation, individual grading (one instructor) is less reliable than combined ratings [18], instructors can be influenced by expectations and other biases, finally as the knowledge level of the instructor is much higher than that of students, it might be difficult for her to distinguish between small scale differences between students.

In the context of MOOCs, Piech et al. [17] address the problem of peer assessment accuracy and suggest that student assessment validity can be significantly improved with a corrective algorithm. They also proposed to use the algorithm for more intelligent distribution of assignments between graders.

2.2 Peer assessment systems

As mentioned above, there are not many available online peer assessment systems [16]. An early, but now extinct, web-based peer grading system, was called PG (Peer Grader) and had already been proposed in 2001 [9].

More recently, researchers at the university of Melbourne introduced PRAZE [15],a dedicated system for peer assessment. It has been successfully evaluated in several courses. Unfortunately, it is currently only available for Australian Universities.

TURNITIN[1] is a plagiarism detection tool, which also provides a module to allow student to comment each other's work. Turnitin provides both an online interface and a mobile app.

Some learning management systems (LMS) provide modules for peer assessment. The self peer assessment (SPA) module in Blackboard is such an example.[2] It is quite rigid and does not allow for late hand ins or proxy submission. WORKSHOP in Moodle[3] is also a sophisticated tool that has been tested in large classrooms [14]. As these modules are strongly coupled to their respective LMS it makes it their adoption difficult beyond the LMS users.

3. CASE STUDY: SOCIAL MEDIA

We performed our evaluation in the Spring Semester 2013 during the Social Media course at EPFL with 60 master students in Computer Science. During this course each student had to hand in a two-page project report to show a long-tail effect in a social media platform chosen by him or her. In a typical example of a project a student took IMDB[4] as social media and inquired whether *"there will be a small collection of very well popular movies [...] which have plenty reviews, against a very large collection of movies with a very low number of reviews (which will represent our long tail)."* In order to fully immerse students into the topic, we decide to use social media platforms to deposit course material (GRAASP), share useful links (TWITTER), and host slides (SLIDESHARE). Furthermore, we used peer assessment for the grading of the student project as social media make extensive use of peer review systems.

Proper instructions are a key success factor for peer evaluation. We made clear to the students at the beginning of the course that their report will be evaluated by peers, and that they will have to evaluated the work of about 20 peers. As the reports were discussing the long tail effect in various social media platforms, we also made clear that their reading was a learning activity. Finally, it was announced that the peer evaluation done by them would only be taken into account if not deviating significantly from the evaluation carried out by the experts to ensure fairness. Since there is a lack of peer assessment tools, we devised a novel extension of GRAASP to accommodate peer assessment.

Further activities including teamwork and live presentations in the classroom were organized but not discussed in this paper.

3.1 Graasp

GRAASP is a flexible social media platform that was initially developed to support communities of practice and extended to support collaborative learning activities [2, 11, 12] as well as online science labs for inquiry-based learning at school [10]. GRAASP can be freely exploited by schools, universities and non-profit organizations for learning activities and knowledge management. In GRAASP, people organize their personal and shared projects, interests and activities into public or private contextual spaces, where they share relevant resources and necessary apps with invited members. For example, for the Social Media course, the instructor created a dedicated space, named Social Media, where students would find lecture material, such as slides or videos and administrative information.

The design of GRAASP follows a flexible bottom-up permissions management approach when it comes to joint projects. Instead of having a top-level administrator in control of all project spaces, everything is managed at the space level. Thereby in the case of the Social Media course, students were able to create a space for their group project, invite their group members, close the space for others and drop documents together with other resources needed for the project.

3.2 Peer assessment using Graasp

For the course we have developed an extension of GRAASP as a proof-of-concept where the reviewer tasks are automated. We are currently working on further improvements, which will allow creating dedicated peer assessment spaces, where a summary of all peer assessment tasks will be visible. We plan to roll out this functionality in the coming months.

To setup the peer assessment activity, students were instructed to create a space in GRAASP, drop their work in it, invite the instruc-

[1] turnitin.com

[2] www.niu.edu/blackboard/assess/spa.shtml

[3] docs.moodle.org/23/en/Workshop_module

[4] http://www.imdb.com

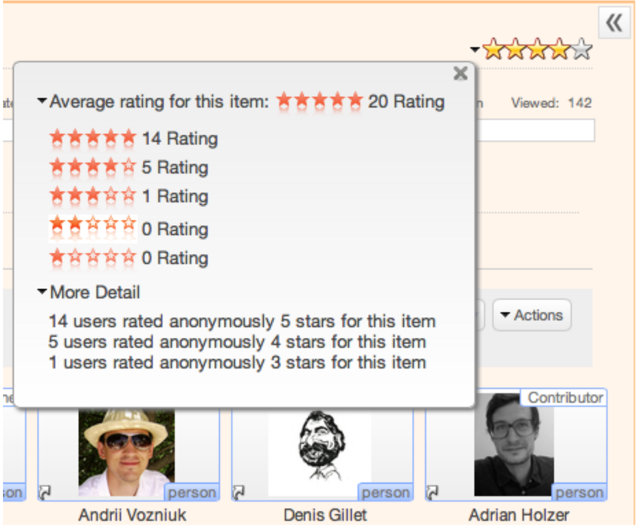
tor, and leave the space in order to be not able to see the reviewers. Then the instructor invited randomly assigned peers into the space to perform the evaluation. All participants of the evaluation were instructed to use a 5-point Likert-type scale for grading the reports. Since the course was about social media, the grade scale was inspired by the ratings system often found in social media (e.g. it is used for product review on Amazon.com):

- 5 stars: I love the report
- 4 stars: I like the report
- 3 stars: The report is OK
- 2 stars: I don't like the report
- 1 star: I hate the report

Peers were instructed to focus on the following list of criteria when evaluating the reports:

- Is the report about a social media?
- Is the long tail hypothesis clear?
- Is the collected dataset representative?
- Is the report technically advanced?
- Is the report well written?
- Is it interesting and creative?

We did not collect the grades regarding each of the criteria independently, but recorded one final grade per report. Once the reviewers completed their task, the instructor was able to see the average result of the peer evaluation as well as the detailed grades distribution as depicted in Figure 1.


Average rating for this item: ★★★★★ 20 Rating

Viewed: 142

★★★★★ 14 Rating

★★★★☆ 5 Rating

★★★☆☆ 1 Rating

★★☆☆☆ 0 Rating

★☆☆☆☆ 0 Rating

More Detail

14 users rated anonymously 5 stars for this item

5 users rated anonymously 4 stars for this item

1 users rated anonymously 3 stars for this item

Figure 1: Rating details for a report in Graasp

3.3 Reviewers

The peer assessment activity was performed by three groups of reviewers: instructors, students and kids. The group of instructors consisted of the two lecturers and the two teaching assistants. Each of the instructors graded all the reports. The group of students consisted of all 60 students in the Social Media course (age 23-28). Each of the students was instructed to grade 20 randomly assigned reports. The group of kids consisted of 40 schoolchildren (age 10-12), each kid graded 15 randomly assigned reports. This assignment procedure ended in each report being assigned to 4 instructors, 20 students and 10 kids.

Note that the idea behind assigning a peer assessment task to kids was to see whether they could predict the grades of instructors and students by the appearance alone without reading them and understanding the content. To perform this boundary case experiment, the kids did not use GRAASP to rate the two-page reports, instead they were each given 15 small 5.83 x 4.13 inch (14.8 x 10.5 cm) cards with the reports printed on them as shown in Figure 2. Most of the text of in the reports was too small to be read by the kids and most of them did not understand English.

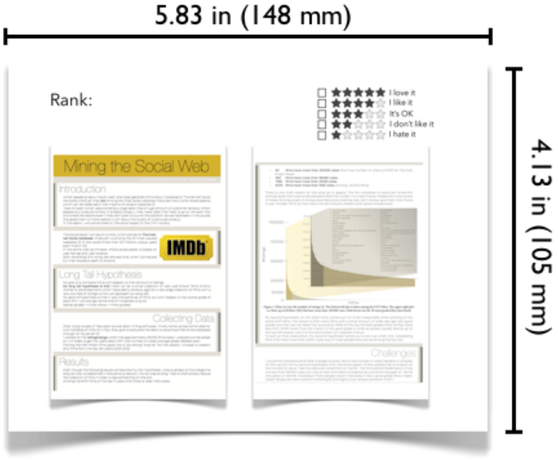


Figure 2: Example of a report card for kids

4. EVALUATION METHODOLOGY

To evaluate the validity of the peer assessment we analyzed the level of agreement in grades between instructors and students.

Measuring group consensus. In the analysis we focused on the level of agreement between groups. Unfortunately, reliable intergroup agreement measures are still lacking [25]. Hence, a common practice is to first define a consensus in each group of raters and in this way to reduce the problem of measuring inter-group agreement to the problem of measuring inter-rater agreement. The consensus is often defined as the median or the mean for ordinal scales or as the mode (modal category) for nominal scales [25]. Since in our study we employed five-star grade ordinal scale, we selected the mean as consensus value. In the context of the Social Media course, this consensus value also had a pedagogical sense as it could be seen as an instance of the *wisdom of the crowd*, a frequently used concept in social media (e.g. Wikipedia).

Measuring correlation between group consensuses. In order to get a first visual idea of the agreement between groups, consensus values can be represented on scatter plots together with regression lines with 90% confidence intervals for the regression slope.

Then to get a numerical figure of the dependency of the consensus values, we computed the Pearson correlation and the Spearman rank correlation [21] coefficients. In general correlation does not show the level of agreement, it just shows the level of association between variables. It is possible that grades of two raters are strongly correlated (Pearson or Spearman), but at the same time they can have little agreement. Thus correlation is not suitable for measuring the level of inter-rater agreement. Based on the literature, Cohen's kappa [5] is the adequate tool for our purpose as it is frequently used in behavioural sciences to compute inter-rater agreement.

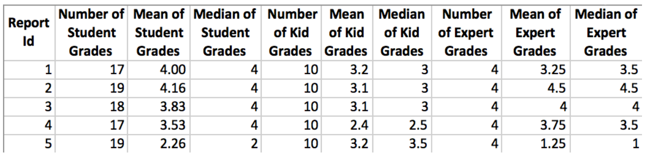
Measuring agreement between group consensuses.

The original version of Cohen's kappa targets nominal variables and assumes that all disagreements are equal. This is not the case in our study, where disagreement can differ in severity. Therefore we use Cohen's weighted kappa [4], a statistic suitable for measuring the level of inter-rater agreement on ordinal scale. Note that as a measure of reliability, the weighted kappa is equal [8] to the intraclass correlation coefficient (ICC), a common measure of reliability of either different raters, or different items on a scale [20].

5. EVALUATION RESULTS

We followed the evaluation methodology presented in the previous section to measure level of agreement between two groups of reviewers: (1) instructors and students, as well as, (2) kids and students.

A sample of five reports from the dataset is presented in Figure 3 and shows the per-report consensus values. Note that the discrepancy between the number of assigned students per report (20) and the actual number of grades is due to the fact that some students did not complete the evaluation assignment. In the end, more than 92% evaluation assignments were completed. We are currently preparing the dataset used in this study and plan to publish it online together with the analysis so it can be repeated.



Report Id	Number of Student Grades	Mean of Student Grades	Median of Student Grades	Number of Kid Grades	Mean of Kid Grades	Median of Kid Grades	Number of Expert Grades	Mean of Expert Grades	Median of Expert Grades
1	17	4.00	4	10	3.2	3	4	3.25	3.5
2	19	4.16	4	10	3.1	3	4	4.5	4.5
3	18	3.83	4	10	3.1	3	4	4	4
4	17	3.53	4	10	2.4	2.5	4	3.75	3.5
5	19	2.26	2	10	3.2	3.5	4	1.25	1

Figure 3: data excerpt representing consensus grades

5.1 Students vs instructors

Hereafter, we compare the student and the instructor assessment in terms of correlation and agreement.

Correlation. The visual representation of the data on the scatter plot in Figure 4 indicates that grades of students for reports increase with the grades of instructors.

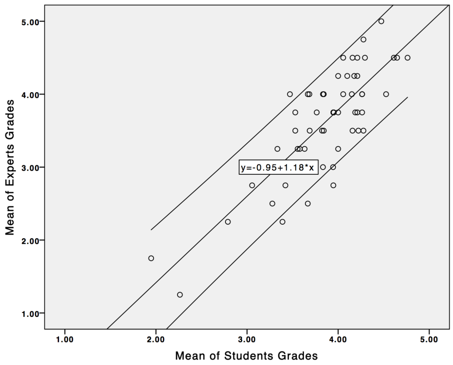
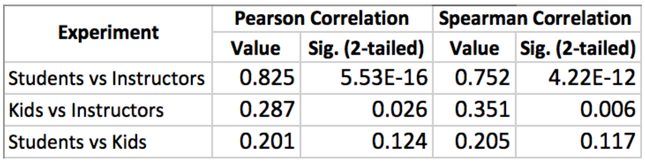


Figure 4: A scatter plot of instructor grades vs student grades with a linear regression line and 90% slope confidence intervals

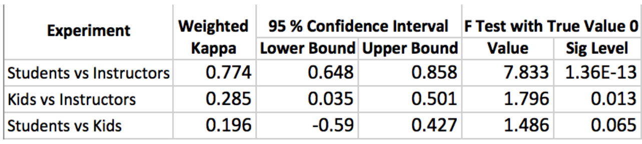
Indeed, the results of the correlation analysis presented in Figure 5 point out that there is a strong (0.825) linear dependence between grades of students and instructors with a low significance level of the null hypothesis (independence of grades), which allows to reject it. The strong positive Spearman correlation (0.752) indicates that students tend to give higher grades for better reports, i.e., reports that received higher grades from instructors. Due to the strong correlation, grades provided by students can be used to predict the grades of instructors.



Experiment	Pearson Correlation		Spearman Correlation	
	Value	Sig. (2-tailed)	Value	Sig. (2-tailed)
Students vs Instructors	0.825	5.53E-16	0.752	4.22E-12
Kids vs Instructors	0.287	0.026	0.351	0.006
Students vs Kids	0.201	0.124	0.205	0.117

Figure 5: Pearson and Spearman correlations together with significance levels of the null hypothesis

Agreement. To measure the level of agreement, we computed Cohen's weighted kappa, presented in Figure 6. When comparing students and instructors, the weighted kappa equals to 0.774, which can be interpreted as a strong agreement. The significance level (1.36E-13) indicates that the null hypothesis (that there is no agreement) can be rejected and the agreement between students and instructors is statistically significant. The 95% confidence interval is rather narrow and in the high agreement area. Since the agreement is high, we can use just one rater (in our case a consensus decision of students) to provide the grade.



Experiment	Weighted Kappa	95 % Confidence Interval		F Test with True Value 0	
		Lower Bound	Upper Bound	Value	Sig Level
Students vs Instructors	0.774	0.648	0.858	7.833	1.36E-13
Kids vs Instructors	0.285	0.035	0.501	1.796	0.013
Students vs Kids	0.196	-0.59	0.427	1.486	0.065

Figure 6: Weighted Kappa, 95% confidence intervals and significance levels of the null hypothesis

5.2 Students vs Kids

Hereafter, we compare the student and the kid assessment in terms of correlation and agreement.

Correlation. The scatter plot in Figure 7 indicates that linear dependency between the grades of students and kids is weak.

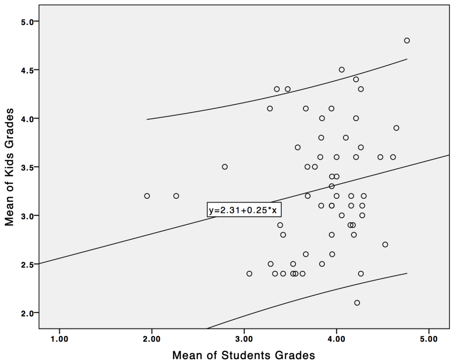


Figure 7: Scatter plot of kid grades vs student grades with a linear regression line and 90% slope confidence intervals

For instance, the graph shows that reports that received grade close to 4 from students, received anything from 2 to 4.5 from kids. The results of the correlation analysis presented in Figure 5 indicate that there is a weak (0.201) linear dependency between grades of students and kids. The significance level (0.124) indicates that the null hypothesis (independence of grades) cannot be rejected. A similar picture is observed for Spearman rank correlation. The low value of the Spearman rank correlation (0.205) indicates that the monotonic relation between grades of students and kids is weak. That is when students give a high grade for a report it is probable that kids can give low grade for the same report. So their grades are not reliable. The weak correlation can also be interpreted as an indicator that students pay attention to the content of reports and are able to evaluate it. They do not base their conclusion solely on the visual representation. In order to confirm this hypothesis additional scrutiny is required to be able to remove other factors that could have influenced the grading process.

Agreement. When comparing students and kids in Figure 6, the weighted kappa equals 0.196, which indicates no agreement or an agreement by chance. The significance level (0.065) of the null hypothesis (independence of grades) indicates that it cannot be rejected. Thus there is significant evidence that the grades of students and kids are not in a agreement.

6. CONCLUSION

Peer assessment is a timely and challenging topic. In this paper, in order to tackle the lack of existing peer assessment systems, we presented a novel extension of GRAASP, devised for that purpose. We also discussed our experience using GRAASP for peer assessment in a Social Media course.

To see whether the assessment of students and instructors was mostly based on appearance of project reports and could be replicated without reading the content of the reports, we conducted an experiment with kids who assessed the reports solely based on their look. Our analysis showed that there is a strong agreement between grades assigned by students and instructors and little agreement between grades by students and kids. These results were obtained from a well-defined evaluation scenario and a large number of reviewers for a single report (20 peers). This number will however be reduced to 15 in the next academic year to take into account students' feedback. These results encourage us to further extend the peer assessment features in GRAASP and to further evaluate the impact of peer assessment in our courses, not only as an evaluation mean, but also as a learning activity.

Acknowledgement

The authors would like to thank Alexandra Holzer for carrying out the field experiment with her pupils.

7. REFERENCES

[1] J. Biggs and C. Tang. *Teaching for quality learning at university*. McGraw-Hill International, 2011.

[2] E. Bogdanov, F. Limpens, N. Li, S. E. Helou, C. Salzmann, and D. Gillet. A social media platform in higher education. In *EDUCON'12*, pages 1–8. IEEE, 2012.

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

[4] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[5] J. Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[6] S. Cooper and M. Sahami. Reflections on stanford's moocs. *Commun. ACM*, 56(2):28–30, Feb. 2013.

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

[8] J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 1973.

[9] E. F. Gehringer. Electronic peer review and peer grading in computer-science courses. In *ACM SIGCSE Bulletin*, volume 33, pages 139–143. ACM, 2001.

[10] S. Govaerts, Y. Cao, A. Vozniuk, A. Holzer, D. G. Zutin, E. S. C. Ruiz, L. Bollen, S. Manske, N. Faltin, C. Salzmann, E. Tsourlidaki, and D. Gillet. Towards an online lab portal for inquiry-based stem learning at school. In *ICWL'13, 244–253*, pages 244–253. 2013.

[11] N. Li, S. E. Helou, and D. Gillet. Using social media for collaborative learning in higher education: a case study. In *ACHI'12*, pages 285–290, 2012.

[12] N. Li, C. Ullrich, S. E. Helou, and D. Gillet. Using social software for teamwork and collaborative project management in higher education. In *ICWL'10*, pages 161–170. Springer, 2010.

[13] B. P. Mathews. Assessing individual contributions: Experience of peer evaluation in major group projects. *British J. of Educational Technology*, 25(1):19–28, 1994.

[14] 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*, pages 1–13, 2012.

[15] R. A. Mulder and J. M. Pearce. Praze: Innovating teaching through online peer review. In *Ascilite'07*, pages 727–736, 2007.

[16] J. Pearce, R. Mulder, and C. Baik. *Involving students in peer review: Case studies and practical strategies for university teaching*. Centre for the Study of Higher Education, University of Melbourne, 2009.

[17] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in moocs. In *EDM'13, 153-160*.

[18] R. Rosenthal and R. Rosnow. Essentials of behavioural research. *McGraw-Hill*, 1991.

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

[20] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[21] C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

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

[23] K. J. Topping. Peer assessment. *Theory into Practice*, 48(1):20–27, 2009.

[24] M. van Zundert, D. Sluijsmans, and J. van Merrienboer. Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, 20(4):270 – 279, 2010.

[25] S. Vanbelle. *Agreement between raters and groups of raters*. PhD thesis, Université de Liège, 2009.