Self-evaluation in Advanced Power Searching and Mapping with Google MOOCs

Julia Wilkowski
Google
1600 Amphitheatre Parkway
Mountain View, CA 94304
Wilkowski@Google.com

Daniel M. Russell
Google
1600 Amphitheatre Parkway
Mountain View, CA 94304
DRussell@Google.com

Amit Deutsch
Google
1600 Amphitheatre Parkway
Mountain View, CA 94304
AmitDeutsch@Google.com

ABSTRACT
While there is a large amount of work on creating autograded massive open online courses (MOOCs), some kinds of complex, qualitative exam questions are still beyond the current state of the art. For MOOCs that need to deal with these kinds of questions, it is not possible for a small course staff to grade students' qualitative work. To test the efficacy of self-evaluation as a method for complex-question evaluation, students in two Google MOOCs have submitted projects and evaluated their own work. For both courses, teaching assistants graded a random sample of papers and compared their grades with self-evaluated student grades. We found that many of the submitted projects were of very high quality, and that a large majority of self-evaluated projects were accurately evaluated, scoring within just a few points of the gold standard grading.

Author Keywords
MOOCs; Google; Assessment

ACM Classification Keywords
K.3.1 Computer Uses in Education: Distance learning

INTRODUCTION
Instructors have several ways to assess how well students have learned course material: exams with either multiple choice, short-answer, or essay questions; projects; labs. Online courses can take advantage of automatic grading for multiple choice and short answer questions for instant feedback to the student. To assess more in-depth work, many MOOCs have implemented peer review/peer grading as a way for students to receive feedback on qualitative projects [10]. While progress is being made to improve automated grading systems [2], we wanted to explore how well student self-evaluation would work in the context of a MOOC as a practical method of grading complex assignments.

In the Advanced Power Searching (APS) and Mapping with Google (MWG) courses, we tested a self-evaluation process following students' completion of final projects. In both cases, the final projects were sufficiently complex and sophisticated that course developers could not (at this point in time) create an automatic grading tool.

Grading exams is a useful tool for developing metacognitive skills about a topic area [13]. Self-evaluation is an important meta-cognitive skill for students to learn [11], so this seemed like the ideal chance to test out how reliable and accurate self-evaluation would be in a MOOC, where students mostly do not meet face-to-face and the social pressures to create a plausible evaluation are not present.

Self-grading appears to result in increased student learning when compared with peer grading [12]. Self-evaluation also helps build students' metacognition that they will use when applying the skills from the class [5]. Google course developers, for example, wanted students to acquire the meta-cognitive skill of reflective design practice for mapmaking. Ideally, after taking this course, students would stop and reflect about the qualities of an effective Google Map when creating a map. This skill is taught explicitly in the class and assessed in the final project by asking students to review their work with a rubric that asks them to evaluate whether they added key map visualization features (e.g. labeling all points and providing relevant descriptions).

In a similar way, for the final project in the APS course, students wrote and submitted case studies of how they used Google tools to solve a complex research problem. In their final self-evaluation task, they reflected on how well they implemented aspects of the research process, such as assessing the credibility of a website, one of the skills addressed during the course. When they conduct research outside of the class, Google course developers intend for students to assess the credibility of websites.

In the rest of this paper, we will describe each of the two MOOCs we used in our analysis, first detailing how the

L@S 2014, March 4–5, 2014, Atlanta, Georgia, USA.
ACM 978-1-4503-2669-8/14/03.
http://dx.doi.org/10.1145/2556325.2566241

MOOC was built, its goals and general design. We then describe the final projects for each MOOC, telling how the self-evaluation process worked for each (they were very different in their details). We then turn to describing the methods we used for collecting the data, describe the data collected, followed by an analysis and discussion of the data. We conclude the paper with a summary of lessons learned.

MOOC #1: ADVANCED POWER SEARCHING (APS)

This course was designed to help members of the general public use Google tools (such as Advanced Search and Google Scholar) to solve complex research questions. The course was built using Google's open-source Course Builder platform [4] (with modifications to add a challenge-based template and a skill summary page) [1]. Registration opened on January 8, 2013; students could access the first six challenges and one final project January 23, 2013. The second set of six challenges and the remaining final project were released on January 30, 2013. A total of 38,099 people registered for the course.

The course consisted of four introductory lessons (How the Course Works, Sample Challenge, Research Process, and Solving the Sample Challenge). Following these lessons, students could select one of twelve complex search challenges. The course authors define complex as problems that require multiple steps, have more than one correct answer, or have multiple ways to achieve the answer. Figure 1 demonstrates one of the sample challenges presented in the course. Students could attempt the challenge, explore related skills, review how experts solved similar problems, get hints, and check their final answer. Students could attempt as many challenges as they wished before attempting two case study projects as their final exam requirement.

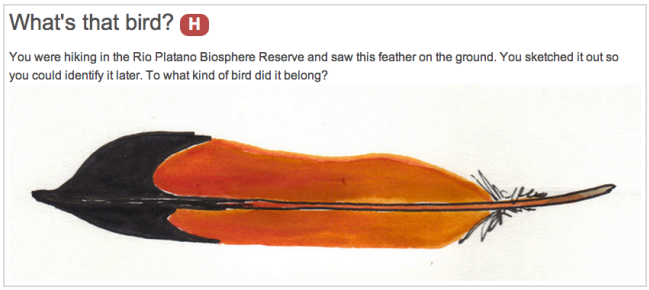

Figure 1. Sample challenge: ***You were hiking in the Rio Platano Biosphere Reserve and saw this feather on the ground. You sketched it so you could identify it later. To what kind of bird did it belong?***

Certificates of completion were awarded to students who completed and scored both projects as well as submitted the correct answer to an auto-graded final exam search challenge.

Case Study Projects

The case study projects asked students to describe how they solved a search problem, either for a problem in the list, or one drawn from their lives:

1. Solve one of the example problems below or select one that relates to your life experiences. Your problem should be complex enough to require at least three Power Search skills.
2. Record your experience using one of the provided templates or choose your own format (document, spreadsheet, slideshow, video, etc)

Example problems:

- Plan a trip for a friend who will be visiting your area. Is she interested in ethnic food, local history, natural wonders, sports, or something else? Select a theme and create an itinerary composed of five unusual destinations that fit that theme.
- Propose a new World Heritage site in your country. What are the criteria for becoming a World Heritage site? What are the existing locations near you? Prepare to argue what qualifies the location you selected to become a World Heritage site.
- Suggest a new word you've encountered this year that you think should be added to dictionaries in your language. What are the criteria for adding a word to your local dictionary? What new words were added in 2012? Prepare to make an argument about why the word you suggest qualifies to be in the dictionary.
- Conduct some genealogical research to locate the origin of your last name. What does it mean? Who was a notable member of your family from at least three generations ago? If your name has its origins in another country, what town might have members of your extended family?

Students were then presented with the evaluation criteria, submitted their assignment by either filling in text boxes or supplying a link to a Google document (for which we had provided a template asking the same questions as the text fields within the course). Questions they answered and the evaluation criteria are shown in Table 1.

After submitting each case study, for training purposes, students evaluated a sample assignment using the same checklist that they would later use to evaluate their own work. The goal of this exercise was to give the students practice in using the checklist and to develop their metacognitive skills. We then provided feedback showing how an expert would have graded the sample assignment.

After this training, students proceeded to evaluate their own work. The evaluation checklist consisted of fourteen yes/no

Assignment questions	Evaluation checklist questions
What is your research goal? What will you do with the information you gather?	Is the goal written as a complete sentence or phrased as a question? Does the description include why this research is important to you and what you will do with the information?
What questions do you need to answer in order to achieve your research goal?	Are there at least three smaller or related questions? Are the steps sequenced appropriately so that information gathered leads toward the end goal? Are the questions directly related to the goal of the research?
What queries did you type in during your research (either to Google or databases/sites you discovered)?	Are there three queries you used when searching? Do the queries relate to the questions above? Do the queries demonstrate advanced power searching skills?
What specific websites did you use when gathering information? How did you know they were credible?	Are there URLs of at least three specific websites? Are the listed websites credible?
What was your final result?	Does this answer the question you set out to solve? Does the research end at an appropriate point, even if the stated goal was not reached?
What did you learn while conducting your research?	Is there at least one interesting factor insight?
What Advanced Power Searching skills did you apply during this assignment? *(multiple-select from a list)*	Are there three skills identified?

Table 1. APS case study questions and evaluation checklist

questions. Each question was worth one point except for the last one, which was worth three points, for a total of sixteen points. The checklist was presented to the right of the student's submission (see Figure 4, which shows the top part of the evaluation form).

Methods

After the course closed, course administrators provided researchers with an anonymized sample of assignment submissions. Thirteen members of the course staff (including instructors, teaching assistants, content experts and instructional designers) graded seventeen percent of the scored, accessible assignments. To ensure consistent interrater correlation before grading the sample set, graders trained together, independently evaluating assignments until they reached a point of being able to replicate the grading score across all of the graders. (It took five sample practice assignment-grading sessions to train to this level of consistency.)

Data

Students submitted a total of 3,948 assignments. Out of this entire set of assignments, students chose not to score 95 (2.4%). Another 672 (17.0%) of assignments were submitted as links to Google Documents but were not marked as "Shared" with course staff (making them effectively ungradable). This left a total of 3,181 that could be scored by course staff. Of these, course staff graded a random sample of 535 (17%) that were both accessible and self-graded by students.

The mean student score of the graded assignments was 15.2 (standard deviation = 2.2); the mean TA score of the graded assignments was 13.3 (standard deviation = 3.5). Of these assignments, 295 (55.1%) had student and TA scores within one point of each other (out of a total sixteen points). 368 (69.0%) had student and TA scores with two points of each other. 338 (63.2%) received TA scores of fourteen or above out of sixteen, while 392 (73.3%) received TA scores of thirteen or above.

Out of the 3,853 assignments where students graded themselves, 2,708 (70.3%) awarded themselves full credit. 267 (9.9%) of the full credit submissions were blank or nonsense (e.g. ffwevrew).

We also assessed how many of the full credit submissions were copies of other assignments and found that 231 (8.5%) of full credit submissions were duplicates of others. Of these duplicates, 143 consisted of three assignments that appeared over 40 times each. We later discovered that these had been either posted on the Internet by students or were merely copies of examples provided in class. 54 of the duplicates appeared between 3 and 9 times each. 34 of the duplicates copied one other assignment, which likely resulted from one student submitting the same assignment for both projects.

In addition to grading student work, we assessed how worthwhile students found the self-graded assignments via an anonymous post-course survey. We sent the survey to the 1645 people who completed the course. Of 651 students

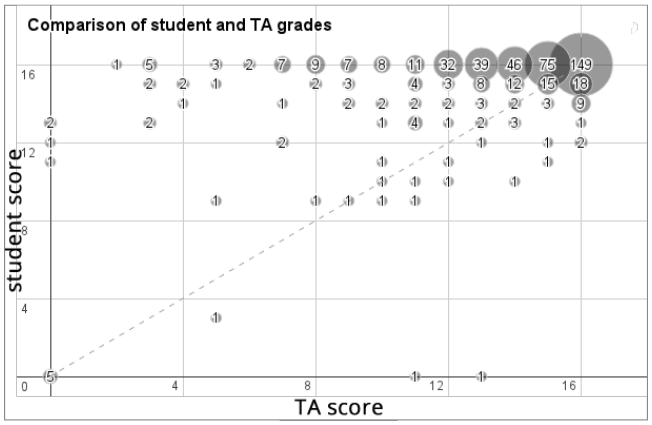

Figure 2. Student and TA scores for the APS MOOC

who responded to the post-course survey, 306 (47.0%) found the case study assignment very worthwhile; 299 (45.9%) found the project somewhat worthwhile.

Analysis

There is a moderate yet statistically significant correlation (Pearson r=0.44) between student scores and TA scores. The majority of students graded themselves within two points of how an expert grader would assess their work. The overall quality of valid self-graded assignments was high, with nearly three-quarters receiving at least a B average (73% of graded assignments received thirteen out of sixteen or better, or 81.3%).

Most students submitted two assignments. The number of blank or duplicate assignments that were submitted that received full credit was 498. If all students submitted two assignments, then this corresponds with 249 students. A total number of 1,874 students submitted two assignments. Therefore a moderate number of students (13%, or 249 out of 1,874) took advantage of the system by plagiarizing or submitting blank assignments but giving themselves full credit.

Discussion

Self-grading seems to be an effective alternative to multiple-choice assessments for in-depth, qualitative student work in low-stakes massive open online courses. The lower than expected correlation we found likely corresponded to a lack of training students how to evaluate their own work, vagueness in the evaluation checklist, and the ability for students to reward themselves for submitting low quality work.

Previous studies in which self-grading was successful included an in-depth training process that involved students co-creating the rubrics as well as discussion during the grading about elements of specific assignments [12]. Although this course provided a sample assignment for students to grade, it appears that this was not sufficient for students to truly understand all of the criteria. Future work may explore a more comprehensive training process for grading calibration similar to assessing the "ground truth" on several assignments prior to grading students' own work [7] or a gating process that required students to reach the same scores as experts on sample assignments before they could score their own work.

Students who completed all course requirements earned a printable certificate but could not necessarily receive university credit. Based on conversations between course staff and students, some students appeared to be motivated by the mistaken belief that earning this certificate would automatically get them a job at Google. This could have provided an incentive for students to take shortcuts. This problem could be resolved by having the course assignment system check for valid work in text entry boxes as well as reject duplicate submissions.

MOOC #2: MAPPING WITH GOOGLE

The MWG course [8] was created to teach the general public how to use Google's Maps, Maps Engine Lite, and Google Earth products more efficiently and effectively. The course was announced when registration opened on May 15, 2013; students could access instructional materials from June 10 through June 24. The course was created using Google's open-source Course Builder platform [4] with minor modifications to improve usability (we slightly changed the standard registration questionnaire, and the final project self-assessment interface to support the self-evaluation options).

In addition to standard video and text lessons, the course offered application activities for a variety of skills (such as using Google Maps to find directions between two points on a map, creating a customized map, using Google Maps Engine Lite to import a csv file of locations for display on a map, and using Google Earth to create a tour with audio, images, videos, and panoramic views).

Based on our observations with self-grading in APS, we implemented a self-evaluation system for two final projects in this course. Students could choose to complete a Google Maps project, a Google Earth project, or both. As before, we awarded certificates of completion to students who completed and scored final projects. We required students to turn in and score themselves on the final projects in order to receive the certificate.

Final Projects

In contrast to the APS MOOC (which asked students for a case study), students in this course could complete a final project that involved creating two online maps that would meet established criteria. They were asked to "Create a map that communicates geographical information using Maps Engine Lite. Meet all of the basic criteria and select one or more advanced features from the list [of maps features]." Students were given an evaluation rubric before completing their task. They submitted their assignment by supplying a link to their Map as well as by answering additional questions about their project, each intended to facilitate their metacognitive design practice as shown in Table 2.

Assignment questions	Evaluation rubric
1. What story are you telling with your map? 2. Did you change the base map? If so, why? If not, why not? 3. What advanced feature(s) skills did you apply to your map? *(multi-select from a list)*	• Does your map have a title? (*Yes/No*) • Does your map have a description? (*Yes/No)* • How many points are in your map? *(0, 1, 2, 3, 4, 5 or more*) • How many points have titles? *(0, 1, 2, 3, 4, 5 or more*) How many points include a relevant description? *(0, 1, 2, 3, 4, 5 or more*) • How well does the styling enhance the distinction between map points? *(score between 0-5, from none to very-well)* • How well do the advanced features included enhance the clarity of the map? (*score between 0-5, from none to very-easy-to-understand)*

Table 2. MWG project questions and rubric

As in the APS MOOC, after submitting their final assignment, students were guided to grade two sample assignments using the same rubric that they would later use to evaluate their own work. Based on experiences in Advanced Power Searching, course designers believed that one sample assignment may not have been sufficient to train students how to evaluate their work. The Course Builder system provided feedback based on how an expert would have graded the assignment. Students then proceeded to evaluate their own work. The rubric consisted of the seven questions listed in Table 2. The two yes/no questions were worth one point each, and each subsequent question was worth five points for a total possible score of twenty-seven points.

Methods

After the course closed, course administrators provided researchers with an anonymized sample of assignment submissions. Three members of the course staff (teaching assistants and content experts) graded ten percent of submitted assignments. As before, course staff calibrated scoring by reviewing several sample assignments together until they achieved consistent scores on several assignments.

Data

Students submitted 5,160 Google Maps projects. Out of this entire set of projects, students scored 5,058 (98.0%). Course staff sampled 285 projects and found that about one-third (34.7%) of the maps (99 out of 285) were inaccessible because students did not choose to make their maps public or share them with course staff. We therefore extrapolated that 1,755 out of the self-scored 5,058 projects would also be inaccessible for a total of 3,303. Course staff graded a random sample of 384 of these 3,303 projects (11.6%). The mean student score of the graded assignments was 25.7 (standard deviation of 2.03); the mean TA score of the graded assignments was 24.9 (standard deviation of 2.79).

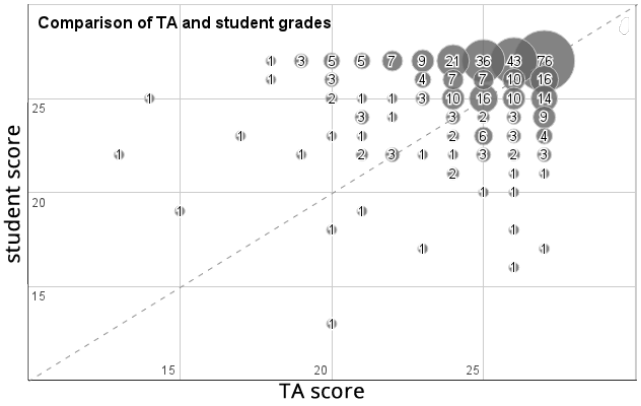

Figure 3. Score differences between students and course staff for the MWG MOOC.

Of these assignments, 201 (52.3%) had student and TA scores within one point of each other (out of a total twenty-seven points). 275 (71.6%) had student and TA scores with two points of each other. 340 (88.5%) had student and TA scores within five points of each other. 359 (93.5%) received TA scores of 21 or above (out of 27, a B average). Out of the 5,058 assignments where students graded themselves, 2,605 (51.5%) awarded themselves full credit. Oddly, 73 (2.8%) of the full credit submissions were blank (and were the only submissions by those users). We assessed how many of the full credit submissions were duplicates, finding that 9 (0.3%) of full credit submissions were duplicates of other submissions. No students with the same UserID submitted two duplicate assignments.

In addition to grading student work, we assessed how worthwhile students found final projects via a post-course survey. Of 1901 students who completed the final project and responded to a post-course survey 1407 (74.0%) found the Maps project very worthwhile; 475 (25.0%) found the project somewhat worthwhile.

Discussion

We found significantly better results with the self-grading experience in this course than in the APS MOOC. Similar to other online courses, the primary challenge in this self-evaluation process seemed to be the difficulty students had in precisely interpreting the rubric [6]. Even TAs who graded the students' work encountered confusion about how to apply the rubric. We further developed the rubric during the grading process. In retrospect, we should have

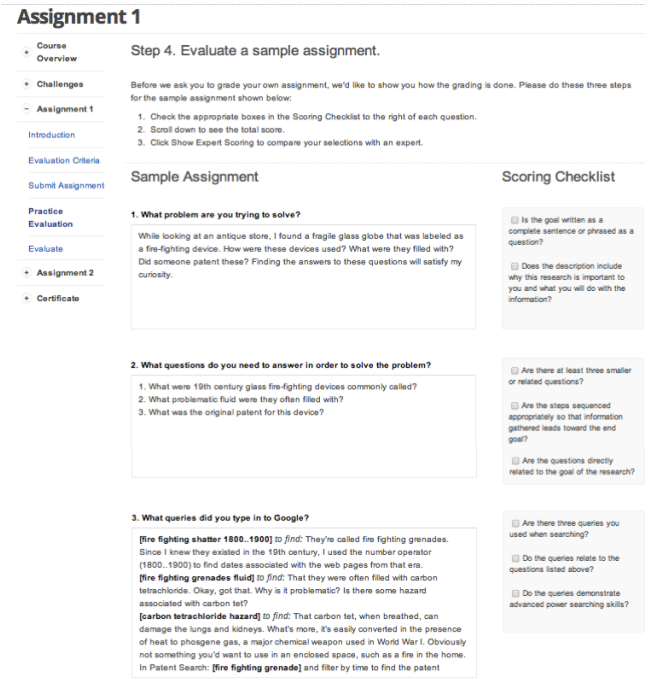

Figure 4: Sample of grading practice, with sidebar Scoring Checklist. (Note that there are 14 questions in the entire form, here for space reasons we only show the top 8.)

done this at the outset (although we did not have a large sample set of the maps to predict how students would be applying the skills). If we taught the course again, we anticipate that publicizing the more detailed rubric earlier in the course would increase the correlation between student and TA grades. As graders, we also discovered that five points of grading on subjective questions was too many. Future rubrics might try using just three points of quality to see if this would increase student accuracy.

We surmise from comments in the open-ended questions on the two course surveys that the large number of students in APS who submitted blank or duplicate assignments but graded themselves full credit had to do with the level of difficulty of the assignment. Students perceived the MWG course assignments as relatively easy, therefore there may have been reduced incentive to cheat. Other differences between the two assignments that may explain the discrepancy include the fact that students in Advanced Power Searching were asked to submit two assignments instead of one. There may also have been a perception that earning an Advanced Power Searching certificate would help students obtain a job at Google. Although we work hard to be clear about such things, misconceptions occasionally persist.

We also find it interesting that significantly more students rated the MWG course projects as *very worthwhile* compared with the Advanced Power Searching case studies. Assignments in both courses were designed to be relevant to students' lives, show the application of skills gained in the course, and create an artifact they could use after leaving the course.

	APS	MWG
TA/student scores within 6% of each other	55.1%	71.6%
TA/student scores within 12% of each other	69.0%	88.5%
assignments that received over 80% (B average) by TAs	73.3%	93.5%
blank assignments that were scored full credit by students	9.9%	2.8%
duplicate assignments that were scored full credit by students	8.5%	0.3%
survey respondents indicating the final projects were very worthwhile *(5 on a scale of 1 to 5)*	47.0%	74.0%
survey respondents indicating the final projects were somewhat worthwhile *(4 on a scale of 1 to 5)*	45.9%	25.0%

Table 3. Comparison of two courses

An additional difference between the two courses is that APS students could score their projects anything (including zero) in order to receive credit for completing the project. MWG students were required to score their work anything greater than zero. This may have caused students to be more thoughtful about the scores they gave themselves, or it may have discouraged students who were simply trying to earn credit without doing the work.

Likewise, the discrepancy between the fractions of duplicate assignments submitted between the two courses begs further investigation. We could not determine why these two MOOCs would be so different in duplicate final project submission rates.

FUTURE WORK

This work suggests several directions for future studies. Given the issues that arose with creating and using effective rubrics for self-evaluation, in future courses, authors could explore adjusting rubrics and clarifying grading criteria as the course progresses. In addition, courses could spend more time training students how to evaluate their work. In theory this is a separate skill from the skills of doing or completing activities [3] and merits a separate part of the course content. Students might practice grading several standardized assignments until they reach alignment with the gold standard scores. Once they have achieved this alignment they could proceed to grading their own assignments. As we saw from the number of duplicate and blank or nonsense submission, developing technology to prevent students from submitting and scoring blank, nonsensical, or duplicate assignments should also be in the near term planning horizon.

CONCLUSIONS

Self-grading seems to be an effective alternative to multiple-choice assessments for in-depth, qualitative student work in low-stakes massive open online courses. It is a simple and effective way to create direct student engagement in their learning, while *not* requiring the development of very sophisticated autograding systems.

In looking back at our experience with these two MOOCS, several points come to mind.

First, as is well known in the education literature, writing rubrics for anyone to use in performance assessment is difficult [9].

Yet we know that the process of answering the questions on the rubric is valuable to students [9]. A rubric helps communicate to students the specific requirements, expectations, and acceptable performance standards for an assignment. The can help students monitor and assess their progress as they work toward clearly indicated goals. By making the objectives of the course clear, students can more easily recognize the strengths and weaknesses of their work and direct their efforts accordingly.

But unlike most classroom settings, MOOCS are often composed of a wide variety of students, often from many different educational backgrounds, with widely varying language abilities, and dramatically differing degrees of practice in learning in online settings.

With this in mind, we recommend not only developing the clearest and simplest rubrics possible, but also user-testing them before the MOOC is offered. This is often difficult pragmatically, as the student composition is often not known ahead of time, but we have found that even limited user testing of self-evaluation rubrics to be of enormous help.

As we found with our own experience of creating a panel of experts to consistently grade the sample set of student assignments, practice is key. We also suggest that every self-evaluation method also come paired with enough practice (and sufficient evaluation of *that* skill as well) to ensure that consistent evaluations take place for all students.

Finally, while we were pleased with the overall correlation between self-evaluations and the gold standard of expert assessments, the number of bogus submissions was somewhat troubling, and suggests that for online classes where evaluation has a higher stakes consequence, robust checking of assignments for blanks, nonsense entries, and duplicates is well worth the effort.

ACKNOWLEDGEMENTS

We thank Alfred Spector, Maggie Johnson, and Google's MOOC development team for their advice, feedback, and support. The Mapping with Google course used Course Builder 1.4.0. [4] We thank Saifu Angto, Pavel Simakov, and John Cox for continuous support, customizations and code.

REFERENCES

1. Advanced Power Searching course. 2013. http://www.powersearchingwithgoogle.com/course/aps
2. Balfour, S.P. 2013. Assessing writing in MOOCS: Automated essay scoring and Calibrated Peer Review. *Research & Practice in Assessment* 8.1 (2013): 40-48.
3. Black, P. & Wiliam, D. 1998. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7.
4. Course Builder platform. 2013 http://code.google.com/p/course-builder
5. Eslinger, E., White, B., Frederiksen, J., & Brobst, J. 2008. Supporting Inquiry Processes with an Interactive Learning Environment: Inquiry Island. *Journal of Science Education and Technology*, *17*(6), 610–617. doi:10.1007/s10956-008-9130-6
6. Kulkarni, C., and S. R. Klemmer. 2012. Learning design wisdom by augmenting physical studio critique with online self-assessment. *Technical report, Stanford University.*
7. Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Koller, D., Klemmer, S.R. 2013. Scaling self and peer assessment to the global design classroom. *Transactions on Computer-Human Interactions Journal, 20*(6) *in publication.*
8. Mapping with Google course. 2013 http://mapping.withgoogle.com
9. Moskal, B. M. 2000. Scoring rubrics: what, when and how? *Practical Assessment, Research & Evaluation,* 7(3).
10. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (n.d.). Tuned Models of Peer Assessment in MOOCs. Retrieved (11/7/13) http://www.stanford.edu/~cpiech/bio/papers/tuningPeerGrading.pdf
11. Rivers, W. P. 2001. Autonomy at All Costs: An Ethnography of Metacognitive Self-Assessment and Self-Management among Experienced Language Learners. *The modern language journal* 85.2 (2001): 279-290.
12. Sadler, P., & Good, E. 2006. The Impact of Self- and Peer-Grading on Student Learning. *Educational Assessment*, *11*(1), 1–31.
13. Veenman, M., Van Hout-Wolters, B., & Affleerbach. P. 2006. Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning, 1*(1):3, (Apr 14, 2006).