Grading by Experience Points: An Example from Computer Ethics

Edward Gehringer, Barry Peddycord III
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
{efg, bwpeddyc}@ncsu.edu

***Abstract*—In most of education, courses are graded based on percentages—a certain percentage is required for each letter grade. Students often see this as a negative, in which they can only lose points, not gain points, and put their class average at risk with each new assignment. This contrasts with the world of online gaming, where they gain "experience points" from each new activity, and their score monotonically increases toward a desired goal. In Fall 2012, the lead author switched to grading by experience points in his Ethics in Computing class. Students earned points for a variety of activities, mainly performing ethical analyses of various issues related to computing, and participating in debates on ethics-related topics. The students appreciated the ability to earn extra points by performing extra activities. But they were less likely to complete analyses after signing up to do them than were students in a traditionally-graded class. At semester's end, the number of peer reviews increased, as students strove to top off their point total. The grade distribution was bimodal, with clusters at both ends (A+ and F). Students' greatest concern was rapid grading turnaround, so they would know where they stood in the class at all times.**

Keywords—gamification, experience points, XP, computer ethics, assessment (*key words****)***

I. INTRODUCTION

In most of education, courses are graded based on percentages—a certain percentage is required for each letter grade. Students often see this as a negative, in which they can only lose points, not gain points, and each new assignment puts their class average at risk. This contrasts with the world of online gaming, where one only gains, never loses, "experience points" from each new activity, causing their score to increase monotonically toward a desired goal.

In the last two or three years, several instructors have moved to grading by experience points (XP). The strategy was showcased by Lee Sheldon in his 2012 book, *The Multiplayer Classroom: Designing Coursework as a Game* [1]. It offers a way to motivate students in almost any course, without the need to design actual games to teach different units of the course material. It seemed appropriate for our Ethics in Computing course, which is not always a favorite among our technically oriented students. We thought it might boost the level of student interest and engagement.

II. MOTIVATION

The motivation for using experience points stems from a number of shortcomings of more traditional, percentage-based grading schemes. In such schemes, a finite number of potential points are set at the outset of the course, and students strive to obtain 100% of the potential points. Because the number of points is fixed, it can be difficult for students to recover after poor performance on a single, heavily-weighted assignment.

When assignments are weighted heavily, students are under intense pressure to perform well [9] on them. An unintended corollary is to incentivize behavior that does not adequately promote learning [4]. To gain insight into how students are prioritizing their time and energy, listen to questions that students ask about assignments, e.g., How many pages should the essay be? Or, will the exam be mostly multiple choice or mostly essay? The guiding principle from this insight is that *high-stakes assessment incentivizes performance over learning*.

Gamification [3] is used in business and education to reward positive behavior. In the context of the classroom, gamification would be used to incentivize student actions associated with learning. By setting a goal and allowing students to choose their own way to reach that goal, it allows students more agency in the tasks they perform. The hope is that because they define their own tasks, they will choose those those in which they take a personal interest, rather than performing them simply "for the grade." One of the simplest ways to gamify a course is to grade it by a monotonically increasing number of *experience points*, rather than weights and percentages for each assignment.

The principal detriment of percentage-based grading is that it is difficult for students to recover from mistakes. In the event of failure, experience points are more forgiving than percentages. This is desirable, since failure can serve as an excellent learning experience. While percentage grades equate success with near-perfect performance, experience points can reward deep and thorough learning.

978-1-4673-5261-1/13/$31.00 ©2013 IEEE

In addition to reducing the pressure on students, points can also be used to reward socially positive activities that not only advance the learning of the individual, but also improve the learning experience for the rest of the class [5]. Students can be awarded points for holding study sessions, office hours, or participating in peer reviews. In this way, students help with the logistics of the course, improve the experience for their peers, and absorb the positive message that teaching others is a valid approach to demonstrating mastery of course material. One manifestation of peer teaching in computer science is pair programming, which is often used to facilitate a transfer of skills between students [2].

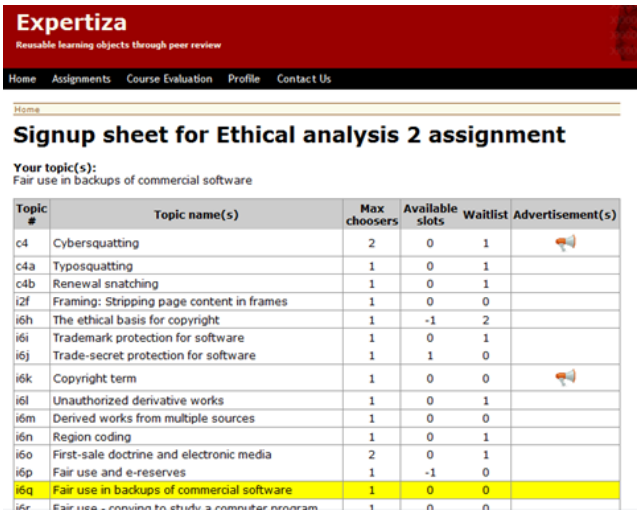
III. IMPLEMENTATION

Our Ethics in Computing course is a one-hour course required for all Computer Science students. Students were awarded points for each activity in the class:

- weekly quizzes over the reading material for the class (each worth 50 points maximum),
- ethical analyses (done alone or in pairs, each worth 400 points maximum),
- peer reviews of other students' ethical analyses (each worth 50 points maximum),
- evaluations of one's partners' contribution to an ethical analysis (each worth 50 points), and
- in-class debates (each worth 200 points maximum).

Except for partner evaluations, all activities were graded in some way, and the students' points were scaled by the grade they received. For example, if they scored 80% on a weekly quiz, they earned 50 × 80% = 40 points. For evaluating their partner, they received the full 50 points automatically. This provided an incentive to do the ethical analyses in pairs, since one could earn an easy 50 points just by filling out the partner survey. Some students asked if teams of more than two would be allowed. We decided that students could work in larger teams, but regardless of how many were on a team, the maximum score earned by all team members would be 400 x 2 = 800. So each student on a 3-member team could earn only 800/3 = 266 2/3 points, even if (s)he got a perfect score on the analysis. Of course, they could earn 100 more points for evaluating both of their partners, but even so, the maximum point total was less than for a solo or 2-member analysis. Only 6 out of 174 analyses during the semester were written by 3-member teams; all the rest had solo or pair authors.

The course Web site specified grading thresholds like this: "To get an A in the class requires 2600 points; a B requires 2300 points, etc." The meaning of "etc." was intentionally left vague because this was the instructor's first time grading by experience points, and he wanted the flexibility to adjust cutoffs if necessary to achieve a "reasonable" grade distribution. In any case, if students needed to earn extra points to reach their desired grade, they would merely need to perform additional activities, like analyzing more issues, or



Topic #	Topic name(s)	Max choosers	Available slots	Waitlist	Advertisement(s)
c4	Cybersquatting	2	0	1	
c4a	Typosquatting	1	0	1	
c4b	Renewal snatching	1	0	1	
i2f	Framing: Stripping page content in frames	1	0	0	
i6h	The ethical basis for copyright	1	-1	2	
i6i	Trademark protection for software	1	0	1	
i6j	Trade-secret protection for software	1	1	0	
i6k	Copyright term	1	0	0	
i6l	Unauthorized derivative works	1	0	1	
i6m	Derived works from multiple sources	1	0	0	
i6n	Region coding	1	0	1	
i6o	First-sale doctrine and electronic media	2	0	1	
i6p	Fair use and e-reserves	1	-1	0	
i6q	Fair use in backups of commercial software	1	0	0	
i6r	Fair use - copying to study a computer program	1	0	0	

Fig. 1. Signup sheet for ethical-analysis topics

reviewing analyses done by other students. The A threshold was set so that students who (i) took all the weekly quizzes, (ii) did three ethical analyses during the semester, (iii) and two reviews per analysis, and (iv) participated in two debates, and received an average of 93% on all their work, would collect just enough points for an A. However, students were permitted to do as many as six analyses (there were six two-week analysis rounds during the semester), and also to participate in extra debates if space was available. They could thus "make up for" poor performance on other assignments, or for missing those assignments altogether. In effect, all work was treated as if it was "extra credit," counting for a particular number of points. No one needed special permission to do extra work.

Ethical analyses were reviewed by a combination of peer review and instructor/TA review. The process is similar to the practice that we earlier described for writing sections for a student-authored wiki textbook [8]. Using our Expertiza system [6,7], students first select a topic from a list presented by the instructor. Students are allowed to sign up individually (Figure 1) or in pairs, but usually, only one student or team will be allowed to sign up for each topic. This is to ensure that all the topics will chosen by someone. Students then write up their work in a Google doc (ultimately to be included in a Google site) and submit the link to Expertiza. Other students review it using a rubric that asks them to comment and rate approximately ten aspects of the work. The student authors then have an opportunity to respond to their reviewers (also using a rubric). Based upon reviewers' comments, they revise their work. Their final submission is again assessed by their peer reviewers. The instructor and TA review the work and assign the final grade, but are often influenced by the reports of peer reviewers.

To give students an incentive to do careful reviews, the instructor and TA "metareviewed" the reviews done by the students. We looked over all the reviews done by the student during a particular round of ethical analyses, and scored the student's *average* review on a scale from 1– to 3+. (This was faster than assigning a score to each review.) The 1– to 3+ scale was translated to points (1– = 10 points; 3+ = 50 points), and the student was awarded this many experience points for each review done in this round (Figure 2).

1=vacuous, 3=good, 2=in between

David	# rev	qual	points	Ed	# rev	qual	points	Shea	# rev	qual	points	
Student 1	8	2	240	Student 28	1	2	30	Student 57	0			
Student 2	6	2	180	Student 29				Student 58	4	3-	160	
Student 3	2	2	60	Student 30				Student 59	0			
Student 4	2	2+	70	Student 31	2	3–	80	Student 60	7	3+	350	
Student 5				Student 32				Student 61	0			
Student 6	2	3-	80	Student 33	1	3	45	Student 62	0			
Student 7	1	2+	35	Student 34				Student 63	0			
Student 8				Student 35	7	2	210	Student 64	0			
Student 9	1	2+	35	Student 36	15	2	450	Student 65	9	2+	315	
Student 10	3	2	90	Student 37				Student 66	0			
Student 11	2	2	60	Student 38				Student 67	6	2+	210	
Student 12	1	2	30	Student 39				Student 68	1	3-	40	scale for reviewing:
Student 13	1	1	15	Student 40				Student 69	0			4 or 3+: 50 points
Student 14	3	1+	60	Student 41	1	1+	20	Student 70	0			3: 45 points
Student 15				Student 42	13	1	195	Student 71	27	2-	675	3–: 40 points
Student 16				Student 43	2	2	60	Student 72	0			2+: 35 points
Student 17	4	3-	160	Student 44	10	2–	250	Student 73	0			2: 30 points
Student 18	1	1+	20	Student 45	9	3–	360	Student 74	9	2+	315	2–: 25 points
Student 19	6	1	90	Student 46	10	3+	500	Student 75	2	3	90	1+: 20 points
Student 20	9	1+	180	Student 47	4	2+	140					1: 15 points
Student 21	13	1	195	Student 48	2	2+	70					1–: 10 points

Fig. 2. Metareview scores assigned by instructor and TA

IV. Results

As it turned out, the most common grade in the course was A+. The A+ threshold was set at 97/93 of the A threshold, or 2712 points (typical university cutoffs are 97 for an A+, 93 for an A). More than 1/4 of the students (21 out of 77) received an A+. Some of them were very excited about the course. Five students wrote the maximum 6 ethical analyses, and five of these 6 earned an A+ in the course. Of the A+ students, 8 of them participated in at least 3 debates (one took part in 4 debates). These students accumulated 2782 to 3694 experience points. However, there was more than one path to an A+. One of our A+ students wrote only one analysis and received almost 60% of his points for reviewing.

The second most common grade in the course would have been F—if we had extrapolated the grading scale linearly (2600 for an A, 2300 for a B, 2000 for a C, 1700 for a D). Fifteen students accumulated fewer than 1600 XP, with totals that ranged from 152 to 1586. Figure 3 shows the grades that would have been awarded without any "curve." To avoid failing a large number of students (in an ethics course nonetheless!), we loosened the scale so that only 1600 points were needed for a C and 900 for a D. With these cutoffs, there were only 5 Fs. But this meant that an A required nearly three times as many points as a D; on Sheldon's scale [1], by contrast, an A can be earned with only 50% more points than a D.

The idea that grading was based just on points meant that we could not really require students to complete any course activity. According to the syllabus, all students were compelled to participate in two debates. However, there was no way to enforce this requirement, since students could instead earn points for writing or reviewing extra ethical analyses. Even among the A+ students, 7 of them failed to do

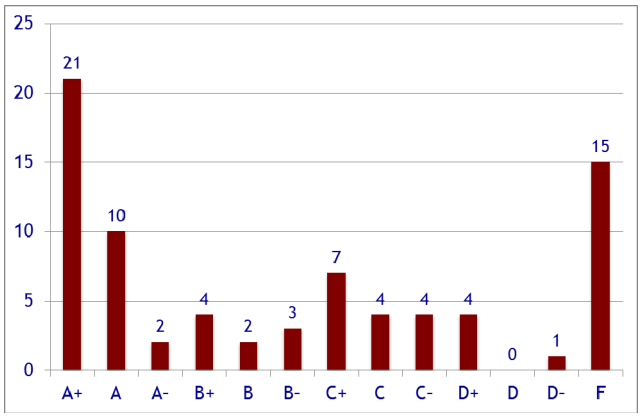


Fig. 3. Grade distribution with unadjusted grading scale

two debates; one of them didn't do any. In future semesters, we will consider making debates an achievement that all students must attain in order to pass the course.

Figure 4 compares the results in the ethics course with the results in the author's other Fall 2012 course, a course in object-oriented design, which was not graded by experience points. The column marked "% of teams submitting" refers to the percentage of teams that chose a topic and then followed through with a submission. In the ethics class, about 2/3 of the teams submitted an analysis of their chosen topic in the first round of ethical analyses. This percentage declined slightly in the next few rounds, then rose as the end of the course approached and opportunities for procrastination evaporated. In the OOD course, students were required to submit a wiki chapter for either assignment 1a or 1b, and again for either 2a or 2b. The percentage actually submitting in the OOD class was lower than the ethics course in the "a" rounds, but much higher in the "b" rounds. This may be a manifestation of the fact that the ethics students feel less pressure to do any particular assignment, since there are always other ways to earn points. This is especially true early in the semester, when the other opportunities are numerous.

Now look at the row titled, "# responses/student". This is the number of reviews done per student for the submissions in

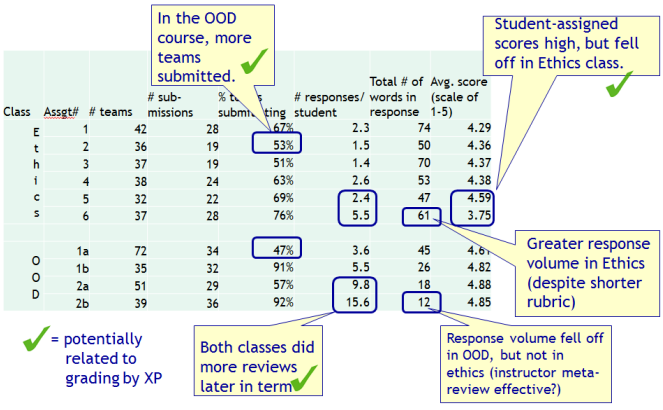


Fig. 4. Comparison of peer reviewing in ethics and another course

the different rounds. Note that it rises in both classes as the end of the semester approaches. This is probably a consequence of the fact that students in the OOD class could earn extra credit for doing extra reviews (though there was a cap on the amount of extra credit that could be earned in this way). Thus, students in both classes were incentivized by earning extra points for extra reviews, and this inducement seemed to grow as the semester wound down and the other opportunities for earning credit dried up.

The next column tells how many words of prose were included in reviews submitted by the students. Note that this is much higher in the ethics class than in the OOD class. However, the OOD class was using a much more detailed rubric than the ethics class (18 criteria vs. 6). Thus, it took reviewers longer to rate the OOD submissions on the required criteria, which seemed to subtract from the energy that the student reviewers invested in writing prose. It is probably also true that student authors derived more guidance from the more detailed rubric, so prose comments were somewhat less necessary.

The volume of prose comments falls off sharply toward the end of the semester in the OOD class, but not in the ethics class. This is likely due to the fact that students (rather than instructors) were serving as metareviewers in the OOD class. Students tend to be much less demanding of the reviewers that they rate than the instructor/TA would be. This suggests that instructor/TA metareviewing served to "keep the ethics students honest," doing reviews that were just as careful at the end of the semester as at the beginning.

On the other hand, when one looks at the average score assigned by the reviewers (on a Likert scale of 1 to 5), it falls off sharply at the end of the semester in the ethics class, but not in the OOD class. This suggests that perhaps the quality of ethical analyses fell at the end, with students attempting to earn points via quantity rather than quality of submitted work. However, if one compares the average score awarded (by the instructor and TA) to analyses in round 6 (75.3) with the average score in round 5 (76.2), there isn't much difference. So it's not clear that quality did fall. Of course, one would hope to see quality increase from one round to the next, but we didn't see that happen.

A. Student Responses

Students were surveyed on their reaction to the review process and grading by experience points. Results are shown in Table 1. Students were most positive (4.02 on a Likert scale of 1 to 5) on having the chance to do extra work to achieve the grade they desired. They were least positive (2.11 on a scale of 5) on the grading process giving them a clear indication of where they stood. As one student put it,

> "I believe this system could really work..the problem was this semester is that the grading was unacceptably late. I can't even describe how infuriating this was as a student. I didn't even start knowing what my XP points were until mid November."

TABLE I. RESULTS FROM STUDENT SURVEY

Survey Question (Likert scale of 1 = strongly disagree to 5 = strongly agree)	**Average** (n = 44)
1. I had trouble understanding what was expected of me in writing an ethical analysis.	3.36
2. I put a lot of effort into writing my ethical analyses for the Ethics in Computing	3.83
3. The material I read in order to write my ethical analyses gave me new insight into the topic I was writing on.	3.57
4. The ethical analyses I wrote would be credible entries for an undergraduate textbook	2.81
5. I am proud of my contributions to the Ethics site.	3.34
6. Having students write analyses for an ethics Web site in a course like CSC 379 is a good idea.	3.05
7. I clearly understood what was expected of me in reviewing an ethical analysis.	2.95
8. The analyses I read that were authored by other students gave me new insight into the material they covered.	3.21
9. I read the content of the ethical analyses more critically than I read the content of other textbooks I have used.	2.98
10. I often questioned the information and ideas in the ethical analyses I read..	3.29
11. I tried to decide if I agreed with the ethical analyses when I read them.	3.76
12. When I read the ethical analyses I considered what other opinions and ideas might exist on the topic.	3.67
13. As I read, I often thought about my own ideas and opinions on the topic.	3.83
14. I looked to see if the claims made in the ethical analsyes were well supported.	3.57
15. When I read the ethical analyses, I often thought about information that seemed to be missing or what else could be included.	3.48
16. The reviews I received helped me to improve my work.	2.95
17. The scores assigned by the reviewers were fair.	3.07
18. There was too much rating required for this class.	3.55
19. I had trouble determining how to carry out the assigned activities in Expertiza.	3.33
20. Grading by experience points (XP) motivated me to work more on this course.	2.83
21. Grading by XP motivated me to write more than the required number of ethical analyses.	3.40
22. Grading by XP gave me a clearer idea of where I stood in the course"	2.11
23. I appreciated being able to do more work to earn the extra points that I needed to achieve a certain grade	4.02

This can be attributed to two factors, one logistical and one inherent to the review process. Early in the semester, the course staff *was* slow in posting grades. Ethical Analysis 1 grades were not finalized until after Ethical Analysis 2 was due. In part, this reflects the large amount of work required to write up the topics (52 topics for rounds 1 and 2). It should get easier in the future; next time, we can use a revision of this year's topic descriptions. However, it also reflects the time needed to take the submissions through the required rounds of

review. The final review deadline was 11 days after submission, and final grading/metareviewing could not start until after that deadline. In later rounds, we caught up and were posting final grades within a week of the final review deadline.

Complaints about the speed of grading were the most frequent prose comment (10 out of 44 respondents). Next most frequent (8 respondents) were favorable reactions to grading by XP:

> "I think it is an interesting and novel idea. Perhaps the greatest benefit is I could participate in as many debates, ethical analyses, and reviews available and not be required to perform any one task. This allowed me to strengthen my skills in an area of my choosing. However, I did have to spend some time deciphering the allocation and accumulation of my experience points."

We thought our grading scheme was simple, but not all students agreed. The third most frequent comment (6 respondents) said it was hard to figure out:

> "Grading by XP gave me no idea about where in the class I stood or how well I was doing the entire semester. I still have no idea how well I've done this semester because the system was so convoluted and hard to figure out."

Some students (4 respondents) thought the course was too much work:

> "Also please assign less work for a 1 credit hour course, its not helpful to have a course that counts for so little of my time take up so much of it. A passing grade for this course should be out of 1600 not 2600 by my estimate of the workload appropriate for a 1-hour course."

Two students thought the system rewarded quantity, not quality:

> "Grading by XP encourages quantity of work, not quality. I didn't care if i got a C on an assignment, because if I got enough C's I would have well over an A."

V. Conclusion

In 2012, we switched to grading by experience points in our Ethics in Computing course. The course has often been plagued by low attendance. We wanted an approach that would make students delve into the very relevant and often very interesting issues of law, public policy, and personal morality. The change was successful in many respects. It did motivate a large number of students to go beyond what was required. Nearly a quarter of the class earned a grade of A+. These students, by and large, wrote more ethical analyses, and participated in more debates, than any student had in previous classes. Some of them told us how much they enjoyed the debates. Students in the 77-member class submitted a total of 174 ethical analyses, on about 150 different topics. This, too, is more topics than we had been able to cover in any previous running of the class.

On the other hand, a significant fraction of the students thought that the class was too much work for one credit. Many students did less than was expected. Even after curving, 10 students received Ds or Fs. This too was a much higher number than in previous semesters. Activity rose markedly near the end of the semester, as student strove to earn the points that they needed for the next higher grade. However, the quality of the ethical analyses was not as high as desired. This may indicate that many students thought it was better to write many analyses, rather than write a fewer number of good analyses. On the other hand, the quality of peer reviews remained high, probably reflecting the effort that the instructor and TA were investing in assessing the reviews and rewarding good quality.

The most important lesson learned is that speed of feedback is critical in motivating students. The peer-review process makes it difficult to finalize grades quickly. But scores should be recorded almost immediately after the final review deadline. Students need a single location where they can go to view the number of points they have accumulated thus far. There is ample opportunity for us to improve the ethical analyses by (i) using good analyses from this semester as examples for later students, and (ii) assigning some students to improve this semester's analyses in specific ways identified by the instructor. As time goes along, the analyses should only get better. This will show the students what a good analysis looks like, and motivate them to do work of similar quality.

Acknowledgments

We acknowledge the efforts of Shea McIntee, the teaching assistant in Ethics in Computing in 2012-13, and Dr. David Wright, the instructor in Spring 2013, who also assisted with the course the previous fall. We also acknowledge the National Science Foundation's support of the Expertiza system, which was used to peer-review the ethical analyses.

References

[1] Sheldon, L. (2012). The Multiplayer Classroom: designing coursework as a game. Course Technology.

[2] Nagappan, N., Williams, L., Ferzli, M., Wiebe, E., Yang, K., Miller, C., & Balik, S. (2003, February). Improving the CS1 experience with pair programming. In ACM SIGCSE Bulletin (Vol. 35, No. 1, pp. 359-362). ACM.

[3] Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011, September). From game design elements to gamefulness: defining gamification. In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (pp. 9-15). ACM.

[4] Birney, R. C. (1964). The effects of grades on students. The Journal of Higher Education, 35(2), 96-98.

[5] Decker, A., & Lawley, E. L. (2013, March). Life's a game and the game of life: how making a game out of it can change student behavior. In

Proceeding of the 44th ACM technical symposium on Computer science education (pp. 233-238). ACM.

[6] Gehringer, E.F. Expertiza: information management for collaborative learning. In Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-Learning Support, A. A. Juan Perez [ed.], IGI Global Press, 2009.

[7] Gehringer, E.F., Ehresman, L.M., Conger, S.G., and Wagle, P.A. Reusable learning objects through peer review: The Expertiza approach. Innovate—Journal of Online Education 3:6 (August/September 2007).

[8] Gehringer, E.F., Kadanjoth, R., and Kidd, J. Software support for writing wiki textbooks. Frontiers in Education 2010, Arlington, VA, October 27−30, 2010.

[9] Rau, W., & Durand, A. (2000). The academic ethic and college grades: Does hard work help students to" make the grade"? Sociology of Education, 19-38.