RESEARCH ARTICLE

Peer Assessment Enhances Student Learning: The Results of a Matched Randomized Crossover Experiment in a College Statistics Class

Dennis L. Sun[1]*, **Naftali Harris**[2], **Guenther Walther**[2], **Michael Baiocchi**[2,3]*

1 Department of Statistics, California Polytechnic State University, San Luis Obispo, CA, 93407, United States of America, **2** Department of Statistics, Stanford University, Stanford, CA, 94305, United States of America, **3** Prevention Research Center, Stanford School of Medicine, Stanford, CA, 94305, United States of America

* dsun09@calpoly.edu (DLS); baiocchi@stanford.edu (MB)

Citation: Sun DL, Harris N, Walther G, Baiocchi M (2015) Peer Assessment Enhances Student Learning: The Results of a Matched Randomized Crossover Experiment in a College Statistics Class. PLoS ONE 10(12): e0143177. doi:10.1371/journal.pone.0143177

Editor: Andrew R. Dalby, University of Westminster, UNITED KINGDOM

Received: February 25, 2015

Accepted: November 1, 2015

Published: December 18, 2015

Data Availability Statement: Data are available, with restrictions, from the authors who may be contacted at baiocchi@stanford.edu or dsun09@calpoly.edu. Restrictions are in place to protect against person-level identification, as required by the IRB and the informed consent process. Before data can be accessed, privacy agreements will need to be in place.

Funding: Stanford University supplied a grant to fund the development of our online homework platform (< $10,000USD). This platform is being used on an ongoing basis by the department, separately from the

Abstract

Feedback has a powerful influence on learning, but it is also expensive to provide. In large classes it may even be impossible for instructors to provide individualized feedback. Peer assessment is one way to provide personalized feedback that scales to large classes. Besides these obvious logistical benefits, it has been conjectured that students also learn from the practice of peer assessment. However, this has never been conclusively demonstrated. Using an online educational platform that we developed, we conducted an in-class matched-set, randomized crossover experiment with high power to detect small effects. We establish that peer assessment causes a small but significant gain in student achievement. Our study also demonstrates the potential of web-based platforms to facilitate the design of high-quality experiments to identify small effects that were previously not detectable.

Introduction

Feedback is one of the single most important factors influencing student learning [1, 2]. However, it is often not possible to provide feedback that is both detailed and prompt, thus limiting its effectiveness in practice [3]. In large college classes and massively open online courses (MOOCs), providing personalized feedback to students is especially challenging.

While automated feedback can be an adequate substitute in some cases [4], many concepts and skills are still challenging for a machine to evaluate. Consider the following question about the interpretation of a p-value, an important concept in introductory statistics.

> Josh flips a coin 100 times. The coin comes up heads 60 times. He calculates the p-value to be about 2% for testing the null hypothesis that the coin is fair. Explain what this 2% means in the context of this problem.

study. A grant from the Agency for Health Research & Quality provided support for some of Professor Baiocchi's time (KHS022192A).

Competing Interests: The authors have declared that no competing interests exist.

The correct answer is that the 2% represents the probability of observing a result at least this large if the coin were fair. However, a common misconception among students is that it represents the probability the coin is fair. Even with state-of-the-art semantic parsing, machines cannot accurately discriminate incorrect answers from correct ones [5]. On the other hand, a human who understands the concept would have little difficulty distinguishing the two.

Thus, the problem of providing feedback falls into the large class of tasks that are relatively easy for a human but challenging for a machine. Such tasks are fertile ground for crowdsourcing, which has been applied to otherwise intractable problems with surprising success [6, 7]. Feedback can be "crowdsourced" by having students grade one another, a practice known as *peer assessment*. Peer assessment provides as many graders as students, enabling more timely and thorough feedback [8]. It has already made personalized feedback feasible in a number of settings—most notably MOOCs—where it otherwise would be impossible [9].

Instructors often have two main concerns about peer assessment. The first is whether students can be trusted to grade accurately. This question has been extensively studied in the literature, and the consensus is that peer grades are at least comparable to instructor grades [10, 11]. The second is logistics; peer assessment is logistically complicated if students have to exchange papers in person. However, web-based tools have largely solved this problem. Most modern learning management systems (LMS) come with a built-in peer assessment tool that automatically distributes student responses to peer graders.

Therefore, peer assessment is a workable solution to the problem of feedback; it reduces the burden to the instructors with minimal sacrifice to quality. On top of this, it has been conjectured that students also learn in the process of providing feedback. If true, then peer assessment may be more than just a useful tool to manage large classes; it can be a pedagogical tool that is both effective and inexpensive [8, 12, 13].

This claim was perhaps most visibly advanced in the U.S. Supreme Court case *Owasso v. Falvo* (2002) [14]. Although the case was primarily concerned with whether peer assessment violated students' privacy, Justice Anthony Kennedy praised peer assessment in his majority opinion, saying

> Correcting a classmate's work can be as much a part of the assignment as taking the test itself. It is a way to teach material again in a new context, and it helps show students how to assist and respect fellow pupils [11].

However, to date, there has been scant empirical evidence for this claim. The evidence is based mostly on surveys of student and teacher perceptions [8, 15]. Only a few studies have attempted to quantify the effect on an objective criterion such as achievement, but most have been correlational studies. A representative study in this latter category is [16], which examined whether peer assessment improved students' writing abilities. However, the study lacked a control group, so it is not possible to know whether students improved any more with peer assessment than they would have otherwise. Furthermore, the study measured achievement using the students' own peer grades, rather than an objective measure (e.g., scores given by a third-party observer who was blinded to the treatment). To our knowledge, only one randomized experiment has ever been conducted to measure the effect of peer assessment on achievement, but the study lacked statistical power to reach a conclusion either way [17]. This gap in the literature has been noted by several researchers, who have suggested this as an important direction of future research [17, 18].

Materials and Methods

We were interested in whether peer assessment could aid conceptual understanding and problem solving, two skills that are especially relevant in science, technology, engineering, and mathematics (STEM) classes. We conducted a randomized controlled trial (RCT) in a large introductory statistics class, using a crossover design to enhance precision. Written, informed consent was obtained from all participants, and the study was approved by the Stanford University Institutional Review Board. The ten-week course was divided into an introductory unit and four main units. The introductory unit was excluded from the study so that students had time to become acclimated to peer assessment. Each student was then assigned to participate in peer assessment for exactly two of the main units and to the control group for the other two. In all, there were four treatment arms, each one receiving the peer assessment treatment (T) and control (C) in a different order over the four main units (TCTC, CTCT, TCCT, and CTTC).

This crossover design controls for all differences between students, since each student participates both as a treatment and a control subject at different times in the course. This eliminates what is arguably the largest source of variability in educational studies: variation between students. To further ensure against baseline differences between the treatment and control groups, we augmented this design with matched-pairs randomization. Students were matched based on covariate information (e.g., class year, previous statistics experience), and the students within each pair were assigned to complementary treatments in all four of the main units (i.e., TCTC vs. CTCT). Fig 1 shows the result of the matched pairs design, alongside a random pairing for comparison. The resulting covariate balance for the actual randomization is provided in (S1 File), showing that the matched-pairs design produced indistinguishable groups.

The treatment was defined as follows: in weeks that students were required to participate in peer assessment, they provided scores and comments on the homework responses of three peers. In turn, they received feedback on their homework from three peers. All homework responses and peer assessments were submitted through an online platform, and all responses and reviews were anonymized before distribution to ensure privacy. Students in the control group also submitted their homework responses online but did not participate in peer assessment and had their homework graded by instructors. In order to control for the possible effect of feedback timing, feedback was delivered to the two groups simultaneously. Also, the students in the control group were provided the same solution key as the one provided to the peer graders. For more details about the specific implementation of peer assessment, please refer to (S1 File).

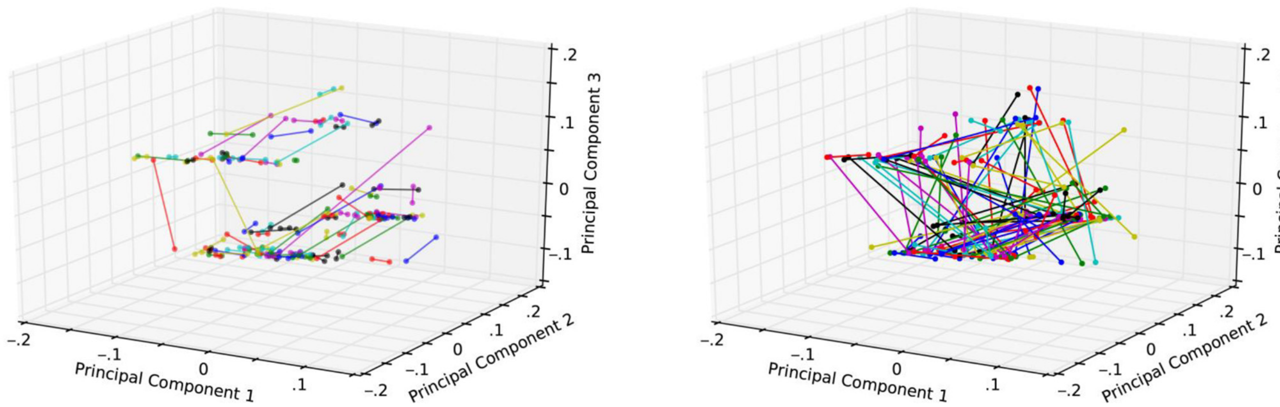

Fig 1. Two plots showing the effect of the matched pairs randomization design (1A) as compared with complete randomization (1B). Each point represents a student's covariate information, and each connecting edge indicates that those students have been assigned to opposite treatment groups. The edges in the matched pairs design are much shorter than under complete randomization, confirming that matching produces more similar randomizations.

doi:10.1371/journal.pone.0143177.g001

To measure achievement, students completed a quiz after each unit that measured the short-term effect of peer assessment. The students also took a comprehensive final exam that measured longer-term learning. These assessments consisted entirely of free-response questions that required explanations or calculations. The instructors, who were blinded to the treatment groups of the students, graded all assessments to ensure consistency for the purposes of the study.

Finally, the study was fully replicated in a different academic term with a different instructor. The same homework questions were used in the two terms, but different exam questions were used. In all, 148 students participated in the study during the first term (autumn) and 239 students during the second (winter). Because the crossover design should eliminate any student or instructor effects, we pooled the data from the two terms to obtain a single sample of 387 students. However, in the analyses, we excluded any students who failed to comply with the peer assessment treatment or to complete the assessments. This left us with 299 students in the analysis of the unit quiz scores and 320 students in the analysis of the final exam scores. Although excluding non-compliers can sometimes bias the treatment effect, our crossover design ensures that non-compliers are excluded from both the treatment and control groups. As a result, we obtain an unbiased estimate of the treatment effect on the subpopulation of students who would be affected by the peer assessment intervention. A further discussion of this and the definition of compliance can be found in (S1 File).

Results

The students who participated in peer assessment during a given unit performed significantly better on the unit quizzes (Cohen's d = .115, t(298) = 2.92, p = .002), as compared with students who did not. Students participating in peer assessment also did better on the corresponding questions on the final exam (d = .122, t(319) = 3.03, p = .001), which suggests that the benefits of peer assessment persist over time. These results are summarized in Table 1. In the context of our course, where the standard deviations of exam scores ranged from 15 to 25 percentage points, an effect size of .122 would translate to a 2 to 3 percentage point increase in the average exam score. Fig 2, which depicts the actual distribution of scores for one of the unit quizzes, shows that a modest increase in average score can be practically important.

To understand the magnitude of this effect size, we also calculated the "effect sizes" of well-known achievement gaps: between males and females, between underrepresented minorities and others, between students with more math background and students with less, etc. These achievement gaps are reported in Table 2. We estimated the gap twice, once at the beginning of the course (using quiz 1 scores, which was administered prior to randomization) and again at the end of the course (using final exam scores). We see that the effect size of peer assessment, which is .122, represents about 40% of the gender achievement gap and about 20% of the racial achievement gap, which are persistent challenges in college science classes [19, 20]. By comparing the gaps before and after the course, we see also that the course tended to reduce achievement gaps, although we do not have enough evidence to attribute this to the peer assessment intervention.

Table 1. The effect sizes of peer assessment in the short term and long term. (Standard errors are shown in parentheses.)

Type of Effect	Effect Size
Short Term (as measured by unit quizzes)	0.115 (0.04)
Long Term (as measured by final exam)	0.122 (0.04)

doi:10.1371/journal.pone.0143177.t001

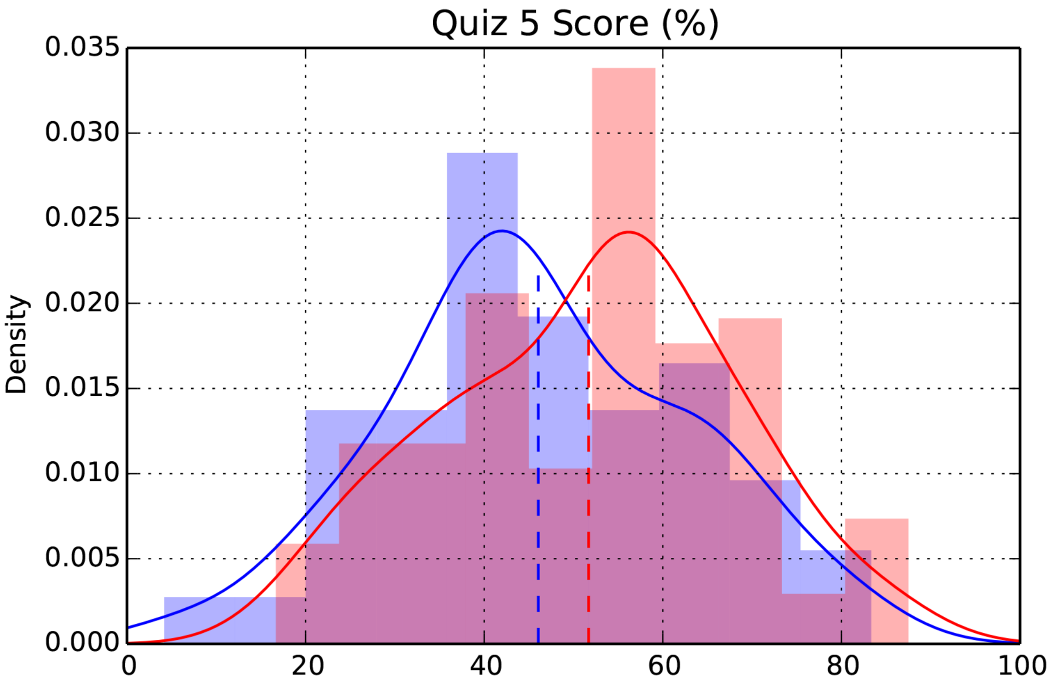

Fig 2. Distribution of scores for the control (blue) and treatment (red) groups on quiz 5 in the winter quarter. The dashed vertical lines designate the means. (The difference in means on this quiz was 5.9.) Similar plots for all of the quizzes and final exam may be found in (S1 File).

doi:10.1371/journal.pone.0143177.g002

We also surveyed students on their perception of the benefit of peer assessment on a scale from 1 to 5, with 1 indicating “not helpful at all" and 5 indicating “extremely helpful". Although the median student reported finding peer assessment only “somewhat helpful," there was virtually zero correlation (r = .01, p = .94) between a student's perception of the benefit and the estimated benefit. This lends credence to our concern that surveys may not be the best measure of student learning. A further analysis of student reaction to peer assessment can be found in (S1 File).

Discussion

This study has established that peer assessment produces concrete gains in student achievement, above and beyond the effect of receiving feedback. Thus, peer assessment is unique among educational interventions in that the usual cost-benefit tradeoff seems not to apply: it saves instructors time, while also benefiting students. This suggests that peer assessment should not be limited to MOOCs and large classes where it is the only option, but that it has a place even in smaller settings where it is not strictly needed.

Table 2. Achievement gaps in our population of students, reported as an effect size. We show the gap before and after the course. (Standard errors are shown in parentheses.) The “before” numbers were calculated using scores on a pre-quiz administered prior to the randomization. The “after” numbers were calculated using scores on the final exam.

Achievement Gap	Difference before Course	Difference after Course
Gender achievement gap (1 = male)	0.32 (0.12)	0.13 (0.12)
Racial achievement gap (1 = underrepresented minority)	−.61 (0.14)	−.42 (0.13)
Statistics background (1 = passed AP stats)	0.54 (0.11)	0.59 (0.12)
Math background (1 = course beyond calculus)	0.68 (0.10)	0.54 (0.11)
Class year (1 = upperclassman)	0.22 (0.12)	0.07 (0.11)

doi:10.1371/journal.pone.0143177.t002

This study is also a demonstration of the role that web-based platforms, such as learning management systems and MOOCs, can play in education research. The ability to personalize an individual's experience using such platforms is easier than in traditional formats. This makes individual-level randomizations, the core of high quality RCTs, much easier. While we have focused on peer assessment specifically, a similar study design could be used to investigate other questions as well. The transformational impact of web-based educational tools may be their ability to facilitate experiments in the classroom, enabling us to obtain unprecedented insight into the learning process.

Supporting Information

S1 File. Supplementary Information for *Peer assessment enhances student learning*.
(PDF)

Acknowledgments

We are grateful to Sergio Bacallado and Renjie You for discussions, as well as to Zhou Fan, Scott Freeman, Max Grazier-G'sell, and Carl Wieman for their careful reading of our paper and helpful suggestions.

Author Contributions

Conceived and designed the experiments: DLS MB GW. Performed the experiments: MB DLS GW NH. Analyzed the data: DLS MB. Contributed reagents/materials/analysis tools: DLS GW. Wrote the paper: DLS MB GW.

References

1. Bangert-Drowns RL, Kulik C-LC, Kulik JA, Morgan M (1991). The instructional effect of feedback in test-like events. Review of Educational Research, 61, 213–238
2. Hattie J, Timperley H. The Power of Feedback Review of Educational Research, 77, 1 (2007).
3. Kulik JA, Kulik CC (1988). Timing of feedback and verbal learning. Review of Educational Research, 58(1), 79–97.).
4. Linn MC, Gerard LF, Ryoo K, Liu L, Rafferty AN (2014). "Computer-guided inquiry to improve science learning." Science, 344, 155–156. doi: 10.1126/science.1245980 PMID: 24723599
5. Leacock C, Chodorow M. Computers and the Humanities. 37, 389 (2003).
6. von Ahn L, Maurer B, McMillen C, Abraham D, Blum M. reCAPTCHA: human-based character recognition via Web security measures. Science 321, 1465 (2008). doi: 10.1126/science.1160379 PMID: 18703711
7. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, et al. Predicting protein structures with a multiplayer online game. Nature 466, 756 (2010). doi: 10.1038/nature09304 PMID: 20686574
8. Falchikov N. Improving assessment through student involvement (Routledge, 2005).
9. Kulkarni C, Wei KP, Le H, Cia D, Papadopoulos K, Cheng J, et al. Peer and self assessment in massive online classes. ACM Trans. Comput.-Hum. Interact. 20 (2013).
10. Topping K. Peer assessment between students in colleges and universities. Rev. Ed. Res. 68, 249 (1998).
11. Falchikov N, Goldfinch J. Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. Rev. Ed. Res. 70, 287 (2000).
12. Bloom BS. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educ. Res. 13, 4 (1984).
13. Topping KJ. Peer Assessment. Theory into Practice 48, 20 (2009).
14. Supreme Court of the United States, Owasso v. Falvo (2002).
15. Tinapple D, Olson L, Sadauska J. CritViz: Web-based software supporting peer critique in large creative classrooms. Bull. IEEE Tech. Com. on Learn. Tech., 15, 1 (2013).

16. Gerdeman RD, Russelll AA, Worden KJ. Web-based student writing and reviewing in a large biology lecture course. J. Coll. Sci. Teaching, 36, 5 (2007).

17. Sadler PM, Good E. The Impact of Self- and Peer-Grading on Student Learning. Ed. Assess. 11, 1 (2006).

18. Freeman S, Parks JW. How Accurate is Peer Grading? CBE Life Sci. Ed. 9, 482 (2010).

19. Miyake A, Kost-Smith LE, Finkelstein ND, Pollock SJ, Cohen GL, Ito TA. Reducing the gender achievement gap in college science: a classroom study of values affirmation. Science 330, 1234 (2010). doi: 10.1126/science.1195996 PMID: 21109670

20. Haak DC, HilleRisLambers J, Pitre E, Freeman S. Increased structure and active learning reduce the achievement gap in introductory biology. Science 332, 1213 (2011). doi: 10.1126/science.1204820 PMID: 21636776