Nonbinding Peer Review and Effort in Teams

Evidence from a Field Experiment

Kristian Behrens
Matthieu Chemin

ABSTRACT

Individuals tend to free-ride in teams, thus providing inefficiently low effort. We implement a system of confidential peer review in a randomly selected set of teams, whereby teammates complete an online survey to review the effort of their peers. These reviews are not linked to any rewards or sanctions, thus making them nonbinding. We find that nonbinding peer reviews increase effort and team productivity and do not decrease worker morale. The effects are stronger for low-ability individuals in low-ability teams, where the traditional forces of peer effects may be absent.

I. Introduction

Effort is usually provided at inefficiently low levels in teams when there is a common payoff. The impossibility of monitoring effort and the existence of a common payoff create an incentive problem (see, for example, Kandel and Lazear 1992; Ledyard 1995; Chaudhuri 2011). One way to increase effort is to reward or punish individuals on the basis of evaluations by teammates, who are, after all, ideally placed to observe individual effort. The issue in practice is that peer monitoring may crowd

Kristian Behrens is a professor of economics in the Department of Economics, Université du Québec à Montréal, Canada; National Research University Higher School of Economics, Russian Federation; and CEPR, UK (behrens.kristian@uqam.ca). Matthieu Chemin is a professor of economics in the Department of Economics, McGill University, Canada; CIREQ, Canada; and CIRANO, Canada (matthieu.chemin@mcgill.ca). *They thank seminar participants at McGill University and UBC for helpful comments and suggestions. Théophile Bougna, Mehnaz Rabbani, and Veronique Russell provided excellent research assistance. Behrens and Chemin gratefully acknowledge financial support from the Social Sciences and Humanities Research Council (SSHRC) of Canada for the funding of the Canada Research Chair in Regional Impacts of Globalization. They also acknowledge funding from the Standard Research Grants program of SSHRC, and from FRQSC Québec. The study has been funded by the Russian Academic Excellence Project '5-100.' The data used in this article can be obtained at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LIKUGT.*
[Submitted July 2017; accepted November 2018]; doi:10.3368/jhr.55.4.0717-8907R2
JEL Classification: J24 and I20
ISSN 0022-166X E-ISSN 1548-8004

out intrinsic motivation (as in Benabou and Tirole 2003; Falk and Kosfeld 2006). Corgnet (2012) shows, for example, that when teammates can evaluate the relative contribution of their peers—and when these peer evaluations are used to split profits unequally—motivation and team productivity do not increase, but instead *decrease*. One explanation is that the unequal split of team output essentially breaks up the team structure: individuals have to work in teams, but are rewarded according to their individual contribution. Low-ability individuals may particularly suffer compared to the case of equal payoffs, since they work with high-ability individuals but cannot benefit from their effort. This may generate negative feelings and loss of motivation.

To avoid this issue, nonbinding peer review may be used. In this case, teammates formally evaluate the performance of their peers and share this information with the principal, but the reviews are not explicitly linked to any rewards or sanctions. Compared to the case of no such peer reviews, the feeling of being observed by peers or by the principal is heightened, which should increase effort if people dislike being observed as shirkers (as shown in Mas and Moretti 2009; Corgnet, Hernan-Gonzales, and Rassenti 2015; Falk and Ichino 2006). As explained above, nonbinding peer reviews are also less likely to generate negative feelings, since they are not used to reward or punish. Yet—being nonbinding—these reviews do not fundamentally alter incentives. It is thus a priori unclear whether they will increase effort.

To evaluate the impact of nonbinding peer reviews on effort, we undertake a randomized field experiment (Chemin 2018). We study 739 students *who are randomly assigned* to teams of four to work on a teamwork task that lasted from two to six weeks. There is a common team grade that counted for 30 percent of the final grade, the difference between A– and F, in a real university course. All students had to fill out an online "opinion survey" three days after the beginning of the teamwork and at the end of the teamwork. These surveys contained fairly innocuous questions on students' experience in the teams (for example, "Do you think the work done helped you develop your teamwork skills?"). The peer review intervention—the treatment—consisted in adding a new section to these opinion surveys in a randomly chosen half of the teams. In those teams, all teammates were asked who in their team: (i) was on time to all team meetings, (ii) respected the deadlines set by the team, and (iii) contributed a fair share to the teamwork. This information was collected once three days after the beginning of the teamwork and once at the end of the teamwork. Thus, all students in the treatment group knew that their teammates were rating their performance during the teamwork. These peer reviews were nonbinding in the sense that we never explicitly linked them to any rewards or sanctions. In fact, we were not even allowed to look at them during the semester because it would have been unethical to have different evaluation criteria for different students in the same classroom (we obtained ethical approval for this research from the relevant ethics review board). In both the treatment and control groups, students could and did complain directly to the principal about free-riders. If no work was provided by a teammate, and that teammate had no valid explanations for his/her behavior, the principal gave them a grade of zero. This happened for two students. It is important to understand that this sanctioning system was the *same* in the treatment and control groups. Hence, the only difference between the treatment and the control groups was an extra online survey of peer ratings three days after the beginning of the teamwork, which created a heightened feeling in the treatment group of being observed and evaluated by peers during the teamwork.

Finally, all students completed a nonbinding peer review at the end of the course after the teamwork grades had been given to them. This ex post review allows us to construct measures of effort for all students. We define several such measures: adherence to "three rules of good behavior" (being on time, respecting deadlines, and contributing a fair share) according to peers; individual contributions to the teamwork on a scale of 0–100, rated by self and by peers; the number of online posts on team forums where students could communicate about the teamwork; the number of posts trying to set up an appointment for the team; the number of posts being proactive about the teamwork; and the number of direct reports by students to the professor about the shirking behavior of their teammates. Our aim is to investigate whether nonbinding peer reviews have a significant effect on these measures of effort and how this effect varies with the ability of the students and of their teammates. We also collected data on team grades and individual test grades to look at the effect of nonbinding peer reviews on productivity.

Previewing our first key result, we find that nonbinding peer reviews: (i) increase effort and team productivity, (ii) increase individual academic performance on later tests, and (iii) do not decrease worker morale. These results are not obvious given the nonbinding nature of the peer reviews: their mere presence—the feeling of being watched—increased effort. They are also important given the ease with which nonbinding peer reviews can be implemented: online surveys are easy to collect even on a large scale, and these peer reviews were not used to sanction students, which could have generated complaints. The practical implications of our findings are that managers should only loosely link peer ratings—already implemented in almost half of U.S. firms (Fisher 2013) under the names of "360 degree performance review," or "multisource feedback," or "crowdsourced performance appraisal" (see Smither, London, and Reilly 2005 for a literature survey)—with actual rewards or sanctions. Nonbinding reviews seem to promote social pressure without being plagued with standard issues associated with consequential reviews, such as manipulability. They can be implemented with much less resistance than binding peer reviews in an organizational setting.

Previewing our second key result, we find significant heterogeneity in the effects: nonbinding peer reviews have stronger effects on low-ability students in low-ability teams. For example, nonbinding peer reviews increase effort by one standard deviation for the lowest-ability students in the lowest-ability teams. The effect is smaller for higher-ability students since they already provide a higher level of effort, which is unlikely to increase further as a result of these nonbinding peer reviews (there are natural "ceiling effects" for effort and performance). Similarly, the effect is smaller in higher-ability teams since in those teams the traditional forces of peer effects—role modeling, leading by example, learning, peer pressure—from high-ability peers are already increasing effort. The fact that nonbinding peer review affects low-ability members in low-ability teams more strongly is important because high-ability peers are scarce in areas with low human capital, in firms relying strongly on low-skilled labor, or in low-quality universities or schools. Even when high-ability peers are present, it can be difficult to make them interact with low-ability students. Carrell, Sacerdote, and West (2013) show that an algorithm producing teams with many high- and low-ability students but few middle-ability students actually had harmful effects on low-ability students: high-ability students sorted themselves into subgroups, leaving low-ability students together in other subgroups. The performance of low-ability students actually decreased

compared to groups with a more even mix (a result confirmed by Booij, Leuven, and Oosterbeek 2017). Page, Putterman, and Unel (2005) and Bandiera, Barankay, and Rasul (2013) also show that individuals tend to sort into teams based on ability, thereby creating both high- and low-ability teams. The issue with low-ability teams is that the traditional forces of peer effects are, at best, absent or, at worst, working in reverse. For example, Gunnthorsdottir, Houser, and McCabe (2007) find that low contributors contribute even less to public goods when assigned to low-contributing teammates. Our paper suggests a mechanism to increase effort even in low-ability teams.

Our results complement an extensive literature on peer effects—in the lab, the workplace, or the classroom—that has found that high-ability peers motivate others through a variety of channels: peer pressure (Mas and Moretti 2009), leading by example (Jack and Recalde 2015), pro-social behavior (Fischbacher, Gächter, and Fehr 2001; Falk and Ichino 2006), learning (Sacerdote 2001; Zimmerman 2003; Foster 2006; Stinebrickner and Stinebrickner 2006; Kang 2007; Lyle 2007; Carrell, Fullerton, and West 2009; Azoulay, Graff, and Wang 2010; Waldinger 2011; Arcidiacono et al. 2012), or any combination of these. We find that nonbinding peer reviews increase the effort of low-ability students in low-ability teams; that is, provide a mechanism that elicits more effort even in contexts where the traditional forces of peer effects may be absent.

The remainder of the paper is organized as follows. Section II details the experimental design of the study and explains how we measure our key variables of interest. Section III lays out our empirical methodology. Section IV provides an overview of the randomization and shows that the observables are balanced across both treatment and control groups and across low- and high-ability teams. Section V describes the effects of nonbinding peer review on various measures of effort, worker morale, team productivity, and individual performance. Finally, Section VI concludes.

II. Experimental Design

Our experiment was implemented over three years in a three-month undergraduate course on economic development taught by the same professor to 739 students in four different classrooms. This course was mandatory for a noneconomics program on international development studies. As such, the students had diverse backgrounds, with 47 percent of the students from humanities (anthropology, English literature, etc.), 13 percent from science, 13 percent from economics, 14 percent from business, and 14 percent from political science. The important part for this paper is the teamwork element. Within each classroom, students were randomly assigned to teams of four, and those teams were randomly assigned to the treatment and the control groups.[1] The randomized assignment to teams is important for two reasons. First, it avoids any problem of self-selection into teams that could influence the outcomes. For

1. The overwhelming majority of the students (89.4 percent) were in teams of four students. When the total number of students was not divisible by four, and the remainder was three, we created one team of three students. When the remainder was two, we created one team of two students. When the remainder was one, we created one team of five students. Because some students dropped out of the course, 9.3 percent of students ended up being in a team of three students. 0.9 percent of students ended up in a team of five students, and 0.4 percent of students ended up in a team of two students. Considering 89.4 percent of students were in a team of four, the typical team is best thought of as having four students. Nonetheless, we control in the subsequent analysis for the size of the team.

example, given that students come from different fields of study, they would tend to sort into groups based on the other students they know from their other courses in their field. Second, it generates substantial variation in the "quality" of teams. The latter is important to test whether there are differential effects of nonbinding peer reviews on effort depending on the quality of the teams and the ability of the students.

We now explain in detail the tasks to be performed (the teamwork), the treatment (nonbinding peer reviews), the way we measure our outcome variable (effort), and the way we account for heterogeneity (ability of the students and the quality of the teams).

A. Teamwork Tasks

There were two teamwork tasks in the course. They were used to increase observations but were otherwise unrelated. In the first one, each team had to download data from the web, analyze it statistically in Excel, and use concepts from the lectures to test theories explained in the course. In 2010, for example, Question 1 involved downloading poverty and foreign aid data to compare the magnitude of both phenomena. Question 5, for example, involved calculating the correlation between GDP and the Human Development Index. This teamwork task, which we refer to as the "statistical teamwork," lasted for two weeks and counted for 15 percent of the final grade.

After this first teamwork task, students were randomly reassigned to new teams of four for a second teamwork task, which consisted of a presentation of an applied academic paper related to topics seen in the lectures. In 2010, 22 papers published since 2004 were selected from *Econometrica*, the *American Economic Review*, the *Quarterly Journal of Economics*, the *Journal of Development Economics*, and the *American Economic Journal: Applied Economics*. All papers involved a randomized experiment such that students from all backgrounds—including nonquantitative ones—could more easily understand and interpret the results than with other more sophisticated econometric techniques. Each team had to present the article in front of the class. This second teamwork task, which we refer to as the "presentation teamwork," lasted between two and six weeks, depending on the order in which the teams had to present.

B. Nonbinding Peer Review Treatment

All students had to fill out an "opinion survey" at the beginning and at the end of the teamwork task. This could be done online on any device. The survey contained fairly innocuous questions on the students' perception of their experience in the teams (for example, "Do you think the work done helped you develop your teamwork skills?"). The responses of each student to the surveys could not be seen by the teammates.

The peer review intervention—the treatment—consisted of adding a new section to these opinion surveys in a randomly chosen half of the teams. In these teams, all teammates were asked who in their team: (i) was on time to all team meetings, (ii) respected the deadlines set by the team, and (iii) contributed a fair share to the teamwork. This information was collected once three days after the beginning of the teamwork task and once at the end of the teamwork task. Thus, all students in the treatment group knew that their teammates were rating their performance during the teamwork task.

As explained above, these peer reviews were never explicitly linked to any rewards or sanctions. One may thus wonder how such "nonbinding" peer reviews may affect

behavior since they do not change incentives. It is known that the mere presence of others—and the associated monitoring and evaluation—is enough to affect behavior and elicit more effort. This phenomenon—that people tend to perform better when being watched or when competing with others—is well studied in the social psychology literature and referred to as "social facilitation" (see, for example, Allport 1920, 1924; Zajonc 1965) or "evaluation apprehension" (see, for example, Cottrell et al. 1968; Henchy and Glass 1968). Zajonc (1965) shows, for example, that audience effects (the observation of behavior as it occurs) or co-action effects (the presence of others engaged in the same activity) can increase performance by stimulating arousal.[2] We thus conjecture that nonbinding peer review may increase effort.

C. Measures of Effort

To obtain measures of effort, we collected another opinion survey from all students at the end of the course, after the final teamwork project grades were submitted. This final survey included the peer review section for all students. We thus use the information collected there to create various measures of effort.

First, we create a measure that captures "good behavior in teams." Students were repeatedly told in class that they need to adhere to three "rules of good behavior": (i) being on time to team meetings, (ii) respecting team deadlines, and (iii) contributing a fair share to the teamwork. We create a dichotomous variable equal to one if all three teammates report that the student respected these three rules, and zero otherwise. This is our simplest measure of effort.[3] The issue with this measure is that it is binary and may thus appear quite rough: it equals zero if a single of the three teammates answers negatively to a single question of the three questions about good behavior. To refine that measure, we also create a more continuous version that takes values on a scale from 0 to $3 \times$ (team size – 1), that is, 9 in a team of four, which reports the number of positive answers to these three questions given by the teammates.

Second, we measure effort by the individual contributions to the teamwork on a scale from 0 to 100, rated by self and by peers in the final opinion survey. We asked students to evaluate their contribution, and that of their team members, with a number ranging from 0 to 100 percent. For example, if everybody contributed equally, we instructed students to report: 25, 25, 25, 25. If the third student did everything, for example, we instructed students to report: 0, 0, 100, 0. This provides a continuous measure of effort as viewed by the student, or by their teammates. In the latter case, we use the average reported by the teammates as our measure of effort for the student.

Third, we use the number of online posts on team forums where students could communicate about the teamwork. We collected this objective measure of effort by opening online forums for each team, gathering all messages on these forums, and matching them with their authors. A distinct advantage of this measure is that it is not

2. Social facilitation theory states that an audience increases performance—via enhanced arousal—but impedes learning—by distracting. This would increase performance on simple tasks, but decrease performance on complex tasks (Zajonc 1965). Though interesting, it is unclear in our study which of the two tasks—the presentation teamwork or the statistical teamwork—is "simple" or "complex." Analyzing data in Excel and reading regression results in a scholarly journal are probably equally challenging for students.

3. The rationale for such a binary measure is that the three rules of good behavior were explicitly told and repeated to students. Thus, even one peer who mentions that a student was not on time, did not respect the deadlines, or did not contribute a fair share of work is indicative of an effort problem.

top-coded. Since some of the online posts are only loosely related to effort, we code the content of the 81,888 posts. We count the sum of posts per student that try to set up an appointment for the team and the posts signaling a "proactive attitude" about the teamwork.[4] We take the number of such posts as our measure of effort.

Last, we also use direct reports by students to the professor about the shirking behavior of their teammates. A number of students directly reported "extreme shirkers" to the professor, that is, individuals not responding to emails, not coming to meetings, not respecting the deadlines, and not contributing any work.[5] We create a dummy variable equal to one if the student was reported as extreme shirker by his teammates, and zero otherwise.

D. Student and Team Abilities

The last piece of information we need—in order to assess the heterogeneity in treatment across individuals and teams—is a measure of individual and team abilities. To measure ability, we conducted individual tests. Five such tests, counting for three points each (together 15 percent of the final grade) were undertaken before and after the teamwork tasks. Test 1 was undertaken before any teamwork. Hence, Test 1 serves as the baseline for assessing the ex ante ability of students before the statistical teamwork. Test 2 took place two weeks after Test 1, that is, immediately after the statistical teamwork. Test 3 took place immediately before the presentation teamwork (hence serving as the baseline for assessing the ex ante ability of students before the presentation teamwork). Test 4 occurred two weeks after the start of the presentation teamwork. Test 5 occurred at the end of the course after all presentations were completed.[6] The questions in the five individual tests were related to the course material.[7]

4. A research assistant coded the contents of these online posts according to the two criteria. We found 170 posts trying to set up an appointment in the group. For example, "Hey guys! When do you want to get together to work on the project? When is everyone available? I'm free every day weekday after 5:30 and this weekend but I'm going home for Thanksgiving next Friday (I'm guessing you all are too!) so hopefully we can get this done either before or after the holiday!" We found 423 posts where the student was being proactive, that is, was trying to take active steps to complete the teamwork. Examples are: "Hello all, I hope you don't mind but I booked a room for us until 9:00 pm. It would be good for us to all come prepared, that is, having read the paper thoroughly and roughly answered the questions from the questionnaire sheet. This way we could be extra productive during the meeting." Other students organized a Dropbox, created Facebook groups, discussed the work, or emailed the teaching assistant. We counted the sum of posts setting up appointments or being proactive per student.

5. There is ample anecdotal evidence that effort provision was low in our experiment: 24 teams (that is, 8 percent of all teams) formally complained of extreme shirkers. Less anecdotally, our estimates—explained in greater detail below—show that, according to their peers, only 60 percent of individuals respected the three rules of good behavior. This means that 40 percent of students were either not on time for meetings, did not respect the deadlines set by the team, or did not contribute a fair share of the teamwork. In other words—and as expected in a teamwork setting with a common payoff—free-riding was endemic in our experiment.

6. Students with a valid reason (for example, health reasons as certified by a doctor's note, work-related reasons as certified by an employer's note, etc.) were exempted from these tests. This happened for 1 percent of the students. Students without a valid reason got a grade of zero.

7. For example, the questions of Test 1 in 2010 were:

1. What is the implication in terms of convergence of the Solow model (without technical progress)?
2. Today is the United Nations Millennium Development Goals (MDG) Summit. Cite one of the eight MDGs.
3. A country has no population growth rate, no depreciation, a propensity to save of 10 percent, and a capital output ratio of 2.

To obtain a measure of the ability of student *i*, we use the student's score on the individual baseline test that predates the teamwork (Test 1 for the statistical teamwork, and Test 3 for the presentation teamwork). Using measures that predate the teamwork allows us to sidestep the reflection problem that usually plagues the empirical estimation of peer effects (Manski 1993). Last, the ability of the team to which student *i* is assigned is defined as the average baseline ability of all teammates, excluding *i*.

III. Estimating Equation

To identify the causal effect of nonbinding peer reviews on effort, we estimate the following specification:[8]

$$\begin{aligned} Effort_{ig} = {} & \beta_0 + \beta_1 PeerReview_g + \beta_2 BaselineAbility_{i,t-1} \\ & + \beta_3 PeerReview_g \times BaselineAbility_{i,t-1} + \beta_4 TeamAbility_{-i,t-1} \\ & + \beta_5 PeerReview_g \times TeamAbility_{-i,t-1} \\ & + \beta_6 PeerReview_g \times TeamAbility_{-i,t-1} \times BaselineAbility_{i,t-1} \\ & + \beta_7 TeamAbility_{-i,t-1} \times BaselineAbility_{i,t-1} + X_i\zeta + \varepsilon_i \end{aligned}$$

where *ig* stands for individual *i* in team (group) *g*; $PeerReview_g$ is a dichotomous variable equal to one if the individual is in a team doing peer review, and zero otherwise; $BaselineAbility_{i,t-1}$ is individual ability on the baseline Test 1 or Test 3 (done before each teamwork, hence the index $t-1$); $TeamAbilty_{-i,t-1}$ is baseline ability of the team excluding oneself (hence the index $-i$); and where $PeerReview_g \times BaselineAbility_{i,t-1}$, and $PeerReview_g \times TeamAbility_{-i,t-1}$, and $TeamAbility_{-i,t-1} \times BaselineAbility_{i,t-1}$, and $PeerReview_g \times TeamAbility_{-i,t-1} \times BaselineAbility_{i,t-1}$ are interaction terms between the peer review indicator, the team's baseline ability, and the individual's baseline ability. The coefficient β_1 tests whether nonbinding peer review increases effort. Since $BaselineAbility_{i,t-1}$ and $TeamAbility_{-i,t-1}$ are included in the regression, the interpretation of β_1 is at $BaselineAbility_{i,t-1} = 0$ and $TeamAbility_{-i,t-1} = 0$, that is, low-ability students in low-ability teams. The coefficient β_3 measures the interaction between peer review and individual ability. A negative coefficient means that the effect of nonbinding peer review is stronger for low-ability students. The coefficient β_4 measures peer effects, that is, the effect of being grouped with high-ability teammates on individual effort. Turning to β_5, a negative coefficient means that the effect of nonbinding peer review is weaker in high-ability teams. Finally, for completeness, we also run fully saturated specifications including the remaining two interactions that are measured by β_6 and β_7.

---

a. According to the Harrod–Domar model, what is the growth rate of capital stock ($\Delta k/k_{t-1}$)?
b. What if the propensity to save goes up to 12 percent?
c. What if the capital output ratio goes down to 1 (and propensity to save still at 10 percent)?
d. What is the policy recommendation for this country?
e. How, in practice, could *s* be increased?
f. Why might these kinds of calculations fail to capture important elements of growth?

8. We estimate the model by ordinary least squares. Although we could use a model with a limited dependent variable, for example, a probit model, incorporating and interpreting interaction effects in those models is quite complicated (Greene 2010).

Table 1
Balance of Observable Characteristics

	Control vs. Peer review		
	Control	Peer review	Difference (p-Value)
Gender: Female (0, 1)	0.65	0.64	0.01 (0.75)
Field of study			
Humanities	0.47	0.48	–0.01 (0.78)
Business	0.06	0.06	0.00 (0.80)
Political science	0.14	0.15	–0.01 (0.57)
Economics	0.13	0.15	–0.02 (0.37)
Science	0.13	0.14	–0.00 (0.87)
Ethnic group			
East Asian	0.08	0.11	–0.03 (0.14)
Black	0.04	0.03	0.00 (0.78)
South Asian	0.05	0.06	–0.01 (0.39)
Caucasian	0.77	0.72	0.05 (0.10)
Middle Eastern	0.06	0.07	–0.01 (0.61)
Year of study			
Second year	0.19	0.23	–0.03 (0.25)
Third year	0.55	0.50	0.06 (0.13)
Fourth year	0.24	0.27	–0.03 (0.34)
Baseline ability	1.61	1.58	0.03 (0.64)
Team ability	1.60	1.58	0.02 (0.69)

Notes: Equality-of-means tests for the balancing of treatment and control groups for 739 students. p-values are given in parentheses.

As explained above, to solve the self-selection problem students were randomly allocated to teams. To control for any differences between control and treatment groups that could remain despite this random allocation, we finally also include a vector X_i of control variables. Our controls include variables related to the student's gender, field of study, ethnic group, and year of study (see Table 1 for a full list of our individual controls). We also include controls at the team level, namely measures of diversity in the team based on these four factors: the number of females in the team, the number of different fields of study present in the team, the number of different ethnic groups in the team, and the number of third- or fourth-year students in the team.[9] We also include the variance in team ability (to control for the presence of "star students" in the team) and a dummy for the statistical teamwork to control for different effort levels and difficulty in both teamwork tasks. Last, a set of classroom dummies, to control for the fact that this course was taught in four different classrooms, and a dummy equal to one if the group was of three students or fewer, and zero otherwise, complete our controls. In all regressions, we report robust standard errors clustered at the team level.

IV. Randomization and Balance of Observable Characteristics

Figure 1 shows the distribution of team abilities ($TeamAbilty_{-i,t-1}$) in our randomly generated teams. As shown, there is substantial variation in team ability due to the random assignment. For example, for six students, the average ability of their teammates was zero at the baseline test, thus placing them in very low-ability groups.

Table 1 summarizes the balance of all observable characteristics across control and treatment groups. As shown, the proportions of females in the control and treatment groups are 65 and 64 percent (see Columns 1 and 2). These proportions are not statistically different, as indicated by the p-values of a T-test (in parentheses) in Column 3. Put differently, the proportion of females is balanced across control and treatment groups, so any effect on outcomes is not due to varying gender composition.

As explained above, the students are from a diverse set of study fields. Their proportions in the different teams do not differ significantly between the control and the treatment groups. Neither do the ethnic and year-of-study compositions. Seventy-seven percent of the class is Caucasian in the control group, slightly more than in the treatment group. Fifty-five percent of students are in their third year, compared to 50 percent in the treatment group. In our analysis below, we control for these factors to ensure that our results are not driven by these small differences.

Last, we verify that the control and treatment groups have no significant difference in the students' abilities. As Table 1 shows, average baseline ability and team ability are similar in all groups. They are, on average, 1.61 and 1.58 (out of 3) across the control and treatment groups, respectively. Hence, control and treated students were placed on average in similar teams.

9. Because of confidentiality reasons, we had no access to the sociodemographic data of the students. Hence, we coded their ethnic background from their pictures. The pictures were coded by two independent raters, and the two series have a correlation of 0.8.

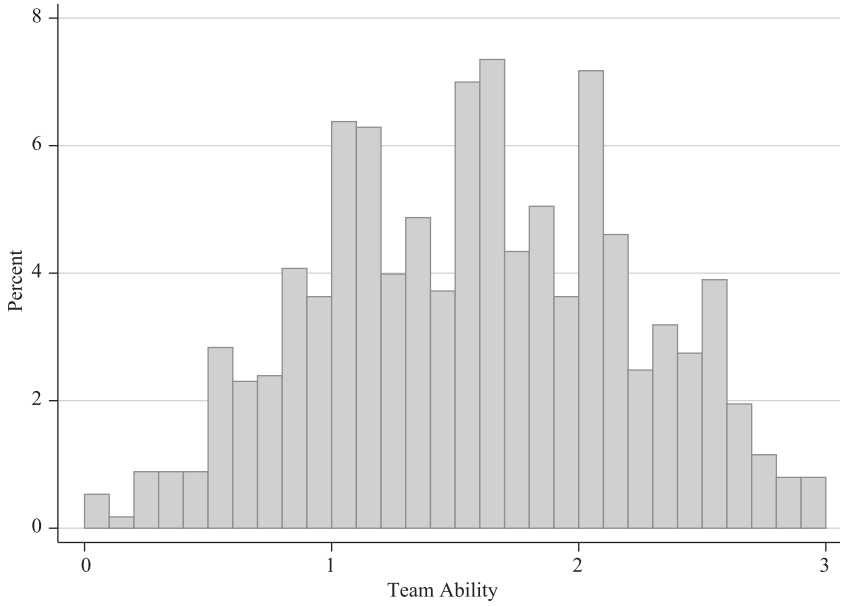

Figure 1

Distribution of Team Ability

Notes: "Team ability" for student *i* is computed as the simple average of the other team members' individual scores on the baseline test, which takes values from zero to three.

The random allocation of individuals to low- and high-ability teams is as important as the random allocation of teams to the control and treatment groups. Table 2 reports the statistical test for random assignment to groups based on Guryan, Kroft, and Notowidigdo (2009). This test is an ordinary least squares (OLS) regression of individual *i*'s predetermined characteristics on the average ability of *i*'s peers (conditional on any variable on which randomization was conditioned) and conditional on the mean ability of all individuals in the classroom, excluding individual *i*. The last point controls for the fact that sampling of peers is done without replacement, so that the peers for high-ability individuals are chosen from a group with a slightly lower mean ability than the peers for low-ability individuals. Table 2 reports small and insignificant conditional correlations between the individual's predetermined characteristics—including our two measures of the own predetermined ability from Test 1 and Test 3 in Columns 16 and 17—and partners' abilities, consistent with students being randomly assigned to teammates of different ability. We also test the random assignment of three additional variables that we will not use in what follows ("English name," "French name," and "Smile"; see notes for Table 2) to check for additional potentially unobserved differences in the randomization.

Given that the observable characteristics are similar across control and treatment groups—and because of the random assignment of individuals to low- and high-ability teams, including in terms of individual ability—we can simply compare control and treatment groups to isolate the causal impact of nonbinding peer review on effort.

Table 2
Test of Random Assignment to Groups

	Field of Study					Gender	Ethnic Background				
	Humanities	Business	Political science	Economics	Science	Female	East Asian	Black	South Asian	Caucasian	Middle Eastern
Team ability	–0.04	0.02	0.01	0.01	0.00	–0.03	0.00	–0.01	0.02	–0.01	–0.01
	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
Leave-me-out mean class score	0.06	0.01	0.08	–0.05	–0.03	–0.05	0.10	–0.04	–0.01	–0.07	0.02
	(0.11)	(0.06)	(0.08)	(0.08)	(0.08)	(0.11)	(0.10)	(0.07)	(0.06)	(0.11)	(0.08)
Observations	729	729	729	729	729	729	729	729	729	729	729
R^2	0.021	0.013	0.004	0.025	0.006	0.017	0.021	0.022	0.011	0.038	0.022

(*continued*)

Table 2 (*continued*)

	Year of study				Baseline ability	Additional		
	First year	Second year	Third year	Fourth year	Test 1&3	French name	English name	Smile
Team ability	–0.01	0.01	–0.01	0.01	–0.08	0.01	–0.02	–0.03
	(0.01)	(0.02)	(0.02)	(0.02)	(0.06)	(0.02)	(0.03)	(0.02)
Leave-me-out mean class score	0.00	0.03	0.02	–0.05	0.77***	–0.03	0.06	0.00
	(0.04)	(0.11)	(0.13)	(0.11)	(0.24)	(0.08)	(0.14)	(0.11)
Observations	729	729	729	729	729	729	729	729
R^2	0.005	0.016	0.031	0.010	0.211	0.015	0.011	0.046

Notes: Based on Guryan, Kroft, and Notowidigdo (2009). Robust standard errors in parentheses, clustered at the team level. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The dependent variables are all the variables of Table 1, as well as three predetermined variables: "French name" equals one if the student has a French sounding name, zero otherwise, "English name" is defined similarly. "Smile" is a dichotomous variable equal to one if the student has a large smile on the class picture, zero if no smile or a normal smile. The main explanatory variable is "Mean group score." "Leave-me-out mean class score" is the mean of the class score excluding individual *i*. "Ability" is the individual score on the baseline individual Test 1 (done before the statistical teamwork) and on the baseline individual Test 3 (done before the presentation teamwork).

V. Results

A. Good Behavior in Teams

We first present results for our dichotomous measure of effort, defined as the adherence to the three rules of good behavior. Column 1 of Table 3 shows that nonbinding peer review has a positive, yet insignificant, effect on effort in the simplest specification. As shown in Column 2 and as expected, the coefficient of "Baseline ability" is positive and significant, indicating that individual ability is correlated with effort. The interaction term between "Baseline ability" and "Peer review" is negative, but insignificant.

In Column 3, we control for team ability and include an interaction term between peer review and team ability. Peer review increases effort of students in a team of zero ability by 28 percentage points. The interaction term is negative at –0.09, which shows that the effect of nonbinding peer review decreases with team ability. Hence, the effect of nonbinding peer review is stronger for students in low-ability teams. In fact, the coefficient of the interaction term is approximately equal in absolute value to the coefficient of team ability.[10] One way to understand this result is as follows. Teammates' ability increases effort exerted by low-ability students in the control group through peer effects. In teams not doing peer review, a one standard deviation increase in teammates' ability increases effort of low-ability students by nine percentage points.[11] In teams doing peer review, a one standard deviation increase in teammates' ability increases effort of low-ability students by $-0.09 + 0.09 = 0$, that is, by zero percentage points. This shows that peer effects crowd out the effectiveness of peer review. Overall, the effect of peer review is stronger for low-ability students in low-ability teams.

In Column 4, we include both individual and team abilities, and all interaction terms. Nonbinding peer review increases effort of students with zero ability in teams with zero team ability by 45 percentage points, a full standard deviation of effort. The effect of peer review decreases with individual and team ability, since the coefficients of "Peer review × Baseline ability" and "Peer review × Team ability" are both negative, albeit insignificant.

Another way to see the stronger impact of peer review on low-ability students is to split the sample into low- and high-ability students. In Column 5, the sample is restricted to low-ability individuals, defined as individuals scoring below the median on the baseline test. Peer review increases effort by 37 percentage points and is crowded out by team ability as before. In Column 6, the sample is restricted to high-ability individuals, defined as individuals scoring above the median on the baseline test. For these students, the effect of peer review is weaker, only 23 percentage points (not significantly different from zero). Thus, peer review disproportionately affects low-ability

10. The sum of coefficients is not significantly different from zero, as indicated by the *p*-value of the *T*-test "Team ability" + "Peer review × Team ability" = 0, called "*p*-val sum. coeff. peer review."

11. This estimate is consistent with the literature on peer effects. Considering the standard deviation of effort is 0.48, a nine percentage point increase means a (0.09/0.48 = 0.19) standard deviation increase of effort. Mas and Moretti (2009) find that a one-unit increase in coworkers' productivity increases speed of supermarket cashiers by 0.15 units. Falk and Ichino (2006) find that a one-unit increase in coworkers' effort increases output—measured by the number of stuffed envelopes—by 0.14 units.

Table 3
Impact of Nonbinding Peer Review on Effort (Good Behavior in Teams)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Peer review	0.06	0.13	0.28**	0.45*	0.37**	0.23	0.47*	0.50**
	(0.05)	(0.08)	(0.13)	(0.24)	(0.15)	(0.18)	(0.24)	(0.24)
Baseline ability		0.08***		0.12			0.12	0.12
		(0.03)		(0.09)			(0.09)	(0.09)
Peer review × Baseline ability		–0.05		–0.11			–0.12	–0.13
		(0.04)		(0.13)			(0.14)	(0.13)
Team ability			0.09**	0.12*	0.11**	0.07	0.14**	0.14**
			(0.04)	(0.07)	(0.05)	(0.05)	(0.07)	(0.07)
Peer review × Team ability			–0.09*	–0.14	–0.11**	–0.07	–0.15*	–0.15*
			(0.05)	(0.09)	(0.05)	(0.07)	(0.09)	(0.09)
Team ability × Baseline ability				–0.02			–0.02	–0.02
				(0.03)			(0.03)	(0.03)
Peer review × Team ability × Baseline ability				0.03			0.04	0.04
				(0.05)			(0.05)	(0.05)

(*continued*)

Table 3 (*continued*)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Observations	730	719	730	719	417	302	719	719
R^2	0.007	0.018	0.019	0.032	0.045	0.017	0.051	0.067
Mean control group[a]	0.64	0.64	0.64	0.64	0.61	0.68	0.64	0.64
SD control group	0.48	0.48	0.48	0.48	0.49	0.47	0.48	0.48
p-val. sum. coeff. peer review			0.98	0.83	0.99	0.91	0.87	0.91

Notes: OLS regressions. Robust standard errors in parentheses, clustered at the team level. $*p<0.1$, $**p<0.05$, $***p<0.01$. The dependent variable is effort, a dichotomous variable equal to one if all teammates report that individual *i* adhered to the three rules of good behavior. "Peer review" is a dichotomous variable equal to one for teams that could use the peer review system, zero otherwise. "Baseline ability" is individual ability on the baseline test. "Peer review × Baseline ability" is an interaction term between peer review and baseline ability. "Team ability" is the average ability of teammates (excluding oneself) on the baseline test. "Peer review × Team ability" is an interaction term between peer review and team ability. "Team Ability × Baseline Ability" is an interaction term between team ability and baseline ability. Last, "Peer review × Team ability × Baseline ability" is the triple interaction term. In all columns, the regressions include a dummy for the statistical teamwork, and three classroom dummies to control for the fact that this course was taught in four different classrooms. In Column 5, the sample is restricted to low-ability individuals, defined as individuals scoring below the median on the baseline test. The difference in sample sizes for high- and low-ability is due to the fact that we define "high ability" as being strictly above the median. Since grades are discrete variables (they vary from 0 to 3 in 0.5-point steps), and since students at the median are included in the low-ability sample, this explains the difference. Our results are robust to the redefinition of high-ability that excludes the students at the median. In Column 6, the sample is restricted to high-ability individuals, defined as individuals scoring above the median on the baseline test. In Column 7, the additional control variables are: gender, field of study, ethnic group, and year of study. In Column 8, in addition we include control variables defined at the team level: the variance of baseline ability in the team, the number of females in the team, the number of different fields in the team, the number of different ethnic groups in the team, the number of third and fourth year students in the team, and a dummy equal to one if the group was of three students or fewer, and zero otherwise.

[a]This is the mean value of the dependent variable across all students in the control group, irrespective of the ability of the team they are in. It should thus not be interpreted as the mean value for an individual in a zero-ability team.

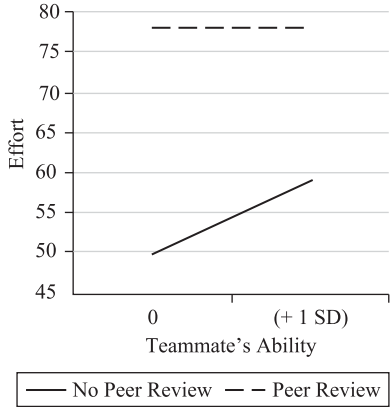

Figure 2

Impact of Nonbinding Peer Review on Effort (Good Behavior in Teams)

Notes: Difference between treatment and control groups for the case of low-ability students, based on the estimates in Column 5 of Table 3. The 37 percentage point difference at "Teammates ability" zero is the effect of peer review. The gap between control and treatment groups shrinks since peer review becomes less efficient as "Teammates ability" increases.

students. This heterogeneity may be due to ceiling effects: it is harder for students who provide effort to increase effort even further. This may also explain why the coefficient in Column 1 is not significant.

Figure 2 depicts the results taken from the sample of low-ability students in Column 5. It illustrates that nonbinding peer review increases effort in all teams. In the absence of peer review, effort increases with the team's ability, the standard result from the peer effects literature. However, peer review is crowded out by peer effects. Thus, the effect of peer review is smaller in higher-ability teams, and the effect of team ability becomes largely unimportant once nonbinding peer review is implemented, as shown by the horizontal upper line.

Our foregoing results are robust to the inclusion of all individual controls from Table 1 (gender, field of study, ethnic group, and year of study), as can be seen from Column 7. They are also robust to the inclusion of team controls (the number of females in the team, the number of different fields in the team, the number of different ethnic groups in the team, the number of third- or fourth-year students in the team, and a dummy equal to one if the group was of three students or fewer, and zero otherwise), as Column 8 shows. We provide the estimated coefficients for the control variables in Table A1 in Appendix 1. The only two significant effects that we find are (i) that female students exert greater effort, everything else equal, and (ii) that the statistical teamwork elicited less effort and was prone to more shirking from the students.[12]

12. One may be worried about the generality of those findings. Indeed, our experiment was implemented in a university ranked among the top 40 in the world (according to the "Center for World Universities" ranking, the "Academic Ranking of World Universities," the Times, and the "QS World Universities" ranking). To verify

B. Other Measures of Effort

Our dichotomous measure—good behavior in teams—may be too coarse and may not adequately capture "effort." Recall that it equals zero if any teammate answers negatively to one of the three questions about good behavior.[13] Hence, a brilliant student who comes late to the team meetings would be coded as zero. We now show that our results (Table 4) are robust to a large range of alternative measures of effort.

Column 1 of Table 4 simply replicates our main result of Column 8 of Table 3. In Column 2 of Table 4, we instead use the simple average for the three questions about good behavior given by the three teammates. The results are very similar and show that nonbinding peer review increases effort of low-ability students in low-ability teams by 23 percentage points.

In Column 3 of Table 4, we present the results of another measure of effort, where we assess quantitatively the individual contribution to the teamwork. The dependent variable is the student's contribution to the teamwork—on a scale from 0 to 100 percent—as rated by self. Peer review increases contribution to the teamwork by 2.60 percentage points for the lowest-ability students in the lowest ability teams, a result that is not statistically significant.

Of course, students could have false self-perceptions. In Column 4 of Table 4, the dependent variable is the student's contribution to the teamwork task—again on a scale from 0 to 100 percent—as rated by the others in the team. We use the simple average of the student's contribution rated by the three teammates. Again, peer review increases this measure of effort by 1.53 percentage points for the lowest-ability students in the lowest-ability teams, though not significantly.

There are two issues with these percentage measures of effort. First, not all students answered the question correctly. In Column 3, there are only 474 answers out of a total of 739 observations. This may be because students were uncomfortable answering the question, had difficulties rating their peers on a scale of 0 to 100, or simply did not understand the question. The second issue is that there is very little variation in the data, with an overwhelming proportion of students answering "25 percent" (for themselves and for the others). As can be seen from the top panel of Figure 3, more than 80 percent of students (who answered the question correctly) rated themselves "25 percent." The bottom figure shows the student's contribution, as rated by their peers. According to

---

the scope of our results, we replicated the experiment in a university below the 400th place in those same rankings. We relegate a more detailed discussion to Appendix 2 (see Table A2 for the results). The key finding is that nonbinding peer review continues to have a positive effect, whereas team ability no longer matters very much. We view these results as confirmation that even when peer effects are not operational—because there are few high-ability peers, that is, the average quality of teams is low—nonbinding peer review still works to increase effort.

13. Another issue with this measure of effort may be top-coding for high-ability students, that is, the existence of ceiling effects. If all high-ability students score high on this measure, this will make the estimation of coefficients for high-ability students impossible. This is not an issue in our case. As can be seen from "Mean control group" in Column 6 of Table 3, the average effort for high-ability students is only 0.68, less than the maximum score of one. In other words, only 68 percent of high-ability students adhere to the three rules of good behavior, according to their peers. Observe that this is not very different from the fraction of low-ability students reported to exert little effort, which stands at 0.61. Thus, there is considerable room for improvement even for high-ability students.

Table 4
Impact of Nonbinding Peer Review on Various Measures of Effort and Other Outcomes

	Effort (1)	Effort: Mean (2)	Contribution (Rated by Self) (3)	Contribution (Rated by Peers) (4)	Postings (5)	Postings: Appointment (6)	Postings: Proactive (7)	Extreme shirker reported (8)	Enthusiasm (9)	Increase in Grade (10)
Peer review	0.50**	0.23*	2.60	1.53	6.14***	0.31*	0.68**	–0.16**	0.10	1.13***
	(0.24)	(0.12)	(2.84)	(1.45)	(1.84)	(0.17)	(0.30)	(0.08)	(0.11)	(0.35)
Baseline ability	0.12	0.08**	1.17	0.20	2.35***	0.19**	0.13	–0.07**	0.00	–0.26
	(0.09)	(0.04)	(1.47)	(0.47)	(0.77)	(0.08)	(0.11)	(0.03)	(0.05)	(0.18)
Peer review × Baseline ability	–0.13	–0.05	–1.77	–1.22*	–2.80***	–0.22**	–0.36**	0.08*	–0.04	–0.52**
	(0.13)	(0.06)	(1.65)	(0.73)	(0.93)	(0.11)	(0.18)	(0.04)	(0.07)	(0.23)
Team ability	0.14**	0.08***	0.09	–0.09	2.18***	0.08	0.12	–0.04***	0.02	0.32***
	(0.07)	(0.03)	(0.82)	(0.40)	(0.61)	(0.06)	(0.09)	(0.02)	(0.04)	(0.12)
Peer review × Team ability	–0.15*	–0.07	–1.07	–0.39	–2.69***	–0.14**	–0.28**	0.07**	–0.03	–0.35**
	(0.09)	(0.04)	(0.91)	(0.53)	(0.66)	(0.07)	(0.13)	(0.03)	(0.05)	(0.13)
Team ability × Baseline ability	–0.02	–0.02	–0.42	0.04	–0.75**	–0.07**	–0.03	0.02*	0.00	–0.15**
	(0.03)	(0.01)	(0.46)	(0.16)	(0.29)	(0.03)	(0.04)	(0.01)	(0.02)	(0.06)
Peer review × Team ability × Baseline ability	0.04	0.01	0.66	0.31	0.97***	0.08**	0.12	–0.04**	0.01	0.16**
	(0.05)	(0.02)	(0.53)	(0.24)	(0.35)	(0.04)	(0.07)	(0.02)	(0.03)	(0.08)

(*continued*)

Table 4 (*continued*)

	Effort (1)	Effort: Mean (2)	Contribution (Rated by Self) (3)	Contribution (Rated by Peers) (4)	Postings (5)	Postings: Appointment (6)	Postings: Proactive (7)	Extreme shirker reported (8)	Enthusiasm (9)	Increase in Grade (10)
Observations	719	719	474	666	729	729	729	729	611	720
R^2	0.067	0.077	0.098	0.067	0.165	0.036	0.084	0.185	0.043	0.441
p-val. sum. coeff. peer review	0.91	0.80	0.02	0.20	0.16	0.11	0.08	0.31	0.82	0.81
Mean control group[a]	0.64	0.88	25.90	24.83	4.26	0.17	0.43	0.03	0.91	–0.24
SD control group	0.48	0.21	4.46	3.64	4.21	0.41	0.81	0.17	0.28	1.20

Notes: OLS regressions. Robust standard errors in parentheses, clustered at the team level. $*p<0.1$, $**p<0.05$, $***p<0.01$. In Column 1, the dependent variable is a dichotomous variable equal to one if all teammates report that individual *i* adhered to the three rules of good behavior. In Column 2, the dependent variable is the average on the three rules of good behavior given by the three teammates. In Column 3, the dependent variable is the percentage that student *i* reported as being her own contribution to the teamwork. In Column 4, the dependent variable is the percentage that the other team members reported as being the contribution of student *i* to the teamwork. In Column 5, the dependent variable is the number of postings in team online forums. In Columns 6 and 7, we break down the number of postings into those where students organized appointments ("appointment") and in which students took the lead on organizing the teamwork ("proactive"). In Column 8, the dependent variable is a dichotomous variable equal to one if the the student was reported directly to the professor as a shirker by teammates, and zero otherwise. In Column 9, the dependent variable is a dichotomous variable equal to one if the student rated herself as showing enthusiasm and a positive attitude in the team. Finally, in Column 10, the dependent variable is the increase in scores between the individual baseline test (before the teamwork) and the individual endline test (after the teamwork). "Peer review" is a dichotomous variable equal to one for teams that could use the peer review system, zero otherwise. "Baseline ability" is individual ability on a baseline test. "Peer review × Baseline ability" is an interaction term between peer review and baseline ability. "Team ability" is the average ability of teammates (excluding oneself) on the baseline test. "Peer review × Team ability" is an interaction term between peer review and team ability. "Team ability × Baseline ability" is an interaction term between team ability and baseline ability. Last, "Peer review × Team ability × Baseline ability" is the triple interaction term. In all columns, the regressions include a dummy for the statistical teamwork; the variance of baseline ability in the team; three classroom dummies to control for the fact that this course was taught in four different classrooms; individual-level controls (gender, field of study, ethnic group, year of study); and team-level controls (the number of females in the team, the number of different fields in the team, the number of different ethnic groups in the team, the number of third- and fourth-year students in the team; and a dummy equal to one if the group was of three students or fewer, and zero otherwise.

[a]This is the mean value of the dependent variable across all students in the control group, irrespective of the ability of the team they are in. It should thus not be interpreted as the mean value for an individual in a zero-ability team.

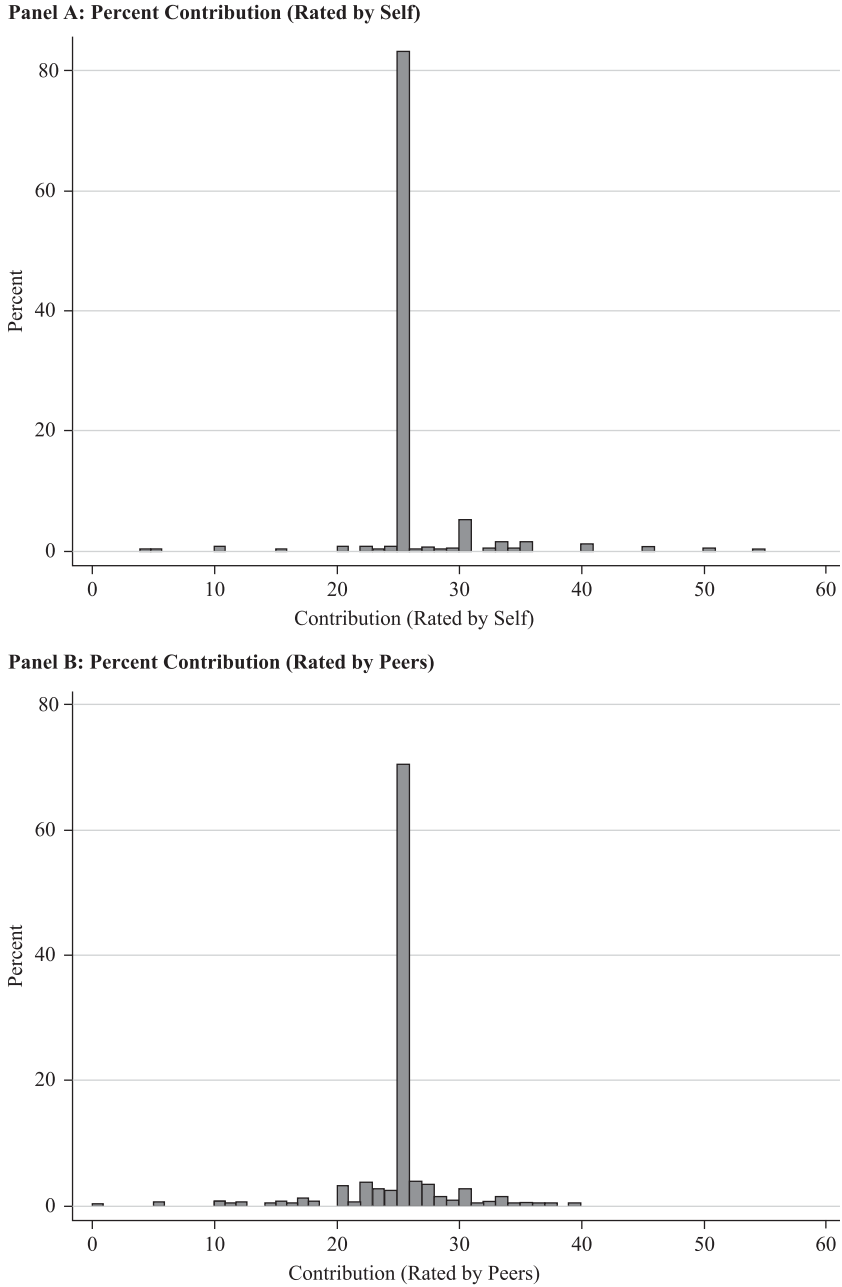

Figure 3

Distribution of Reported Percentage Contributions to the Teamwork

Notes: Each student was asked to report the percentage that each team member (including themselves) has contributed to the teamwork. A number of students did not report sensible answers in terms of percentages that sum to 100 percent. We exclude those from the figure and the analysis when using these reported values as our measure of effort.

their peers, 70 percent of students contributed "25 percent." This illustrates a pitfall of individual ratings of others: students may be reluctant to negatively rate their teammates.[14] By contrast, our basic measure of effort—based on a dichotomous variable equal to one if all three other teammates answer one to the three rules of good behavior—maximizes variation: the mean is 0.64, and the standard deviation is 0.48.

We next present results for another measure of effort where top-coding and reporting bias are not an issue: the number of online postings on team forums where students could communicate about the teamwork. One big advantage of this measure is that it is objective (that is, not reported by peers, and therefore not dependent on peer assessment). We use as our dependent variable the number of postings of each student. As Column 5 of Table 4 shows, we find exactly the same results than when we use our basic measure of effort: nonbinding peer review increases effort, especially for low-ability students in low-ability teams. The results continue to hold when we break down the posts into "setting up appointments" or "being proactive." Columns 6 and 7 of Table 4 show that nonbinding peer review increases the likelihood of posting messages to set up appointments or being proactive, especially for low-ability students in low-ability teams.

Finally, in Column 8 of Table 4, we use yet another measure of effort: whether the student was directly reported to the professor as a shirker by teammates. Peer review decreases the probability to be reported as a shirker, especially for low-ability students in low-ability teams. This indicates again that effort increased for the treated individuals.

C. Worker Morale

One potential downside of peer review is that workers or students dislike being monitored by their peers, especially when the reviews are used to reward or punish them (see, for example, Corgnet 2012). Hence, *binding* peer review may lead to a loss of the "esprit de corps," which could negatively impact subsequent performance. We find that *nonbinding* peer review does not decrease self-reported enthusiasm about the teamwork, as shown in Column 9 of Table 4, where no coefficient is significant.[15]

D. Academic Performance

One important question is whether the temporary increase in effort on the teamwork translates into longer-term gains for individuals, as measured, for example, by their subsequent academic performance on individual tests. As shown in Column 10 of Table 4, nonbinding peer review increases the academic performance of low-ability

14. Holmstrom and Milgrom (1991) show that, with confidential peer evaluations, agents have an incentive to underreport teammates' effort. In the extreme case, all agents report that others exerted no effort, and the peer evaluations cannot be used to assign unequal grades because they carry no information. Though not exactly the same, we also find a substantial amount of "reporting bias" in our data (see Figure 3) when asking students directly to assign shares of contributions to the teamwork.

15. The students were asked to answer the question: "Indicate who showed enthusiasm and a positive attitude in the team: 1 2 3 4". We take the student's self-reported reply—that is, the rating for oneself—as our measure of enthusiasm.

Table 5
Impact of Nonbinding Peer Review on Team Productivity

	Grade Teamwork
Peer review	1.90*
	(1.09)
Observations	191
R^2	0.356
Mean control group	87.33
SD control group	8.81

Notes: OLS regressions. Robust standard errors in parentheses. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The dependent variable is the teamwork grade. "Peer review" is a dichotomous variable equal to one if the team can do peer review, and zero otherwise. Control variables are: average team ability on the baseline test, the variance of baseline ability in the team, a dummy for the statistical teamwork, three classroom dummies to control for the fact that this course was taught in four different classrooms, the number of females in the team, the number of different fields in the team, the number of different ethnic groups in the team, and the number of third- and fourth-year students in the team, and a dummy equal to one if the group was of three students or fewer, and zero otherwise.

students in low-ability teams on individual tests performed after the teamwork by almost one full standard deviation. Thus, extra effort on one task—the teamwork task—does not come at the expense of performance on another task—the individual tests. As before, the effect is larger for lower-ability students in lower-ability teams.

Why do grades increase after peer review in the teamwork task? One explanation may be that higher effort on the teamwork task translates into a better understanding of the material. Another explanation may be that students interacted more together and learned from each other. While we cannot disentangle the precise channels through which the effects operate, it is reassuring to find that effort is rewarded by subsequent productivity gains.

E. Team Productivity

Nonbinding peer review translates into more effort on the teamwork tasks—without adverse effects on the positive attitude of the team members—and into greater individual productivity after the teamwork. How is productivity on the teamwork itself affected?

Table 5 shows that treated teams doing nonbinding peer review scored 1.90 percentage points higher at the teamwork projects (about 0.2 standard deviations) than teams in the control group.[16] This result differs from Corgnet (2012), who finds that

16. One problem may occur if team performance is measured by the professor. Since he knew which teams were assigned to the control and treatment groups, this could have biased the evaluations and, therefore, our estimates. In our case, this is not an issue, since all grading was done by teaching assistants, not the professor. Teaching assistants were not aware of the experiment.

Table 6
Impact of Nonbinding Peer Review on Various Measures of Effort, Excluding Controls Exposed to Previous Treatment

	Effort (1)	Effort: Mean (2)	Postings (3)	Postings: Appointment (4)	Postings: Proactive (5)	Extreme Shirker Reported (6)	Enthusiasm (7)	Increase in Grade (8)
Peer review	0.51**	0.20*	6.07***	0.40**	0.70**	–0.12	0.02	1.05***
	(0.25)	(0.11)	(1.97)	(0.17)	(0.31)	(0.08)	(0.11)	(0.38)
Baseline ability	0.12	0.06*	2.46***	0.25***	0.14	–0.05*	–0.03	–0.26
	(0.10)	(0.04)	(0.87)	(0.08)	(0.12)	(0.03)	(0.06)	(0.19)
Peer review × Baseline ability	–0.13	–0.03	–2.86***	–0.27**	–0.36*	0.06	–0.00	–0.49**
	(0.14)	(0.06)	(1.01)	(0.11)	(0.19)	(0.04)	(0.07)	(0.24)
Team ability	0.15**	0.08***	2.32***	0.11*	0.13	–0.04**	–0.00	0.33***
	(0.07)	(0.03)	(0.67)	(0.06)	(0.10)	(0.02)	(0.04)	(0.12)
Peer review × Team ability	–0.16*	–0.07	–2.76***	–0.18***	–0.30**	0.06**	–0.00	–0.33**
	(0.09)	(0.04)	(0.70)	(0.07)	(0.13)	(0.03)	(0.05)	(0.14)
Team ability × Baseline ability	–0.02	–0.01	–0.80**	–0.09***	–0.04	0.01	0.02	–0.14**
	(0.04)	(0.01)	(0.33)	(0.03)	(0.05)	(0.01)	(0.02)	(0.07)
Peer review × Team ability × Baseline ability	0.04	0.01	1.02***	0.10**	0.13*	–0.03**	–0.00	0.15*
	(0.05)	(0.02)	(0.38)	(0.04)	(0.08)	(0.02)	(0.03)	(0.08)

(*continued*)

Table 6 (*continued*)

	Effort (1)	Effort: Mean (2)	Postings (3)	Postings: Appointment (4)	Postings: Proactive (5)	Extreme Shirker Reported (6)	Enthusiasm (7)	Increase in Grade (8)
Observations	674	674	684	684	684	684	579	676
R^2	0.063	0.067	0.177	0.048	0.095	0.177	0.047	0.457
p-val. sum. coeff. peer review	0.87	0.81	0.22	0.11	0.08	0.34	0.82	0.96
Mean control group[a]	0.64	0.88	4.26	0.17	0.43	0.03	0.91	–0.24
SD control group	0.48	0.21	4.21	0.41	0.81	0.17	0.28	1.20

Notes: OLS regressions. Robust standard errors in parentheses, clustered at the team level. *$p<0.1$, **$p<0.05$, ***$p<0.01$. In Column 1, the dependent variable is a dichotomous variable equal to one if all teammates report that individual *i* adhered to the three rules of good behavior. In Column 2, the dependent variable is the average on the three rules of good behavior given by the three teammates. In Column 3, the dependent variable is the number of postings in team online forums. In Columns 4 and 5, we break down the number of postings into those where students organized appointments ("Appointment") and in which students took the lead on organizing the teamwork ("Proactive"). In Column 6, the dependent variable is a dichotomous variable equal to one if student *i* was reported directly to the professor as a shirker by teammates, and zero otherwise. In Column 7, the dependent variable is a dichotomous variable equal to one if the student rated herself as showing enthusiasm and a positive attitude in the team. Finally, in Column 8, the dependent variable is the increase in scores between the individual baseline test (before the teamwork) and the individual endline test (after the teamwork). "Peer review" is a dichotomous variable equal to one for teams that could use the peer review system, zero otherwise. "Baseline ability" is individual ability on a baseline test. "Peer review × Baseline ability" is an interaction term between peer review and baseline ability. "Team ability" is the average ability of teammates (excluding oneself) on the baseline test. "Peer review × Team ability" is an interaction term between peer review and team ability. "Team ability × Baseline ability" is an interaction term between team ability and baseline ability. Last, "Peer review × Team ability × Baseline ability" is the triple interaction term. In all columns, the regressions include a dummy for the statistical teamwork, the variance of baseline ability in the team, three classroom dummies to control for the fact that this course was taught in four different classrooms, individual-level controls (gender, field of study, ethnic group, year of study), and team-level controls (the number of females in the team, the number of different fields in the team, the number of different ethnic groups in the team, the number of third- and fourth-year students in the team, and a dummy equal to one if the group was of three students or fewer, and zero otherwise. We also replicated Specifications 3 and 4 of Table 4, but the coefficients are all insignificant. We hence do not report these results
[a]This is the mean value of the dependent variable across all students in the control group, irrespective of the ability of the team they are in. It should thus not be interpreted as the mean value for an individual in a zero-ability team.

peer evaluations—used to split unequally profits generated by counting games—actually decrease team productivity. In our experiment, we did not use the peer reviews to assign unequal grades, because the ethics review board explicitly forbids us to look at the peer reviews during the course and because we do not want to hurt intrinsic motivation.

F. Controls Exposed to Previous Treatment

There may be a potential problem with the students assigned to the treatment group for Task 1 (the statistical teamwork) and assigned to the control group for Task 2 (the presentation teamwork). Indeed, these students know that peer review was conducted for Task 1. Hence, even if there is no peer review for them at the beginning of Task 2, they may still believe there will be peer review in Task 2 at some point, given their prior experience in Task 1. This raises the question whether the controls in Task 2 are "contaminated" by their previous exposure to treatment in Task 1—they might provide more effort than they would if they had not been treated in Task 1, thereby biasing our coefficients towards zero.

To deal with this problem, we replicate our results by excluding these students from the analysis. The results are presented in Table 6.[17] As shown, they are very similar to the results in the previous case. One possible explanation for this is that these students correctly concluded that there would be no peer review in Task 2 since there was not peer review three days after the beginning of the presentation teamwork.

VI. Conclusions

We provide experimental evidence showing that nonbinding peer review increases individual effort, team productivity, and individual performance on other tasks. Importantly, this effect works in *all* teams, even teams composed of low-ability individuals, where the traditional forces of peer effects are weaker or simply not operational. Since peer effects crowd out peer review—because individuals already exert more effort in the presence of peer effects—nonbinding peer review works even better in contexts where most individuals are of low ability and exert low levels of effort.

Our findings have implications for our understanding of the success of teams, including low-ability ones. Some of the existing literature suggests that the problem of low effort provision in teams can be solved by peer effects, that is, the exogenous introduction of high-performing individuals into teams (though sorting into subgroups within teams may still be problematic; see Carrell, Sacerdote, and West 2013).[18] We suggest another solution when this is not feasible, for example, when no high-ability

17. We do not replicate the results for the percentage distributions of effort, as reported by self or by others. As explained before, there is not enough variation to identify any effects.

18. Although this is beyond the scope of this paper, it would be interesting to examine if nonbinding peer review works in self-selected groups, given the fact that individuals often choose their teammates. Doing so would also add to our understanding of endogenous group formation.

peers are available: confidential nonbinding peer review of effort, which exploits the unique position of peers to monitor their teammates. Even if low-ability teammates cannot increase effort through role modeling, peer pressure, or knowledge spillovers, they can still monitor other low-ability individuals. Even when the reviews are not directly linked to rewards or sanctions, the mere presence of others who monitor may increase performance, as argued in the psychology literature on social facilitation.

The management literature has looked at confidential peer review, called "360 degree performance review," or "multisource feedback," or "crowdsourced performance appraisal" (see Smither, London, and Reilly 2005, for a literature survey, and Fisher 2013). This literature has traditionally focused on managers, and has found a small and insignificant effect of peer review on leadership skills. In our paper, we argue that peer review can be used to evaluate all team members. On a practical level, conducting confidential peer review on all employees, not just managers, is made possible by the recent advances in electronic surveying. Whereas previous peer reviews involved a complex paper-based effort (Antonioni 1995), conducting online surveys is now free and easy, even with a large number of teams, as in our experiment. In fact, almost half of U.S. firms have already adopted a form of confidential peer review of effort (Fisher 2013). For example, at Google, all employees must confidentially rate their peers every six months, and these evaluations are then forwarded to the manager to determine bonuses and promotions (Homem de Mello 2019). Our recommendation is to link peer reviews to rewards or sanctions only loosely since their mere presence increases effort and productivity and does not generate negative feelings.

Appendix 1

Detailed Results for All Control Variables (Individual, Team, and Classroom) Used in Table 4

Table A1 below shows the coefficients associated with the control variables (individual, team, and classroom) used in Table 4. In terms of the classroom controls, it is interesting to mention that the statistical teamwork elicited slightly less effort, was prone to more "extreme shirking," and had a smaller effect on the subsequent increase in grade (as compared to the presentation teamwork).

Turning to the individual and team controls, almost none of them are statistically significant. The field of study (business, political science, economics, science, as compared to the omitted category humanities) does not affect effort or performance. Furthermore, the year of study, group heterogeneity, and the ethnic group (East Asian, Black, Middle Eastern, compared to the omitted category Caucasian) also do not generally affect effort or performance. The only two effects that we find are (i) that female students exerted more effort, posted more on forums, were more proactive, and were less frequently reported as extreme shirkers (in fact, all extreme shirkers that were signalled to the professors were male), and (ii) the ethnic group South Asian seems to be associated with slightly less effort as measured, for example, by proactive behavior in terms of online posts.

Table A1
Detailed Results for All Control Variables (Individual, Team, and Classroom) Used in Table 4

Variables	Effort (1)	Effort: Mean (2)	Postings (2)	Postings: Appointment (2)	Postings: Proactive (2)	Extreme shirker Reported (8)	Enthusiasm (9)	Increase in Grade (10)
Variance group score	0.11**	0.03	–0.40	–0.02	–0.13*	–0.02*	0.02	–0.16
	(0.06)	(0.03)	(0.50)	(0.03)	(0.07)	(0.01)	(0.03)	(0.10)
Statistical teamwork	–0.15**	–0.06**	–0.47	0.04	–0.02	0.06***	–0.09**	–0.75***
	(0.07)	(0.03)	(0.54)	(0.03)	(0.07)	(0.02)	(0.04)	(0.11)
Fall 2011	0.07	–0.03	0.54	0.02	–0.01	0.02*	–0.02	–0.14
	(0.06)	(0.03)	(0.57)	(0.03)	(0.08)	(0.01)	(0.03)	(0.09)
Fall 2012_001	0.06	0.02	–2.02**	0.00	–0.24**	0.05**	–0.02	0.04
	(0.08)	(0.03)	(0.86)	(0.04)	(0.11)	(0.02)	(0.04)	(0.11)
Fall 2012_002	0.04	–0.05	–2.50***	0.06	–0.38***	0.04**	–0.04	0.61***
	(0.10)	(0.05)	(0.84)	(0.06)	(0.13)	(0.02)	(0.05)	(0.15)
Business	–0.04	–0.04	–0.69	–0.02	–0.10	0.01	–0.08	0.13
	(0.08)	(0.05)	(0.58)	(0.06)	(0.09)	(0.03)	(0.07)	(0.15)
Political science	–0.05	–0.02	–0.75**	0.04	–0.11	–0.00	0.03	–0.02
	(0.05)	(0.02)	(0.33)	(0.05)	(0.08)	(0.02)	(0.04)	(0.10)
Economics	–0.05	–0.02	–0.41	–0.00	0.11	0.05*	0.03	0.11
	(0.06)	(0.03)	(0.38)	(0.05)	(0.10)	(0.03)	(0.04)	(0.12)

(*continued*)

Table A1 (*continued*)

Variables	Effort (1)	Effort: Mean (2)	Postings (2)	Postings: Appointment (2)	Postings: Proactive (2)	Extreme shirker Reported (8)	Enthusiasm (9)	Increase in Grade (10)
Science	–0.04	0.00	0.58	–0.00	0.04	–0.01	0.04	0.12
	(0.05)	(0.02)	(0.40)	(0.05)	(0.10)	(0.01)	(0.03)	(0.13)
Female (0, 1)	0.10**	0.03*	0.51**	0.04	0.28***	–0.04**	0.01	0.07
	(0.04)	(0.02)	(0.24)	(0.04)	(0.08)	(0.02)	(0.03)	(0.10)
East Asian	–0.01	–0.01	–0.57	–0.02	0.07	–0.00	–0.03	0.06
	(0.05)	(0.02)	(0.37)	(0.05)	(0.11)	(0.02)	(0.04)	(0.14)
Black	0.04	0.06*	0.21	–0.04	0.27	–0.01	–0.01	–0.28
	(0.09)	(0.03)	(0.63)	(0.06)	(0.20)	(0.03)	(0.06)	(0.21)
South Asian	–0.14*	–0.08	–0.57	–0.05	–0.32***	0.13***	–0.07	–0.04
	(0.08)	(0.06)	(0.59)	(0.05)	(0.09)	(0.05)	(0.06)	(0.16)
Middle Eastern	–0.03	–0.00	0.65	0.05	–0.08	–0.04**	0.03	–0.18
	(0.09)	(0.03)	(0.53)	(0.08)	(0.11)	(0.02)	(0.05)	(0.19)
First year	0.04	–0.01	0.77	0.01	0.39	0.21	0.08*	0.01
	(0.17)	(0.11)	(1.03)	(0.15)	(0.31)	(0.14)	(0.05)	(0.30)
Third year	–0.00	0.00	0.33	–0.06	–0.07	0.00	–0.01	0.02
	(0.05)	(0.02)	(0.38)	(0.04)	(0.08)	(0.01)	(0.03)	(0.09)

(*continued*)

Table A1 (*continued*)

Variables	Effort (1)	Effort: Mean (2)	Postings (2)	Postings: Appointment (2)	Postings: Proactive (2)	Extreme shirker Reported (8)	Enthusiasm (9)	Increase in Grade (10)
Fourth year	0.03	0.03	0.24	–0.02	–0.11	0.03	–0.04	0.11
	(0.05)	(0.02)	(0.36)	(0.05)	(0.09)	(0.02)	(0.04)	(0.11)
Number female in team	–0.04	0.00	–0.44	–0.02	–0.06	–0.00	0.01	0.02
	(0.03)	(0.01)	(0.28)	(0.02)	(0.04)	(0.00)	(0.01)	(0.04)
Number of different fields in team	0.00	–0.00	–0.32	0.03	–0.01	–0.01	–0.02	–0.00
	(0.04)	(0.02)	(0.38)	(0.02)	(0.05)	(0.01)	(0.02)	(0.05)
Number of different ethnic groups in team	–0.04	–0.02	0.33	–0.00	0.06	0.01	0.00	–0.04
	(0.04)	(0.02)	(0.37)	(0.02)	(0.05)	(0.01)	(0.02)	(0.05)
Number of third and fourth year students in team	–0.01	–0.01	0.05	–0.00	0.02	–0.02***	0.02	–0.03
	(0.03)	(0.02)	(0.30)	(0.02)	(0.04)	(0.01)	(0.02)	(0.04)
Group of three students	–0.06	–0.06	–0.20	–0.05	–0.01	0.01	–0.06	0.13
	(0.08)	(0.04)	(0.68)	(0.04)	(0.10)	(0.03)	(0.04)	(0.12)

Notes: See the notes to Table 4 for additional details. We do not report results for Specifications 3 and 4 of Table 4 since the results are not significant. We also do not report our main coefficients of interest, the number of observations, and the R^2 since they are given in Table 4.

Appendix 2

Replication in a Second University

Our experiment was implemented in a university ranked among the top 40 in the world (according to the "Center for World Universities" ranking, the "Academic Ranking of World Universities," the Times, and the "QS World Universities" ranking). We replicated the experiment at a university below the 400th place in those same rankings. The sample size is 246 students, taught during three years in an introductory economics course in three different classrooms.

Column 1 of Table A2 shows that nonbinding peer review increased effort of low-ability students by 20 percentage points, or 0.41 standard deviations, a larger effect than in the first university. This result is in line with the prediction that nonbinding peer review should work best in lower-ability settings. In the second university, students exerted less effort than in the first: only 36 percent (versus 64 percent in the first university) of low-ability individuals are unanimously rated by their peers as: (i) being on time to meetings, (ii) respecting the deadlines, and (iii) contributing a fair share to the teamwork. As effort is generally lower, nonbinding peer review enhances social facilitation to a larger extent. Column 2 of Table A2 shows that the effect of peer review is similar on low- and high-ability students in this setting.

Peer effects are also weaker in this context, where students have lower average ability. In Column 3 of Table A2, the coefficient of team ability is not significantly different from zero. This illustrates the fundamental advantage of nonbinding peer review over peer effects: even in contexts where peer effects are weaker—for example, due to a lack of high-ability peers—nonbinding peer review increases effort, especially of low-ability students.

The coefficients in Columns 4–7 are very similar to those obtained in the first university in Table 4, albeit less precisely estimated due to smaller sample sizes. Overall, Table A2 confirms our findings pertaining to peer review: peer review increases effort and grades, even in the absence of peer effects.

Table A2
Impact of Nonbinding Peer Review on Various Measures of Effort in the Second University

	Effort (1)	Effort (2)	Effort (3)	Effort (4)	Effort: Mean (5)	Enthusiasm (9)	Increase in Grade (10)
Peer review	0.20**	0.20	0.33*	0.46	0.17	0.20	1.07
	(0.08)	(0.12)	(0.17)	(0.34)	(0.23)	(0.43)	(0.74)
Baseline ability		0.02		0.08	0.03	0.14	–0.73**
		(0.03)		(0.11)	(0.07)	(0.12)	(0.28)
Peer review × Baseline ability		0.00		–0.05	0.02	–0.09	–0.14
		(0.05)		(0.14)	(0.09)	(0.16)	(0.37)
Team ability			–0.03	0.04	0.00	0.16	–0.13
			(0.05)	(0.11)	(0.07)	(0.15)	(0.24)
Peer review × Team ability			–0.05	–0.10	–0.04	–0.03	–0.32
			(0.06)	(0.13)	(0.08)	(0.16)	(0.27)
Team ability × Baseline ability				–0.03	0.00	–0.07	0.15
				(0.05)	(0.04)	(0.06)	(0.13)
Peer review × Team ability × Baseline ability				0.02	0.00	0.05	0.01
				(0.06)	(0.04)	(0.07)	(0.17)

(*continued*)

Table A2 (*continued*)

	Effort (1)	Effort (2)	Effort (3)	Effort (4)	Effort: Mean (5)	Enthusiasm (9)	Increase in Grade (10)
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Team controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	246	246	246	246	249	186	264
R^2	0.101	0.102	0.114	0.116	0.093	0.132	0.347
p-val. sum. coeff. peer review			0.06	0.39	0.54	0.11	0.00
Mean control group[a]	0.36	0.36	0.36	0.36	0.70	0.79	–0.21
SD control group	0.48	0.48	0.48	0.48	0.31	0.41	1.33

Notes: OLS regressions. Robust standard errors in parentheses, clustered at the team level. $*p<0.1$, $**p<0.05$, $***p<0.01$. In Column 1, the dependent variable is a dichotomous variable equal to one if all teammates report that individual *i* adhered to the three rules of good behavior. In Column 2, the dependent variable is the average on the three rules of good behavior given by the three teammates. In Column 3, the dependent variable is the percentage that student *i* reported as being her own contribution to the teamwork. In Column 4, the dependent variable is the percentage that the other team members reported as being the contribution of student *i* to the teamwork. In Column 5, the dependent variable is the number of postings in team online forums. In Columns 6 and 7, we break down the number of postings into those where students organized appointments ("appointment") and in which students took the lead on organizing the teamwork ("proactive"). In Column 8, the dependent variable is a dichotomous variable equal to one if the student was reported directly to the professor as a shirker by teammates, and zero otherwise. In Column 9, the dependent variable is a dichotomous variable equal to one if the student rated herself as showing enthusiasm and a positive attitude in the team. Finally, in Column 10, the dependent variable is the increase in scores between the individual baseline test (before the teamwork) and the individual endline test (after the teamwork). "Peer review" is a dichotomous variable equal to one for teams that could use the peer review system, zero otherwise. "Baseline ability" is individual ability on a baseline test. "Peer review × Baseline ability" is an interaction term between peer review and baseline ability. "Team ability" is the average ability of teammates (excluding oneself) on the baseline test. "Peer review × Team ability" is an interaction term between peer review and team ability. "Team ability × Baseline ability" is an interaction term between team ability and baseline ability. Last, "Peer review × Team ability × Baseline ability" is the triple interaction term. In all columns, the regressions include a dummy for the statistical teamwork; the variance of baseline ability in the team; three classroom dummies to control for the fact that this course was taught in four different classrooms; individual-level controls (gender, field of study, ethnic group, year of study); and team-level controls (the number of females in the team, the number of different fields in the team, the number of different ethnic groups in the team, the number of third- and fourth-year students in the team; and a dummy equal to one if the group was of three students or fewer, and zero otherwise.

[a]This is the mean value of the dependent variable across all students in the control group, irrespective of the ability of the team they are in. It should thus not be interpreted as the mean value for an individual in a zero-ability team.

References

Allport, Floyd H. 1920. "The Influence of the Group upon Association and Thought." *Journal of Experimental Psychology* 3(3):159–82.

———. 1924. *Social Psychology.* Boston, MA: Houghton Mifflin Company.

Antonioni, David. 1995. "Problems Associated with Implementation of an Effective Upward Appraisal Feedback Process: An Experimental Field Study." *Human Resource Development Quarterly* 6(2):157–71.

Arcidiacono, Peter, Gigi Foster, Natalie Goodpaster, and Josh Kinsler. 2012. "Estimating Spillovers Using Panel Data, with an Application to the Classroom." *Quantitative Economics* 3(3): 421–70.

Azoulay, Pierre, Zivin J.S. Graff, and Jialan Wang. 2010. "Superstar Extinction." *Quarterly Journal of Economics* 125(2):549–89.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2013. "Team Incentives: Evidence from a Firm Level Experiment." *Journal of the European Economic Association* 11(5):1079–114.

Benabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *The Review of Economic Studies* 70(3):489–520.

Booij, Adam S., Edwin Leuven, and Hessel Oosterbeek. 2017. "Ability Peer Effects in University: Evidence from a Randomized Experiment." *Review of Economic Studies* 84(2):547–78.

Carrell, Scott E., Richard L. Fullerton, and James E. West. 2009. "Does Your Cohort Matter? Measuring Peer Effects in College Achievement." *Journal of Labor Economics* 27(3): 439–64.

Carrell, Scott E., Bruce I. Sacerdote, and James E. West. 2013. "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica* 81(3):855–82.

Chaudhuri, Ananish. 2011. "Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature." *Experimental Economics* 14(1):47–83.

Chemin, Matthieu. 2018. Data from: "Non-Binding Peer Review and Effort in Teams: Evidence from a Field Experiment." Version 1. Harvard Dataverse. https://doi.org/10.7910/DVN/LIKUGT

Corgnet, Brice. 2012. "Peer Evaluations and Team Performance: When Friends Do Worse than Strangers." *Economic Inquiry* 50(1):171–81.

Corgnet, Brice, Roberto Hernan-Gonzales, and Stephen Rassenti. 2015. "Peer Pressure and Moral Hazard in Teams: Experimental Evidence." *Review of Behavioral Economics* 2(4):379–403.

Cottrell, Nicholas B., Dennis L. Wack, Gary J. Sekerak, and Robert H. Rittle. 1968. "Social Facilitation of Dominant Responses by the Presence of an Audience and the Mere Presence of Others." *Journal of Personality and Social Psychology* 9(3):245–50.

Falk, Armin, and Andrea Ichino. 2006. "Clean Evidence on Peer Effects." *Journal of Labor Economics* 24(1):39–57.

Falk, Armin, and Michael Kosfeld. 2006. "The Hidden Costs of Control." *The American Economic Review* 5(96):1611–630.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr. 2001. "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." *Economics Letters* 71(3):397–404.

Fisher, Anne. 2013. "Should Performance Reviews Be Crowdsourced?" *Fortune*, October 8. http://fortune.com/2013/10/08/should-performance-reviews-be-crowdsourced/ (accessed February 24, 2020).

Foster, Gigi. 2006. "It's Not Your Peers, and It's Not Your Friends: Some Progress toward Understanding the Educational Peer Effect Mechanism." *Journal of Public Economics* 90(8):1455–475.

Greene, William. 2010. "Testing Hypotheses about Interaction Terms in Nonlinear Models." *Economics Letters* 107(2):291–97.
Gunnthorsdottir, Anna, Daniel Houser, and Kevin McCabe. 2007. "Disposition, History and Contributions in Public Goods Experiments." *Journal of Economic Behavior & Organization* 62(2):304–15.
Guryan, Jonathan, Kory Kroft, and Matthew J. Notowidigdo. 2009. "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments." *American Economic Journal: Applied Economics* 1(4):34–68.
Henchy, Thomas, and David C. Glass. 1968. "Evaluation Apprehension and the Social Facilitation of Dominant and Subordinate Responses." *Journal of Personality and Social Psychology* 10(4):446–54.
Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, & Organization* 7(2):24–52.
Homem de Mello, Francisco. 2019. "Google's Performance Management Practices." https://qulture.rocks/en/blog/googles-performance-management-practices-part-1/ (accessed February 27, 2020).
Jack, B. Kelsey, and Maria P. Recalde. 2015. "Leadership and the Voluntary Provision of Public Goods: Field Evidence from Bolivia." *Journal of Public Economics* 122(C):80–93.
Kandel, Eugene, and Edward P. Lazear. 1992. "Peer Pressure and Partnerships." *Journal of Political Economy* 100(4):801–17.
Kang, Changhui. 2007. "Classroom Peer Effects and Academic Achievement: Quasi-Randomization Evidence from South Korea." *Journal of Urban Economics* 61(3):458–95.
Ledyard, John. 1995. "Public Goods: A Survey of Experimental Research." In *Handbook of Experimental Economics*, ed. John H. Kagel and Alvin E. Roth, 111–94. Princeton, NJ: Princeton University Press.
Lyle, David S. 2007. "Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point." *Review of Economics and Statistics* 89(2):289–99.
Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies* 60(3):531–42.
Mas, Alexandre, and Enrico Moretti. 2009. "Peers at Work." *American Economic Review* 99(1):112–45.
Page, Talbot, Louis Putterman, and Bulent Unel. 2005. "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency." *Economic Journal* 115(506):1032–53.
Sacerdote, Bruce. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics* 116(2):681–704.
Smither, James W., Manuel London, and Richard R. Reilly. 2005. "Does Performance Improve Following Multisource Feedback? A Theoretical Model, Meta-Analysis, and Review of Empirical Findings." *Personnel Psychology* 58(1):33–66.
Stinebrickner, Ralph, and Todd R. Stinebrickner. 2006. "What Can Be Learned about Peer Effects Using College Roommates? Evidence from New Survey Data and Students from Disadvantaged Backgrounds." *Journal of Public Economics* 90(8):1435–54.
Waldinger, Fabian. 2011. "Peer Effects in Science: Evidence from the Dismissal of Scientists in Nazi Germany." *Review of Economic Studies* 79(2):838–61.
Zajonc, Robert. 1965. "Social Facilitation." *Science* 149(3681):269–74.
Zimmerman, David J. 2003. "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment." *Review of Economics and Statistics* 85(1):9–23.