SHOULD SCHOOLS RECOGNIZE OR AWARD ACHIEVEMENT?*

Leonardo Bursztyn
Anderson School of Management, UCLA
and
NBER

Robert Jensen
Wharton School, University of Pennsylvania
and
NBER

Abstract: Awards, honors and other forms of public recognition for good performance are used in many academic settings. Though they are intended (at least in part) to encourage effort, there have been few empirical tests of whether they actually do. Further, there are reasons why public recognition could in fact adversely affect performance, such as a fear of peer social sanctions. We test this hypothesis using a natural experiment that introduced a point system and leaderboard into computer-based high school remedial courses. The change was unannounced and unanticipated, and occurred after students had already been using the system for over a month. Prior to the change, students would answer questions and receive private feedback on whether they were correct. Subsequently, points were assigned for correct answers and a leaderboard revealed the top three performers to the entire class. We find that this change led to a 13 percent decline in performance for students overall. Further, we find evidence suggesting that students reduced performance specifically in order to avoid appearing on the leaderboard, such as due to a fear of peer social sanction. In particular, students who were performing at the top of the class prior to the change, i.e., those most "at risk" of being in the leaderboard, had a 24 percent decline in performance. By contrast, those at the bottom of the pre-change distribution had a slight improvement in performance.

* We would like to thank Stefano Fiorin and Craig Jones for excellent research assistance, and the California Center for Population Research and the UCLA Anderson Price Center for financial support.
I. INTRODUCTION

Most academic settings make use of awards or other forms of recognition for good performance. These include gold stars, certificates, prizes, honors, honor rolls and selection for advanced track or honors courses and "gifted" programs. Awards and recognition are generally intended at least in part to encourage effort and reward good performance.\(^1\) However, there is little empirical evidence that such policies actually do so. And there are several reasons why awards might in fact adversely affect performance. For example, unlike grades, which are typically reported only to the student, awards or honors are often observable to peers, such as through announcements in class or at ceremonies, or through the classes that a student takes. Students revealed to have performed well may then incur social sanctions from peers or classmates, such as being teased, bullied, made fun of or ostracized. When faced with the tradeoff between the future benefits of academic performance and the present social costs, some students may therefore deliberately choose to reduce effort or performance. An example of such behavior is the "Acting White" hypothesis in which minorities may face social sanctions from peers for, among other things, doing well in school (see Fordham and Ogbu 1986, Austen-Smith and Fryer 2005 and Fryer and Torelli 2010). Alternatively, environments with awards or honors may create pressure, or promote competition instead of cooperation, which may adversely affect some students. For example, many studies have explored whether women perform worse under competition, or attempt to avoid it altogether (Flory, Leibbrandt and List, forthcoming, Gneezy, Niederle and Rustichini 2003 and Niederle and Versterlund 2007, 2010, 2011; Bertrand 2011 has a summary).

In this paper, we test how the introduction of a system that recognized top performers affected subsequent student performance. The natural experiment we consider was applied to an in-class, computer-based learning system in use in almost 200 high schools, located predominantly in one large American state. The system is used for a range of classes, though primarily remedial English and math, particularly in preparation for a statewide high school exit exam. It consists of a large database of questions that students access when they log into personal accounts, typically along with guided instruction in the classroom.\(^2\) Prior to the change, students

---

\(^1\) They may also be intended as a signaling device for when students apply for higher education or enter the labor market, which again in turn may also encourage effort.

\(^2\) Students can log in from anywhere else as well, including at home. But over 95% of access is at school.
would answer multiple-choice questions and after each one receive private feedback on whether their answer was correct. Approximately one month into the 2011–12 school year, and without any advanced notice or explanation, the company operating the software introduced a points and leaderboard system. Students were now assigned a fixed number of points for each question, based on how many tries it took them to answer correctly. Simultaneously, home screens provided tabs that showed the names of the top three scorers within the classroom, in the school and among all users of the system, as measured by cumulative points received for the past week, month and all time (separate tabs for each). Each tab also showed students their own rank (again, in the classroom, school and among all users, and for the past week, month and all-time). There were no other changes to the system.

Using administrative data on the universe of all questions answered, we find that the introduction of the system led to a 13 percent decline in performance (number of questions answered correctly per day). This result in itself has important implications for school policy, since it is the opposite of the intended effect of the policy change, and of awards and recognition programs more generally.

To test the hypothesis that these results are driven by concerns over peer stigma, we consider how the effect of the policy change varies by student performance prior to the change. Because students had already been using the system for over a month before the change, they would have already had some private information about their own performance (plus additional feedback from teachers), and thus a signal, albeit imperfect, about their risk of showing up on the leaderboard if they continued with performing at their previous level. If a fear of being revealed as a top performer is driving our results, we would expect to observe the biggest performance declines for those students who were performing the best prior to the change.

We find results consistent with this hypothesis. Students in the top quartile of the pre-change distribution of performance, those most at-risk of showing up on the leaderboard, on average had a striking 24 percent decline in the number of questions answered correctly (primarily through attempting fewer questions, as opposed to getting fewer questions correct). By contrast, students in the bottom quartile of the pre-change performance distribution actually did slightly better following the introduction of the system. The pattern of the effect across the distribution of pre-change performance is monotonic, with an initially negative but increasing effect, turning positive for the worst performers. These results are consistent with a concern over
peer social sanctions; for example, Bursztyn and Jensen (2014) find that high school students' decision to take-up free access to a commercial, online SAT prep course is affected by whether the decision to enroll might be made public to their classmates; the effect is particularly large for students in non-honors classes, who sharply reduce enrollment.³

A limitation of our empirical strategy is that the system was implemented in all classrooms at the same time. Therefore, our first empirical analysis consists of simply looking at how performance changed after the points and leaderboard system was introduced. Because of the lack of any cross-student variation in exposure to identify these effects, we conduct a series of placebo tests using alternate dates for the change to rule out that these results could have occurred by chance. In particular, we find that out of 217 alternative placebo start dates, none yield a decline in performance for previously high performing students that is as large as what we observe around the date the points system and leaderboard were introduced. Exploring heterogeneous responses based on pre-change performance in this way also allows us to control for any common trends, shocks or other changes. Thus, for example, the worst-performing students actually improved after the change, suggesting that the decline for the top students was not simply due to any confusion created by the change. It is precisely that the declines occur for the group of previously highest performing students that make the results more likely attributable to fear of being on the leaderboard. Finally, we also note that the declines for the previously high-performing students occur on the very first day of the change (and persisted). This helps rule out alternative explanations, such as student learning about their relative performance and adjusting effort accordingly (e.g., students at the top cutting back when they learn that they are doing well), since there would have been too little time to learn about relative performance on the first day, much less enough time to do well and then still cut back enough so that the net change on the first day is negative.

Though our results are consistent with a fear of being revealed to be at the top of the distribution, we cannot rule out that competition plays a role as well (though below, we do present some arguments against this and other alternatives).⁴ For example, unlike top performing

---

³ The current results complement our other study in several ways, including by having schools across a wider geographic area, a larger sample, and focusing on changes to an existing classroom-based activity.
⁴ We also cannot directly test the Acting White hypothesis because we have do not have student-level data on ethnicity or race. There are administrative data on total ethnic composition for each school, but given the well-documented performance disadvantages of minority students, the school's composition
students, those who know that they are poor performers don't face any additional pressure because they know they will not be competing for the top slots. However, in terms of policy implications, we note that in general, most awards or recognition will have both a public revelation of performance (and thus scope for peer sanctions) as well as competition. So even without separately testing these mechanisms, we feel that a primary contribution of the paper is to show that public recognition or awards, which are commonly used throughout academic settings, can have mixed effects, including worsening overall performance, and particularly that of students most likely to be revealed to be at the top of the class. Though awards or honors may still be valuable, our results show the potential adverse consequences.

Several studies have found that financial academic achievement awards perhaps at best modestly improve student performance (Angrist and Lavy 2009, Angrist, Oreopoulos and Williams 2014, Angrist, Lang and Oreopoulos 2009, Fryer 2011, Kremer, Miguel and Thornton 2009 and Leuven, Oosterbeek and van der Klaauw 2010). However, these studies did not explore whether the effect depends on whether the award was public or private. Further, they all used financial rewards, which may alleviate social stigma; students who win them can claim that they were just doing it for the money. Levitt, List, Necerkmann and Sadoff (2012) explore the effects of non-financial awards (trophies) as well as financial awards for improvements in performance on an exam. They find that the trophies improved performance for elementary school students, but not high school students. For high school students, financial rewards improved performance (though only when provided immediately, rather than at a future date). However, this study also did not explore the effects of making awards public or private. Overall, we are not aware of any previous studies that explicitly address the question of how public recognition affects school performance. An additional advantage of our study is that it involved

---

5 Though in generalizing our results, the natural experiment we consider added not just a leaderboard, but points and rank as well, whereas some other awards may not involve changes to the latter two. Below, we also argue that rank itself is unlikely to play a role in explaining our results.

6 By limiting the ability of teachers to accurately assess student learning, such effects could further lead to a misallocation of attention and resources across students (giving more attention to students who don't need it), or overall (e.g., reviewing material that most students already actually understand).

7 Though the latter study finds evidence of declines among worse performing students (alongside gains for top performing students; with a net effect close to zero).

8 The closest possible exception is Tran and Zeckhauser (2012). In their experiment, Vietnamese undergraduates in an English language class were given information on their class rank either privately
a small change to an existing program that was already in place and that students were using and familiar with, and continued to use. Thus, the classroom setting and all other aspects of the program were exactly as they occur in their real-world environment.

A number of studies have examined the effects of public recognition or other forms of non-financial awards on performance in work environments. For example, Kosfeld and Neckermann (2011) find that students hired for a database project increased performance if they were told in advance that the top performers would be publicly recognized and given a thank you card by the managing director. The effects were biggest among those who were most likely to win the award. Bradler et al. (2013) also find that publicly distributed thank you cards for top performers improve the subsequent performance of workers hired for a data entry task, though the gains come primarily from those who did not win the award. Ashraf, Bandiera and Lee (2014) find that providing information on rank and the top and most improved performances to Zambian health workers in a training program reduces test performance; however, the effect is not statistically different from a treatment that just provided information on rank, suggesting that the additional information on top performers may not have had in independent effect. By contrast, possible public recognition for leaders in the form of a letter of congratulations or being featured in a newsletter improves performance (or rather, offsets the negative effects of just being given information on rank and the names of top and most improved performers). In all cases, the effects are larger for the worst performing trainees.

Finally, we note some caveats and limitations. First, the classes we studied were all remedial, and such students may be more susceptible to peer sanctions than other students. Further, the "reward" offered by being featured on the leaderboard may have had very little value to students; students might be more willing to endure social sanctions when the rewards have greater value or consequences (such as improving the chances of being accepted to a more prestigious university or of receiving merit-based financial aid). On the other hand, the costs of

(by phone) or publicly (posted on a university notice board and website). However, the public treatment did not consist of revealing top performers, but instead of revealing the performance of the randomly selected students (who should fall throughout the distribution). Further, the study was limited to four classrooms, which would have limited the ability to randomize revelation of top performers at the classroom level. Finally, the authors cannot draw strong inferences from their public treatment, which they argue arises from a lack of statistical power arising from the small sample size.

9 The fact that their experiment also recognized or rewarded most improved performers somewhat limits the comparison to our results, since low performers have a greater chance of receiving recognition compared to when it is only for the top performers.
reducing performance in order to avoid the leaderboard are not trivial, since some schools used performance on these exercises as part of the student's course grade. Further, as noted, most of the decline in performance comes from attempting fewer questions, depriving students of the opportunity to learn and test their knowledge of critical material, which may hurt their chances of passing the high school exit exam.

Despite these caveats, we feel that the natural experiment considered here provides a unique opportunity to examine this important topic, within the natural classroom setting. Given how widespread the use of awards and recognition is and, to the best of our knowledge, the lack of any previous empirical evidence of the mixed or even adverse consequences of such practices, we also hope our results stimulate further research on this topic.

The remainder of this paper proceeds as follows. In Section II, we provide background and more detail about the point system and leaderboard. Section III discusses the empirical strategy and data. Section IV presents the results and Section V concludes.

II. BACKGROUND AND POLICY CHANGE

Many schools throughout the state(s) that we examine use in-class, computer based learning materials prepared by private companies that create and operate the system. These companies will occasionally, either on their own or in consultation with school districts, implement changes to the system. The company responsible for the software that we examine was operating in over 200 high schools, across several states (though the vast majority were in just one large state). The sample of schools is not random; they are simply the ones that chose to purchase software from this company. Students in the schools using this software are on average poorer and more likely to be minorities when compared to other schools in the large state where the system is most widely used.

Though many courses are available, the vast majority of use is 10th and 11th grade remedial English and Math courses, including courses designed for statewide high school exit exams. In schools using the software, students are required to take these courses if they scored in the lowest proficiency categories on the previous year's statewide standardized test, or if they had failed the high school exit exam in their first attempt. The fact that students taking these courses were low performing is a relevant consideration for generalizability, since these students are negatively selected; these students might for example be the most sensitive to potential peer
social stigma, which might explain part of the poor performance that led them to be assigned to these classes in the first place.

Students are given individual online accounts. When logged in, they have access to a large database of questions. Questions were multiple choice, and after each question, students would receive private feedback on whether their answer was correct. The questions were organized into modules that typically followed along with in-class instruction. But students had some discretion in how many questions they chose to answer, and the database was sufficiently large that students were unlikely to run out of questions.

On September 20, 2011, without any prior notice, the company introduced a point system and a series of rankings and "leaderboards,"\textsuperscript{10} intended to encourage and motivate students. Students were now awarded 1,000 points for answering a question correctly on the first try, 325 points for a correct answer on the second or third tries, and no points after that. There was no penalty for incorrect answers; thus students could increase their score by getting more questions correct on the first, second or third tries, or by attempting more questions.

The second aspect of the change is that students could access a series of tabs on their homepage that showed the full first name and last initial of the top three scorers (based on cumulative points) in their class, school and among all users of that course. These leaderboards were updated in real time, and were separately available for the past week, month and all-time. Finally, the third change is that students could also access in real time their own personal rank (also for their class, school and all users, and for the past week, month and all-time). However, below we will argue that the information on rank itself is unlikely to have an independent effect on our results, after accounting for the leaderboard.

Beyond assigning points and creating the ranking and leaderboards, the system was otherwise completely unchanged during the period of our analysis. Thus, there was no change in the question database or the difficulty of questions, or any other change to the system that would lead us to expect changes in student performance.

III. EMPIRICAL STRATEGY AND DATA

III. A. Regression Specification

\textsuperscript{10} We were not involved in this change. Our evaluation relies on analyzing data collected by the company, which we obtained approximately two years after the change was introduced.
The point system and leaderboard were introduced at the same time to all students, across all classes and schools. We therefore have no cross-student or -school variation in exposure that can be used to identify the effects of the change. Instead, we start by just examining how performance changed upon the introduction of the system on September 20, 2011. Since we have data on the same students over time, we also include student fixed effects. Thus, we estimate:

\[ Y_{i,t} = \beta_0 + \beta_1 \text{Post}_t + \alpha_i + \epsilon_{i,t} \]

where \( Y_{i,t} \) is an indicator of performance such as the number of questions answered correctly by individual \( i \) on day \( t \), \( \text{Post} \) is an indicator for before vs. after the leaderboard was introduced, and \( \alpha_i \) is the student fixed effect.\(^{11}\) We trim the post period to one month to match the (approximately) one month available prior to introduction (in the appendix, we also consider one and two week intervals instead of one month, since less should have changed in student performance over shorter intervals). The identifying assumption is that had it not been for the change in the system, there would have been no change in student performance around this date. Since this is a strong assumption, we also conduct placebo tests using all other dates in place of September 20, 2011.\(^{12}\)

The above empirical strategy will allow us to identify the net effect of introducing the points and leaderboard system, which may have implications for guiding school policy decisions. In order to test the potential role of peer sanctions, we then exploit the fact that the potential for social sanctions will be greater for certain students than others. As noted above, because students had been using the system for more than a month before the change, they would have had an estimate of their own ability, and possibly even relative performance. For example, in the extreme cases, a student getting almost no answers correct will infer that they are likely to be near the bottom of the distribution, whereas one getting most or all correct will infer that they are likely to be near the top. Therefore, students would have an approximate sense of whether, if they continued their performance unchanged, they were likely to be among the top performers in the class and having this information revealed to others through the leaderboard. These are the students that we predict will be most likely to reduce their effort if fear of peer social sanctions are operative.

\(^{11}\) The results are robust to controlling for the day of the week and the difficulty of questions answered.

\(^{12}\) We also control for separate time trends before and after the change.
Even though there were no points or leaderboard before the change, the company still captured data on all questions answered by each student, including whether they answered correctly. We therefore construct a measure of pre-program performance and "leaderboard risk" by examining the number of correct answers a student had in the month prior to the change, and look for differential responses to the policy change by estimating separate regressions for each quartile of the distribution (computed within each classroom).\footnote{Students also were given leaderboards (and rank) for their school and the universe of all users. However, we believe the classroom contains the most relevant set of peers that students are concerned about. And for a student to be a top three performer in the school or among all users, they have to be a top three performer in their class. So the school-wide and all users leaderboards can only further reinforce how well a student is performing, and our measure is therefore the weakest of the three signals.}

Beyond examining the potential role of a fear of peer sanctions, comparing across students based on pre-change performance also allows us to control for any other factors that may have changed that also affect performance (or if there was something else unusual about this specific day), provided those effects are common across students.

\textit{III.B. Data}

We have data for the universe of students using the software, dating back to 2010. The data are "click-based," meaning that all questions attempted by all students are recorded. We restrict the sample to the 2011–12 academic and to the most widely used courses, remedial English and Math, including preparation for the statewide high school exit exam.\footnote{Other courses include English as a second language and others which on average have few students enrolled per school. In these cases, a leaderboard within the class is unlikely to be meaningful.} This leaves us with a data file containing over 3 million questions answered by approximately 13,000 students, across almost 200 schools.\footnote{In a number of cases, teachers and employees of the software company created accounts for themselves in order to test the system. This was particularly common at the start of the school year or just before. These cases create extreme outliers, such answering over 300 questions correctly in a single day (again, often on a Sunday or the day before the school year began), when the median for the full sample is around 5, and often never answering questions on any other day. There is unfortunately no variable in our dataset that allows us to identify these cases. We therefore trim the top one percent of observations, in terms of correct answers on a given day. However, the results are robust to including these observations. We also drop all classrooms in that only use the platform after the introduction of the new system. Finally, we also drop all classrooms with strictly less than 5 students. For these classrooms, it is not possible to create quartiles, and the information in the leaderboard with the top 3 students might not be very meaningful. Our results are robust to including these classrooms.}
Each student is uniquely identified in the data by an ID code. However, we unfortunately have no other data on students other than the school they attend, which course they are taking and their first and last name. But below in considering heterogeneity, we will match student names to gender frequencies to estimate student gender.

Because there is some bunching in the number of questions answered correctly, and due to indivisibilities in the number of students per classroom, the sample sizes are not balanced across quartiles, with the first quartile having more than 25% of the observations and the fourth quartile having less than 25% of them.

IV. RESULTS

IV.A. Main Effects of the Point System and Leaderboard

We first provide visual evidence of effects of the policy change in Figure 1. Each one of the four panels focuses on students from a different quartile of the within-classroom distribution of performance in the month before the system was introduced. In all four figures, we plot the average number of correct answers per day. We also fit linear trends for the pre- and post-change periods separately, along with a 95 percent confidence band. The results are clear and striking. For the high performing students (quartiles 3 and 4), performance declines sharply on the very day that the leaderboard is introduced (these effects are statistically significant in regressions for quartile 4 if we include just the day before and day of the change). The number of correct answers then stays lower for the remainder of the period (this persistence is confirmed by regressions excluding the first day or first few days after the change). By contrast, there is no decline for the students in quartiles 1 and 2. There is a slight decline on the day of the change for those in quartile 2, but the decline is not persistent. For students in quartile 1, the effect is a persistent increase, which again even shows up on the very first day. This differential response across quartiles suggests that the declines for the previously top performing students are not driven by general confusion over the system, change in teacher behavior, a day-of-the-week effect, or any other factor that would equally affect higher and lower performing students. We also note that any effect already observable on the first day is unlikely to be driven by students updating their beliefs about own ability (or teachers redirecting time and attention away from higher performing students), since there would not yet have been enough information to update priors.
Table 1 shows regression results that confirm the visual evidence (robust standard errors in parentheses). Column 1 shows that the effect of the program across all students was negative. After the system is introduced, on average students answer about one less question correctly per day (statistically significant at the 1 percent level). This is over a 13 percent decline from the baseline of 8.5. Thus overall, the points and leaderboard system reduced performance.

Columns 2–5 provide the results for each quartile of pre-change performance separately, which again allows us to both explore how the response to the leaderboard varies with threat of being in the top performers as well as net out other changes common to all students (results are similar if we pool the regressions and interact the Post indicator with quartile indicators). For students in the top quartile (Quartile 4, in column 5), the change was associated with answering 3.02 fewer questions correctly per day (significant at the 1 percent level). This is a very large, 24 percent decline from the pre-change baseline of 12.6.

As we move from the top quartile down in the pre-change performance distribution (reducing the risk of being in the leaderboard), the effects on performance become less and less negative in absolute terms (though the percent decline is actually greater for quartile 2), and eventually for the bottom quartile becomes positive (in all cases, the results are statistically significant at the one percent level). Again, these results are suggestive of a role for social sanctions, since it is the students who likely perceive the greatest risk of being in the leaderboard, and thus having their high performance publicly revealed, who cut back most. Students at the bottom of the distribution know they have little risk of being in the top (unless they believe that those in the top will cut back enough to move them to the top), so they do not have to cut back.

IV.B. Placebo Tests and Robustness Checks

In this section, we present a series of robustness checks and consider how unique the changes observed around the date the points and leaderboard system were introduced are. First, rather than looking at the period one month before and one month after the change, we consider one and two weeks before and after. Though this will tell us less about whether the effect persists, we might expect that much less should have changed in student performance over these shorter intervals. Appendix Tables A.1 and A.2 show that the estimates are very similar when looking at either of these alternative intervals.
An additional concern is whether there may have been a secular trend over time. For example, students may perform better on average over time, since our month-before period includes the start of the term when students may be less prepared or engaged, or less familiar with the system. Alternatively, any initial enthusiasm may wear off as the school year drags on, or attention may wane as the vacation break gets closer. Of course, if this is common across all students, it would not affect our comparison across quartiles. However, different students may be affected differently. Additionally, we want to estimate the effects of the system change net of any trend which may cause us to over- or under-state the magnitude of the change. Appendix Table A.3 presents results from regressions like those above, but where we also include a linear time trend, and the trend interacted with the Post dummy variable. The effects for both the top and bottom quartiles approximately double. Net of trend, the introduction of the leaderboard resulted in a 42 percent decline in performance for the top performing students (5.3 fewer questions correct per day, out of a base of 12.6). Students at the bottom improve by 0.4, out of a base of 4.6, yielding a 9 percent improvement.

So far, our analysis primarily looks at a simple before vs. after analysis around the specific date of the change. Although comparing across quartiles deals with any common factors affecting all students equally at the time of the system change, it is possible that other factors that changed around this same time could explain the big declines for the best performing students, and improvements for the worst students.\textsuperscript{16} Though we could not identify any obvious examples (school vacations, major sporting or entertainment events, etc.), it would be much less likely that either our pattern emerges by chance or that some other factor besides the leaderboard can explain our results, if we see don't see similar changes at any other date in our sample.

Therefore, in our first placebo test, we run the same regressions as above, but assign the introduction of the point and leaderboard system to every other date, starting one month after the true date of the change,\textsuperscript{17} and ending one month before the end of the school year (there are 217 such days in our data). For quartile 4 on its own, none of the other 217 dates yield a greater decline in performance than the −3.02 found for the true change date. In fact, none of the other dates yields a point estimate greater than 3.02 either. So even in a two-sided test, no other date in our sample yields as large a decrease (or increase) in performance for this previously top-

\textsuperscript{16} As noted, the results are robust to controlling for day of the week.
\textsuperscript{17} We start one month after the true date because our regressions use a one month interval around the change date, so any date less than one month after the true date would capture the effects of that change.
performing set of students as the day of the leaderboard introduction. Appendix Figure A.1 provides a histogram with the distribution of placebo treatment effects, which shows that the estimated decline around the true change date is an outlier in terms of sustained changes in student performance around any specific date.

And in fact, such large and sustained increases and decreases in performance never occur for any other quartile either. Running the placebo tests for quartiles 1, 2, and 3 yields the same results; no alternative date yields an estimated decrease in performance of more than $-3.02$ or an increase greater than $3.02$.

**IV. C. Alternative Mechanisms**

*Competition*

One potential alternative explanation of our results aside from the threat of peer sanctions is that students may be performing less well due to the pressure or competition created by the point system and leaderboard. To the extent that we believe that competition affects all students equally, or perhaps affects stronger students less (since they know they are good performers), we might not predict the same pattern we observe. However, students in the top may face the most competitive pressure, such as to maintain their position in the top, and if competitive pressure adversely affects performance, we might expect the same pattern as we observe (though all students taking these courses are poor performing, they are only high performing relative to other remedial students). Of course, most awards or recognition will include the threat of peer sanctions as well as competitive pressure, so the policy implications will still hold.\(^\text{18}\)

However, we also note that the decline in the number of correct answers per day for the top quartile is primarily due to a decline in the number of questions attempted per day, not the percent answered correctly. Table 2 shows results from the same regressions as above, but using the percent of questions answered correctly on a given day as the dependent variable. The results show for example that the top quartile had a 2 percentage point decline in the percent of their answers that were correct, from a base of 64 percent. Combining the results of Tables 1 and 2, the top quartile answered 19.7 questions on average per day before the change, getting 64 percent correct, for an average of 12.6 questions correct per day. After the change, this same

\(^{18}\) Though some awards are not zero-sum games in the same way. Awards that create a fixed number of winners, as here, will create more competitive pressure than those that simply establish a threshold without limiting the number of winners.
group tried on average only 15.5 questions per day, getting 62 percent correct, for an average of 9.6 correct per day. So the overall decline in the number of correct answers is due primarily to attempting fewer questions rather than performing less well on those they answered; the former accounts for approximately 87 percent of the total decline. The fact that performance on questions attempted did not decline significantly, but students instead just chose to answer fewer questions seems more consistent with fear of being at the top, rather than an effect of competitive pressure; however, students may respond to competition by withdrawing (they cannot opt out of the system, but they could just not answer many questions), or may answer fewer questions because they freeze up under competitive pressure, so we cannot completely dismiss the competition hypothesis.

Rank

An additional issue to consider is that along with the point system and leaderboard, the software also told each student their rank. This information might itself have an independent effect on performance. For example, students who find that they are at the top of the leaderboard may not have known before how well they were doing relative to classmates, and may then decide that they don't need to work as hard. Alternatively, they may feel that their hard work is paying off, and decide to try harder. Students at the bottom of the distribution might also not have known how far behind they were, and thus tried harder when they learned their rank, or alternatively they could become demoralized and stop trying.

However, we believe rank is unlikely to have an independent effect in the present case, at least not in explaining the decline for the previously top-performing students. First, since the change took place about a month into the school year, students would have already had some

---

19 Some studies have documented effects of rank. As noted above, Tran and Zeckhauser (2012) find that providing private information on rank improved the performance of Vietnamese undergraduates. Barankay (2012) finds that eliminating information on rank improves performance among furniture salesperson, with additional evidence suggesting that there is a demoralization effect whereby those receiving a lower than expected rank reduce performance. As noted above, Ashraf, Bandiera and Lee (2012) find that the provision of information on rank decreases performance, particularly for those at the lower end of the performance distribution; further, this effect holds even before students are informed of their rank. This is potentially consistent with "self-handicapping" (Benabou and Tirole 2002), where students intentionally underperform or not exert effort in order to not learn about their own ability.

20 Azmat and Irriberi (2010) exploit a natural experiment in a Spanish high school where for one year, students were given information not just on their own performance as usual (an absolute measure, since there is no curve), but the average for the class; they find that both high and low performing students improved their performance.
information on their own performance, such as through grades on exams or assignments or from feedback from their teachers. Second, as noted above, the decline in performance for students at the top of the pre-change distribution occurred on the very first day the system was in place. Within the scope of that first session, there would not have been enough opportunity for these previously top-performing students to perform well, learn enough about their own performance to decide that they are at the top of the distribution (over and above whatever information they had about their own performance prior to the change) and that they don't need to keep trying harder, and then still have time cut back enough (including overcoming their original high performance at the start that put them at the top) that we would still see a net decline in their performance for that very first session (much less such a large net decline). Such learning on the first day would have been even more challenging given that all students would be starting with no points, and rank would change rapidly and in real time, potentially with every question answered. For similar reasons, we believe it is unlikely that rank gave teachers enough new information on student performance to allow them to redirect attention and resources away from students performing well and towards those who need more help (or the reverse) on the very first day. The same arguments (in reverse) would hold for explaining the increases for the bottom performing students on the first day.

Crowding Out of Intrinsic Motivation

A number of studies have considered whether rewards may reduce intrinsic motivation, and thus performance (Camerer and Hogarth 1999 and Bénabou and Tirole 2003, 2006). Thus, one might argue that the potential "reward" of being in the leaderboard may reduce the intrinsic motivation students had to work hard and perform well. One could argue that any such loss of motivation would affect all students and thus reduce performance across all quartiles. And in fact, most treatments of intrinsic motivation would argue for the opposite pattern of results than we observe; for example, if students previously believed that learning was a goal in itself or something to be done for self satisfaction, and the leaderboard crowds out that motivation and leads them to believe that the goal is instead to win recognition, then the lower performing students should suffer the biggest declines because they will have lost their original intrinsic

---

21 Some studies have argued that rewards may crowd out intrinsic motivation and change behavior in ways that are not as relevant here; for example, incentives for pro-social behaviors may reduce the signal others receive from one's behavior (e.g., donating blood for the money, rather than out of altruism).
motivation, and will not have had it replaced with extrinsic motivation, since they know they are unlikely to be in the top 3. However, it is possible that a loss of intrinsic motivation affects better students more, since they are the ones likely to be appear on the leaderboard and thus be rewarded for their efforts; observationally, such an effect would be similar to what we find.

Though we cannot rule this out completely, we believe it is unlikely to explain our results. First, there is very little actual reward associated with the leaderboard. Most studies of intrinsic motivation crowd out have focused on financial incentives, rather than just recognition or acknowledgment. We believe it is unlikely that students will believe their effort is "cheapened" by the fact that by doing well, they might appear on the leaderboard. It also seems likely that students performing well in a course that is used as part of their class grade and designed to help them pass the statewide high school exam were not motivated much by intrinsic motivation, or, to the extent that they were, that a small recognition of being recognized by having others know that you are in the top 3 would reduce that intrinsic motivation. Second, most studies in both academic and workplace settings have found that adding awards, financial or otherwise, either improve or do not affect performance. And Kremer, Miguel and Thornton (2009), who attempt to explicitly measure intrinsic motivation through surveys, find no evidence that academic financial rewards reduced such motivation for girls in Kenya.

Finally, we note that to the extent that reduced intrinsic motivation can explain our results, as with competition, most awards or recognition would include these effects as well, so the policy implications again still hold.

Novelty Effects

Finally, we can rule out alternative explanations such as the decline in performance simply being due to the sudden change or the newness of the system and a lack of familiarity with it. First, the decline was not common to all students, and in fact the decline was among better performing students, while worse performing students actually improved. Second, the effects appear to persist beyond the first day's decline (as evidenced in Figure 1, and by the fact

An exception is Leuven, Oosterbeek and van der Klaauw (2010), who find that financial incentives decreased performance for low performing students and increased it for high performing students. They argue that these results may be due to a reduction in intrinsic motivation. Also, the seminal study of Deci (1975) finds that college students paid to work on a puzzle were less likely to work on the puzzle later than students who were not paid.

Though it is perhaps possible that previously higher performing students were more distracted by the leaderboards or spent more time checking them and therefore had less time to answer questions.
that the results are robust to excluding the first day or first few days after the change), and over
time we might expect that students would become more familiar with the system and improve.

**Mean Reversion**

The general pattern of declines in performance for previously top-performing students
alongside gains for previously poor-performing students is also consistent with some form of
mean reversion. We note first however that there is no explicit design that would lead to such
reversion. For example, questions are drawn from the database at random, so subsequent
questions are not contingent on past performance, either on a given day or over time. Top
performing students do not receive more difficult questions and poor performing students less
difficult ones. There could still be statistical or incidental mean reversion, but we believe this is
unlikely to be the case. First, the pre-change quartile is based on over a month of performance, so
any randomness or luck in getting correct answers is likely to have balanced out. Second, as
noted, the biggest changes observed are in the number of questions attempted, not the percent
correct; this likely reflects a conscious choice of effort, whereas a student simply on a lucky
(unlucky) streak would likely experience a decline (increase) in the percent answered correctly.
Finally, the fact that we find no other changes this large in our placebo test suggests that there is
not just some general tendency towards mean reversion, such as students who perform well
reducing effort or students performing poorly working harder to improve.

**IV.D. Heterogeneity by Gender**

As noted above, we unfortunately have little data on students other than their names.
However, we can assign a likely gender to each student by matching our data to the Social
Security Administration's (SSA) database of gender frequencies by first name, drawn from the
universe of social security card applications. Using SSA applications filed within a five year
interval around the birth year for students in our sample (January 1, 1993 – December 30, 1997),
we assign a student as male or female if at least 80 percent of children born in those years was of
that sex. Doing so yields a likely gender assignment for approximately 95 percent of students.
Appendix Table A.4 shows that the interaction between gender and the Post dummy variable are small and not statistically significant.\textsuperscript{24} Though we cannot rule out that some students’ gender is misclassified (the 80 percent frequency threshold means that as many as 20 percent of students with some names may be assigned the wrong gender, biasing the coefficient towards zero), overall we find no evidence of a gender difference in response to the change.

\textit{IV.E. Caveats and Limitations}

There are some important caveats and limitations to generalizability worth noting. First, the students taking these courses were low performing, and the fear of peer sanctions may be greater for these students. In fact, this fear may have been a contributor to the performance that caused them to be in a remedial course in the first place, leading us to overestimate the effects of the fear of peer sanctions for the average student.

Further, the leaderboard is a zero-sum award. Some other awards are given to everyone passing a certain threshold (such as GPA), and it is unclear if the same effects would still hold. A zero-sum award creates more competition among classmates, and does not encourage collaboration or helping. Peer penalties may even be different for the two kinds of awards, since one student getting an award prevents another from getting it.

\textbf{V. CONCLUSION}

We find that a point system and leaderboard designed to encourage and motivate students to perform better caused previously higher-performing students to do significantly worse, with smaller gains for previously low performing students. The overall effect across all students was negative. Though there may still be some value to using awards or honors, it is important to recognize these potential adverse consequences. It may also be worthwhile to consider making awards private, at least in certain settings, in an effort to retain benefits (such as signaling quality for college admissions or the job market) while trying to minimize adverse consequences.

The main results and additional supportive evidence suggest that the fear of peer social sanctions is likely to be the mechanism behind reductions for previously high performing students, though we are unable to perfectly rule out other effects such as competition. However,\textsuperscript{24} These regressions exclude the 10 percent of students with names with less than an 80 percent match to one gender.
given that many other awards will also have the same joint effects, and given the widespread use of public awards and recognition in academic settings, we believe the results have significant implications for school policy.

The natural experiment we consider is valuable because it provides an example of a clear change in the obervability of performance, within the natural classroom setting and with only a small change to a course that students had already been using. As we are not aware of any previous research documenting the adverse consequences of such academic awards, an additional hope is that our results can stimulate further research, including experimental designs that further explore peer social sanctions and other mechanisms through which awards may affect behavior. Similarly, it would be worth exploring whether these effects hold in other settings, or how award systems might better be designed to compensate for these effects.

REFERENCES


FIGURES AND TABLES

Figure 1: Average Number of Correct Answers per Day: One Month Before vs. One Month After Introduction of the New System

Quartile 1

Quartile 2

Quartile 3

Quartile 4

Notes: These figures plot, for each day in the period between thirty days before and thirty days after the introduction of the new system, the average number of correct answers per day (dots). Each figure does it for a different quartile of the within-classroom distribution of the total number of correct answers during the month prior to the introduction of the new system. The figures also fit linear trends separately before and after the introduction of the new system, and the 95% confidence interval associated with the trends. Finally, the red line corresponds to the day of the introduction of the new system, September 20, 2011.
Table 1: Effects of Points and Leaderboard System: One Month Before vs. One Month After Introduction

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-system change dummy</td>
<td>-1.0371</td>
<td>0.9459</td>
<td>-0.9132</td>
<td>-2.1380</td>
<td>-3.0247</td>
</tr>
<tr>
<td></td>
<td>[0.083]***</td>
<td>[0.116]***</td>
<td>[0.161]***</td>
<td>[0.174]***</td>
<td>[0.232]***</td>
</tr>
<tr>
<td>Constant</td>
<td>8.0043</td>
<td>4.5622</td>
<td>7.4817</td>
<td>9.4896</td>
<td>12.5929</td>
</tr>
<tr>
<td></td>
<td>[0.061]***</td>
<td>[0.078]***</td>
<td>[0.116]***</td>
<td>[0.130]***</td>
<td>[0.175]***</td>
</tr>
<tr>
<td>Observations</td>
<td>28,869</td>
<td>9,385</td>
<td>7,057</td>
<td>7,039</td>
<td>5,388</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.422</td>
<td>0.393</td>
<td>0.380</td>
<td>0.386</td>
<td>0.387</td>
</tr>
</tbody>
</table>

Sample: FULL QUARTILE 1 QUARTILE 2 QUARTILE 3 QUARTILE 4

Notes: This table presents OLS regressions of the number of correct answers per day on a dummy on whether the date is after the introduction of the points and leaderboard system. All columns restrict the analysis to the time window between one month before the introduction and one month after it. Column 1 presents the results for the entire sample. Columns 2-5 present results by quartile of the within-classroom distribution of the total number of correct answers during the month prior to the introduction of the new system. All regressions include student fixed effects. Robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Table 2: Effects of Points and Leaderboard System: Percentage of Correct Answers per Day

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-system change dummy</td>
<td>0.0118</td>
<td>0.0634</td>
<td>0.0043</td>
<td>-0.0215</td>
<td>-0.0206</td>
</tr>
<tr>
<td></td>
<td>[0.003]***</td>
<td>[0.006]***</td>
<td>[0.007]***</td>
<td>[0.006]***</td>
<td>[0.007]***</td>
</tr>
<tr>
<td>Constant</td>
<td>0.5158</td>
<td>0.4115</td>
<td>0.5078</td>
<td>0.5667</td>
<td>0.6378</td>
</tr>
<tr>
<td></td>
<td>[0.002]***</td>
<td>[0.005]***</td>
<td>[0.005]***</td>
<td>[0.004]***</td>
<td>[0.005]***</td>
</tr>
<tr>
<td>Observations</td>
<td>28,869</td>
<td>9,385</td>
<td>7,057</td>
<td>7,039</td>
<td>5,388</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.422</td>
<td>0.395</td>
<td>0.384</td>
<td>0.404</td>
<td>0.409</td>
</tr>
</tbody>
</table>

Sample: FULL QUARTILE 1 QUARTILE 2 QUARTILE 3 QUARTILE 4

Notes: This table presents OLS regressions of the percentage of correct answers per day on a dummy on whether the date is after the introduction of the points and leaderboard system. All columns restrict the analysis to the time window between one month before the introduction and one month after it. Column 1 presents the results for the entire sample. Columns 2-5 present results by quartile of the within-classroom distribution of the total number of correct answers during the month prior to the introduction of the new system. All regressions include student fixed effects. Robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1
Notes: This histogram displays the distribution of placebo treatment effects estimated for quartile 4 of the within-classroom distribution of the total number of correct answers during the month prior to the introduction of the new system. We run the same regressions as in our main specification, but assign the introduction of the point and leaderboard system to every other date, starting one month after the true date of the change, and ending one month before the end of the school year (there are 217 such days in our data). The dashed line represents our estimated treatment effect for quartile 4 (−3.02).
### Appendix Table A.1: Effects of Points and Leaderboard System: 
One Week Before vs. One Week After Introduction

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Number of correct answers per day</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Post-system change dummy</td>
<td>-1.4591</td>
</tr>
<tr>
<td>[0.146]***</td>
<td>[0.211]***</td>
</tr>
<tr>
<td>Constant</td>
<td>8.1261</td>
</tr>
<tr>
<td>[0.090]***</td>
<td>[0.123]***</td>
</tr>
<tr>
<td>Observations</td>
<td>9,503</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.643</td>
</tr>
</tbody>
</table>

Sample: FULL QUARTILE 1 QUARTILE 2 QUARTILE 3 QUARTILE 4

Notes: This table presents OLS regressions of the number of correct answers per day on a dummy on whether the date is after the introduction of the points and leaderboard system. All columns restrict the analysis to the time window between one week before the introduction and one week after it. Column 1 presents the results for the entire sample. Columns 2-5 present results by quartile of the within-classroom distribution of the total number of correct answers during the month prior to the introduction of the new system. All regressions include student fixed effects. Robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

### Appendix Table A.2: Effects of Points and Leaderboard System: 
Two Weeks Before vs. Two Weeks After Introduction

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Number of correct answers per day</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Post-system change dummy</td>
<td>-1.0801</td>
</tr>
<tr>
<td>[0.102]***</td>
<td>[0.141]***</td>
</tr>
<tr>
<td>Constant</td>
<td>7.9103</td>
</tr>
<tr>
<td>[0.067]***</td>
<td>[0.087]***</td>
</tr>
<tr>
<td>Observations</td>
<td>17,622</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.503</td>
</tr>
</tbody>
</table>

Sample: FULL QUARTILE 1 QUARTILE 2 QUARTILE 3 QUARTILE 4

Notes: This table presents OLS regressions of the number of correct answers per day on a dummy on whether the date is after the introduction of the points and leaderboard system. All columns restrict the analysis to the time window between two weeks before the introduction and two weeks after it. Column 1 presents the results for the entire sample. Columns 2-5 present results by quartile of the within-classroom distribution of the total number of correct answers during the month prior to the introduction of the new system. All regressions include student fixed effects. Robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1
**Appendix Table A.3: Effects of Points and Leaderboard System: One Month Before vs. One Month After Introduction - Including Time Trends**

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-system change dummy</td>
<td>-1.9221</td>
<td>0.4041</td>
<td>-0.8294</td>
<td>-3.3057</td>
<td>-5.3332</td>
</tr>
<tr>
<td></td>
<td>[0.350]***</td>
<td>[0.501]</td>
<td>[0.675]</td>
<td>[0.729]***</td>
<td>[0.986]***</td>
</tr>
<tr>
<td>Constant</td>
<td>8.1748</td>
<td>4.6459</td>
<td>7.5279</td>
<td>9.6407</td>
<td>13.2135</td>
</tr>
<tr>
<td></td>
<td>[0.220]***</td>
<td>[0.298]***</td>
<td>[0.392]***</td>
<td>[0.461]***</td>
<td>[0.650]***</td>
</tr>
<tr>
<td>Observations</td>
<td>28,869</td>
<td>9,385</td>
<td>7,057</td>
<td>7,039</td>
<td>5,388</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.422</td>
<td>0.393</td>
<td>0.380</td>
<td>0.386</td>
<td>0.388</td>
</tr>
</tbody>
</table>

Sample: FULL QUARTILE 1 QUARTILE 2 QUARTILE 3 QUARTILE 4

Notes: This table presents OLS regressions of the number of correct answers per day on a dummy on whether the date is after the introduction of the points and leaderboard system, a linear time trend, and the interaction of the time trend with the post-system change dummy. All columns restrict the analysis to the time window between one month before the introduction and one month after it. Column 1 presents the results for the entire sample. Columns 2-5 present results by quartile of the within-classroom distribution of the total number of correct answers during the month prior to the introduction of the new system. All regressions include student fixed effects. Robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

**Appendix Table A.4: Effects of Points and Leaderboard System: Gender Heterogeneity**

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-system change dummy</td>
<td>-0.9889</td>
<td>0.9522</td>
<td>-0.9810</td>
<td>-1.8928</td>
<td>-3.0500</td>
</tr>
<tr>
<td></td>
<td>[0.125]***</td>
<td>[0.174]***</td>
<td>[0.251]***</td>
<td>[0.263]***</td>
<td>[0.338]***</td>
</tr>
<tr>
<td>Post-system*male</td>
<td>-0.0615</td>
<td>-0.1018</td>
<td>0.2154</td>
<td>-0.2819</td>
<td>0.1051</td>
</tr>
<tr>
<td></td>
<td>[0.176]</td>
<td>[0.246]</td>
<td>[0.342]</td>
<td>[0.365]</td>
<td>[0.483]</td>
</tr>
<tr>
<td>Constant</td>
<td>8.0297</td>
<td>4.6336</td>
<td>7.5071</td>
<td>9.4241</td>
<td>12.5818</td>
</tr>
<tr>
<td></td>
<td>[0.064]***</td>
<td>[0.083]***</td>
<td>[0.123]***</td>
<td>[0.135]***</td>
<td>[0.183]***</td>
</tr>
<tr>
<td>Observations</td>
<td>26,137</td>
<td>8,464</td>
<td>6,329</td>
<td>6,445</td>
<td>4,899</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.420</td>
<td>0.390</td>
<td>0.382</td>
<td>0.388</td>
<td>0.381</td>
</tr>
</tbody>
</table>

Sample: FULL QUARTILE 1 QUARTILE 2 QUARTILE 3 QUARTILE 4

Notes: This table presents OLS regressions of the number of correct answers per day on a dummy on whether the date is after the introduction of the points and leaderboard system, and the interaction of the post-system change dummy with a male student dummy. All columns restrict the analysis to the time window between one month before the introduction and one month after it. Column 1 presents the results for the entire sample. Columns 2-5 present results by quartile of the within-classroom distribution of the total number of correct answers during the month prior to the introduction of the new system. All regressions include student fixed effects. Robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1