How Do One's Peers on a Leaderboard Affect Oneself?

Weiwen Leung
University of Toronto
Toronto, Ontario
weiwen@cs.toronto.edu

ABSTRACT

Leaderboards are a workhorse of the gamification literature. While the effect of a leaderboard has been well studied, there is much less evidence how one's peer group affects the treatment effect of a leaderboard. Through a pre-registered field experiment involving more than 1000 users on an online movie recommender website, we expose users to leaderboards, but different sets of users are exposed to different peer groups. Contrary to what a standard behavioral model would predict, we find that a user's contribution increases when their peer's scores are more dispersed. We also find that decreasing average peer contributions motivates a user to contribute more. Moreover, these effects are themselves mediated by group size. This sheds new light on existing theories of motivation and demotivation with regards to leaderboards, and also illustrates the potential of using personalized leaderboards to increase contributions.

ACM Reference Format:
Weiwen Leung. 2019. How Do One's Peers on a Leaderboard Affect Oneself?. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3290605.3300397

1 INTRODUCTION

Recent years have seen an explosion of gamification studies, and the leaderboard is a workhorse of the gamification literature. Indeed, according to the literature review of Hamari et al. [13], hundreds of gamification studies have been published, and the most commonly used gamification technique among the studies in their meta-analysis is the leaderboard. This perhaps reflects the extensive use of leaderboards in online communities such as Wikipedia, Stack Exchanges, educational tools, and games.

However, while the effect of a leaderboard has been well studied, there is much less evidence on how the group of users that one appears with on a leaderboard ("peer group") affects the treatment effect of a leaderboard. For example, Kraut and Resnick's book chapter [20] gives a reasonable view about how social comparison techniques such as leaderboards work: "Comparative performance feedback can enhance motivation, as long as high performance is viewed as desirable and potentially obtainable." Yet, the only study they cited was Chen et al. [8], which did not use leaderboards.

There is also much speculation, even among experts, on how the size of one's peer group (i.e. the number of users that appear on a leaderboard alongside a given user) affects oneself. For example, Karl Kapp says in an Lynda.com tutorial, "Perhaps the best practice in leaderboards is to have a group or team leaderboard. Group leaderboards naturally make a smaller number of teams, so it seems more likely to get to the top of the leaderboard" [19], suggesting that smaller group sizes boost activity. However, little scientific study has directly examined such claims.

The question of how one's peers on a leaderboard affect oneself is important to online communities that use leaderboards, for online communities can personalize the set of peers they display to any given user, and obscure (or completely hide) all other users. One can easily come up with many plausible sets of peers a user can be ranked against: all users, all users in their country, all new users, and so on. Also, organizations often rank their employees against each other, and privately tell them about their rankings [4]; they can also personalize the set of peers one is ranked against.

In this study, we examine more closely how leaderboards work by examining how one's peer group affects one's activity on an online platform. Specifically, we ran a field experiment on the online movie recommender website MovieLens, where we allocated users into different groups. Conditional on one's activity level (activity was measured by the number of movie ratings and tags contributed to the system over the past two weeks, and is henceforth referred to as "contributions"), one's group assignment was random, allowing one

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300397

to draw causal inferences. At the time of allocation, some groups had many users with high contribution levels, while other groups had fewer of such users. This allowed us to see whether having more active peers would result in more contributions to having less active peers. Also, some groups were more diverse in terms of user contribution level than others. In other words, the standard deviation of user contributions differed across groups, allowing us to see the effect of group heterogeneity on contributions. Finally, in an attempt to see how social comparison is affected by group size, some groups were larger than others; some groups had ten users, others had 20 or 50. Users were able to see a leaderboard comparing themselves against their peer groups only. They were not able to see leaderboards of other groups. As a result, different users had different sets of peers.

We chose group size, standard deviation, and mean as experimental manipulations for they are often the subject of speculation as to how they affect user contributions. Also, online administrators who wish to personalize peer groups can easily calculate these variables and either derive the rough impact of a given peer group, or feed these variables into self-learning algorithms.

We find that one's contribution was significantly affected by one's peer group. For example, a user's contribution was on average positively affected by an increase in the standard deviation of peer group contributions, which is somewhat counterintuitive given that increased standard deviation resulted in increased gaps between users, hence increasing the difficulty of climbing in rank. However, this overall effect masks heterogeneity: in large groups, an increase in the standard deviation of peer contributions positively affected users' contributions. The opposite was true in small groups.

Also, within our sample, increasing the mean of peer group contributions ("group mean") on average negatively affected user's contributions. However, the effect of increasing group mean was also heterogenous: in large groups, users' contributions were negatively impacted by increasing group mean (and the opposite was true in small groups). Finally, even though group size had a mediating effect on contributions, it did not have a direct effect.

Our paper makes several contributions. For example, the observed results regarding standard deviation call into question the idea that increased heterogeneity decreases contributions, suggesting that the effect of heterogeneity is not as straightforward as believed. This indicates that designers should not dismiss the use of leaderboards simply because gaps between users appear to be big. Also, our results shed light on optimal group configurations for personalized leaderboards. When using a personalized leaderboard with a small group size, designers can lean towards selecting a group with a high mean. In contrast, designers can lean towards selecting a group with a low mean for large group sizes. We elaborate on other contributions in the conclusion.

2 RELATED LITERATURE

Effects of leaderboard

As previously mentioned, many gamification studies have examined whether people are more motivated in the presence of a leaderboard (compared to without a leaderboard) [e.g. 9, 22, 26, 27, 29, 30]. To elaborate on one example, Landers and Landers [22] randomly assigned students completing an online wiki-based project to a gamified version with a leaderboard, or a non-gamified version without a leaderboard, and found that leaderboards greatly increased the amount of time learners spent on their projects. Studies like theirs are informative regarding the overall effect of the leaderboard, but are less informative about heterogenous effects, or how leaderboards can be personalized to increase activity as much as possible.

Some studies look at heterogenous effects of leaderboards by interviewing or surveying participants at different positions of the leaderboard. For example, Preist et al. [30] interviewed 18 participants who had taken part in a study testing how gamification and financial incentives affected their use of an app designed to encourage shopkeepers to close their doors to save energy during the cold British winter. They found that low scorers who saw the leaderboard appeared demotivated while high scorers who saw the leaderboard appeared motivated. This is consistent with the idea that having a low rank is demotivating, but is not in itself causal, for high scorers could differ from low scorers in many ways. In contrast, our study makes a causal claim by assigning users with the same score (i.e. level of activity) into different peer groups.

Effects of position on leaderboard

Some studies have attempted to manipulate one's position on the leaderboard and examined the associated effects. To give two illustrative examples, Sun et al. [31] had people play a short game. After showing them a simulated leaderboard where their position was randomized, participants were asked hypothetically about their willingness to replay the game. Also, Jia et al. [17] showed subjects mockups of leaderboards which included their name and the names of some of their friends subjects themselves had entered. Thereafter, participants were asked how they felt about the leaderboard, and their willingness to use an application like those shown through mockups, among other questions.

Our study expands on these studies in many different ways. First, the use of a field experiment ensures that participants are not aware (or at least less likely to be aware) that they are part of an experiment; it is well known that

participants may behave differently if they know they are in an experiment [35]. Second, our study had more than 1000 users, which was an order of magnitude larger than many studies which recruited users to use an app which involved a leaderboard [e.g. 29, 30], and at least twice as large as many MTurk studies that elicited hypothetical choices from participants [e.g. 17, 31]. As such, we have much more statistical power to reject null hypotheses that are truly false, giving us more confidence that any statistically significant results we find are not Type I errors (i.e. false rejections of the null hypothesis). Third, participant responses in previous studies about willingness to continue (if asked) are hypothetical, and real choices may be different from hypothetical ones [15]. Indeed, what participants say in a gamification study may not reflect reality; Palacin-Silva et al. [29] shows that participants using a gamified version of an app reported similar levels of engagement with the app compared to a control group which used a non-gamified version of the app; however, *actual* levels of engagement were different.

Moreover, this study captures actual contributions to a real online community over several weeks, which contrasts with many other studies which capture participant intentions over the very short run (e.g. willingness to play one more time [31]). Finally, this study can offer a variety of insights on how one's peer group affects oneself as our various experimental manipulations change peer groups in very different ways.

Other literatures

There are many relevant literatures from other disciplines. Due to space constraints, we'll review only select articles from the most relevant literatures: peer effects, rank concerns, and how group size affects decision making.

Several literatures in social science disciplines study peer effects (i.e. how one's peers affect oneself). The review of Herbst and Mas [14] found that people are generally motivated by higher performing peers (or peers that produce more output). However, there is one key difference between most peer effects studies and our study: in most studies, peers are physically visible to each other (and can usually interact with each other), which is not the case in our (online) study. Since studies such as Mas and Moretti [25] found that a supermarket cashier's effort is positively related to the productivity of workers who see her, but not workers who do not see her, peer effects studies by themselves do not imply that varying a user's peer group in our online movie recommendation website will affect their contributions.

A growing literature shows that people are motivated by rank. For example, Tran and Zeckhauser [32] found that students who were told of their ranks on practice tests did better on the final test, even when ranking information could not be reliably communicated to others, suggesting that people had a inherent preference for high rank. In addition, one's rank in itself may be a motivating (or demotivating) force. While Genakos and Pagliero [11] found that professional weightlifters systematically underperformed when ranked closer to the top, Gill et al. [12] found that laboratory experiment participants were the more motivated when they were ranked close to the top or close to the bottom, and less motivated when they were ranked around the median. These differing results suggest that the effect of rank on performance may be context specific.

Another closely related literature is that of how group size affects decision making. For example, Garcia and Tor [10] found that increasing group size led to decreased effort, likely due to decreased social comparison, and coined this phenomeneon "the N effect". However, this effect is not universal, for Boudreau et al. [6] found that as the number of competitors solving a coding problem increased from 15 to 19, average effort decreased for low uncertainty problems, but increased for high uncertainty problems. Their study also highlights that small changes in group size could have non-negligible effects. Hence, it is reasonable to expect that average contribution levels might be different in groups of 50, 20, and 10, which are the group sizes we use.

3 EXPERIMENT DESIGN

Timeline

MovieLens is an online movie recommendation website where users can browse, rate, and tag movies in return for personalized movie recommendations. Possible movie ratings range from 0.5 stars to 5 stars (in increments of 0.5 stars). Tags refer to words or short phrases that describe a movie. For example, at the time of writing, more than 1000 users have tagged "sci-fi" to the movie Star Wars, making it the most common tag for that movie.

In mid-2018, we started a "Rate-and-Tag" campaign by sending an email to MovieLens users who had logged in within the past six months, and had not opted out of experiments. Users were encouraged to rate and tag movies to provide better recommendations for all users. Ratings and tags to "obscure" movies (those with less than 40 ratings[1]) were especially encouraged, as the system did not know enough about those movies to make accurate recommendations. Crucially, users were not told of any leaderboard at this point. Nor did they know about other users' contribution levels.

[1]Both ratings and tags help the system make accurate recommendations. However, "obscure" movies were defined in terms of ratings only so that participants would not find it too difficult to determine which movies were obscure.

Two weeks after the first email, we sent a second email introducing leaderboards as part of the Rate-and-Tag campaign. Users were told that they were ranked against a carefully curated set of peers based on points given for their contributions over the past two weeks as follows: 1 point per rating or tag, but 3 points if the movie they rated or tagged was obscure. Users were also given 3 points for each movie they added to the database (that was not already in MovieLens's system). Points would also be given likewise for future contributions, and leaderboards would be updated in real-time. Users saw the leaderboard when they first logged in to MovieLens, and were also able to see the leaderboard "on-demand" by clicking a prominent button on the top right hand corner of screen when logged in. Users were assigned pseudonyms through Python's Faker package; the user herself was referred to as "You". Figure 1 shows a screenshot of the leaderboard. Notice that ties for a given rank were broken randomly, which allows us to see the effect of being assigned a given rank.

The first email was sent to around 14,000 users who met the previously mentioned criteria. The second email was sent only to users who had logged in in the two weeks between the first and second email, and hence was sent only to around 1700 users. Having two emails had several benefits. First, the period in between the two emails was long enough to create significant variation in activity levels across users, but short enough to avoid too big of a gap. Second, activity in the period immediately after the second email was due to both the email and the leaderboard. Having a first email allowed us to isolate the effect of an email and hence recover a rough estimate of the leaderboard's effect.

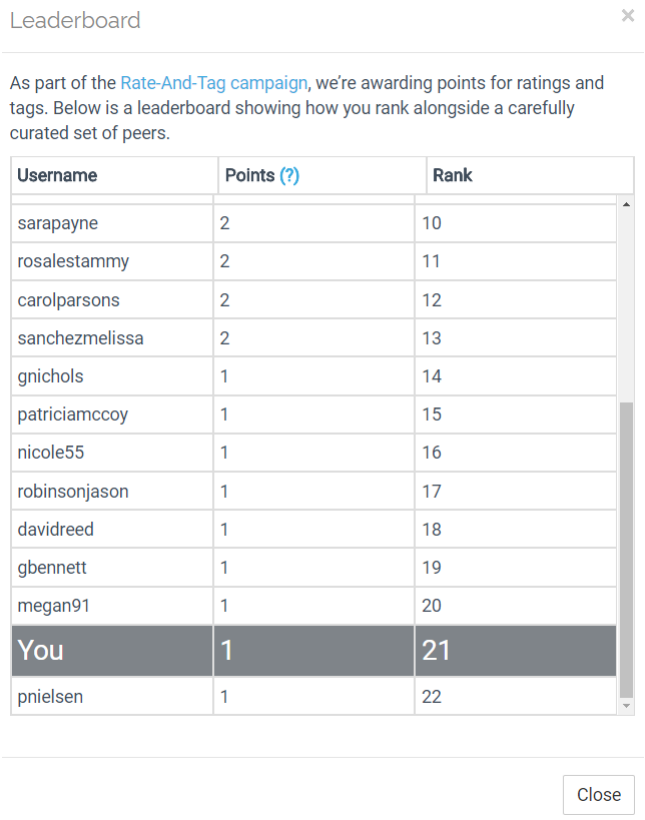

Figure 1: Screenshot of leaderboard

Groupings

Conditional on their contributions between the first and second emails, users were randomly allocated to groups. Hence, one can estimate causal effects by regressing users' post-leaderboard contribution on the group characteristics we manipulated (group size, mean, and standard deviation), provided one controls for users' contributions between the two emails (which we do in our regression).

When allocating users into groups just before the second email, we ranked them in descending order of contributions over the past two weeks[2], breaking ties randomly. We then divided users into deciles, and then adjusted the decile bins slightly so that users with the same score would be in the same bin. Within each bin, allocation to a group was random. Overall allocation was done such that roughly speaking, the experiment design was 3 x 5 x 2: group size could be 10, 20, or 50, group mean could either be *approximately* the 30th, 40th, 50th, 60th, or 70th percentile of activity (between the two emails), while standard deviation of activity could either be "low" or "high". To maximize statistical power, there were around five times as many groups of 10 as compared to groups of 50, and there were around twice as many groups of 20 as compared to groups of 50. More specifically, there were 12 groups of 50, 25 groups of 20, and 60 groups of 10.

Ethics and pre-registration

The IRB approved this study and waived informed consent, thus allowing this to be a field experiment. (However, a post-experiment survey conducted to shed light on possible mechanisms used informed consent.) The study protocol was also pre-registered (see supplement for more details).

4 METHODS

Our baseline OLS[3] regression specification follows:

[2]Contributions were calculated as the number of points a user would have scored based on the points system we were about to introduce. We note that this may introduce measurement error. However, measurement error typically biases coefficients towards zero [2, 34]. Hence if measurement error is a problem, the effects we find likely underestimate the true effect.

[3]See the supplementary materials for why OLS was preferred over Negative Binomial Regression

$$y_i = \beta_0 + \beta_1 Points_i + \beta_2 GroupMean_i + \beta_3 GroupSD_i + \beta_4 1(GroupSize_i = 20) + \beta_5 1(GroupSize_i = 50) + \gamma \mathbf{X}_i + \epsilon_i$$

where

- y_i is outcome for user i *(in a specified period after leaderboards are revealed)*
- $Points_i$ is the *pre-leaderboard* score[4] of user i
- $GroupMean_i$ is the average *pre-leaderboard* score of user i's group, excluding i
- $GroupSD_i$ is the standard deviation of *pre-leaderboard* scores in user i's group, excluding i
- $1(GroupSize_i = Z)$ is a dummy variable for whether or not user i's peer group size is equal to Z
- $\mathbf{X}_i$ is a vector of interactions of experimental treatments (γ is the associated coefficient vector).

Ensuring that the dependent variable only makes use of post-leaderboard contributions while the independent variables only make use of pre-leaderboard contributions helps to avoid the reflection problem coined by Manski [24]. Indeed, Angrist [1] shows that if both the dependent and independent variables use experimental outcomes (in our case, post-leaderboard contributions), a correlation between individual contributions and group contributions is uninformative about whether a causal relationship even exists. One can avoid this problem by making sure that independent variables do not make use of experimental outcomes.

In calculating the dependent variable (post-leaderboard contributions), we use a two week period to measure the short term effect, and a four week period to capture effects over a longer period of time. Though a four week period is somewhat shorter than some field studies, it is still considerably longer than most lab studies, which last around two weeks (if not confined to a single lab session) [13].

Note that the variable $1(GroupSize_i = 10)$ as well as interactions involving this variable are omitted to avoid perfect multicollinearity in the form of the dummy variable trap[5] [2, 34].

A concern regarding our methodology is how outliers can affect our results, for user contributions often follow a power law distribution (and they do in this case). For example, allocating a power user to a group of 10 would greatly increase that group's mean contribution, but have much less of an impact in a group of 50. First, note that outliers by themselves do not affect the unbiasedness of our estimates[6]. Intuitively, even though outliers affect the group mean more when the group size is 10, a group of 50 is five times as likely to have an outlier compared to a group of 10. To be sure, we do examine the effects of outliers in our robustness checks by (1) taking logs of the relevant variables, (2) removing outliers, (3) controlling for the skewness and kurtosis each group to capture more accurately the shape of the distribution of group scores[7] and (4) using median instead of mean, and find that our results are unchanged (refer to supplementary materials for results).

Note that our regressions capture the average effect of changing the group size, mean, and standard deviation. Indeed, some users may not be aware of their group's characteristics (and it would be odd for us to expect our users to be fully aware of these), different users may be affected differently by the same manipulation, and our manipulations may actually work through intermediate channel(s) (e.g. we hypothesize that changing the group mean may change one's rank, which in turn causes a change in contributions). Our estimates thus capture the average effect of changing the three group characteristics we manipulate, including on those who may not be fully aware of their group's characteristics, and so on. Since our primary interest is the average causal effect of an online administrator allocating users into groups with different characteristics, part of which may operate by changing awareness of group characteristics (and so on), our regressions will actually provide us with the desired estimates.

Our initial regressions do not control for rank because rank is itself affected by our experimental variables (especially *GroupMean*), and hence controlling for rank will cause the signal in our experimental variables of interest to be inappropriately absorbed by rank, resulting in a methodological error[8] known as "bad control" [2, 16].

However, in subsequent regressions, we add rank to examine whether it is a channel through which our experimental manipulations affect outcomes. Moreover, we also run t-tests which exploit the fact that whenever multiple users had the

[4]Pre-leaderboard scores are calculated based on contributions between the first and second emails.

[5]To see this clearly, notice that $1(GroupSize_i = 10) = 1 - 1(GroupSize_i = 20) - 1(GroupSize_i = 50)$. The case of the interaction terms is analogous.

[6]In other words, our coefficient estimates would still be correct on average as long as regressors are uncorrelated with the error term, for the mathematical proofs for unbiasedness of regression coefficients hold

[7]While mean and standard deviation are based on the first and second moments of the distribution, skewness and kurtosis are based on the third and fourth moments of a distribution. Skewness measures how symmetric a distribution is, while kurtosis measures how fat the tails are.

[8]Intuitively, suppose a researcher ran an experiment studying if watching comedies affected generosity. In the extreme case where comedies only affected generosity through mood, then a researcher who controlled for mood would draw a completely wrong conclusion that watching comedies has no effect on generosity. Notice also that a manipulation check is only required if the research question is about how *mood* affects generosity.

same number of points, ties for that rank were broken randomly, which allows us to independently examine the effect of rank on contributions.

Hypotheses

We illustrate some plausible hypotheses based on a simple behavioral economics framework. When deciding whether to contribute, a user considers the costs and benefits. Costs of contributing include one's time in rating or tagging a movie, as well as searching for the movie. We assume that the cost of contribution remains fixed regardless of peer group composition, or even whether a leaderboard exists.

There may be many benefits to rating and tagging a movie. However, a key benefit to contributing that changes with peer group composition is the likelihood that one's rank increases. Recall that a growing literature shows that people are concerned about their rank, even when ranking highly does not provide any tangible benefit [7, 21, 32]. This likelihood decreases as the standard deviation of group contributions increase, for higher standard deviations indicate bigger average gaps between users[9]. Hence, we hypothesize that individual contributions decrease if the standard deviation of peer's contributions increases.

In contrast, the effect of group size is ambiguous. While people may derive greater satisfaction from being at the top of a large group, compared to a small group, it may be more difficult to climb to the top of a large group. Hence, we have no strong prior regarding the expected sign of the coefficient of group size. Likewise, the effect of increasing group mean is ambiguous. To give two simple examples, an increase in group mean may motivate those far above the mean, but demotivate those who are already far below the mean. Of course, the opposite could occur depending on the distribution of the group's points. Also, increasing the group mean lowers one's rank, and existing results on how rank affects motivation are not consistent enough to provide a strong prior. To summarize, whether group size and mean affect contributions (and if so, how) are empirical questions.

5 RESULTS

The results suggest that our experimental manipulations had effects throughout the four-week period, some of which were contrary to our hypotheses. Standard deviation had a positive effect on average, group mean had a negative effect on average, but both effects depended on group size.

Table 1 contains the main regression results. Models 1 and 2 illustrate the effect two weeks after leaderboards were introduced, while Models 3 and 4 illustrate the effect after four weeks. Models 1 and 3 do not contain any interaction terms, while Models 2 and 4 contain two-way and three-way interactions between our experimental treatments[10].

In all models, we observe that one's pre-leaderboard contributions (*Points*) strongly predicts one's post-leaderboard contributions, which is not surprising.

Also, the positive coefficient of *GroupSD* in Models 1 and 3 indicates that increasing group standard deviation has a positive effect on user contributions, contrary to what our behavioral economics model predicted. However, Models 2 and 4 show that the effect of *GroupSD* is different across groups of different sizes. The interactions $GroupSD * GroupSize = 20$ and $GroupSD * GroupSize = 50$ are all positive (and also statistically significant in all cases except one), while the coefficient of *GroupSD* is negative and statistically significant in both models. Moreover, the absolute size of the coefficient of $GroupSD * GroupSize = 50$ is larger than the absolute size of the coefficient of *GroupSD* in Models 2 and 4. Taken together, the evidence indicates that increasing group standard deviation has a negative impact in smaller groups (recall that $GroupSize = 10$ is the omitted category), but a positive impact in larger groups.

Group mean appears to have the opposite effect. In Models 1 and 3, increasing group mean on average has a negative effect, though this is not statistically significant at conventional levels in Model 3. However, when one allows for interactions in experimental treatments, one observes clear heterogeneity in the effect of *GroupMean*. For example, in Model 4, increasing group mean has a positive effect when group size is 10 (as evidenced by the positive coefficient of *GroupMean*), but has a non-positive effect when the group size is 20, and a negative effect when the group size is 50 (which follows from the fact that the coefficient of $GroupMean * GroupSize = 50$ is negative, and its absolute value is larger than that of *GroupMean*).

Although group size interacts with both other experimental treatments, we do not observe any main effect of group size in any of our models. This suggests that if the mechanisms discussed in the Hypothesis section work, they cancel each other out (or neither mechanism works).

Finally, the observed effects are still clear even after four weeks: three statistically insignificant coefficients in Models 1 and 2 become significant in Models 3 and 4, but only one coefficient loses significance (which may be a Type II error). Also, coefficients of interest are often larger in Models 3 and 4 compared to Models 1 and 2 (e.g. $GroupSD * GroupSize = 50$ increases from 3.69 in Model 2 to 4.68 in Model 4).

[9]Indeed, in our experiment, the correlation between standard deviation and average gap between users was 0.93, and the correlation between standard deviation and median gap between users was 0.90.

[10]The statistically significant variables remain significant even if we remove the three-way interactions, which is not surprising given that the coefficients on the three-way interactions are small and statistically insignificant

Table 1: Effects of peer group composition at two weeks (Models 1 and 2) and four weeks (Models 3 and 4)

	Model 1			Model 2			Model 3			Model 4		
	Coef	S.E.		Coef	S.E.		Coef	S.E.		Coef	S.E.	
(Intercept)	11.00	13.07		9.86	15.09		19.03	14.90		7.90	17.24	
Points	0.99	0.10	***	1.12	0.11	***	1.39	0.12	***	1.48	0.13	***
GroupMean	-0.98	0.49	**	1.12	0.90		-0.56	0.56		2.39	1.03	**
GroupSD	0.81	0.24	***	-1.42	0.77	*	0.66	0.27	**	-1.71	0.88	*
GroupSize=20	-2.38	17.80		11.53	23.13		-0.69	20.29		20.49	26.42	
GroupSize=50	2.81	17.56		18.46	21.44		5.13	20.03		32.65	24.49	
GroupMean*GroupSD				0.00	0.01					0.00	0.01	
GroupMean*GroupSize=20				-2.26	1.40					-3.30	1.60	**
GroupMean*GroupSize=50				-6.19	1.98	***				-8.41	2.27	***
GroupSD*GroupSize=20				0.79	0.98					1.93	1.12	*
GroupSD*GroupSize=50				3.69	1.16	***				4.68	1.33	***
GroupMean*GroupSD*GroupSize=20				0.01	0.01					0.00	0.01	
GroupMean*GroupSD*GroupSize=50				0.00	0.01					0.00	0.01	
R-squared	0.13			0.15			0.17			0.18		

Note: $***: p < 0.01$, $**: p < 0.05$, $*: p < 0.10$.
Dummy variable notation on group size omitted for simplicity; i.e. $GroupSize = 20$ should read as $1(GroupSize = 20)$, etc.

Effect Size

We use Model 4 of Table 1 to examine effect sizes. In a group of 50, increasing group mean by one point would *decrease* a user's contributions by around 6.02 points (= 2.39 - 8.41). In contrast, in a group of 10, the same change in group mean would increase a user's contributions by around 2.39 points. Given that the average user in our sample obtained 24.4 points in the two weeks after the first email (which, if extrapolated, would be equivalent to 48.8 points in four weeks), that is equivalent to an activity decrease of 12.3% and an increase of 4.9% respectively.

Increasing group standard deviation by one point would decrease a user's contributions by 1.71 points in a group size of 10, but increase contributions in a group of 50 by 2.97 (= 4.68 - 1.71). These would translate into an activity decrease of 3.5% and an activity increase of 6.0% of respectively. Taken together, the effects of peer group composition are non-negligible, and in some cases substantial.

Robustness Checks

Robustness checks are useful to address several possible concerns regarding the model, and to ensure that the results are not driven by the functional form chosen.

A first concern is that the relationship between $Score_i$ (pre-leaderboard contributions) and y_i (post-leaderboard contributions) may not be linear. For example, users that rated all the movies they have watched before the leaderboard was introduced may not have any movies left to rate after the leaderboard is introduced. As such, we add in $Points^2$ to the models of Table 1 to allow for a nonlinear relationship. The results indicate that the concern is not likely to be valid. For example, when $Points^2$ is added to Model 2, its coefficient is -0.0007822, while the coefficient of *Points* is 1.626219. This suggests that the value of y_i increases as the value of *Points* increases, until the value of *Points* reaches 2080; only one contributor had a score higher than this when the leaderboard was released. Results for the other models are similar (see Table 2). In other words, the evidence suggests that this is not too big of a concern.

A second concern is that contribution volume typically follows a power law, and thus results may be driven by outliers. We remove the top 5 contributors, and our results are unaffected. We also take logs of points (adding 1 as log(0) is undefined) and our results are also unchanged. Third, we add the skewness and kurtosis of the distribution of a group's points to our regression to more accurately capture the distribution's shape, and again find that our results are unchanged. Finally, as the median is more robust to outliers than the mean, we replace group mean with the group median, and find that our results are also robust to this change (results included in supplementary materials).

Effect of leaderboard

Although the experiment's focus is on how peer groups affect the treatment effect of a leaderboard, it is still useful to

Table 2: Allowing for y_i to be nonlinear in *Points*

	Model 1			Model 2			Model 3			Model 4		
	Coef	S.E.		Coef	S.E.		Coef	S.E.		Coef	S.E.	
(Intercept)	8.8590	13.1181		8.6989	15.0832		16.4665	14.9580		6.5385	17.2272	
Points	1.4092	0.2718	***	1.6262	0.2822	***	1.8961	0.3100	***	2.0727	0.3224	***
$Points^2$	-0.0007	0.0004	*	-0.0008	0.0004	*	-0.0008	0.0005	*	-0.0009	0.0005	**
GroupMean	-1.2637	0.5188	**	0.8155	0.9131		-0.9027	0.5916		2.0313	1.0429	*
GroupSD	0.8732	0.2419	***	-1.5653	0.7708	**	0.7298	0.2758	***	-1.8771	0.8803	**
GroupSize=20	-0.7718	17.8076		11.7745	23.1020		1.2242	20.3053		20.7697	26.3858	
GroupSize=50	3.4505	17.5552		18.7735	21.4187		5.8979	20.0174		33.0157	24.4632	
GrpMean*GrpSD				0.0047	0.0052					0.0028	0.0059	
GrpMean*GrpSize=20				-2.4895	1.4076	*				-3.5724	1.6077	**
GrpMean*GrpSize=50				-6.2201	1.9819	***				-8.4488	2.2636	***
GrpSD*GrpSize=20				1.0943	0.9943					2.2977	1.1356	**
GrpSD*GrpSize=50				3.8773	1.1639	***				4.8988	1.3293	***
Mean*SD*Size=20				0.0081	0.0085					0.0016	0.0097	
Mean*SD*Size=50				-0.0032	0.0056					-0.0031	0.0064	
R-squared	0.13			0.15			0.17			0.19		

Note: $***: p < 0.01$, $**: p < 0.05$, $*: p < 0.10$
Dummy variable notation on group size omitted for simplicity; i.e. $GroupSize = 20$ should read as $1(GroupSize = 20)$, etc.
To restrict table width, "Group" is abbreviated to "Grp" in two-way interactions, and omitted from the variable name in three-way interactions.

have an idea of whether the leaderboard itself increased contributions (relative to no leaderboard). To roughly estimate the effect of the leaderboard, we compare contributions in the two weeks after the second email was released (where there was a leaderboard and email) to contributions in the two weeks after the first email. Total contributions in the two weeks after the second email was higher by 22%. Note that this estimate could be confounded by factors such as differences in email wording. Also, response to the second email could have been affected by campaign fatigue and email fatigue.

6 LIMITATIONS

Before discussing possible mechanisms and design implications, it is important to note some potential limitations of this study.

Gaming

One concern is that some users may be gaming the system e.g. by rating movies they have never watched or tagging movies inappropriately. We do not know of any test that can formally rule this out, but the available evidence suggests that gaming is unlikely to drive the observed results.

Table 3 illustrates the number of ratings and tags at several points in the experiment. During the course of the experiment, the number of tags increased. In contrast, the number of ratings actually decreased slightly. This suggests that users may have diverted their attention to tags, and it is easy to check whether tags are appropriate. We sampled 20 tags from each of the ten most active users in the experiment, as well as ten tags from each of ten other randomly selected users in the experiment. Of the 300 tags examined, only 12 tags were questionable, and only one was obviously wrong. Almost all of the tags were appropriate, and many were thoughtful. For example, the Indian movie "Spirit" was tagged with "Malayalam", which was not present in the MovieLens page for the movie, but was present in both the IMDB and TMDB descriptions of the movie.

Table 3: Number of tags and ratings per day during the experiment

	Ratings	Tags
Week before first email	3225	447
Week after first email	4007	514
Week after second email	3655	1453

Generalizability

Another concern relates to generalizability: in this experiment, people did not observe other peer groups, and hence did not know the global mean. Would a user at the top of

their group react differently if they knew that their group had a low average contribution?

It is entirely possible that treatment effects may be attenuated if that was the case. However, there are many settings where it is possible to keep the user's focus on the smaller group, or even prevent the user from seeing the larger group. For example, Stackoverflow tailors the social comparison it displays on user's activity dashboard; user contributions can be evaluated relative to other users' contribution for that week, month, year, or all-time. Only if the user clicks on the given social comparison, and then clicks on a dropdown box, can the user "adjust" the social comparison given to them. Systems can likewise be built to get users to focus on a leaderboard with a specially selected peer group (e.g. users in their state) that would likely increase their contributions, and multi-armed bandits can be configured to learn the optimal peer groups to compare the user against. We leave this and other generalizability concerns for future research.

7 POSSIBLE MECHANISMS

The data do not allow us to determine the mechanism(s) that drive our results. However, we offer a speculative explanation consistent with the observed results. When viewing the leaderboard, people first look at their rank, and then compare themselves to others. Increasing the group mean lowers one's rank. Recall that different studies have uncovered different relationships between one's rank and one's motivation. For example, Gill et al. [12] found a U shaped relationship. However, Genakos and Pagliero [11] found that as people ranked closer and closer to the top, their performance systematically worsened, while rank and the degree of risk-taking behavior had an inverted-U shaped relationship. In our study, the relationship between one's absolute rank and one's motivation to contribute could be an inverted U-shaped curve. For example, people may be *more* motivated to contribute when their rank is lowered from 1st to 10th, but *less* motivated if their rank is lowered further. Hence, in a small group of 10 or 20, increasing the group mean would be motivating and hence increase one's contribution, but it would be on average demotivating in a larger group of 50.

The available data are consistent with this explanation. Recall that ties for a certain rank are broken randomly. We use t-tests to examine the impact of "losing" on tiebreak (e.g. bottom of a two-way tie, bottom two of a four-way tie) when a user views the leaderboard, on their contributions over the next 24 hours. Losing on tiebreak increases contributions by 10% when one ends up ranked 2nd to 10th ($p = 0.02$), relative to winning on tiebreak. In contrast, losing on tiebreak results in decreased contributions when one ends up ranked 11th to 50th (-8%, $p = 0.03$). This is consistent with an inverted U-shaped relationship between rank and motivation[11].

Also, when one adds initial rank and its square[12] into the models presented in Table 1, the coefficients are positive and negative respectively. For example, in Model 4, the coefficients are 1.053281 and -0.05044, and both are significant at the 5% level. This suggests that motivation peaks when one is ranked 10th, which is again consistent with the explanation we gave[13]. Moreover, the coefficients of *GroupMean* and its interactions are attenuated, suggesting that group mean affects contributions through rank.

Users may also compare themselves to different sets of users as group size changes. In small groups, the entire group fits on the screen without the need to scroll. Hence users might compare themselves to the topmost user and try to surpass them. In large groups, people may compare themselves to their immediate neighbors, and seek to surpass those immediately above them. When the standard deviation of group contributions increases, the gap between one and one's immediate neighbors grows from small to medium, but the gap between oneself and the leader grows from medium to large. Studies show that making goals more difficult is motivating if goals are not too difficult to begin with, but is demotivating beyond a certain difficulty threshold [5, 23, 33], which could explain the differential effects of increasing group standard deviations in different group sizes. While it is difficult to test this explanation in its entirety, our post-experiment survey[14] found that 66%, 42%, and 22% of users in group sizes of 10, 20, and 50 respectively compared themselves to the topmost user on the leaderboard. Moreover, users in large groups with higher standard deviations reported being more motivated by the leaderboard relative to users in large groups with lower standard deviations, while the opposite was true in small groups.

8 CONCLUSION AND FUTURE WORK

Through a randomized field experiment on MovieLens, we show leaderboards to users on MovieLens, but expose different users to different peer groups. Peer groups differ in group size, mean contribution, as well as the standard deviation of user contributions. Our findings and their implications

[11] However, there is no statistically significant effect at the 10% level when one loses on tiebreak and ends up in the top half of one's group (or in the bottom half), suggesting that rank and not percentile drives the result

[12] The coefficient of initial rank and its square can be interpreted as the *long-term* effect of initial rank, which may act through intermediate channels, such as rank sometime after the leaderboard has been introduced; recall that [1, 24] show that it is inappropriate to add rank sometime after the leaderboard has been introduced.

[13] In contrast, neither initial percentile nor its square are significant at the 10% level when added into the model.

[14] One drawback of the survey was a response rate of only 30%, but response rates across differently-sized groups differed by only 4 percentage points

for the literature as well as designers can be summarized as follows:

First, increasing group standard deviation generally has a positive impact on contributions, contrary to the predictions of a standard behavioral economics model as well as some gamification experts. This suggests that the effects of increased heterogeneity are not as straightforward as believed, and questions the idea that people contribute less as the cost of climbing in rank increases. Hence, designers should not dismiss the use of leaderboards simply because the standard deviation of a group appears to be high (e.g. because gaps between users appear to be big).

Second, the negative effect of increasing group mean on one's own contributions is consistent with the idea that users closer to the bottom of the leaderboard get demotivated.

Third, the effects of mean and standard deviation are themselves affected by group size. This suggests that some caution is needed when generalizing results from small laboratory studies to larger online communities (or from smaller communities to larger ones), and [13] notes that small sample sizes are common in the gamification literature. This finding also has implications for the psychology literature on the effect of group size on competitiveness. Our results differ from existing studies in this literature [e.g. 10] in that we do not show a direct effect of group size, but that group size mediates the effect of mean and standard deviation, suggesting that insights from previous studies may not necessarily apply to environments with leaderboards. Many explanations for this difference are possible; perhaps the simplest explanation is that in previous studies, participants were told directly about the group size, while in this study participants had to find out the group size by themselves. The results we obtained also shed light on optimal group configuration: if designers have good reason to believe our results will generalize to their platform, when using a personalized leaderboard with a small group size, they can lean towards selecting a group with a high mean. In contrast, designers can lean towards selecting a group with a low mean for large group sizes.

Fourth, winning or losing a tiebreak for a certain rank affects a user's contributions. Hence, some tiebreaking mechanisms may be more useful in eliciting contributions than others in online communities. Indeed, designers can consider devising a system which lets the user win on tiebreak in certain situations, but lose on tiebreak in others. For example, if designers have good reason to believe our results will generalize to their platform, personalized leaderboards could show users appearing to lose on tiebreak if they are ranked first to tenth, but appear to win on tiebreak if ranked lower than tenth. This finding also has implications for the growing literature in economics on rank [7, 21, 32] by showing that one's rank can affect one's contributions even in an online community where users are anonymous to one another.

In summary, our paper makes causal claims about how a variety of factors that define peer groups affect behavior over a four week period, and in doing so contributes to several literatures.

Moreover, our paper has implications for leaderboard personalization. Existing research on leaderboard personalization and more generally, gamification personalization has largely focused on personality [e.g. 18, 28]. This requires knowing or predicting user personalities, which can be costly as users may not like to fill in personality questionnaires. Here we show that group size, standard deviation, and mean are predictors of the effect of a personalized leaderboard; these are variables which online administrators can calculate on their own. Online administrators can thus use these variables when considering whether a leaderboard should be displayed to a particular user (and the group the user should be compared against) to increase the chance that the leaderboard will be effective. Of course, online administrators may wish to do experiments of their own to determine the exact effects of these variables on their platform, or even use data-driven algorithms that self-learn which is the optimal peer group to compare a user against in order to increase contributions.

Future work

Future research can build on this paper in many other ways. First, this experiment could be replicated in other contexts, e.g. sites where users know each other. Effects on such sites could be different, though it is possible that effects are bigger when users are friends (e.g. [3] found that co-workers who are friends affect their peers' productivity more than if they were not friends). Second, one can also examine if the effect of group mean and standard deviation continue to change as group size expands beyond 50. Third, one can examine if the observed effects differ if a "global" leaderboard is also displayed (though as mentioned, the global leaderboard can be made obscure or hidden as necessary). Fourth, our experiment was designed such that users in a certain group all saw each other. Future research can examine if bigger gains are possible by relaxing this; to illustrate one of many possibilities, new users from country X can be compared against all new users from that country, while new users from country Y can be compared against all new users if that would increase contributions. Fifth, our experiment was designed to cleanly isolate the impact of group mean, standard deviation, and group size, but in doing so groups were defined arbitrarily. Future research can examine the effect of assigning more meaning to each group (e.g. telling users that they are competing against others from their country,

as opposed to all users). Sixth, our analysis focused on contribution volume; future work can focus on other measures such as contribution quality and participation.

ACKNOWLEDGMENTS

Thanks to Aravind Ramkumar, Joseph Konstan, Haiyi Zhu, Max Harper, Paul Schrater, Hao-Fei Cheng, Qian Zhao, Raghav Karumur, Zachary Levonian, Sarah McRoberts, Yuan Jia, Maria-Palacin Silva, John List, Leonardo Bursztyn, Yan Chen and countless others for their help with this project.

REFERENCES

[1] Joshua Angrist. 2014. The perils of peer effects. *Labour Economics* 30 (Oct. 2014), 98–108.

[2] Joshua Angrist and Jorn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

[3] Oriana Bandiera, Iwan Barankay, and Imran Rasul. 2010. Social Incentives in the Workplace. *Review of Economic Studies* 77, 2 (April 2010), 417–453.

[4] Iwan Barankay. 2012. Rank Incentives: Evidence from a Randomized Workplace Experiment. *Working Paper* (2012).

[5] Gerard Beenen, Kimberly Ling, Xiaoqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, and Robert Kraut. 2004. Using Social Psychology to Motivate Contributions to Online Communities. *CSCW* 6, 3 (November 2004), 212–221.

[6] Kevin Boudreau, Nicola Lacetera, and Karim Lakhani. 2011. Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis. *Management Science* 57, 5 (May 2011), 843–863.

[7] Gary Charness. 2013. The Dark Side of Competition for Status. *Management Science* 60, 1 (2013), 38–55.

[8] Yan Chen, Maxwell Harper, Joseph Konstan, and Xin Li. 2010. Social comparisons and contributions to online communities: A field experiment on MovieLens. *American Economic Review* 100, 4 (Sept. 2010), 1358–1398.

[9] Rosta Farzan, Joan DiMicco, David Millen, Beth Brownholtz, Werner Geyer, and Casey Dugan. 2008. When the experiment is over: Deploying an incentive system to all the users. In *Symposium on Persuasive Technology*.

[10] Stephen Garcia and Avishalom Tor. 2009. The N Effect: More Competitors, Less Competition. *Psychological Science* 20, 7 (July 2009), 871–877.

[11] Christos Genakos and Mario Pagliero. 2012. Interim Rank, Risk Taking, and Performance in Dynamic Tournaments. *Journal of Political Economy* 120, 4 (August 2012), 782–813.

[12] David Gill, Zdenka Kissova, Jaesun Lee, and Victoria Prowse. 2018. First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision. *Management Science* (2018), 1–14.

[13] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does Gamification Work? - A Literature Review of Empirical Studies on Gamification. In *Hawaii International Conference on System Science*. IEEE Computer Society, Hawaii.

[14] Daniel Herbst and Alexandre Mas. 2015. Peer effects on worker output in the laboratory generalize to the field. *Science* 350, 6260 (2015), 545–549.

[15] Charles Holt and Susan Laury. 2002. Risk Aversion and Incentive Effects. *American Economic Review* 92, 5 (2002), 1644–1655.

[16] Solomon Hsiang, Marshall Burke, and Edward Miguel. 2013. Quantifying the Influence of Climate on Human Conflict. *Science* 341, 6151 (Sept 2013).

[17] Yuan Jia, Yikun Liu, Xing Yu, and Stephen Voida. 2017. Designing Leaderboards for Gamification: Perceived Differences Based on User Ranking, Application Domain, and Personality Traits. In *Computer Human Interaction (CHI)*. ACM, Denver, Colorado.

[18] Yuan Jia, Bin Xu, Yamini Karanam, and Stephen Voida. 2016. Personality-targeted Gamification: A Survey Study on Personality Traits and Motivational Affordances. *Computer Human Interaction (CHI)* (2016).

[19] Karl Kapp. 2017. Design Effective Leaderboards. Video. Retrieved July 13, 2018 from https://www.lynda.com/Education-Elearning-tutorials/Design-effective-leaderboards/573400/615940-4.html

[20] Robert Kraut and Paul Resnick. 2011. *Building Successful Online Communities* (1st. ed.). MIT Press, Boston, Chapter 2, 21–76.

[21] Camelia Kuhnen and Agnieszka Tymula. 2011. Feedback, Self-Esteem, and Performance in Organizations. *Management Science* 58, 1 (2011), 94–113.

[22] Richard Landers and Amy Landers. 2014. An Empirical Test of the Theory of Gamified Learning: The Effect of Leaderboards on Time-on-Task and Academic Performance. *Simulation and Gaming* 45, 6 (2014), 769–785.

[23] Edwin Locke and Gary Latham. 1990. *A Theory of Goal Setting and Task Performance.* Prentice-Hall.

[24] Charles Manski. 1993. Identification of Endogenous Social Effects: The Reflection Problem. *Review of Economic Studies* 60, 3 (July 1993), 531–542.

[25] Alexandre Mas and Enrico Moretti. 2009. Peers at Work. *American Economic Review* 99, 1 (March 2009), 112–145.

[26] Elaine Massung, David Coyle, Kirsten Cater, Marc Jay, and Chris Preist. 2013. Using Crowdsourcing to Support Pro-Environmental Community Activism. In *Computer Human Interaction (CHI)*. Paris, France.

[27] Elisa Mekler, Florian BrÃijhlmann, Alexandre N.Tuch, and Klaus Opwis. 2017. Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior* 71, 6 (June 2017), 525–534.

[28] Rita Orji, Gustavo Tondello, and Lennart Nacke. 2018. Personalizing Persuasive Strategies in Gameful Systems to Gamification User Types. *Computer Human Interaction (CHI)* (2018).

[29] Maria Palacin-Silva, Antti Knutas, Maria Ferrario, Jari Porras, Jouni Ikonen, and Chandara Chea. 2018. The Role of Gamification in Participatory Environmental Sensing: A Study In the Wild. In *Computer Human Interaction (CHI)*. ACM, Montreal, Canada.

[30] Chris Preist, Elaine Massung, and David Coyle. 2014. Competing or aiming to be average?: normification as a means of engaging digital volunteers. In *Computer Supported Cooperative Work*. ACM, Baltimore, MD.

[31] Emily Sun, Brooke Jones, Stefano Traca, and Maarten Bos. 2015. Leaderboard Position Psychology: Counterfactual Thinking. In *Computer Human Interaction (CHI) Extended Abstracts*. ACM, Seoul, Korea.

[32] Anh Tran and Richard Zeckhauser. 2012. Rank as an inherent incentive: Evidence from a field experiment. *Journal of PUblic Economics* 96, 9-10 (Oct. 2012), 645–650.

[33] Paul White, Margaret Kjelgaard, and Stephen Harkins. 1995. Testing the contribution of self-evaluation to goal-setting effects. *Journal of Personality and Social Psychology* 69, 1 (July 1995), 69–79.

[34] Jeffrey Wooldridge. 2016. *Introductory Econometrics: A Modern Approach* (6 ed.). South-Western College Publishing, Cincinnati, OH.

[35] Daniel Zizzo. 2010. Experimenter demand effects in economic experiments. *Experimental Economics* 13, 1 (2010), 75–98.