How do one's peers on a leaderboard affect oneself?

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Weiwen Leung

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Paul Schrater

August, 2018

Acknowledgements

I am indebted to Professors Paul Schrater and Haiyi Zhu for constantly supporting me, and more importantly, believing in me. Without them, I would not be who I am today.

I am also thankful to many graduate student colleagues such as Vegard Nygaard, Elena Falcettoni, Hao Fei Cheng, Zachary Levonian, Sarah McRoberts, Colleen Estelle Smith, Bowen Yu, Raghav Karumur, who always made time for me whenever I needed it.

Dedication

To my family and friends, but more importantly to God, who sustained me through the five years of my graduate studies, and through whom all things are possible.

Abstract

Leaderboards are a workhorse within 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 system, 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 suggests that existing theories of motivation and demotivation with regards to leaderboards may need revision, and also illustrates the potential of using personalized leaderboards to increase contributions.

Contents

List of Tables

List of Figures

Chapter 1

Peers and Leaderboards

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. (2014), hundreds of gamification studies have been published, and the most commonly used gamification technique among the studies in their meta-analysis is the leaderboard. Online communities often use leaderboards; Wikipedia has a contribution leaderboard and an anti-vandalism leaderboard, while the various Stack Exchanges have their own leaderboards. Peer production sites are not the only communities that use leaderboards: educational tools, environmental conservation efforts, and games also use leaderboards, to name a few other domains.

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. In their book chapter "Encouraging Contributions to Online Communities" (Kraut and Resnick, 2011), Kraut and Resnick's Design Claim 21 represents 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. (2010), which did not use leaderboards. Other views are reflected in more recent literature. For example, Preist et al. (2014) concludes from a qualitative study that low scorers were often demotivated, while high performers were often motivated. However, a correlation between score and motivation does not by itself imply a causal relationship.

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" (Kapp, 2017), 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 they 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 users of their age group or gender, all new users, and so on. Such personalization is also potentially beneficial in offline communities, for it is now common for private companies, schools, and hospitals to rank their employees against each other, and privately tell them about their rankings (Barankay, 2012).

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 system 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 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, and can also be easily calculated by online administrators.

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 standard deviation of peer group contributions, which is somewhat counterintuitive given that increased standard deviation resulted in increased gaps between users, and hence an increased cost of climbing in rank. However, this overall effect masks heterogeneity: in large groups, an increase in standard deviation of peer contributions positively affected users' contributions. The opposite was true in small groups, where user contributions increased when the standard deviation of peer contributions decreased.

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. However, in small groups, users' contributions were positively affected by increasing the group mean. Finally, even though group size had a mediating effect on contributions, it did not have a direct effect.

We offer one potential rationalization of the differential effects of standard deviation by noting that in small groups, users may compare themselves to the topmost user on the leaderboard, while in large groups, users may compare themselves to their neighbors as the topmost user is not easily visible (which is supported by a post-experiment survey). Hence, while users in small groups may be demoralized as the large distance between them and the leader increases, users in large groups may be energized as the small gap between them and their neighbors increases. Our potential rationalization for the effects of group mean differing across group sizes is based on the differential motivating effects of different ranks, and is elaborated on later in the paper.

Our paper makes several contributions: First, the observed results regarding standard deviation call into question the idea that increased heterogeneity decreases contributions. Second, the mediating effects of group size suggest that some caution is needed when generalizing certain results of small laboratory studies on gamification, which Hamari et al. (2014) notes are commonplace. That said, the fact that group size has no direct effect when it changes from 10 to 50 calls into question viewpoints of

those such as Kapp (2017), who believe that leaderboards are more effective when implemented in smaller groups. Third, since many online communities use leaderboards, they can consider personalizing their leaderboards (although one should avoid directly extrapolating results from this study to other contexts, one can employ self-learning algorithms such as multi-armed bandits to learn the best set of peers to display to a given user).

1.1 Related Literature

1.1.1 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. Farzan et al., 2008; Mekler et al., 2017; Landers and Landers, 2014; Palacin-Silva et al., 2018; Massung et al., 2013; Preist et al., 2014). To elaborate on one example, Landers and Landers (2014) 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.

Some studies look at heterogenous effects of leaderboards by interviewing or surveying participants at different positions of the leaderboard. For example, Preist et al. (2014) 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.

1.1.2 Effects of position on leaderboard

To be sure, 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. (2015) had people play a short game and then showed them a simulated leaderboard where their position was randomized. They were then asked hypothetically about their willingness to replay the game. Also, Jia et al. (2017) 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 the very 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 that they are part of an experiment due to factors such as experimenter demand effects (Zizzo, 2010). Second, more than 1000 users were involved in our study, thus making our sample size at least an order of magnitude compared to many studies which recruited users to use an app which involved a leaderboard (e.g. Preist et al., 2014; Palacin-Silva et al., 2018), and at least twice as large as many MTurk studies that elicited hypothetical choices from participants (e.g. Sun et al., 2015; Jia et al., 2017). 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 (falsely rejecting 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 (Holt and Laury, 2002). Indeed, what participants say in a gamification study may not reflect reality; Palacin-Silva et al. (2018) shows that participants using a gamified version of an app reported similar levels of engagement and experiences with the app compared to a control group which used a non-gamified version of the app; however, *actual* levels were different.

Moreover, many other studies only captured participant intentions over the very short run (e.g. willingness to play one more time Sun et al. (2015)). In contrast, this study captures the real intention to contribute to a real online community over several weeks. Finally, and perhaps most importantly, this study not only manipulates

a person's position on the leaderboard (by experimentally varying the mean of group contributions), it also manipulates the cost of climbing the leaderboard (by varying standard deviation of group contributions), and also group size. Hence, we are able to offer a variety of insights as to how one's peer group affects oneself.

1.1.3 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: rank concerns, peer effects, and how group size affects decision making.

A growing literature shows that people are motivated by rank. For example, Tran and Zeckhauser (2012) 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 (2012) found that professional weightlifters systematically underperformed when ranked closer to the top, Gill et al. (2018) found that laboratory experiment participants were the most motivated when they were ranked closest to the top and bottom, and least motivated when they were ranked around the median. These differing results suggest that the effect of rank on performance may be context specific.

Several literatures in social science disciplines study peer effects (i.e. how one's peers affect oneself). The review of Herbst and Mas (2015) 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 (2009) 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 community will affect their contributions.

Another closely related literature is that of how group size affects decision making. For example, Garcia and Tor (2009) 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. (2011) 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. Interestingly, 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, compared to groups of 10 (or even groups of 10 compared to groups of 20).

1.2 Experiment Design

1.2.1 Timeline

MovieLens is an online movie recommendation community 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 emailer to MovieLens users who had logged in within the past six months, and had not opted out of experiments. This email encouraged users to rate and tag movies to help the system make better recommendations for all. In particular, ratings and tags to "obscure" movies (i.e. movies with less than 40 ratings[1]) were especially encouraged, because the system did not know enough about those movies to make accurate recommendations. Crucially, at this point, users were not told of any leaderboard. Nor did they know about the contribution levels of other users.

Two weeks after the initial email described above, a second emailer was sent, 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

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

if the movie they rated or tagged was obscure. Users were also given 3 points for each movie they added to the database. Points would also be given likewise for contributions from that point onwards, and leaderboards would be updated in real-time. Users saw the leaderboard when they first logged in to the site, and were also able to see the leaderboard "on-demand" by clicking a prominent button on the top right hand corner of the MovieLens website when logged in. Users were assigned pseudonyms through Python's Faker package; the user herself was referred to as "You". Figure 1.1 shows a screenshot of the leaderboard.

The first emailer was sent to around 14,000 users. The second emailer was sent only to users who had logged in in the two weeks between the first and second emailer, and hence was sent only to around 1700 users. Having two emailers had several benefits. First, the period in between the two emailers 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 emailer was due to both the emailer and the leaderboard. Having a first emailer allowed us to isolate the effect of an emailer and hence recover a rough estimate of the leaderboard's effect.

1.2.2 Groupings

Recall that our experimental manipulations were group size, group mean, and group standard deviation. 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 emailers), while standard deviation of activity could either be "low" or "high". Note that 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.

Conditional on their contributions between the first and second emailer, users were randomly allocated to groups. Hence, one can estimate causal effects by regressing users' post-leaderboard contribution on peer group characteristics (group size, as well as mean and standard deviation of activity between the two emailers), provided one controls for users' contributions between the two emailers.

Figure 1.1: Screenshot of leaderboard

1.2.3 Ethics and pre-registration

IRB approval was obtained for the this field experiment. The IRB also waived informed consent, for social science literature shows that behavior could be affected if subjects knew they were part of an experiment (Zizzo, 2010). The experiment was also pre-registered with the American Economic Association's Randomized Controlled Trial Registry (study number: AEARCTR-0002905). Pre-registration of experiments increases the credibility of research; for example, because researchers report the experimental manipulations before the experiment is run, they cannot employ many experimental manipulations and only report the manipulations which had statistically significant effects.

1.3 Methodology

Our baseline specification follows:

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

where

- y_i is outcome for user i *(in a specified period after leaderboards are revealed)*
- $Points_i$ is the *pre-leaderboard* score[2] 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
- $\mathbf{X}$ is a vector of interactions of experimental treatments (γ is the associated coefficient vector).

[2] Pre-leaderboard scores are calculated based on contributions between the first and second emailers

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 (1993). Indeed, Angrist (2014) shows that if the dependent variable and independent variables are both calculated from experimental outcomes (in our case, post-leaderboard contributions), a correlation between individual contributions and group contributions is uninformative about the direction of causality, or 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) (Hamari et al., 2014).

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[3] .

A key concern with regards to 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 in themselves do not affect the unbiasedness of our estimates; in other words, our coefficient estimates would still be correct on average, for the mathematical proofs for unbiasedness of regression coefficients hold as long as regressors are uncorrelated with the error term. 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

[3] To see this clearly, notice that $1(GroupSize_i = 10) = 1 - 1(GroupSize_i = 20) - 1(GroupSize_i = 50)$. Also, $GroupSD_i * 1(GroupSize_i = 10) = GroupSD_i - GroupSD_i * 1(GroupSize_i = 20) - GroupSD_i * 1(GroupSize_i = 50)$. Finally, $GroupMean_i * 1(GroupSize_i = 10) = GroupMean_i - GroupMean_i * 1(GroupSize_i = 20) - GroupMean_i * 1(GroupSize_i = 50)$. Perfect multicollinearity occurs when an independent variable is a linear function of other independent variables in the equation. Hence, for example, $1(GroupSize_i = 10)$ cannot be included if $1(GroupSize_i = 20), 1(GroupSize_i = 50)$, and a constant are included. Note that removing $1(GroupSize_i = 10)$ and all associated interactions does not affect one's ability to make causal inference about the effect of other group sizes, relative to group sizes of 10.

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[4] and (4) using median instead of mean, and find that our results are unchanged.

Note that our regressions capture the overall effect of changing the group mean, standard deviation, and group size. Of course, users may not be aware of their group's characteristics (and it would be odd for us to expect our users to be 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 overall 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. Given that 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[5] known as "bad control" (Angrist and Pischke, 2008; Hsiang et al., 2013).

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

[4] 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.

[5] 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.

1.3.1 Hypotheses

Here, we illustrate some plausible hypotheses based on a simple behavioral economics framework.

When deciding whether to contribute, a user takes into account both 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 (Tran and Zeckhauser, 2012; Kuhnen and Tymula, 2011; Charness, 2013). This likelihood decreases as the standard deviation of group contributions increase, for higher standard deviations indicate bigger average gaps between users[6] . 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. On the one hand, people may derive greater satisfaction from being at the top of a large group, compared to a small group. On the other hand, 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.

[6] 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.

1.4 Results

1.4.1 Effect of Peer Group

Table 1.1 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[7] .

In all models, we observe that one's pre-leaderboard contributions ($Points$) strongly predicts one's post-leaderboard contributions, which is not surprising. Moreover, in Models 1 and 3, we note that increasing group standard deviation has a positive effect on user contributions, as evidenced by the positive coefficient of $GroupSD$. This is the opposite of what our behavioral economics model predicted.

However, Models 2 and 4 show that this result masks significant heterogeneity. 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, there is clear heterogeneity. 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

[7] 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

observe any main effect of group size in any of our models. This suggests that if the mechanisms discussed earlier in the Hypothesis section work, they actually cancel each other out. (Of course, it may be that neither mechanism works.)

Finally, the effect of leaderboards is still clear even after four weeks. In fact, three variables/interactions that were statistically insignificant in Models 1 and 2 become significant in Models 3 and 4, while only one coefficient becomes insignificant (which could be a Type II error). Moreover, in many cases coefficients of interest are larger in Models 3 and 4, compared to Models 1 and 2. For example, $GroupSD * GroupSize = 50$ increases from 3.69 in Model 2 to 4.68 in Model 4.

Table 1.1: 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.

1.4.2 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. To examine if this is a concern, 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 1.2). In other words, the evidence suggests that this is not too big of a concern.

Table 1.2: 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	
GroupMean*GroupSD				0.0047	0.0052					0.0028	0.0059	
GroupMean*GroupSize=20				-2.4895	1.4076	*				-3.5724	1.6077	**
GroupMean*GroupSize=50				-6.2201	1.9819	***				-8.4488	2.2636	***
GroupSD*GroupSize=20				1.0943	0.9943					2.2977	1.1356	**
GroupSD*GroupSize=50				3.8773	1.1639	***				4.8988	1.3293	***
GroupMean*GroupSD*GroupSize=20				0.0081	0.0085					0.0016	0.0097	
GroupMean*GroupSD*GroupSize=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.

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 qualitatively similar. 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 (not shown due to space constraints).

1.4.3 Gaming

Another concern is that the leaderboard may have caused some users to game 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 1.3 illustrates how the number of ratings and tags at several points in the experiment. After the leaderboard was introduced, the number of tags increased, but the number of ratings actually decreased slightly. This suggests that users diverted their attention to tagging, and it is easy to check whether tags are appropriate for a given movie. We sampled twenty tags from each of the ten most active users in the experiment, as well as ten tags from ten other randomly selected users in the experiment. Of the three hundred tags examined, only a dozen 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 1.3: Number of tags and ratings per day during the experiment

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

1.4.4 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 know whether the leaderboard itself increased contributions (relative to no leaderboard). To determine the effect of the leaderboard, we compare contributions in the two weeks after the second emailer was released (where

there was a leaderboard and emailer) to contributions in the two weeks after the first emailer. Total contributions in the two weeks after the second emailer was higher by 22%. This may be a conservative estimate, for response to the second emailer could have been affected by campaign fatigue and email fatigue. Hence, it is reasonable to conclude that leaderboards increased contributions, at least in the short term.

1.5 Discussion

1.5.1 Effect Size

We use Model 4 of Table 1.1 to examine effect size. 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 manipulation 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 emailer (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.

1.5.2 Possible mechanism

The data do not allow us to determine the mechanism(s) that drive our results. However, we offer a speculative explanation that is 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. (2018) found a U shaped relationship, but Genakos and Pagliero (2012) found that as people ranked closer and closer to the top, their performance systematically worsened, while the relationship between rank and risk-taking had an inverted-U shape.

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 20th, 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[8] .

Also, when one adds initial rank and its square 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[9] . Moreover, the coefficients of $GroupMean$ and its interactions are attenuated, suggesting that group mean affects contributions through rank.

People may also compare themselves to different groups of people as group size changes. In small groups, the entire group fits on the screen without the need to scroll, and hence people might compare themselves to the topmost member, and make surpassing the topmost member as their goal. 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

[8] 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 at the bottom half of one's group), suggesting that it is rank and not percentile that is driving the result

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

difficult is motivating if goals are not too difficult to begin with, but is demotivating beyond a certain difficulty threshold (Beenen et al., 2004; White et al., 1995; Locke and Latham, 1990), 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, a post-experiment survey 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[10] .

1.5.3 Generalizability

One important question relates to generalizability. In this experiment, people did not observe other peer groups, and hence did not know have any idea of the global mean (i.e. mean of all users). A question arises as to whether a user at the top of their group would react differently if they knew that their group was relatively mediocre.

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.

1.5.4 Contribution to literature

This paper principally contributes to literature on leaderboards. First, in showing that a higher standard deviation of one's peer group contribution has a positive overall effect on one's own contribution, it questions the idea that people tend to be more

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

demoralized as the cost of climbing in rank increases. However, 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. Second, showing that the effects of mean and standard deviation are themselves affected by group size not only has implications for leaderboard personalization, but also suggests that one should be cautious in generalizing certain results from small laboratory studies to larger online communities (or from smaller communities to larger ones).

This paper also contributes to literatures in other disciplines. First, it contributes to the psychology literature showing that group size can affect competitiveness. Existing literature on the "N effect" shows a decreased propensity to compete as group size increases (Garcia and Tor, 2009). However, the N effect has not been tested thus far in the context of leaderboards, and more generally, in situations where people receive real-time rank feedback. Our results differ from existing studies 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. Future research can build more comprehensive models to reconcile differences in findings. Another discipline this paper contributes to is the growing literature in economics about rank (Charness, 2013; Kuhnen and Tymula, 2011; Tran and Zeckhauser, 2012); this paper shows that group size can mediate rank concerns.

1.6 Conclusion

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. We find that 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. Also, increasing group mean generally has a negative impact on user contributions. However, both the effects of mean contribution and standard deviation are mediated by group size. In small groups, contribution increases when standard deviation decreases, or when mean contribution increases, but the opposite is true for large groups. We then gave a potential rationalization of our findings based on

people comparing themselves to different users as group size increases, and motivation peaking when people are ranked 10th.

The strength of our experiment lies in the ability to make (i) causal claims (ii) about how a variety of factors that define peer groups (iii) affect behavior over the span of four weeks. As such, it is not surprising that our paper contributes to several literatures. Moreover, our paper serves to highlight the value of designing algorithms that learn which is the optimal peer group to compare a user against in order to increase contributions.

Besides self-learning algorithms, future research can build on this paper in many other ways; some obvious examples follow. First, this experiment could be replicated in other contexts. 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; for example, 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, and perhaps most importantly, our experiment was designed to cleanly isolate the impact of group mean, standard deviation, and group size. Future research can examine the effect of telling users that they are competing against users from their country (as opposed to all users).

Chapter 2

References

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

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

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

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.

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.

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

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.

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*.

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

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.

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.

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.

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

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

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

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.

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`

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

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

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.

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

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

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

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.

Elisa Mekler, Florian Brhlmann, 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.

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.

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.

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.

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.

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.

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