Virtual Teams for the Modern Workforce*

Teng Ye, Wei Ai, Yan Chen, Qiaozhu Mei, Jieping Ye

November 10, 2020

Hailed as the future of work, the gig economy provides flexible and low-barrier jobs for millions of workers around the world. However, many gig platforms suffer from high attrition, partially due to the lack of organization identity and social bonds (*1*). Here, we investigate the efficacy of virtual teams on worker productivity and retention in a global ride-sharing platform. Using a large-scale natural field experiment with 27,790 drivers, we organize drivers into virtual teams and randomize these teams into three experimental conditions. Treated drivers receive their team ranking, or their individual ranking within a team, whereas those in the control condition receive individual performance information without social comparison. We find that treated drivers are significantly more productive than those in the control condition. Three months after the experiment ended, drivers in the team leaderboard treatment continue to work longer hours on the platform. Lastly, within virtual teams, laggards benefit the most from team contest.

One sentence summary: Virtual teams increase productivity and retention in the modern workforce.

*Y. Chen: School of Information, University of Michigan, 105 South State Street, Ann Arbor, MI 48109-2112. Email: yanchen@umich.edu.

Introduction

The gig economy provides workers with the benefits of autonomy and flexibility (*2*), but it does so at the expense of work identity and co-worker bonds. Many gig platforms have experienced low engagement and high attrition rates among its workers, who typically work alone with "no interaction or relationship with other colleagues," on jobs "that don't lead to anything" (*1, 3*). In comparison, while many traditional sectors of the economy have been defined by daily in-person interactions, the pandemic upended that as thousands of organizations mandated that their employees work from home. By August 2020, 42 percent of the U.S. labor force work from home full-time (*4*). Given that some of these radical changes are here to stay, how can organizations help their workers create and maintain positive social relations at work while working alone?

To answer this question, we propose to form virtual teams among workers and engage the teams in contests to strengthen team identity. We then evaluate the effects of virtual teams on worker productivity and retention through a large-scale randomized field experiment conducted on the largest ride-sharing platform in the world.

Our research applies insights from the social identity research from psychology (*5, 6*) and behavioral economics (*7, 8*). This research shows that when people feel a stronger sense of common identity with a group, they exert higher effort and make more contributions to improve group outcomes in the laboratory, using either induced (*9–11*) or natural identities (*12, 13*). Moving from the lab to the field, researchers find that identity-based teams are effective to increase pro-social behavior in fruit harvesting (*14*) and online peer-to-peer pro-social lending (*15,16*). By contrast, when workers are paid by piece rate, prior research indicates that providing team ranking information might reduce average worker productivity in a field experiment where teams were not randomly assigned to treatments (*17*). To estimate the causal effects of team

incentives on productivity and retention, we randomly assign teams into different experimental conditions in a real, large-scale gig platform. To our knowledge, this is the first natural field experiment to do so.

One key design question is how to structure team contests. In a prior field experiment in the southern Chinese city of Dongguan in August 2017, we randomly assigned 2,100 drivers to teams of size seven either randomly or based on homophily in age, hometown location, or productivity. We organized the teams to compete for cash prizes for five days. We find that, compared to those in the control condition, treated drivers worked longer hours and earned 12% higher revenue during the contest. We find that teams formed based on age similarity are more productive two weeks post-contest than randomly formed teams (*18*).

Encouraged by the results of this first field experiment, DiDi shipped two of our team-formation algorithms, hometown similarity and age similarity, into production. In 2018 alone, DiDi conducted 1,548 team contests across 180 cities in China, involving over two million drivers. These contests, typically one-week long, helped the platform meet high demands from tourists during national holidays, and increased both driver income as well as retention (*19*). A common feature among the 1,548 team contests DiDi ran in 2018 is that they were all one-week contests with cash incentives, and teams were dismissed immediately afterwards. The design represents a missed opportunity in that team identity should have long-term effects on drivers' identity with the organization and their bonds with teammates.

To investigate this potential, we ran a longer-term team-contest field experiment without any monetary incentives. In October 2018, we conducted a natural field experiment on the DiDi platform involving 27,790 drivers across three cities: Beijing, Kunming, and Taiyuan. The experiment started on October 22, 2018 and ended on December 3, 2018. To evaluate our treatment effect on driver retention, we continued to collect data for three months after our experiment.

We randomly assigned each leaderboard to one of the three experimental conditions - the team, individual leaderboard treatments, and the control condition. In the Team Leaderboard treatment, drivers had access to both team and individual leaderboards. We sent a daily reminder to these drivers to check the rankings of the same five teams within their leaderboard, as well as individual teammate rankings within their team. In the Individual Leaderboard treatment, drivers only had access to the individual leaderboard within their team. Again, we sent a daily reminder to drivers to check their individual rankings. Finally, in the control condition, drivers could not access either leaderboard. However, to maintain the same communication frequency, drivers continued to receive a daily reminder that they could access their own performance statistics in the app. With the exception of the normal piece rate, there was no other monetary incentive in any of the experimental conditions.

During the three-week status contest intervention, we find that drivers in the two virtual teams treatments generated 1.7% higher revenue than those in the control condition. Separately investigating the two treatments, we find that drivers under the team (individual) leaderboard treatment generated 1.8% (2%) higher revenue than those in the control condition. At the city level, the team (individual) leaderboard treatment leads to a 5.3% (2.3%) increase in driver revenue in Taiyuan (Beijing) compared to the control condition, whereas neither treatment has a significant effect in Kunming. The city-level difference is likely due to the fact that both Beijing and Taiyuan could only fulfill 90% of the passenger orders, whereas Kunming drivers could fulfill 98% of the orders prior to our experiment, which did not leave much room for improvement. Three months after the experiment ended, drivers in the team leaderboard treatment continued to work longer hours on the platform. Within virtual teams, laggards benefited the most from team contests.

Our results show that platform designers can leverage team identity and team contests to increase revenue and worker engagement in a gig economy. The virtual team organization form

is now part of DiDi's ecosystem.

Experiment Design

We conducted a natural field experiment on the DiDi platform involving 27,790 drivers across three cities with diverse in demographics, locations, and number of team contests hosted prior to our experiment (see Table S1 for more details). These cities include Beijing, Kunming, and Taiyuan. Our experiment is approved by the University of Michigan IRB (HUM00153090), and pre-registered at the AEA RCT Registry (*20*). The experiment started on October 22, 2018 and ended on December 3, 2018. To evaluate our treatment effect on driver retention, we continue to collect data for three months after our experiment, till March 1, 2019. Our experiment consists of the following stages (see Figure S1 in SI for the experimental process).

Driver recruitment and team formation. The platform sent out a call for participation on October 22, 2018, inviting all drivers each of the three cities to participate in a week-long team contest for a monetary prize. Interested drivers may sign up for the contest and start forming teams. Drivers can create a new team as a captain, invite others to join their team, or join an existing team upon receiving an invitation.

Each team is comprised of seven drivers, but fewer than 40% of the teams achieved the desired size during the team formation period. Those that reached the desired size are referred to as *self-formed* teams. At the end of recruitment stage, the system randomly selects 90% of the drivers in under-sized teams and groups them into full-sized teams, which we refer to as *system-formed* teams. The remaining 10% are not assigned to any team and will not participate in the contest. These drivers are referred to as *solo drivers*. Overall, about 36% of teams are self-formed across the three cities. Note that the system uses one of two algorithms to form teams, including hometown similarity and age similarity which were shown to be the most

successful among the team formation algorithms in our first field experiment on DiDi (*18*).

In addition to driver recruitment, we encourage all teams to conduct a team building exercise to strengthen team identity, i.e., coming up with a team name. More specifically, team captains can change the team name from the default name assigned by the computer ("X_Team" where X is an 8-digit random number) and receive 10 RMB bonus. At the end of this phase, about 63% of the teams came up with a team name. Based on our first experiment, teams with a stronger identity are more productive in the contest (*18*).

Driver in either self-formed or system-formed teams would participate in the multi-phase contests. To assign the teams into contest groups, we first sort all teams decreasingly within each city based on their prior productivity (the sum of individual team member revenue in the two weeks prior to this stage), regardless of its formation method. We then partition every five adjacent teams into a contest group, also referred to as a leaderboard. Each team only competes with other teams in the same leaderboard. Our grouping method ensures that teams in the same leaderboard have similar prior productivity. We now describe the three phases of team contest.

The pre-intervention contest. We first conduct a week-long best-of-five team contest to enhance team identity, starting October 29th. Social science experiments demonstrate that intergroup competition is among the most successful methods used to create a strong sense of group identity (*9*). In this contest, within each leaderboard, the team with the highest cumulative team revenue during the contest week wins a cash prize, whereas the other four teams receive no prize. Following DiDi's current contest practice, we exclude the lowest driver revenue in the team in each day when calculating the cumulative team revenue. This allows one driver to take a day off without affecting team performance. The cash prize is 1,000 RMB (per winning team) for Beijing, 650 for Taiyuan and Kunming, adjusted by the drivers' average hourly revenue in each city. For the winning team, the prize is allocated to team members proportional to their

contributions to the cumulative team revenue, shown to be incentivizing for group contests in the laboratory (*21*), and credited to their driver accounts immediately after the contest.

During this stage, drivers can use the DiDi app to access a team leaderboard and an individual leaderboard for social information, illustrated in Figure S2 in the Supplemental Materials. The team leaderboard shows the cumulative revenue of each of the five teams in descending order. The top three teams are highlighted with badges. The individual leaderboard shows the individual team members' daily revenue in descending order. In addition, we mark the average performance with a line on the individual leaderboard to enhance the effect of ranking (*22, 23*). The team ranking is updated every hour while individual revenue is updated in real time. We send a daily reminder of the contest and the leaderboards at the end of each day. The communication messages can be found in the Supplemental Materials 1.3.

The intervention: A status contest. Immediately after the short-term contest, we conduct a three-week status contest between November 5-25 to examine the effect of team identity on productivity and retention. We randomly assign each leaderboard to one of the three experimental conditions:

- *Team Leaderboard.* In this treatment, drivers continue to have access to both team and individual leaderboards as in the short-term contest. We send out a daily reminder to these drivers to check the rankings of the same five teams within their leaderboard, as well as individual teammate rankings within their team.

- *Individual Leaderboard.* In this treatment, drivers only have access to the individual leaderboard within a team. Again, we send out a daily reminder to drivers to check their individual rankings.

- *Control*. In the control condition, drivers cannot access either leaderboard. However, to

keep the same communication frequency, drivers continue to receive a daily reminder that they can access their own performance statistics in the app.

We announced the status contest on the first day of this stage. The randomization is stratified on the average productivity of the leaderboard prior to the experiment. Kolmogorov-Smirnov tests show that the distribution of pre-experiment revenue, age, gender or DiDi age is not significantly different in pairwise comparisons across the three conditions ($p > 0.10$, see Table S4 in Supplemental Materials). Furthermore, we do not provide monetary incentives for the status contest to focus only on the social information. Communications of this stage are detailed in §1.3 in Supplemental Materials.

The post-intervention contest. After the status contest is over, every team is again invited to participate in a one-week contest for a cash prize from November 27th to December 3rd, 2018. This contest is designed to evaluate the spillover effects of the leaderboard on driver productivity immediately after the intervention. The settings, including the prize, leaderboard, and communications of performance, are identical to as those in the pre-intervention contest. We announce the contest one day after the intervention. Communications related to this contest is again included in Section 1.3 of the Supplemental Materials.

The post-experiment survey. Finally, after the surprise contest, all drivers receive a survey which evaluates their sense of belonging to their team as well as to the organization (DiDi). The survey questions and responses are included in Section 8 of the Supplemental Materials.

Results

We evaluate the effect of virtual teams on driver productivity and retention, during the status contest intervention, as well as up to three months post contest. In addition to our pre-registered

hypotheses on the average treatment effects (*20*), we also explore heterogeneous treatment effects in different cities.

We first examine the average treatment effect on revenue during the experiment period. In Fig. 1a, we plot the weekly average revenue of each condition. To better compare the treatments, we realign the lines based on the pre-experiment period. The y-axis presents the revenue difference between a given week and the baseline week in the pre-experiment period. The three lines coincide up to the pre-intervention contest period. During the status contest intervention, drivers in different treatment conditions receive different social information and the lines start to diverge. Pooling all three cities (1a), we observe that treated drivers are more productive than those in the control condition, both during the status contest and the post-intervention contest.

To quantify the average treatment effects on outcome, Y, we construct the following difference-in-differences models:

$$\Delta \mathrm{Y}_{i,t} = \beta_0 + \beta_1 \cdot \text{Treated} + \alpha_c + \epsilon_{i,t}, \tag{1}$$

$$\Delta \mathrm{Y}_{i,t} = \beta_0 + \beta_1 \cdot \text{Team Leaderboard} + \beta_2 \cdot \text{Individual Leaderboard} + \alpha_c + \epsilon_{i,t},, \tag{2}$$

where $\Delta \mathrm{Y}_{i,t}$ represents the outcome change of the t-th week in the current period compared to the corresponding pre-contest week, and α_c captures the fixed effect of a city. We report the results of these models in Tables 1 to 5 in the main text, and robustness checks in the Supplemental Materials (shortened as SM henceforth). In these and subsequent analyses, we report the false discovery rate adjusted q-values in square brackets to correct for multiple hypothesis testing (*24*) using the Stata code provided by Anderson, 2008 (*25*). We follow the convention of using 5% (respectively 10%) cutoff for p-values (respectively q-values) to claim statistical significance (*26*).

Our first pre-registered hypothesis predicts that treated drivers will generate more revenue

than those in the control condition. That is, being reminded daily of belonging to a virtual team through a leaderboard will lead to higher revenue. This hypothesis implies that $\beta_1 > 0$ in Equation (1).

We test this hypothesis in column (1) of Table 1, which shows that the virtual team treatments increase driver revenue by 34.53 RMB, or 1.66% of the average weekly revenue per driver, during the three-week intervention ($p < .01$). Therefore, we reject the null in favor of $\beta_1 > 0$. Looking at each of the three cities separately, we find that the treatment has a significant effect in Beijing (41.67 RMB, $p < .01$, 1.69% of average weekly revenue), but not in Taiyuan or Kunming. The results become stronger when we additionally control for demographics and team formation method (specifications 4-6). Therefore, we reject the null in favor of $\beta_1 > 0$.

Investigating the two types of interventions separately (Equation 2), we further expect that drivers in the team leaderboard condition will generate higher revenue than those in the individual leaderboard condition, who in turn will generate higher revenue than those in the control condition during our intervention. This hypothesis implies that $\beta_1 > 0$, $\beta_2 > 0$, and that $\beta_1 > \beta_2$ in Equation (2).

We test these hypotheses in column (1) of Table 2, and we find that, during our three-week intervention, drivers in the team leaderboard treatment generate 32.12 RMB marginally higher revenue compared with the control group ($p < .10$). In comparison, drivers in the individual leaderboard condition generate 36.96 RMB higher revenue per week, or a 1.77% increase, compared to those in the control group ($p < .05$). After controlling for demographics and team formation method (column 5), the team and individual leaderboards generate 36.7 RMB and 41.47 RMB more revenue, equivalent to a 1.76% and 1.99% increase in weekly revenue, respectively ($p < .05$ in each case), although the difference between the two treatments is not significant ($p > 0.10$).

Although we did not pre-register any hypothesis on heterogeneous treatment effect, we

present some interesting results on how drivers in each city respond differently to our treatments, as revealed in Figures 1b, 1c, and 1d. In Beijing (column 2 in Table 2), only the individual leaderboard treatment has a significant treatment effect on driver revenue (56.32 RMB per week, or 2.29%, of the weekly revenue of the control group), whereas in Taiyuan (column 3 in Table 2), only the team leaderboard treatment has a significant effect on revenue compared to the control condition (58.49 RMB per week, $p < .05$). By contrast, neither treatment has a significant effect in Kunming. The lack of any treatment effect in Kunming might be explained by the fact that 98% of the passenger orders were fulfilled before the start of our experiment, which does not leave much room for an increasing of productivity. In comparison, 90% of the orders were fulfilled in Beijing and Taiyuan during the same time period (see Table S1 in SM). After controlling for demographics and team formation methods (columns 6-8 of Table 2), the city-specific treatment effects remain statistically and economically significant, with the individual and team leaderboard effect size equal to 59.24 RMB in Beijing and 62.31 RMB in Taiyuan, respectively ($p < .05$ in each case). Furthermore, the difference between the two treatments is in the direction we hypothesized, albeit marginally significant ($p < 0.10$). We summarize our results below.

Result 1 (Virtual teams and productivity). During the status contest intervention, (1) drivers in virtual teams generate 1.7% higher revenue than those in the control condition; (2) drivers under the team (individual) leaderboard treatment generate 1.8% (2%) higher revenue than those in the control condition. (3) At the city level, the team (individual) leaderboard treatment leads to a 5.3% (2.3%) increase in driver revenue in Taiyuan (Beijing) compared to the control condition, whereas neither treatment has a significant effect in Kunming.

For information interventions, such as the status contest in this experiment, an open question is whether and how long the effect persists when the intervention is over. To evaluate its short-

term spillover effect, we implemented a one-week best-of-five contest with monetary rewards immediately after the intervention. We expect that the treatment effects will persist during the post-intervention contest (pre-registered Hypothesis 2).

Table 3 presents our post-intervention contest analysis testing Hypothesis 2. As shown in column (1), drivers in the team leaderboard treatment generate 49.91 RMB higher weekly revenue, or a 2.49% increase, compared to those in the control group ($p < 0.05$). By contrast, drivers in individual leaderboard treatment do not differ significantly from those in the control condition ($p > 0.10$). Although the coefficient for the team leaderboard dummy is greater than that for the individual leaderboard, this difference is not significant ($p = 0.114$). These results become stronger after we control for demographics and team-formation method (column 5).

At the city level, we find that Beijing drivers generate 59.89 RMB marginally higher revenue than those in the control condition ($p < .10$, column 2). This effect becomes stronger and statistically significant after we control for demographics and team formation method (67.2 RMB, $p < .05$, column 6). By contrast, the individual leaderboard treatment does not have significant spillover effects compared to the control. In Taiyuan, drivers treated under the team leaderboard condition do not differ significantly from those under the control condition ($\beta_1 = 58.03$ RMB, $p = .12$), but generate significantly higher revenue than those under individual leaderboard condition ($\beta_1 \neq \beta_2$, $p < 0.01$ in columns 3 and 7). It is worth noting that the individual leaderboard treatment leads to a 68.26 RMB reduction in average weekly revenue, or a 6% drop, during the post-intervention contest compared to the control. In comparison, we do not observe any significant treatment effect in the post-intervention contest in Kunming, which again could be due to the near equilibrium level of demand and supply for rides prior to our experiment (Table S1). Again, these results become stronger after we control for demographics and team-formation method (column 5 in Table 3).

To investigate the extent to which various treatment effects are driven by team captains,

we re-run all analysis by excluding team captains, and find that our results are robust to this specification (see Tables S5 and S6, respectively). We summarize the results below.

Result 2 (Productivity spillover). During the one-week post-intervention contest, treated drivers in the team (individual) leaderboard condition continue to generate 2.49% more (the same) weekly revenue compared to those in the control condition.

As discussed in the introduction, ride-sharing platforms across the globe encounter the issue of driver retention. As such, an important goal for our intervention is to evaluate the effects of virtual teams on driver retention. We expect that drivers randomly assigned to a virtual team are more likely to stay on the platform than those in the control condition (our pre-registered Hypothesis 3).

As most drivers would take a vacation during the Spring Festival (also known as the Chinese New Year), starting on February 5, 2019, we had originally set a cutoff date on January 18, 2019, the last Friday before the beginning of Spring Festival Travel Season (Chunyun), to observe retention during the six weeks post experiment in our pre-analysis plan. It turns out that we are able to evaluate retention during the one-week, six-week, and three-month intervals post intervention. Unlike in the traditional sector where a worker who quits a job tends to completely disappear from the workplace, gig workers rarely delete the app. Instead, they just become more inactive. Therefore, we measure retention as the number of days a driver works in a week and separately analyze retention during the one week (Table S7), six weeks and three months (Table 5) after the post-intervention contest.

As shown in Fig. 2, drivers in the team leaderboard treatment consistently exhibit higher retention than those from either of the other experimental conditions after the experiment. Table 5 presents treatment effects on retention three months post our experiment (Equation 2). We find that drivers in team leaderboard treatment on average work 0.11 days (or one hour) more

than those in the control group ($p < 0.01$) in the week after the experiment (Table S7 column 1). The treatment effect is remarkably stable after six weeks ($\beta = 0.11$, $p < 0.01$, Table 4 column 1), and three months after the experiment ($\beta = 0.10$, $p < 0.01$, Table 5 column 1). Furthermore, drivers in the team leaderboard treatment also outperform those under the individual leaderboard treatment in each of the three periods ($p = 0.0004$ in the one-week window, Table S7 column 1, and $p = 0.0161$ in the three-month window, Table 5 column 1). And we observe no significant difference between the individual-leaderboard and the control. These results are robust after controlling for demographic covariates and team characteristics, such as winning the surprise contest or not. We also verified the robustness by repeating the regressions excluding the team captains.

At the city level, we again observe considerable heterogeneity in retention (Tables S7 to 5 columns 2 - 4). Indeed, only in Taiyuan do we see a consistent positive treatment effect of the team-leaderboard treatment (0.39 days $p < 0.01$ for the one-week window, and 0.33 days $p < 0.01$ for the three-month window). We observe significant and similar-sized effect between the team and individual leaderboard treatments. In Kunming, team leaderboard affects retention only during the one-week window (0.22 days, $p < 0.05$) and neither differs significantly from the control condition. We did not observe any significant difference between treatments in Beijing. We summarize the results below.

Result 3 (Virtual Teams and Retention). For up to three months after the experiment, drivers in the team leaderboard treatment work an average of one hour longer per week than those in the control condition. At the city level, Taiyuan drivers in the team leaderboard treatment work three hours longer per week, whereas treated drivers in Beijing and Kunming do not behave differently from those in the control condition.

Discussion and Conclusion

While the gig economy provides flexible jobs for millions of workers around the world, many gig platforms suffer from high attrition, partially due to the lack of organization identity and social bonds (*1*). In this paper, we investigate the efficacy of virtual teams on worker productivity and retention in a global ride-sharing platform. Using a large-scale natural field experiment with 27,790 drivers, we organize drivers into virtual teams and randomize these teams into three experimental conditions. Treated drivers receive their team ranking, or their individual ranking within a team, whereas those in the control condition receive individual performance information without social comparison. We find that treated drivers are significantly more productive than those in the control condition. Three months after the experiment ended, drivers in the team leaderboard treatment continue to work longer hours on the platform, indicating that virtual teams have the potential to increase worker identity with the platform and bonds with co-workers, which in turn increases productivity and worker retention.

References

1. A. J. Ravenelle, *Hustle and Gig: Struggling and Surviving in the Sharing Economy* (University of California Press, Oakland, California, 2019), first edition edn.

2. M. K. Chen, P. E. Rossi, J. A. Chevalier, E. Oehlsen, *Journal of Political Economy* **127**, 2735 (2019).

3. N. Heller, *The New Yorker* (2017).

4. J. M. Barrero, N. Bloom, S. J. Davis, Covid-19 is also a reallocation shock, *Tech. rep.*, National Bureau of Economic Research (2020).

5. H. Tajfel, J. Turner, *The Social Psychology of Intergroup Relations*, S. Worchel, W. Austin, eds. (Nelson- Hall, Chicago, 1986).

6. M. B. Brewer, *Journal of Social Issues* **55**, 429 (1999).

7. G. A. Akerlof, R. E. Kranton, *The Quarterly Journal of Economics* **115**, 715 (2000).

8. G. A. Akerlof, R. E. Kranton, *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being* (Princeton University Press, Princeton, New Jersey, 2010).

9. C. C. Eckel, P. J. Grossman, *Journal of Economic Behavior & Organization* **58**, 371 (2005).

10. G. Charness, L. Rigotti, A. Rustichini, *The American Economic Review* **97**, 1340 (2007).

11. R. Chen, Y. Chen, *The American Economic Review* **101**, 2562 (2011).

12. L. Goette, D. Huffman, S. Meier, M. Sutter, *Management Science* **58**, 948 (2012).

13. Y. Chen, S. X. Li, T. X. Liu, M. Shih, *Games and Economic Behavior* **84**, 58 (2014).

14. I. Erev, G. Bornstein, R. Galili, *Journal of Experimental Social Psychology* **29**, 463 (1993).

15. W. Ai, R. Chen, Y. Chen, Q. Mei, W. Phillips, *Proceedings of the National Academy of Sciences* **113**, 14944 (2016).

16. G. Charness, Y. Chen, *Annual Review of Economics* (2020).

17. O. Bandiera, I. Barankay, I. Rasul, *Journal of the European Economic Association* **11**, 1079 (2013).

18. W. Ai, Y. Chen, Q. Mei, J. Ye, L. Zhang, *Working paper* (2019).

19. T. Ye, *et al.*, *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2020, to appear).

20. T. Ye, W. Ai, Y. Chen, M. Qiaozhu, J. Zhang, *AEA RCT Registry* .

21. R. M. Sheremeta, *Journal of Economic Surveys* **32**, 683 (2018).

22. Y. Chen, F. M. Harper, J. Konstan, S. X. Li, *Amer. Econ. Rev.* **100**, 1358 (2010).

23. Y. Chen, F. Lu, J. Zhang, *Journal of Public Economics* **155**, 11 (2017).

24. Y. Benjamini, Y. Hochberg, *Journal of the Royal Statistical Society, Series B (Methodological)* **57**, 289 (1995).

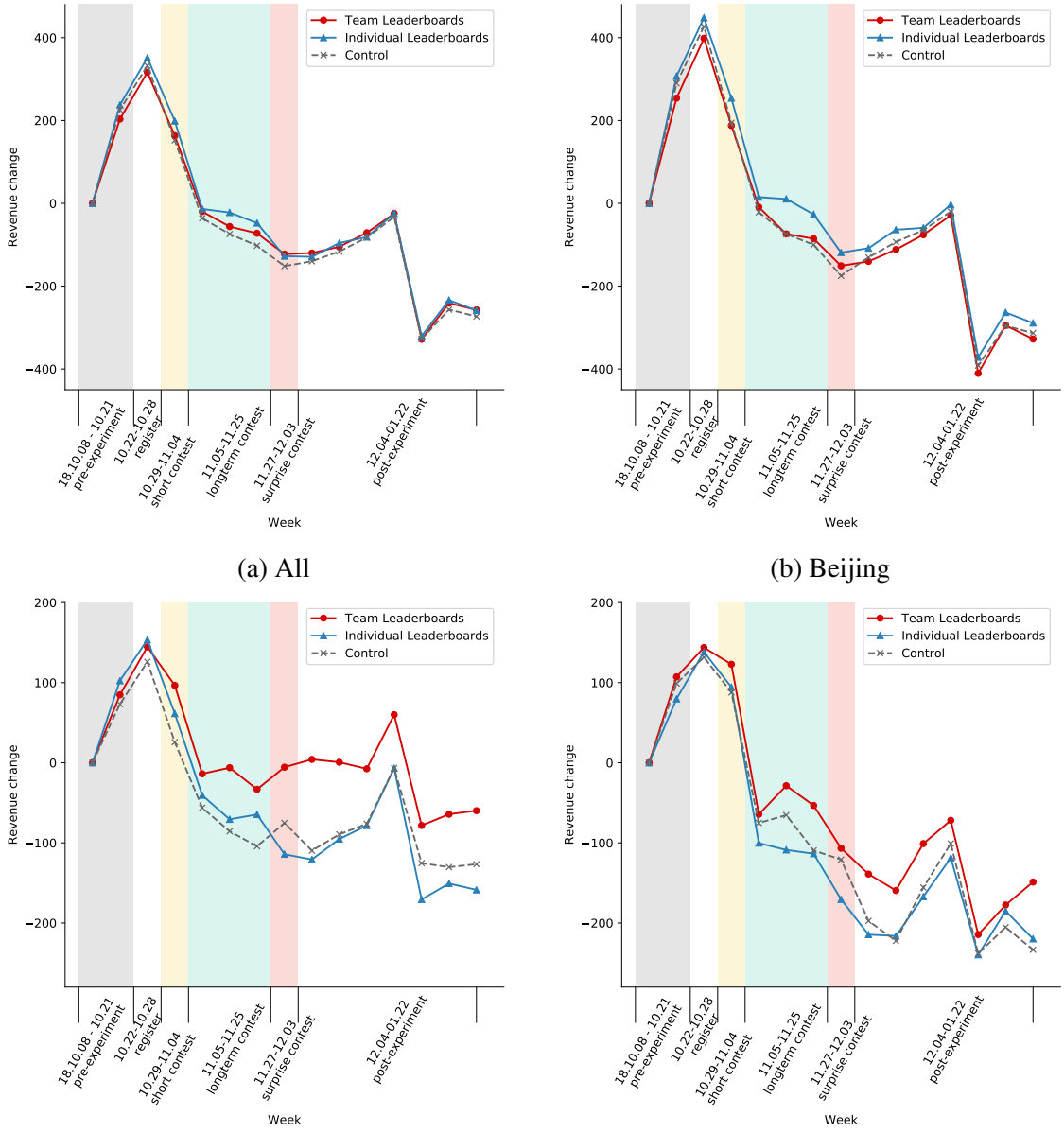
25. M. L. Anderson, *Journal of the American Statistical Association* **103**, 1481 (2008).

26. B. Efron, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction* (Stanford University Online Manuscript, 2010).

27. C. Bellemare, L. Bissonnette, S. Kröger, *Journal of the Economic Science Association* **2**, 157 (2016).

Acknowledgments

We thank Alain Cohn, Steve Leider and Tanya Rosenblat for helpful discussions and comments. The research has been approved by the University of Michigan IRB (HUM00153090) and pre-registered at AEA RCT registry (AEARCTR-0003537).



(a) All

(b) Beijing

(c) Taiyuan

(d) Kunming

Figure 1: Average revenue of each condition over week. To better visualize the change over time, we rescale the revenue of each condition in reference to its pre-experiment average weekly revenue from two weeks before the experiment to seven weeks after the experiment. For example, each point in the treatment line = the weekly average revenue per driver of treatment group - the mean of pre-experiment weekly average revenue per driver of treatment group.

To examine the general effect of having a leaderboard, we coded the binary variable has_a_leaderboard

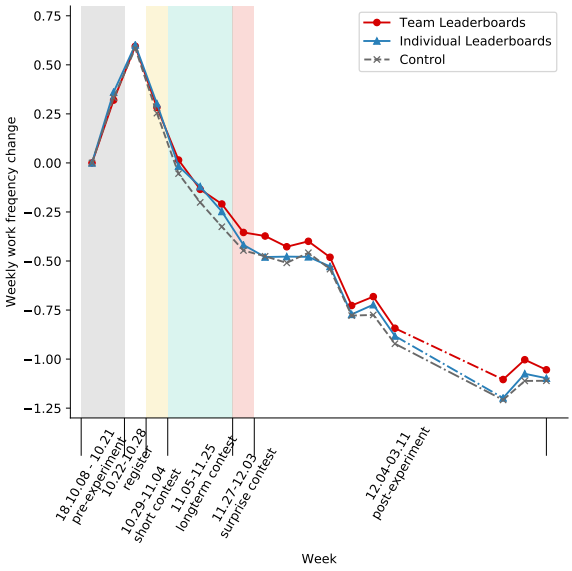


Figure 2: Average work frequency of each condition over week (all cities). To better visualize the change over time, we scale each condition by taking a difference of the average weekly days of driving during the week before the experiment. For example, each point in the treatment line = the weekly average working days per driver of treatment group - the mean of pre-experiment weekly average working days per driver of treatment group. The month of Spring Festival is omitted where the temporary retention (compared to that of the week before the experiment) ranges from -3.40 to -1.54 across different conditions.

as 0 if the driver is in the control group and as 1 if the driver is in the team-leaderboards or individual-leaderboards condition. We use models represented by Equation S2 and Equation S3 to capture the effect with and without controlling driver individual heterogeneity.

$$\Delta y_i = \beta_0 + \beta_1 \text{has\_a\_leaderboard} + \gamma_c + \epsilon_i \tag{3}$$

$$\begin{aligned}\Delta y_i =& \beta_0 + \beta_1 \text{has\_a\_leaderboard} + \beta_2 \text{age} + \beta_3 \text{DiDi\_age} \\ &+ \beta_4 \text{hometown\_distance\_to\_contest\_city} + \beta_5 \text{self\_formed} + \gamma_c + \epsilon_i\end{aligned} \tag{4}$$

Table 1: Average treatment effects on weekly revenue during the intervention – Difference-in-differences analysis with standard errors clustered at different levels. The coefficient represents weekly revenue (GMV) difference.

	Outcome variable: Δ of Weekly Revenue (CNY)							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Treated	34.53**	41.67**	33.99	8.25	39.08**	45.82**	38.40	14.53
(In a virtual team)	(15.37)	(21.01)	(23.86)	(24.97)	(15.31)	(20.93)	(23.69)	(24.81)
	[0.03]	[0.17]	[0.18]	[0.33]	[0.01]	[0.09]	[0.12]	[0.23]
Age					6.98***	7.47***	1.90	8.39***
					(0.83)	(1.17)	(1.37)	(1.27)
DiDi age					32.16***	40.85***	3.64	3.43
					(7.47)	(9.59)	(11.53)	(13.39)
Hometown distance					-0.02	-0.01	-0.12**	-0.03
to contest city					(0.02)	(0.02)	(0.05)	(0.02)
Self-formed team					-45.25***	-60.09***	-24.18	-4.10
					(16.09)	(21.59)	(27.49)	(26.90)
City fixed effect	yes	-	-	-	yes	-	-	-
# of clusters	11,890	8,100	1,625	2,165	11,890	8,100	1,625	2,165
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

Standard errors in parentheses are clustered at team level for ranking conditions and at individual level for control condition. False Discovery Rate q-values are calculated separately for All cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square bracket.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Effect on weekly revenue during the long term ranking period – Difference-in-differences analysis with standard errors clustered at different levels. The coefficient represents weekly revenue (GMV) difference.

	Outcome variable: Δ of Weekly Revenue (CNY)							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team L. (β_1)	32.12*	27.03	58.49**	30.54	36.70**	32.40	62.31**	33.81
	(17.97)	(24.61)	(26.60)	(29.91)	(17.90)	(24.50)	(26.57)	(29.69)
	[0.08]	[0.44]	[0.09]	[0.44]	[0.04]	[0.33]	[0.06]	[0.34]
Individual L. (β_2)	36.96**	56.32**	8.81	-14.50	41.47**	59.24**	13.68	-5.18
	(17.90)	(24.49)	(28.76)	(28.03)	(17.82)	(24.37)	(28.43)	(27.86)
	[0.08]	[0.09]	[0.86]	[0.86]	[0.04]	[0.06]	[0.61]	[0.62]
Age					6.98***	7.47***	1.91	8.35***
					(0.83)	(1.17)	(1.37)	(1.28)
DiDi age					32.15***	40.77***	3.57	3.29
					(7.46)	(9.59)	(11.57)	(13.39)
Hometown distance					-0.02	-0.01	-0.12**	-0.03
to contest city					(0.02)	(0.02)	(0.05)	(0.02)
Self formed					-45.22***	-59.76***	-23.62	-3.96
					(16.10)	(21.59)	(27.40)	(26.91)
City fixed effect	yes	-	-	-	yes	-	-	-
H0: $\beta_1 = \beta_2$ (p-value)	0.7933	0.2487	0.0782*	0.1274	0.7954	0.2877	0.0828*	0.1832
# of clusters	11,890	8,100	1,625	2,165	11,890	8,100	1,625	2,165
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

Standard errors in parentheses are clustered at team (individual) level for ranking (control) conditions. False Discovery Rate q-values are calculated separately for All cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square bracket.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Treatment effects on weekly revenue in the surprising short contest with(out) controlling individual heterogeneity – Difference-in-differences analysis results overview of Hypothesis 2 testing using 2nd week of pre-experiment (2018.10.15-10.21) data as baseline.

	Outcome variable: Δ of Weekly Revenue (CNY)							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team L. (β_1)	49.91**	59.89*	58.03	6.05	55.75**	67.20**	59.78	11.14
	(23.80)	(32.49)	(37.50)	(39.57)	(23.44)	(31.92)	(36.92)	(39.00)
	[0.08]	[0.32]	[0.32]	[0.56]	[0.04]	[0.27]	[0.27]	[0.39]
Individual L. (β_2)	11.75	38.98	-68.26*	-30.36	17.55	42.82	-65.75*	-18.30
	(24.30)	(33.12)	(39.25)	(39.52)	(23.84)	(32.42)	(38.27)	(39.01)
	[0.46]	[0.32]	[0.32]	[0.36]	[0.30]	[0.27]	[0.27]	[0.39]
Age					10.56***	11.31***	4.72***	11.57***
					(1.07)	(1.50)	(1.70)	(1.68)
DiDi age					84.14***	97.94***	38.20**	38.55**
					(9.62)	(12.33)	(15.49)	(17.20)
Hometown distance					-0.03	-0.04	-0.16**	0.02
to contest city					(0.02)	(0.03)	(0.06)	(0.03)
Self formed					-20.55	-39.15	23.93	28.60
					(21.57)	(28.73)	(38.61)	(37.24)
City fixed effect	yes	-	-	-	yes	-	-	-
H0: $\beta_1 = \beta_2$ (p-value)	0.1141	0.5277	0.0015***	0.3320	0.1078	0.4532	0.0014***	0.4248
# of clusters	3,970	2,700	545	725	3,970	2,700	545	725
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

Standard errors in parentheses are clustered at team level.
False Discovery Rate q-values are calculated separately for all cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square bracket.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Treatment effects on weekly number of work days during the week before Chunyun (till Friday: 2019.01.12-01.18), which is about six weeks after the experiment.

	Outcome variable: Δ of weekly # of work days							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team L.(β_1)	0.11**	0.08	0.24**	0.11	0.12***	0.10*	0.26**	0.12
	(0.04)	(0.05)	(0.11)	(0.10)	(0.04)	(0.05)	(0.11)	(0.10)
	[0.02]	[0.43]	[0.18]	[0.56]	[0.01]	[0.18]	[0.11]	[0.39]
Individual L.(β_2)	0.03	0.05	-0.08	0.03	0.04	0.06	-0.05	0.06
	(0.04)	(0.05)	(0.11)	(0.10)	(0.04)	(0.05)	(0.11)	(0.10)
	[0.36]	[0.57]	[0.63]	[0.98]	[0.20]	[0.39]	[0.71]	[0.71]
Age					0.03***	0.03***	0.01***	0.03***
					(0.00)	(0.00)	(0.01)	(0.00)
DiDi age					0.19***	0.22***	0.03	0.18***
					(0.02)	(0.02)	(0.05)	(0.04)
Hometown distance					-0.00***	-0.00**	-0.00**	-0.00
to contest city					(0.00)	(0.00)	(0.00)	(0.00)
Self formed					-0.05	-0.10**	-0.09	0.18**
					(0.04)	(0.05)	(0.10)	(0.09)
Team won in					0.69***	0.73***	0.58***	0.60***
surprise short contest					(0.04)	(0.05)	(0.11)	(0.10)
City fixed effect	yes	-	-	-	yes	-	-	-
H0: $\beta_1 = \beta_2$ (p-value)	0.0522*	0.5185	0.0033***	0.3690	0.0523*	0.4273	0.0036***	0.5289
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

False Discovery Rate q-values are calculated separately for All cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square bracket.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Treatment effects on weekly number of work days during the week after Chunyun (From Monday: 2019.03.04-03.10), which is about three months after the experiment.

	Outcome variable: Δ of weekly # of work days							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team L.(β_1)	0.10**	0.06	0.33***	0.05	0.11**	0.08	0.33***	0.06
	(0.05)	(0.06)	(0.12)	(0.10)	(0.04)	(0.05)	(0.11)	(0.10)
	[0.06]	[1.00]	[0.03]	[1.00]	[0.02]	[0.50]	[0.02]	[1.00]
Individual L.(β_2)	-0.01	-0.01	-0.02	0.01	0.01	-0.00	-0.01	0.04
	(0.05)	(0.06)	(0.12)	(0.10)	(0.04)	(0.05)	(0.12)	(0.10)
	[0.70]	[1.00]	[1.00]	[1.00]	[0.77]	[1.00]	[1.00]	[1.00]
Age					0.03***	0.03***	0.02***	0.03***
					(0.00)	(0.00)	(0.01)	(0.00)
DiDi age					0.22***	0.24***	0.08	0.18***
					(0.02)	(0.02)	(0.05)	(0.05)
Hometown distance					-0.00***	-0.00***	-0.00**	-0.00
to contest city					(0.00)	(0.00)	(0.00)	(0.00)
Self formed					-0.07*	-0.16***	0.10	0.16*
					(0.04)	(0.05)	(0.10)	(0.09)
Team won in					0.66***	0.68***	0.63***	0.61***
surprise short contest					(0.05)	(0.06)	(0.12)	(0.11)
City fixed effect	yes	-	-	-	yes	-	-	-
H0: $\beta_1 = \beta_2$ (p-value)	0.0161**	0.1656	0.0026***	0.6569	0.0179**	0.1229	0.0026***	0.8537
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

False Discovery Rate q-values are calculated separately for All cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square bracket. The results hold if we alternatively control number of wins in the two short contests in stead of team won in the surprise short contest.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Supplementary Materials for

paper title

author

1 Experiment Design Details

1.1 Selected Cities

Table S1: Characteristics summary of cities selected

City	Location	# of historical contests	Order-response rate	# of registered drivers	# of drivers in teams
Beijing	North	17	0.90	21,126	18,900
Taiyuan	Central	14	0.90	4,648	3,815
Kunming	Southeast	5	0.98	5,776	5,075

1.2 Experimental process

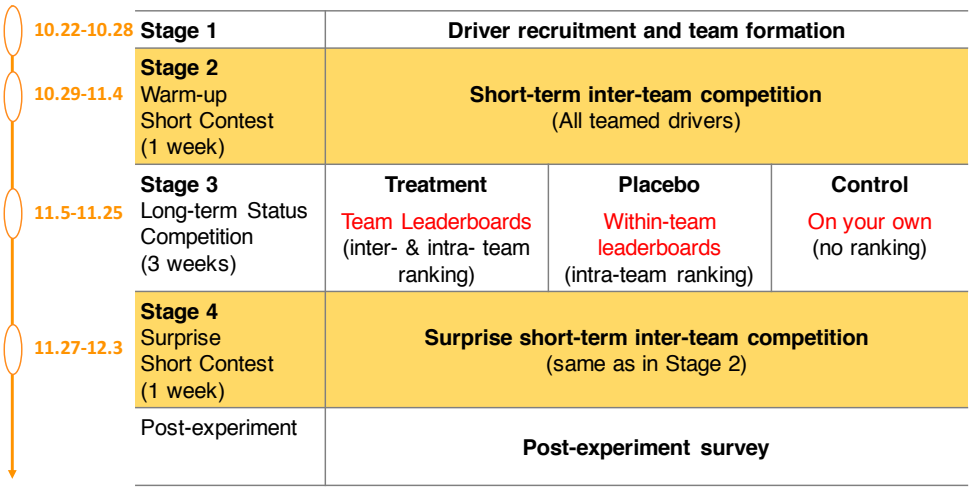


Figure S1: Experiment process

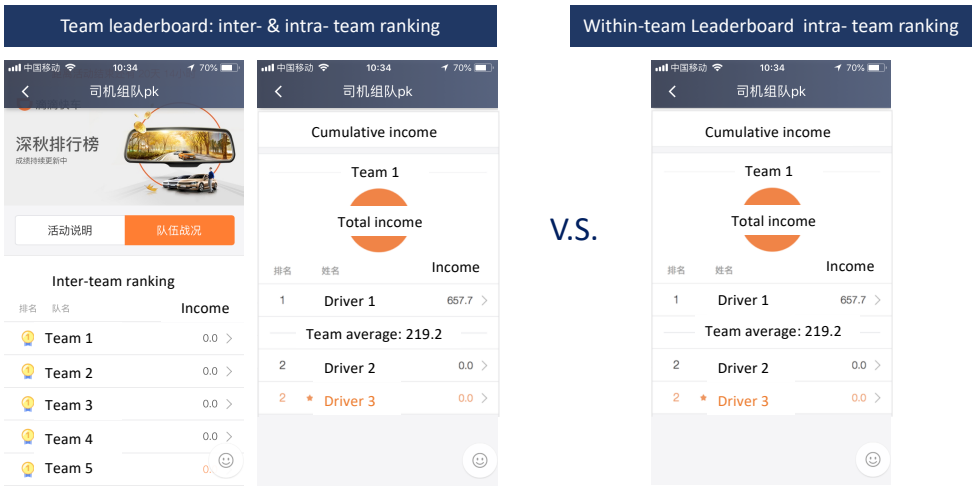


Figure S2: APP interface of team and individual leaderboards

APP interface of team and individual leaderboards

1.3 Messages sent in each stage

Stage 1: Driver recruitment and team formation

Our collaborators at DiDi sent out both text message and in-app push[1] to inform all drivers in the experimental cities that they could participate in a team contest.

The English translation of the call for participation to the drivers reads as follows:

> DiDi driver team contest is about to start soon! Say goodbye to the lonely driving work on your own. Getting to know new driver friends and compete for rewards with your teammates! Click here to register for the contest. Please keep up your good service and drive safely.

Stage 2: Warm-up rewarded short contest

we sent the following reminder by text message and in-app push to every driver every evening during the contest.

[1]Text message refers to the normal message sent out by DiDi. In-app push refers to the message popping up within DiDi app.

> The driver team contest has become more intense! Want to know your team's ranking? Want to check your teammates' performance? Want to know your competitors' performance? Click this link and you can access all the above information. Please keep up your good service and drive safely.

Stage 3: Long-term status competition

The corresponding notification and reminder of the three conditions in the long-term status competition include:

1. **Team leaderboards condition.**

   At the beginning of this stage, drivers in team leaderboards condition are notified by text message that:

   > The team contest is over. The ranking information will continue to be updated during November. Please pay your attention to the performance of your team and your teammates. DiDi is amazing because of you!

   The following reminder is sent by text message and in-app push once a day during the evening:

   > Latest performance just came out! Want to know your team's and teammates' performance? Click this link and you can access all the information. Please keep up your good service and drive safely.

2. **Individual leaderboards.** At the beginning of this stage, we send the following text message to notify drivers in individual leaderboards condition that:

> The team contest is over. The ranking information will maintain updated during November. Please keep your attention on the performance of your teammates. DiDi is amazing because of you!

The following individual performance reminders are sent every evening by both text message and in-app push:

> Latest performance just came out! Want to know your teammates' performance? Click this link and you can access all the information. Please keep up your good service and drive safely.

3. **Control.**

   At the beginning of this stage, we send the following text message to drivers in control group that:

   > The team contest is over. Please pay your attention to your performance. DiDi is amazing because of you!

   Individual performance update reminder is sent every evening by text message and in-app push as follows:

   > Latest performance just came out! Want to know the your outcome? Click this link and you can go to the your revenue page. Please keep up your good service and drive safely.

Stage 4: Surprise contest

The following text message announcement is sent to all drivers in the three conditions on the day before surprise contest to notify the surprise contest:

Here comes the driver team contest again (from 2018.11.27 to 2018.12.3)! You don't need to form the team again. Team members and competitor teams will remain the same teammates as in the last contest. Please contact your team members and get ready to compete for the cash prize!

1.4 Prize determination across cities

To make the experiment in each city most comparable, we determine the bonus volume for the winner team by keeping the rate of bonus over drivers' hourly earning the same across the experimental cities. Specifically, we first calculated the average hourly pay with 30-day data before DiDi set up the experiment. We carefully excluded the national holiday period (2018/10/01 - 2018/10/07) since it had various effects the ride sharing business but we focused more on normal patterns. As a result, we calculated the average hourly pay based on data from 2018/09/10 - 2018/09/29 and 2018/10/08 - 2018/10/17. The details of financial reward for each city are reported in Table S2.

Table S2: Details of prize in each city (money in CNY)

City	Calculated team prize	Rounded team prize	Team leader extra prize
Beijing	1000	1000	10
Taiyuan	654.21	650	10
Kunming	663.02	650	10

2 Power analysis

We use a subset of the experimental data from our 2017 field experiment conducted among DiDi drivers in the city of Dongguan to generate an estimated effect size and variance parameters for the power analysis and sample size calculation. For our experiment, we would like to have a

sample size large enough to obtain 90% power.

In the 2017 experiment, drivers are randomized into the treatment and control conditions. Among the treated drivers, some teams are responsive, measured by whether the team captain submitted a questionnaire before the start of the contest, while others are not. We use the unresponsive teams as an approximation for the control condition of this new experiment, and the responsive teams as an approximation for the treatment. We decide not use the 2017 placebo drivers as an approximation to our current control because they did not form teams at all. We use the five contest days as 5 periods. With this setting, we run the following fixed effects panel regression:

Table S3: Panel analysis with 2017 experiment data by fixed-effects (within-subject) regression

	Δ of Daily Orders
Game day	-1.35**
	(0.29)
Responsive	2.81 **
	(0.37)
# of observations = 17,500; # of groups = 250;	
σ_u = 4.01; σ_e = 12.10; ρ = 0.10;	
Standard errors in parentheses.	
$^{*}\ p < 0.1$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$	

According to the results in Table S3, we use the PowerBBK package (*27*) to compute the power of the new design, assuming similar behavioral responses as in the 2017 experiments.

The parameters are determined based on the following considerations (see Table S3 for statistics):

- budget = 125 teams per condition $\times$ 2 experimental conditions $\times$ 5 contest periods $=$ 1250.

- beta $= (15.24, 2.8)$ since (1) $15.24 = 16.582 - 1.347$ is the daily number of trips of the

unresponsive teams during the contest, (2) whereas 2.8 is the treatment effect of responsiveness.

- muvar $= \sigma_u^2 = 16$.
- espva $= \sigma_e^2 = 144$.
- panel allocation $= 0.4$ since 40% of the teams were unresponsive.

This command outputs power equals 0.896. As we plan to have three experimental conditions, we need 375 teams.

Increasing budget by 1.5 (from 250 to 375 teams in two conditions) would give us a power of 0.982. In this case, having 564 teams (4,000 drivers) would be sufficient. The caveat is that we do not know the potential treatment effect in the leader board phase, and therefore, cannot account for that in our power calculation.

3 Randomization check

Table S4: Randomization check and summary of statistics

	Beijing			Taiyuan			Kunming		
	Team	Individual	Control	Team	Individual	Control	Team	Individual	Control
Daily Revenue	381.97	381.71	381.70	171.64	180.54	176.09	212.71	214.21	218.03
before experiment	(215.35)	(216.15)	(213.83)	(126.25)	(129.99)	(125.88)	(144.30)	(143.99)	(144.56)
Age	36.82	36.91	37.35	36.53	36.63	36.86	36.49	35.91	37.02
	(8.12)	(8.09)	(8.28)	(8.26)	(8.22)	(8.34)	(8.50)	(8.58)	(8.81)
Male	0.97	0.97	0.97	0.97	0.95	0.96	0.93	0.92	0.93
	(0.17)	(0.16)	(0.17)	(0.18)	(0.22)	(0.19)	(0.26)	(0.26)	(0.26)
DiDi age (month)	24.69	24.97	24.86	24.08	23.88	24.55	15.06	14.70	14.51
	(13.19)	(13.11)	(13.01)	(11.13)	(11.62)	(11.04)	(11.04)	(11.01)	(10.98)
# of leaderboards	180	180	180	37	36	36	49	48	48
# of drivers	6,300	6,300	6,300	1,295	1,260	1,260	1,715	1,680	1,680

Standard deviation in parentheses

4 Treatment effect on driver revenue

Table S5: (H1) Removal of captain: Effect on weekly revenue during the long term ranking period – Difference in difference analysis with standard error clustered at different levels. The coefficient represents weekly revenue (GMV) difference.

	Outcome variable: Δ of Weekly Revenue (CNY)							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team L. (β_1)	35.90*	26.81	69.93**	43.43	41.13**	32.82	73.10**	47.72
	(19.30)	(26.54)	(28.34)	(30.87)	(19.24)	(26.47)	(28.36)	(30.58)
	[0.03]	[0.46]	[0.04]	[0.27]	[0.02]	[0.27]	[0.03]	[0.19]
Individual L. (β_2)	48.39**	65.11**	25.95	2.55	52.92***	68.21***	30.54	11.61
	(19.12)	(26.14)	(31.12)	(29.97)	(19.07)	(26.06)	(30.96)	(29.89)
	[0.02]	[0.04]	[0.47]	[0.87]	[0.01]	[0.03]	[0.31]	[0.48]
Age					6.54***	7.13***	1.31	7.75***
					(0.90)	(1.26)	(1.54)	(1.37)
DiDi age					30.20***	38.62***	7.69	-1.25
					(8.05)	(10.32)	(13.22)	(14.06)
Hometown distance					-0.01	-0.00	-0.13**	-0.02
to contest city					(0.02)	(0.03)	(0.05)	(0.03)
Self formed					-43.83**	-54.71**	-19.99	-21.97
					(17.19)	(23.11)	(29.20)	(28.05)
City fixed effect	yes	-	-	-	yes	-	-	-
H0: $\beta_1 = \beta_2$ (p-value)	0.5241	0.1557	0.1419	0.1825	0.5464	0.1878	0.1541	0.2343
# of clusters	10,570	7,200	1,445	1,925	10,570	7,200	1,445	1,925
# of drivers	23,820	16,200	3,270	4,350	23,820	16,200	3,270	4,350

Standard errors in parentheses are clustered at team (individual) level for ranking (control) conditions. False Discovery Rate q-values are calculated separately for All cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square bracket.

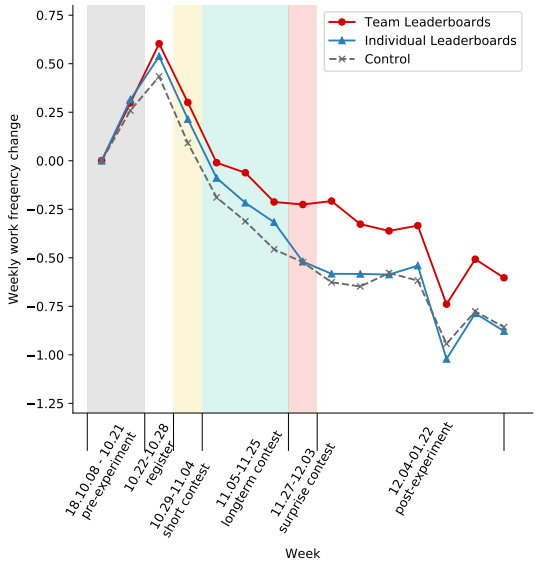
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table S6: (H2) Removal of captains: Effect on weekly revenue during the surprising short contest – Difference in difference analysis results overview of Hypothesis 2 testing using 2nd week of pre-experiment (2018.10.15-10.21) data as baseline

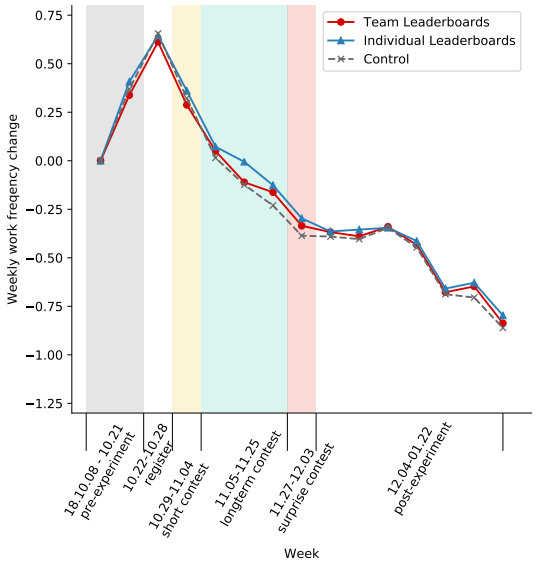
	Outcome variable: Δ of Weekly Revenue (CNY)							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team L. (β_1)	56.43**	61.42*	72.75*	24.86	64.07***	70.28**	74.40**	32.40
	(24.94)	(34.15)	(38.10)	(40.85)	(24.58)	(33.61)	(37.54)	(40.17)
	[0.05]	[0.28]	[0.28]	[0.35]	[0.02]	[0.17]	[0.17]	[0.34]
Individual L. (β_2)	21.20	47.30	-55.35	-19.28	27.71	51.81	-52.66	-7.36
	(25.36)	(34.63)	(39.99)	(40.84)	(24.95)	(34.02)	(39.08)	(40.50)
	[0.25]	[0.28]	[0.28]	[0.35]	[0.15]	[0.21]	[0.22]	[0.40]
Age					10.60***	11.55***	4.08**	11.27***
					(1.14)	(1.60)	(1.77)	(1.79)
DiDi age					81.98***	95.10***	29.31*	45.30**
					(10.34)	(13.22)	(16.78)	(18.50)
Hometown distance					-0.02	-0.03	-0.15**	0.04
to contest city					(0.02)	(0.03)	(0.07)	(0.03)
Self formed					-28.48	-45.97	18.55	9.83
					(22.57)	(30.18)	(39.59)	(38.17)
City fixed effect	yes	-	-	-	yes	-	-	-
H0: $\beta_1 = \beta_2$ (p-value)	0.1638	0.6845	0.0016***	0.2597	0.1446	0.5889	0.0016***	0.3021
# of clusters	3,970	2,700	545	725	3,970	2,700	545	725
# of drivers	23,820	16,200	3,270	4,350	23,820	16,200	3,270	4,350

Standard errors in parentheses are clustered at team level.
False Discovery Rate q-values are calculated separately for All cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square bracket.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

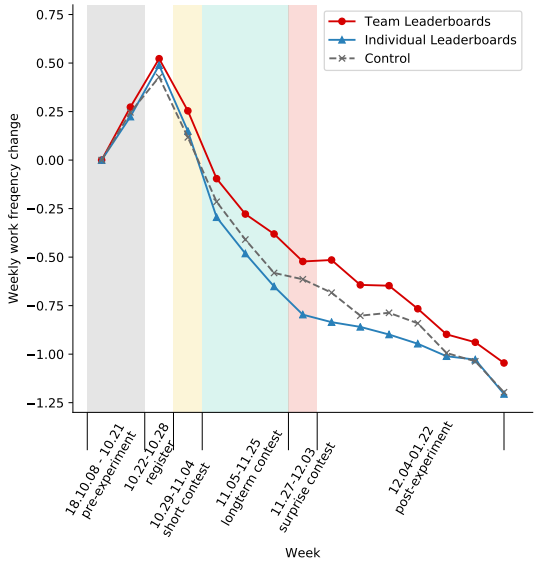
5 Treatment effect on driver retention



(a) Taiyuan



(b) Beijing



(c) Kunming

Figure S3: Average work frequency of each condition over week. To better visualize the change over time, we scale each condition by taking a difference of the average weekly days of driving during the week before the experiment. For example, each point in the treatment line = the weekly average working days per driver of treatment group - the mean of pre-experiment weekly average working days per driver of treatment group.

Table S7: Treatment effects on weekly number of work days during the week after the contest (2018.12.05-12.11).

	Outcome variable: Δ of weekly # of work days							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team L.(β_1)	0.11***	0.05	0.39***	0.14	0.12***	0.06	0.41***	0.15
	(0.04)	(0.05)	(0.11)	(0.09)	(0.04)	(0.05)	(0.11)	(0.09)
	[0.01]	[0.61]	[0.002]	[0.46]	[0.004]	[0.43]	[0.001]	[0.34]
Individual L.(β_2)	-0.03	-0.01	-0.01	-0.12	-0.02	-0.00	0.02	-0.09
	(0.04)	(0.05)	(0.11)	(0.09)	(0.04)	(0.05)	(0.11)	(0.09)
	[0.30]	[0.90]	[0.90]	[0.55]	[0.48]	[0.86]	[0.86]	[0.51]
Age					0.02***	0.02***	0.02***	0.03***
					(0.00)	(0.00)	(0.01)	(0.00)
DiDi age					0.14***	0.15***	0.01	0.15***
					(0.02)	(0.02)	(0.05)	(0.04)
Hometown distance					-0.00	-0.00	-0.00	0.00
to contest city					(0.00)	(0.00)	(0.00)	(0.00)
Self formed					-0.08**	-0.13***	-0.16*	0.19**
					(0.03)	(0.04)	(0.09)	(0.08)
Team won in					0.86***	0.91***	0.77***	0.71***
surprise short contest					(0.04)	(0.05)	(0.11)	(0.10)
City fixed effect	yes	-	-	-	yes	-	-	-
H0: $\beta_1 = \beta_2$ (p-value)	0.0004***	0.2409	0.0002***	0.0054***	0.0003***	0.1813	0.0002***	0.0090**
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

False Discovery Rate q-values are calculated separately for All cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square bracket. The results hold if we alternatively control number of wins in the two short contests in stead of team won in the surprise short contest.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table S8: Removal of captains: Treatment effects on weekly number of work days during the week after the contest (2019.12.05-12.11)

	Outcome variable: Δ of weekly # of work days							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team L.(β_1)	0.13***	0.04	0.47***	0.22**	0.15***	0.05	0.49***	0.23**
	(0.04)	(0.05)	(0.11)	(0.10)	(0.04)	(0.05)	(0.11)	(0.10)
	[0.01]	[0.94]	[0.001]	[0.09]	[0.002]	[0.59]	[0.001]	[0.06]
Individual L.(β_2)	-0.00	-0.01	0.09	-0.05	0.01	0.00	0.10	-0.02
	(0.04)	(0.05)	(0.11)	(0.10)	(0.04)	(0.05)	(0.11)	(0.10)
	[0.95]	[1.00]	[0.94]	[1.00]	[0.66]	[1.00]	[0.59]	[1.00]
Age					0.02***	0.02***	0.02***	0.03***
					(0.00)	(0.00)	(0.01)	(0.00)
DiDi age					0.13***	0.15***	-0.02	0.15***
					(0.02)	(0.02)	(0.05)	(0.05)
Hometown distance					-0.00	-0.00	-0.00*	0.00
to contest city					(0.00)	(0.00)	(0.00)	(0.00)
Self formed					-0.07*	-0.13***	-0.09	0.16*
					(0.04)	(0.04)	(0.10)	(0.09)
Team won in					0.87***	0.95***	0.74***	0.71***
surprise short contest					(0.04)	(0.05)	(0.12)	(0.10)
City fixed effect	yes	-	-	-	yes	-	-	-
H0: $\beta_1 = \beta_2$ (p-value)	0.0024***	0.4099	0.0006***	0.0094***	0.0017***	0.3173	0.0006***	0.0123**
# of drivers	23,820	16,200	3,270	4,350	23,820	16,200	3,270	4,350

False Discovery Rate q-values are calculated separately for All cities (1) & (3) and for individual cities (2-4) & (5-8) and are reported in square bracket. The results hold if we alternatively control number of wins in the two short contests in stead of team won in the surprise short contest.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table S9: Removal of captains: Treatment effects on weekly number of work days during the week after the contest (2019.03.04-03.10)

	Outcome variable: Δ of weekly # of work days							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Team L.(β_1)	0.09*	0.04	0.38***	0.06	0.11**	0.06	0.37***	0.07
	(0.05)	(0.06)	(0.12)	(0.11)	(0.05)	(0.06)	(0.12)	(0.11)
	[0.15]	[1.00]	[0.01]	[1.00]	[0.06]	[1.00]	[0.02]	[1.00]
Individual L.(β_2)	-0.03	-0.05	0.04	0.00	-0.01	-0.03	0.03	0.03
	(0.05)	(0.06)	(0.13)	(0.11)	(0.05)	(0.06)	(0.13)	(0.11)
	[0.42]	[1.00]	[1.00]	[1.00]	[0.78]	[1.00]	[1.00]	[1.00]
Age					0.03***	0.03***	0.02***	0.03***
					(0.00)	(0.00)	(0.01)	(0.01)
DiDi age					0.21***	0.24***	0.08	0.16***
					(0.02)	(0.02)	(0.06)	(0.05)
Hometown distance					-0.00***	-0.00***	-0.00**	-0.00
to contest city					(0.00)	(0.00)	(0.00)	(0.00)
Self formed					-0.07	-0.17***	0.19*	0.15
					(0.04)	(0.05)	(0.11)	(0.10)
Team won in					0.64***	0.68***	0.56***	0.54***
surprise short contest					(0.05)	(0.06)	(0.13)	(0.11)
City fixed effect	yes	-	-	-	yes	-	-	-
H0: $\beta_1 = \beta_2$ (p-value)	0.0174**	0.1570	0.0056***	0.6067	0.0171**	0.1104	0.0055***	0.7364
# of drivers	23,820	16,200	3,270	4,350	23,820	16,200	3,270	4,350

False Discovery Rate q-values are calculated separately for All cities (1) & (3) and for individual cities (2-4) & (5-8) and are reported in square bracket. The results hold if we alternatively control number of wins in the two short contests in stead of team won in the surprise short contest.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

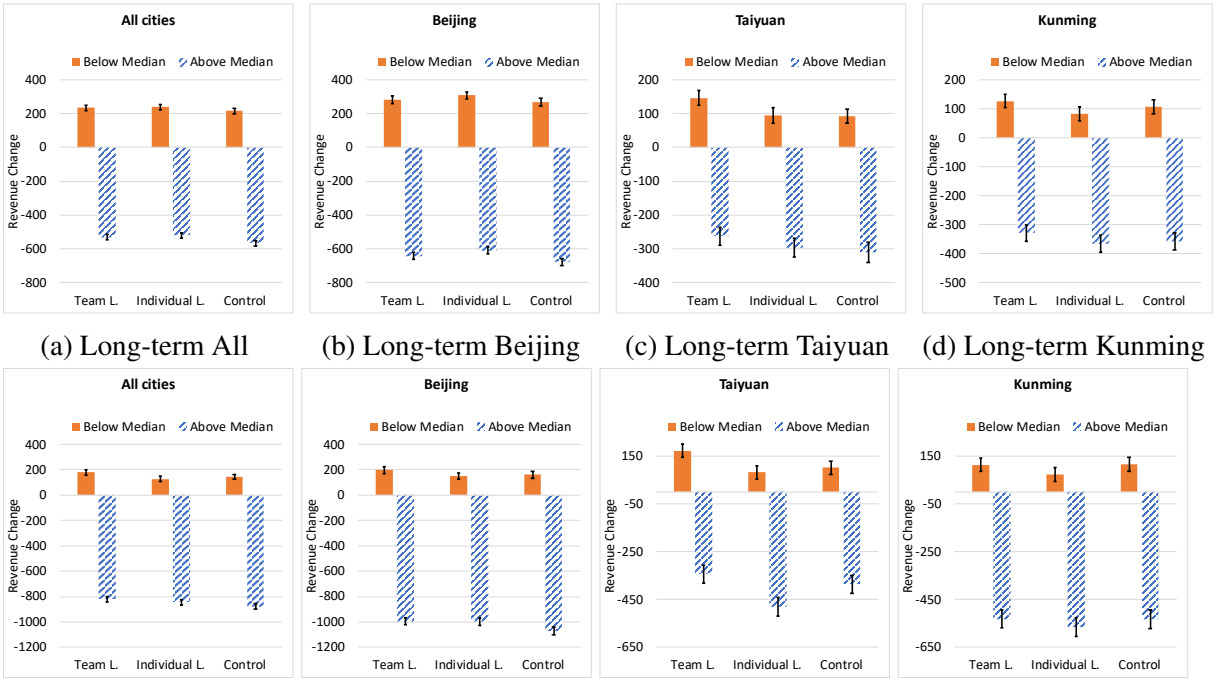
6 Heterogeneous analysis on driver pre-contest revenue

Focusing on how social comparison help individuals, we then ask who benefits more from the team identity and social information. We additionally test the heterogeneous treatment effects on drivers of different levels of pre-experiment revenue. We differentiate drivers by whether their pre-experiment revenue is below the city median. As shown in Fig. S4, drivers whose revenue fall in the lower half in their city consistently generate higher revenue increase than their counterparts, in both the longer-term status competition and the surprise short contest, and across all cities.

Specifically during the longer-term contest, pooling drivers in all cities (table S10 (1)), drivers whose pre-experiment revenue are below the city median generate 782.07 Yuan more than drivers whose pre-experiment revenue are above the city median ($p < .01$), accounting for about 37.53% average weekly revenue. This pattern is consistent in each of the three cities, with a revenue increase of 943.36 Yuan in Beijing (38.32% of Beijing average weekly revenue, $p < .01$), 401.99 Yuan in Taiyuan (36.08% of Taiyuan average weekly revenue, $p < .01$), and 462.37 Yuan in Kunming (33.19% of Kunming average weekly revenue, $p < .01$). No interaction effect is identified across cities and treatments. Additional tests show that below-median drivers in team-leaderboards condition (H0: $\beta_3 + \beta_4 = 0$) and individual-leaderboards (H0: $\beta_3 + \beta_5 = 0$) condition both have higher revenue increase during the long-term status competition overall and in each of the three cities.

According to Table S11, drivers with below-median revenue also benefit more in rewarded surprise contest: they generate a higher revenue of 1024.83 Yuan ($p < .01$) than the above-median drivers overall, which accounts for 51.16% average weekly revenue of all drivers in the control groups in three cities. Among them, below-median drivers in Beijing get a higher increase of 1232.91 Yuan (52.42% of Beijing average weekly revenue, $p < .01$), while drivers

in Taiyuan and Kunming generate 488.47 Yuan (43.58% of Taiyuan average weekly revenue, $p < .01$) and 647.77 Yuan (47.78% of Kunming average weekly revenue, $p < .01$) respectively than the above-median drivers. Results of additional tests (H0: $\beta_3 + \beta_4 = 0$ and H0: $\beta_3 + \beta_5 = 0$) confirm that below-median drivers in both team-leaderboards condition and individual-leaderboards condition have higher revenue increase during the surprise contest period in overall and each of the three cities.



(a) Long-term All (b) Long-term Beijing (c) Long-term Taiyuan (d) Long-term Kunming

(e) Short-term All (f) Short-term Beijing (g) Short-term Taiyuan (h) Short-term Kunming

Figure S4: The effect of team and individual leaderboards for drivers with below and above median pre-contest revenue with standard error as error bars.

Table S10: Heterogeneous treatment effects of pre-contest revenue levels on weekly revenue during the long term ranking period. The coefficient represents weekly revenue (GMV) difference.

	Outcome: Δ of Weekly Revenue (CNY)			
	(1)	(2)	(3)	(4)
	All	Beijing	Taiyuan	Kunming
Team L. (β_1)	35.43	34.86	47.50	28.68
	(23.86)	(31.53)	(40.82)	(43.01)
Individual L. (β_2)	46.50**	67.87**	14.69	-8.29
	(23.59)	(31.28)	(42.13)	(42.99)
Below median (β_3)	782.07***	943.36***	401.99***	462.37***
	(23.15)	(31.32)	(36.21)	(38.36)
Team L. * Below median (β_4)	-17.03	-20.60	6.94	-7.68
	(34.34)	(46.24)	(51.32)	(53.54)
Individual L. * Below median (β_5)	-22.48	-27.47	-13.00	-14.47
	(34.15)	(45.60)	(52.96)	(54.99)
City fixed effect	yes	-	-	-
H0: $\beta_1 = \beta_2$ (p-value)	0.6527	0.3087	0.4301	0.4030
H0: $\beta_3 + \beta_4 = 0$ (p-value)	0.0000***	0.0000***	0.0000***	0.0000***
H0: $\beta_3 + \beta_5 = 0$ (p-value)	0.0000***	0.0000***	0.0000***	0.0000***
H0: $\beta_1 + \beta_4 = 0$ (p-value)	0.4584	0.6771	0.0865*	0.5554
H0: $\beta_2 + \beta_5 = 0$ (p-value)	0.3318	0.2320	0.9594	0.5024
H0: $\beta_1 + \beta_4 = \beta_2 + \beta_5$ (p-value)	0.8291	0.4644	0.1272	0.2105
# of drivers	27,790	18,900	3,815	5,075

Standard errors in parentheses are clustered at team level.

Table S11: Heterogeneous treatment effects of pre-contest revenue levels on weekly revenue during the surprise short term ranking period. The coefficient represents weekly revenue (GMV) difference.

	Outcome: Δ of Weekly Revenue (CNY)			
	(1)	(2)	(3)	(4)
	All	Beijing	Taiyuan	Kunming
Team L. (β_1)	59.42*	76.64*	42.60	1.55
	(33.43)	(43.98)	(59.80)	(58.05)
Individual L. (β_2)	32.66	74.63*	-94.56	-31.43
	(33.90)	(44.79)	(61.10)	(59.24)
Below median (β_3)	1024.83***	1232.91***	488.47***	647.77***
	(31.18)	(40.74)	(53.34)	(47.48)
Team L. * Below median (β_4)	-27.95	-41.66	28.43	-2.42
	(44.57)	(58.94)	(70.18)	(68.47)
Individual L. * Below median (β_5)	-48.25	-84.93	75.07	-9.57
	(44.16)	(58.13)	(72.48)	(69.90)
City fixed effect	yes	-	-	-
H0: $\beta_1 = \beta_2$ (p-value)	0.4301	0.9645	0.0222**	0.5695
H0: $\beta_3 + \beta_4 = 0$ (p-value)	0.0000***	0.0000***	0.0000***	0.0000***
H0: $\beta_3 + \beta_5 = 0$ (p-value)	0.0000***	0.0000***	0.0000***	0.0000***
H0: $\beta_1 + \beta_4 = 0$ (p-value)	0.2976	0.3960	0.0888*	0.9838
H0: $\beta_2 + \beta_5 = 0$ (p-value)	0.6027	0.7999	0.6336	0.3079
H0: $\beta_1 + \beta_4 = \beta_2 + \beta_5$ (p-value)	0.1188	0.2703	0.0267**	0.3491
# of drivers	27,790	18,900	3,815	5,075

Standard errors in parentheses are clustered at team level.

6.1 Preference towards captain

We also conduct analysis to understand drivers' preference to be a team captain.

Hypothesis 1 (Captain) *Drivers with higher productivity prior to our experiment and longer Didi age, and who have served as a team captain in previous contests, will be more likely to*

volunteer to be team captains.

We use a Logistic regression model (eq. S1) to understand how driver's past experience on DiDi affects driver's choice to be a captain (H4), where V refers to the indicator function which equals 1 if a driver volunteers to be a team captain, and *Pre_Experiment_Productivity* is operationalized as the driver revenue in two weeks before our experiment. *Served_as_Captain_Before* is a binary variable that shows whether the driver had been a captain before he participated in the current team contest. We include γ_c to control city specific characteristics.

$$Pr(V = 1) = \Phi(\beta_0 + \beta_1 \text{Pre\_Experiment\_Productivity} + \beta_2 \text{Served\_as\_Captain\_before} + \beta_3 \text{Didi\_Age} + \gamma_c) \quad \text{(S1)}$$

The results (Table SS12) show that drivers with higher performance prior to the experiment and having served as captains before are significantly more likely to volunteer to be a captain in both three cities overall and separately. However, the effects of DiDi age are more complicated. DiDi age is positively correlated with captain preference overall and in Beijing (with $\beta = 0.0137$, $p < .01$ and $\beta = 0.0205$, $p < .01$, respectively), while it is negatively related to captain preference in Taiyuan (with $\beta = -0.0141$, $p < .05$) and has no significant relationship with captain preference in Kunming (with $\beta = 0.0011$, $p = 0.8262$)

Table S12: Results overview of Hypothesis 4 testing with logistic regression with all teamed drivers

	Outcome: Whether drivers volunteer to be captains			
	(1)	(2)	(3)	(4)
	All	Beijing	Taiyuan	Kunming
Pre Experiment Productivity	0.0044***	0.0040***	0.0066**	0.0084***
(in 1000 RMB)	(0.0007)	(0.0008)	(0.0030)	(0.0023)
	[0.001]	[0.001]	[0.012]	[0.001]
Served as captain before	0.2232***	0.2205***	0.2336***	0.2172***
(Binary indicator)	(0.0039)	(0.0041)	(0.0168)	(0.0111)
	[0.001]	[0.001]	[0.001]	[0.001]
DiDi age	0.0137***	0.0205***	-0.0141**	0.0011
(in years)	(0.0020)	(0.0023)	(0.0060)	(0.0052)
	[0.001]	[0.001]	[0.009]	[0.102]
City fixed effect	yes	-	-	-
# of drivers	27,790	18,900	3,815	5,075

Average marginal effect with *delta-method* SE in parentheses. False Discovery Rate q-values are calculated separately for All cities (1) and for individual cities (2-4) and are reported in square bracket.

$^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$

Table S13: Summary of heterogeneity across cities and conditions

	(1) Beijing			(2) Kunming			(3) Taiyuan		
	mean	sd	count	mean	sd	count	mean	sd	count
Age (driver-level)									
Team L.	37.83	8.12	6300	37.49	8.50	1715	37.53	8.26	1295
Individual L.	37.91	8.09	6300	36.91	8.57	1680	37.63	8.22	1260
Control	38.35	8.28	6300	38.02	8.81	1680	37.86	8.34	1260
Total	38.03	8.17	18900	37.47	8.64	5075	37.67	8.27	3815
Hometown distance to the contest city (driver-level)									
Team L.	451.93	399.53	6,300	249.72	374.89	1,715	114.50	192.78	1,295
Individual L.	465.04	426.74	6,300	293.28	478.17	1,680	121.68	228.02	1,260
Control	463.66	396.32	6,300	289.13	501.09	1,680	109.06	200.75	1,260
Total	460.21	407.78	18,900	277.19	454.54	5,075	115.08	207.61	3,815
DiDi age (driver-level)									
Team L.	2.05	1.07	6,300	1.25	0.91	1,715	2.00	0.91	1,295
Individual L.	2.08	1.07	6,300	1.22	0.90	1,680	1.98	0.95	1,260
Control	2.06	1.06	6,300	1.21	0.90	1,680	2.04	0.90	1,260
Total	2.07	1.07	18,900	1.23	0.90	5,075	2.01	0.92	3,815
# of drivers	18,900			5,075			3,815		

7 The Effect of Being Treated (Has a Leaderboard) on Driver Revenue Change

To examine the general effect of having a leaderboard, we coded the binary variable has_a_leaderboard as 0 if the driver is in the control group and as 1 if the driver is in the team-leaderboards or individual-leaderboards condition. We use models represented by Equation S2 and Equation S3 to capture the effect with and without controlling driver individual heterogeneity.

$$\Delta y_i = \beta_0 + \beta_1 \text{has\_a\_leaderboard} + \gamma_c + \epsilon_i \tag{S2}$$

$$\Delta y_i = \beta_0 + \beta_1 \text{has\_a\_leaderboard} + \beta_2 \text{age} + \beta_3 \text{DiDi\_age} + \beta_4 \text{hometown\_distance\_to\_contest\_city} + \beta_5 \text{self\_formed} + \gamma_c + \epsilon_i \quad \text{(S3)}$$

As shown in Table 1, we find that the treatment of having a leaderboard improving drivers revenue by 34.53 RMB ($p < .01$, 1.66% of average weekly revenue) during the long term overall, and by 41.67 RMB ($p < .01$, 1.69% of average weekly revenue) in Beijing, while we don't observe significant effect in Taiyuan and Kunming. These results hold when we additionally control driver individual heterogeneity.

During the surprise short-term contest, according to Table S14, the treatment of having a leaderboard marginally significantly improving drivers revenue by 49.44 RMB ($p < .10$, 2.10% of average weekly revenue) in Beijing, while we don't observe significant effect overall, or in Taiyuan and Kunming. Controlling individual heterogeneiry, overall in three cities having a leaderboard improves drivers revenue by 36.72 RMB (1.83% of average weekly revenue) with marginal significance ($p < .10$) and by 55.00 RMB (2.34% of average weekly revenue) with marginal significance ($p < .05$).

Table S14: Additional analysis for short term: the teams that had leaderboard vs control. Effect on weekly revenue during the surprise short term period – Difference-in-differences analysis with standard errors clustered at different levels. The coefficient represents weekly revenue (GMV) difference.

	Outcome variable: Δ of Weekly Revenue (CNY)							
	Treatment effects				Control individual heterogeneity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Beijing	Taiyuan	Kunming	All	Beijing	Taiyuan	Kunming
Has a leaderboard	30.90	49.44*	-4.25	-11.97	36.72*	55.00**	-1.94	-3.42
	(20.81)	(28.32)	(33.04)	(34.82)	(20.45)	(27.77)	(32.22)	(34.38)
	[0.16]	[0.32]	[1.00]	[1.00]	[0.08]	[0.17]	[1.00]	[1.00]
Age					10.57***	11.31***	4.67***	11.61***
					(1.07)	(1.50)	(1.70)	(1.68)
DiDi age					84.07***	97.86***	38.37**	38.65**
					(9.62)	(12.33)	(15.43)	(17.18)
Hometown distance					-0.03	-0.04	-0.16**	0.02
to contest city					(0.02)	(0.03)	(0.06)	(0.03)
Self formed					-20.27	-38.85	22.50	28.50
					(21.57)	(28.71)	(39.11)	(37.24)
City fixed effect	yes	-	-	-	yes	-	-	-
# of clusters	3,970	2,700	545	725	3,970	2,700	545	725
# of drivers	27,790	18,900	3,815	5,075	27,790	18,900	3,815	5,075

Standard errors in parentheses are clustered at the team level. False Discovery Rate q-values are calculated separately for All cities (1) & (5) and for individual cities (2-4) & (6-8) and are reported in square bracket.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table S15: Logistic regression results of driver tendency to complete the survey

	Outcome: Dummy variable of survey completion			
	(1)	(2)	(3)	(4)
	All	Beijing	Taiyuan	Kunming
Is captain	0.0801***	0.0931***	0.0560***	0.0594***
	(0.0050)	(0.0064)	(0.0150)	(0.0130)
Team won in second short contest	0.1207***	0.1108***	0.1301***	0.1326***
	(0.0042)	(0.0055)	(0.0122)	(0.0105)
Pre-contest Avg. daily gmv	0.0067***	0.0004	0.0224***	0.0210***
in 100 Yuan	(0.0011)	(0.0012)	(0.0045)	(0.0035)
gender	0.0588***	0.0305*	0.1086***	0.0962***
	(0.0121)	(0.0164)	(0.0381)	(0.0241)
Hometown distance to contest city	-0.0026***	-0.0021***	-0.0014	-0.0018
in 100 km	(0.0005)	(0.0007)	(0.0028)	(0.0012)
Age	0.0027***	0.0027***	0.0026***	0.0023***
year	(0.0002)	(0.0003)	(0.0007)	(0.0006)
DiDi age	0.0055***	0.0062**	-0.0064	0.0024
year	(0.0020)	(0.0025)	(0.0064)	(0.0056)
# of drivers	34,335	18,900	3,815	5,075

Average marginal effect with *delta-method* SE in parentheses.

The results hold if we alternatively control the number of wins of the two short-term contests.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

8 Survey and Results

After the experiment, we sent the survey to every teamed-up driver in the contest and 4,295 drivers completed our survey in Beijing, Taiyuan and Kunming, which covered about 15.46% out of 27,790 teamed drivers.

To understand driver's tendency to complete the survey, we additionally conducted the analysis with the following logit model:

1. To what extent do you like recent team contest from 2018.10.29-2018.12.3? Please rate on a scale between 0 (I don't like it at all) and 6 (I like very much). Depending on your answer, choose either question #3 or #4.

   (a) I don't like it at all. (201 out of 4,295, 4.68%)

   (b) I don't like it a moderate amount . (53 out of 4,295, 1.23%)

   (c) I don't like it a little. (75 out of 4,295, 1.75%)

   (d) Neither like nor dislike. (196 out of 4,295, 4.56%)

   (e) I like it a little. (152 out of 4,295, 3.53%)

   (f) I like it a moderate amount. (245 out of 4,295, 5.70%)

   (g) I like it very much. (3,373 out of 4,295, 7.85%)

   [Branch: for who choose like]

2. Why do you like this team contest? (Please check all that apply.)

   (a) Because I like the sense of team belonging. (2,601 out of 3,966, 65.58%)

   (b) Because I like the fun and excitement of the contest. (2,025 out of 3,966, 51.06%)

   (c) Because I got to know more friends during the contest. (2,025 out of 3,966, 51.06%)

   (d) Because winning the contest gave me a sense of honor. (2,417 out of 3,966, 60.94%)

   (e) Because I won the monetary bonus. (2,196 out of 3,966, 55.37%)

   (f) Other reasons. Please specify ___.

   [Branch: for who choose dislike]

3. Why do you dislike this team contest? (Please check all that apply.)

   (a) Because my team members were not collaborative or united enough. (118 out of 330, 35.76%)

   (b) Because my team was not active enough to justify its existence. (121 out of 330, 36.67%)

   (c) Because the captain did not have good leadership or management skills. (83 out of 330, 25.15%)

   (d) Because the contest rules were too complicated to understand. (77 out of 330, 23.33%)

   (e) Because the contest rules were unfair. (106 out of 330, 32.12%)

   (f) Because the financial bonus was not large enough to attract me. (172 out of 330, 52.12%)

   (g) Other reasons. Please specify ___.

4. As a team member, what did you get from this team contest? (Please check all that apply.)

   (a) I got to know more friends. (2,749 out of 4,295, 64.00%)

   (b) I improved my leadership skills. (1,443 out of 4,295, 33.60%)

   (c) I improved my communication skills. (2,067 out of 4,295, 48.13%)

   (d) I improved my collaboration skills with other drivers. (2,541 out of 4,295, 59.16%)

   (e) I became more experienced and skillful about taking DiDi orders. (2,452 out of 4,295, 57.09%)

   (f) I received emotional support from my teammates when I was down. (1,516 out of 4,295, 35.30%)

(g) Other reasons. Please specify ___.

5. During this event, which option best describes how your team members got along with each other?

   (a) Our team shared commonalities and common interests. (586 out of 4,295, 13.64%)

   (b) Although team members each had our own personalities, we got along well. (683 out of 4,295, 15.90%)

   (c) Everyone contributed for our team honor during the contest. (2,377 out of 4,295, 55.34%)

   (d) Inactive team members influenced others' enthusiasm for the contest. (649 out of 4,295, 15.11%)

   (e) Other reasons. Please specify ___. (0)

6. To what extent do you agree that you have developed deep friendship with your teammates? (from 0 being strongly disagree to 6 being strongly agree)

   (a) 0 - Strongly disagree. (288 out of 4,295, 6.71%)

   (b) 1 - Disagree. (49 out of 4,295, 1.14%)

   (c) 2 - Somewhat disagree. (100 out of 4,295, 2.33%)

   (d) 3 - Neither agree nor disagree. (268 out of 4,295, 6.24%)

   (e) 4 - Somewhat agree. (203 out of 4,295, 4.73%)

   (f) 5 - Agree. (264 out of 4,295, 6.15%)

   (g) 6 - Strongly agree. (3,123 out of 4,295, 72.71%)

7. (A reverse coding question) To what extent do you not believe that you are a part of your team? (from 0 being not agree at all to 6 being agree very much)

(a) 0 - Strongly disagree. (1,481 out of 4,295, 34.48%)

(b) 1 - Disagree. (312 out of 4,295, 7.26%)

(c) 2 - Somewhat disagree. (236 out of 4,295, 5.49%)

(d) 3 - Neither agree nor disagree. (255 out of 4,295, 5.94%)

(e) 4 - Somewhat agree. (177 out of 4,295, 4.12%)

(f) 5- Agree. (94 out of 4,295, 2.19%)

(g) 6 - Strongly agree. (1,740 out of 4,295, 40.51%)

8. Which option do you prefer if you participate in a team contest again?

(a) I prefer to be a team captain. (2,648 out of 4,295, 61.65%)

(b) I prefer to be a team member. (1,647 out of 4,295, 38.35%)

[Branch: if choose team member]

9. Why did you choose NOT to be a team captain? (Please check all boxes that apply.)

(a) I don't want to initiate communications with strangers. (146 out of 1,647, 8.86%)

(b) I don't know how to lead a team. (519 out of 1,647, 31.51%)

(c) The extra bonus for a captain was not enough. (196 out of 1,647, 11.90%)

(d) I was concerned that being a captain would entail a lot of extra work. (257 out of 1,647, 15.60%)

(e) I was inexperienced with team leadership and needed more practice in the first place. (1,053 out of 1,647, 63.93%)

(f) Other reasons. Please specify ___.

[Branch: if choose team captain]

10. What do you think a team captain should do? (Please check all boxes that apply.)

    (a) A captain should be a good example for other teammates. (2,351 out of 2,648, 88.78%)

    (b) A captain should be positive and energetic. (2,093 out of 2,648, 79.04%)

    (c) A captain should help his teammates to become more active. (2,108 out of 2,648, 79.61%)

    (d) A captain should help his team win the contest. (1,940 out of 2,648, 73.26%)

    (e) A captain should provide feedback and suggestions to the Didi platform on behalf of team members. (1,621 out of 2,648, 61.22%)

    (f) Other. Please specify ___.

11. Through which approach do you prefer to build your team?

    (a) I prefer to wait for others' phone call to invite me to join a team. (480 out of 4,295, 11.18%)

    (b) I prefer to call other people and ask if I can join their team. (2,983 out of 4,295, 69.45%)

    (c) I prefer to join a team without prior communication and then contact teammates online. (832 out of 4,295, 19.37%)

    (d) Other. Please specify ___.

12. What do you hope would happen to your team?

(a) I hope it was a temporary team and I might be able to join a different team next time. (3,457 out of 4,295, 80.49%)

(b) I hope it is a long-lasting team and team members will keep in touch after the contest. (838 out of 4,295, 19.51%)

13. How do you communicate with your teammates during the contests?

(a) WeChat (3,372 out of 4,295, 78.51%)

(b) phone calls (2,300 out of 4,295, 53.55%)

(c) text messages (1,363 out of 4,295, 31.73%)

(d) face to face (966 out of 4,295, 22.49%)

14. How often do you communicate with your teammates during the first-week contest? During the three weeks in between the contests and during the last contest?

(a) Never (First short term: 712 out of 4295, 16.58%; Longer-term: 717 out of 4,295, 16.69%; Post-intervention contest: 755 out of 4,295, 17.58%)

(b) Once a week (First short term: 725 out of 4295, 16.88%; Longer-term: 796 out of 4,295, 18.53%; Post-intervention contest: 757 out of 4,295, 17.63%)

(c) Multiple times a week, but not every day (First short term: 1,142 out of 4295, 26.59%; Longer-term: 1,153 out of 4,295, 26.85%; Post-intervention contest: 1,097 out of 4,295, 25.54%)

(d) At least once per day (First short term: 1,716 out of 4295, 39.95%; Longer-term: 1,629 out of 4,295, 37.93%; Post-intervention contest: 1,686 out of 4,295, 39.25%)

15. (Ranking groups only) During the long-term ranking period from 2018.11.5 to 2018.11.25, do you hope to see your team are ranking top? (from 0 being not at all to 6 being very

much so)

(a) 0 - Not hope so at all (47 out of 2,824, 1.66%)

(b) 1 - Not hope so (10 out of 2,824, 0.35%)

(c) 2 - Somewhat not hope so (28 out of 2,824, 0.99%)

(d) 3 - Neither hope nor not hope (73 out of 2,824, 2.58%)

(e) 4 - Somewhat hope so (50 out of 2,824, 1.77%)

(f) 5 - Hope so (73 out of 2,824, 2.58%)

(g) 6 - Hope so very much (2,543 out of 2,824, 90.05%)

16. During the long-term ranking period from 2018.11.5 to 2018.11.25, which statement(s) about the ranking leaderboard will you agree with? Please check all that apply.

    (a) Although there was no team bonus, keeping the team relationship makes me feel not lonely anymore. (1,813 out of 2,824, 64.20%)

    (b) Although there was no team bonus, I was curious about my ranking within my team members. (1,459 out of 2,824, 51.66%)

    (c) (Condition of both inter-team and intra-team rankings only.) Although there was no team bonus, I was curious about my team ranking among our competitor teams. (694 out of 1,390, 49.93%)

    (d) The ranking was meaningless since there was no monetary bonus, so I didn't care about the ranking and team. (561 out of 2,824, 19.87%)

17. On a scale of 0 to 6, 0 being not at all, and 6 being very much so, how would you evaluate your sense of belonging to your team?

(a) Very not strong (207 out of 4,295, 4.82%)

(b) Not strong (70 out of 4,295, 1.63%)

(c) Somewhat not strong (92 out of 4,295, 2.14%)

(d) Moderate (257 out of 4,295, 5.98%)

(e) Somewhat strong (205 out of 4,295, 4.77%)

(f) Strong (296 out of 4,295, 6.89%)

(g) Very strong (3,168 out of 4,295, 73.76%)

18. On a scale of 1 to 7, 1 being not at all, and 7 being very much so, how would you evaluate your sense of belonging to DiDi?

(a) Very not strong (237 out of 4,295, 5.52%)

(b) Not strong (74 out of 4,295, 1.72%)

(c) Somewhat not strong (91 out of 4,295, 2.12%)

(d) Moderate (237 out of 4,295, 5.52%)

(e) Somewhat strong (187 out of 4,295, 4.35%)

(f) Strong (256 out of 4,295, 5.96%)

(g) Very strong (3,213 out of 4,295, 74.81%)

19. To what level do you believe that your DiDi income is the primary source of income for your household?

(a) Yes, it's the only source of income for our household. (2,076 out of 4,295, 48.34%)

(b) It's the primary source of income, but not the only one. (1,110 out of 4,295, 25.84%)

(c) It's a good amount of income, but not the primary income of the household. (660 out of 4,295, 15.37%)

(d) It's just an additional source of income. We don't depend on DiDi's income to live a life at all. (449 out of 4,295, 10.45%)

20. Why do you want to be a DiDi driver?

(a) I would like to be a full-time DiDi driver for a long time. (3,188 out of 4,295, 74.23%)

(b) I am and will be a full-time DiDi driver until I find the next job. (406 out of 4,295, 9.45%)

(c) I have another job. I regard DiDi revenue as my extra pocket money in addition to my job. (375 out of 4,295, 8.73%)

(d) I want to kill time by driving. It doesn't matter too much for me whether I make money from it. (77 out of 4,295, 1.79%)

(e) Simply driving is my habit. I like driving. (249 out of 4,295, 5.80%)

21. What suggestions do you have for future team activities?

22. Please fill out the phone number which you use to log into the DiDi driver APP: ___.