General comments on the paper, its limits, and general issues

Methodological issues of IO experiments; what can we learn about the real world?
The paper is framed as largely having methodological implications. The focal question is whether play in laboratory "industrial organization" experiments can be seen as descriptive or predictive for real world behavior. This is, of course a very important issue. There are many differences between lab and field that may prove significant. The issue of other-regarding preferences is an important one, and it is not clear whether this is exclusively important in the lab, important in the lab and in the business world, nor where it is more important. An important insight is that where there is the opportunity for in-group other-regarding behavior, there seems to be less other-regarding (postive or negative) behavior towards the player in the other group.

Team identification
The impact of "minimal group identification" has been explored in many papers in a variety of settings. If it is true that firms do maximize (as theory predicts, but not much empirical evidence is cited in the paper) in the way that individual lab participants do not, one candidate explanation is team identification. Firms *do* try to instill a sense of identity, and an innate affinity and greater sense of caring for the people in one's firm may in fact substitute for empathetic (or spiteful) behavior towards people in other firms. With a strong sense of within-firm identity, the firm may in fact be more shrewd towards other firms.

However, there are many other explanations and stories why firms may behave differently than individual lab participants, including many theories of why firms should optimize more and be shrewder with their competitors. The teams and groups aspect may contribute to this in ways other than the social preferences story. There may be explicit quid pro quo or repeated game cooperative equilibria within a team, in which team members reward those who are most successful in dealing with outsiders and punish those who are too sympathetic to outsiders, and this is not picked up explicitly in the experiment.

Carrying field norms into the lab: However if lab participants are familiar with this logic, they may be carrying over their familiar behavior from playing on teams into the lab setting. Thus it need not be driven by other-regarding behavior towards members of the team. It could be through this heuristic or porting the field norms into the lab.

Key findings: In either case, what the lab experiment offers evidence for, is that there is some aspect of playing in a team that does not involve actual team deliberation that leads to shrewder or better calculated behavior towards outsiders. (Those sorts of deliberation may lead to the "most profitable argument wins" in a firm team setting, but this is not driving the lab results of course, as the second player is clearly passive.)

It is interesting that (if I understand it correctly) in teams, both the leaders' and followers' behavior is closer to profit-maximizing behavior. This suggests that in the latter case, decision-makers substitute some of their other regarding preferences towards the follower player in favor of their other regarding preferences towards their team partner. On the other hand, followers also behave closer to the prediction, presumably producing lower quantities, when they are a member of a team. Here they are substituting their spite, perhaps, towards the leader, in favor of their altruism towards their fellow teammate. So altruism to the teammate seems to dampen both directions of social preferences towards an outsider! Perhaps this idea should be crafted into the narrative.

Limits
This paper brings some evidence that being part of a team that shares in the profits leads to shrewder behavior towards outsiders, in a particular context that may be like a relevant real-world situation, and this may be driven by substitution of other-regarding preferences, or by some other motive, such as a reflex anticipation of punishment by a team member. However, it does not go very far towards explaining whether this is an important factor in the decision-making process in large and small real-world firms. It lends some credibility to this idea, and if this were *not* found in such an experiment it might be

interpretable as evidence against this idea (although it would be very weak evidence, as the lab partners are not nearly as close nor interdependent as the members of the firm).

Methodological implications; a *negative* result?
In terms of methodological implications, it might suggest that it would be preferable at least marginally, when trying to examine the plausibility of a theory about firm behavior in the lab, to make players part of teams whose payoffs are the same. However, having a "dumb" inactive player, would add considerable expense, and it is not clear that the benefit is worth the cost. Even if this is perhaps a step towards realism, there are innumerable ways in which the experiment is still very far from having the relevant features of the real world situation, and many of these differences seem likely to have a significant impact on the results. If anything, this might be seen as a *negative* result for industrial organization experiments: if we make one small change in the direction of "realism" the results change substantially; therefore, we do not expect our lab results to be robust or tell us something that is invariant across environments.

It's also interesting that the Loaded frame seems to have similar effects as the Team setup. Ultimately, the bulk of the results seem to suggest that the Loaded treatment as a similar effect to this Team set up, so perhaps for future work it is not worth putting in the expense of including the team set up. On the other hand, it might be worth using *both* (perhaps this is worth testing).

Important limits to the lab/ differences from the field
Other important differences, to name a few, include:

- In the real world firms have a long time to deliberate
- The stakes are much higher in the real world, and the larger material gains may increase more than the potential other-regarding gains
- Competition decisions in the real world generally involve a great deal of uncertainty and ambiguity, and are much more complex problems.
- Perhaps most importantly, managers in firms are selected for their success in making such decisions, and the firms themselves can be seen to have survived an evolutionary process of selection.
- There is a great deal of communication between members of a team, and sometimes, perhaps between members of different teams (firms) – Note that in many collective action experiments, this communication proves to be impactful.

Because of all of these important differences, experimental results may be argued to be much more useful when they reveal something that is *consistent* across variations, rather than when they reveal the effects of variations. (See argument in previous paragraph).

Is lab work useful for IO (my opinion)
I also think that experiments are useful for testing or giving evidence for the basic plausibility of certain theoretical, perhaps intuitive arguments. They can be useful in finding evidence for individual-level reactions, sometimes in group settings, and some strategic interactions. Making an analogy to real-world environments, such as the competition between quantity-setting firms, is thought provoking, and somewhat useful, but must be done very carefully and cautiously. In order to establish a link that would be worthy in making predictions, elements from the field need to be brought in, and lab results need to be compared and tied to field results both from experiments and observational data. This is the challenge that the practitioners of industrial organization experiments should focus on. This will of course be very hard to do. Experiments involving participants with field experience, such as managers, may be helpful. This might be can combined with "softer" data, such as interviews with industry professionals. Evolutionary type selection process, survival of the fittest simulations and competitions involving real money may also be helpful.

Specific comments

Need for more theoretical specifics
In the example given (section 3), it is not clear what fairness preferences are behind the results in the

second paragraph – these should be specified more precisely. The variety of possible results described in table 2 are interesting but need to be better explained. For example, what does it mean to be "fair and interested in profit?" Also note that the cases described presumably depend on common knowledge of one another's preferences; something which may not be expected in the lab or field.

Replicability issues
Treatment "Loaded" was designed to be identical to Huck et al's Stack Rand treatment. However, some of the results are found to be different. This should be discussed further. Does it imply that the results in such experiments are not robust and are very sensitive to small details of the environment? Can you identify or speculate about any details of the environment which might be driving this, such as difference in the composition of the subject pool? This lack of replication in experimental results is troublesome; it would probably be better if experimental economists were required to the show the replicability of their own results before being able to publish them. This tends to be the case in the physical and natural sciences.

Behavior as in exact predictions
The table looking at the presence and frequency of behavior *exactly* as predicted according to one of the three series is somewhat interesting. But what are we to make of this? Are we to assume that these individuals exactly calculated their best responses according to this model? Was such behavior falling exactly according to a prediction more likely among economist or others who might be expected to do so? On the other hand, would it be more reasonable to consider "near Stackelberg" or "near Cournot" behavior?

Table 5 presentation; direction of differences? I like the presentation in table 5, but shouldn't we look not at the absolute value difference but say something about the direction of the difference? There are several explanations for why followers might produce more than predicted, such as inequality aversion, fairness considerations, spite, but it is unclear what might get a follower to produce *less* than predicted. This would likely be due to a calculation error, one presumes, although perhaps other explanations are possible.

Later versus earlier stages, and learning
It would be wise to specifically examine whether behavior differs in later stages, presumably after learning.

(Imperfect) Stranger matching
The authors should consider robustness checks and controls to deal with the lack of perfect stranger matching and the potential effects of these on responses.

Motivate predictions: For hypothesis 2b, what motivates the prediction for followers? It seems reasonable, and I suspect there are several theories and evidence that support it; however, the cited Hoffman et al paper finds the exact same or approximately the same rejection rate for followers in each framing.

Evidence “rejecting” a broad theory
This statement seems too strong: "the neoclassical theory of the firm as a theory of all firms is expected to be rejected again." Does this possible “strawman” of a theory mean to be making a prediction for individuals in very small-scale lab experiments? One must be careful about making sweeping statements about rejecting a very broad theory with narrow evidence. There is an excellent discussion of how theory should or should not be used and how experiments relate to this, in the book "Rethinking the Rules.” Note, that in fact the author here does not find a significant difference in followers' behavior between the frames. However, the tests given here do not seem to be directly testing closeness to best responses, but just looking at followers’ actual contribution.

Language of hypothesis testing
I note here that the language of "rejecting or confirming" hypotheses is a bit confusing, and perhaps misleading. In some cases, we see perhaps an effect that is in the direction predicted by the hypothesis, but this effect is not statistically significant given the size of the sample, etc. Should that necessarily cause us to "reject" this hypothesis? I more comfortable usually with the language "fails to reject" or" "rejects the null hypothesis in favor of a significant difference between treatment and control"; this either provides evidence, or fails to provide the evidence consistent with the underlying *theoretical* hypothesis.

Eliciting expectations?
It is a good point that the difference in leader behavior might be caused by anticipated differences in follower behavior. This could potentially follow up on if you tried to elicit expectations, of course making the experiment more complicated.

How well does the lab represent the Demsetz model?
I am skeptical that "requirements A to F and H are met" (of the Demsetz model of the firm) in the lab treatment "team". Is there really, for example, "joint output production" in any meaningful way? In a firm, don't all of the owners have some power to exert over the firm's decisions, which they may, for example, delegate to a manager? And in this experimental environment in what sense does the owner "have the right to sell his central contractual residual status"?

Why do teams maximize "more"?
Also the statement "teams are expected to maximize their profits because decision-makers are predicted to maximize their outcomes" does not really predict why these decisions will be closer to the self-interested prediction then the other treatments; wouldn't the same prediction about decision makers hold there?

Statistical dependence issues Note that using each player's "mean quantity choice" is perhaps a more conservative way of dealing with repeated measurement then using each round is an observation, however it does not fully deal with possible feedback from a lack of a perfect strangers matching. For this one might have to observations at the session level, and even then there could be issues that a subject's behavior could be aimed to influence the person one might meet in future rounds of the experiment.

Sizes of effects: It should also be noted that although playing in a team does seem to have a significant effect, it only moves us about one third of the way towards the predicted Stackleberg equilibrium.

(Table 8) It is worth noting that even in the Loaded and Team cases, these response functions are significantly weaker than predicted by Stackelberg. The difference between Loaded and Neutral or between Team and Neutral only takes us again, about one third of the way towards the predicted response function.

Interact the treatments: There was no combined loaded and team treatment; this might have been worth trying, or in future experiments.

Response functions, overstating results
They do note that the interaction of the leaders and followers responses, with the leaders anticipating the followers response, leads to slightly more profit maximizing choices in the team framing and in the loaded framing, but the difference is not large, and I think they are overselling this point. Surely it is coincidental that the leaders' quantity in Team is exactly equal to the optimal one given the response function estimated in table 8. And as noted, the followers' responses are not significantly better in Team than Loaded.

Note that this is also a very rough definition of profit maximization, assuming that one's own partner in a particular's stage have the exact response function of the average person in that treatment. There are also clearly issues of coordination and learning. For this, again, it may make more sense to look at later stages.

Speculative explanations: The discussion of the small discrepancy between the leader's behavior and their so-called optimal behavior is interesting but fairly speculative. And one should consider motivations in general beyond inequality aversion, such as minimax strategies and ambiguity aversion and even risk aversion. Nevertheless, coordination and specific differences in expectations are more likely to be the reason for this, if a single reason can be stated.

The followers' reaction does divert from the prediction in the direction of minimizing inequality. Of course, other explanations for this behavior are possible, including reciprocity motives.

Overstating results using weak evidence to reject a general theory about real world behavior
Again, I think the final sentence in section 3 is overstating the results: "an individual under an IO framing is

less of a profit maximize under a two member team organized…" This results is *not* strongly supported by the statistical analysis, which finds *mixed* results.

As noted before, I think the conclusions need to be stated more cautiously, particularly as pertains a general theory of the profit maximizing firm. This also holds in the conclusion.

Test one variation at a time: But, there is a difference relative to the individual neutral framing; perhaps this is a better approach. When comparing a change into things at the same time it is hard to know what is causing the results.

But it is not an either/or choice: What is more, this does not need to be an either/or choice in an experimental context: one could use both an IO framing and a two member team. So I don't know why it is necessary to "horserace" these anyways.

Repetitive conclusion: The conclusion is largely repetitive of material in the abstract introduction and elsewhere. This does not need to be restated, in my opinion.

Somewhat misleading statement of results, overstatement
The statement "followers' motivations are not affected by such a framing: the followers in Loaded behave the same way as followers in Neutral" is somewhat misleading. This does not appear to be the case for the response functions estimated in table 8. In this table the slopes for loaded and team are nearly the same, and both are somewhat more steeply sloped than Neutral. Perhaps I am misinterpreting this, but if this is the case it needs to be explained better. The chosen statistical results might be alleged to be shoehorned into the conclusion that is desired.

Contract of employment?
In the final paragraph there is a mention for the first time of the idea that "the contract of employment makes people more selfish." This theory was not explored nor explain perform. Furthermore, it is hard to see in what sense the "contract of employment" is being meaningfully implemented in this laboratory experiment in a way that would reflect similar motivations in the field.

Minor points and presentation issues

Fewer tables: In illustrating the results, there are probably too many tables and types of analyses. Since both types of analyses reveal approximately the same results, it would be easier to read if they only reported one of the to, and mentioned the consistency of the other set of results, putting the latter in an appendix.

Clarification: It should be clarified what is meant by "followers choose a higher quantity than predicted"; is this higher conditional on the leader's choice? (clarified only later)

Minor: "Followers' responses appear to be farther away from the Stackelberg prediction". This appearance is been corrected and reversed in the analysis in the next paragraph or two. Why not just give the right answer first, otherwise it is confusing.

Minor: The paragraph beginning "what is surprising" is a bit confusing. Why should this be surprising? Do we expect play always exactly in accordance with in equilibrium for selfish players?

The final paragraph of section 3.2 is a bit thrown away. You test for difference, but then don't say anything about it; what is the meaning here? This there must be some point to doing this replication exercise.

Minor: The framing differs, but I wouldn't necessarily call it an "organizational" framing. The word "organizational" can mean a lot of things, not necessarily the same thing as saying "firms". Perhaps just call this a "firm" framing. Saying "IO" framing is also acceptable.

Minor: This is a bit picky, but hypothesis 2a perhaps does not need to be considered as a separate hypothesis.

Clarification needed: Table 6 could be better described in the table itself. It is obvious perhaps from context what Q_a and Q_b represent, but I am not sure what Q^a_b represents.

It is also not clear why the author chose to present the data as in table 6: he doesn't seem to do anything with the round specific behavior or the trend over time, etc., and the tests seem to be based on average across all of the rounds. It would be easier to read if the data/statistics presented were consistent with the tests reported in the text.

Clarification: The last paragraph on page 13 is a bit messy and confusing. I cannot follow the discussion of things like "even nine rounds."

Clarification: For table A, please remind us what the predicted response function is according to the basic Stackleberg model (this is given later).

Clarification: Table 7 needs to be labeled better. I shouldn't have to refer to the text understand what it means.

Clarification: Table 8 is again helpful, estimating the response functions, but could be better labeled as well