REPLY TO EDITOR AND REFEREES

Editor decision letter

22nd August 2019

Dear Prof. Camerer,

Thank you once again for your manuscript, entitled "General Economic Principles of Bargaining and Trade: Evidence from 2,000 Classroom Experiments", and for your patience during the peer review process. Please accept my sincere apologies for the delay in getting back to you with a decision. I know how important timeliness is, and I am very sorry I failed to provide you with a decision and reviewer feedback sooner.

Your Article has now been evaluated by 4 referees. You will see from their comments copied below that, although they find your work of potential interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to (1) identifying key priorities that should be addressed in revision and (2) overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritised set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

As you will see from their comments, the reviewers are divided in their opinions on the scale of the manuscript's contribution to the literature. Although Reviewers 1 and 2 believe that the dataset alone constitutes a significant advance, we remain concerned by the points raised by Reviewers 3 and 4 about the study's conceptual contribution. In order to consider your manuscript further, we therefore ask that you carry out some of the additional analyses suggested by the reviewers to more fully leverage the power of the dataset and to provide more conceptual insight. We also ask that you more fully situate the current findings in the existing literature.

Additionally, the reviewers call for more work on the existing assessment of zero intelligence models, including a clearer discussion of predictions and a more thorough assessment of the role of autocorrelation.

Reply: We have added material here (denoted by ** below in our reply to referees).

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. I have also attached a template manuscript file that exemplifies our policies and formatting requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript we would hope to receive it within 6 months. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Human Behaviour or published elsewhere. Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date would be that of the revised, not the original, version.

With your revision, please:

- Include a “Response to reviewers” document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the reviewers along with the revised manuscript.

- Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

https://mts-nathumbehav.nature.com/cgi-bin/main.plex?el=A6Co7Czd1A1Djz3J5A9ftdmb46RuIY8W5IvSLxM2AXeQZ

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,
Aisha

Aisha Bradshaw
Editor
Nature Human Behaviour

Reviewer #1 (experimental economics, price formation, trade)

Remarks to the Author:

This paper addresses some old and important research questions regarding market and bargaining outcomes. The dataset is unique and extremely rich, allowing the authors to shed new light on these questions. Market outcomes are remarkably robust across the many countries and regions included, and they are consistent with limited rationality. (This latter result is not terribly surprising given the small number of trading periods conducted in most sessions, usually just 1 or 2, and apparently all classes used inexperienced subjects.) The bargaining data reveal greater geographic variation, and also uncover an interesting discontinuity in acceptance behavior for exactly equal offers.

The biggest limitation of the data is that they were collected in classroom settings, with lower control than usual laboratory experiments. Moreover, the economic incentives vary (and are unobserved) across observations. I believe it may be worth giving up some control in exchange for the other advantages of this unique dataset, such as its large size and geographic diversity.

Suggestions

1. One aspect of the market data that could potentially be exploited, but is not apparently considered in this draft, is the variation in the number of traders across markets—and the associated variation in the number of units traded in equilibrium. The most common market has 10 traders (5 buyers and 5 sellers), but Table B.4 indicates that the number of traders was quite dispersed (mean of 14 and standard deviation of 15, with a range of 2 to 318).

The market size could affect market outcomes and performance because, presumably, the price taking assumption underlying competitive equilibrium theory is less appropriate as the market size decreases. The simplifications used in the studied models of price formation (Wilson, and especially Friedman) may also apply more or less well in small markets—especially those markets with 6 or fewer traders (3 buyers and 3 sellers). Most previous double auction experiments do not vary market size at all, and have rarely considered very large markets. It would be interesting to investigate whether performance is different between the very small and the very large markets.

Reply: This is an excellent suggestion that we very much appreciate. Our first cut was to use the most popular design to gain power, but upon reflection it is obvious that the smaller-N designs may be where the mechanics of the Wilson and Friedman theories shine. We analyze the market with less than or equal to 6 traders and the markets with greater than or equal to 36 traders. We find that small market performance is closer to the predictions of Wilson and Friedman, but large market performance is more consistent with the ZI predictions. We have added a

summary of this analysis to the main text on pages 9 and 10 and the details can be found in SOM B6.

2. The last presented result documents a negative correlation across regions between the average ultimatum proposal offers and conditional acceptance rates. I found this confusing. In Section C.2 I see the model of conditional acceptance rates, but that analysis concludes that these rates show insignificant between-region heterogeneity. Is this the model that this result in the text is referring to? Or something else? The statement in the text is too terse to understand or evaluate.

Reply: We apologize for the lack of clarity. We now clarify in the text on page 19 that this rank-order correlation coefficient supports the shared-norm hypothesis.

3. I do not think the motivation concerning round numbers for market prices is very strong. Moreover, the authors do not really follow up on this in the analysis of the double auction data—although round numbers are prominent in the ultimatum bargaining data.

Reply: We understand your skepticism about the weak motivation for this small speculative point. This, though, was included because the NHB readership ranges across all social sciences. Furthermore, the idea that "efficient coding" can generate round number effects has gotten traction in economics (Dehaene and Mehler, 1992; Frydman and Jin, 2018). This was meant to potentially inspire application of those theories by others.

4. Also in the introduction, the authors point to the value of a "large-scale picture of reproducible ultimatum game behavior … to provide a benchmark that is statistically sound…" In my view, however, this benchmark may not be too useful since it is based on unincentivized decisions.

Reply: To address the concern about incentive effects, we searched exhaustively for syllabi of all the classes in our Ultimatum Game data set and we found them for 58 sessions (out of 490 sessions). We categorize the incentive schemes to four types: no incentive, participation, course points and real money. We find that the patterns of behavior in the Ultimatum Game are similar across these incentive schemes. We discuss the analysis in the main text on page 16 and in more detail in SOM A6.

5. Related to the previous point—it might be overstating things to claim that "Attempts to conduct the ultimatum game without monetary incentives, also find roughly similar results as when financial incentives are used." While it is true that ultimatum game results are less sensitive to the presence of financial incentives than results from dictator games, they are certainly not invariant. For example, Forsythe et al. (1994; the authors' cite #56) show that ultimatum games without financial incentives are not reproducible across time and they differ (modestly) from games with financial incentives. Better

evidence exists that the size of the stakes does not matter much, as long as some reasonable stakes are used.

Reply: We clarify what we meant by the phrase “roughly similar results” on page 4. We meant that average offers are much greater than zero but many offers are less than half, and conditional acceptance rates increase in offer size. The main differences from no incentives are likely to be an increase in “costless rejections” by responders (a downward shift in the acceptance curve), and perhaps more aggressive offers, leading to higher overall unconditional rejection rates. We do see higher overall rejection rates, but there are also substantial rejections of 0 offers. Our previous Reply just above also notes that we were able to collect data on incentives from a subsample of ultimatum games and the results are either similar, or a bit different in magnitude but with too little power to have confidence about differences.

6. If the authors need to reduce the paper’s length, I found the analysis of reaction time less interesting than other presented results. In my opinion this could be moved to the online appendix.

Reply: At least one other referee found the reaction time results to be interesting and revealing. We believe that they contribute significantly enough to the growing reaction time literature to warrant inclusion in the main text.

7. Most market experiments were conducted for 1 or 2 periods. I think it would be useful to indicate the number of separate markets somewhere in Figures B.8 and B.9 since the N seems to fall off rather quickly.

Reply: Thank you for this suggestion. The change has been implemented in those Figures.

Reviewer #2 (markets, price formation, reciprocity)

Remarks to the Author:
The authors use a large new data set (2000 classroom experiments using MobLab software) to revisit and extend two classic analyses. First, the seminal 1991 AER paper by Roth et al found differences across four countries in behavior in Ultimatum game bargaining conducted by hand with paid subjects. The classroom data use an on-line version of the same game, with a more geographically diverse and much larger group of unpaid student subjects. The results include (for the first time I am aware of) reaction time data as well as choice data, and extend the 1991 results in useful ways. Second, a 1993 volume by Friedman and Rust, and a set of related papers mostly of the same era, compare various theories of price formation in double auction markets. Again, the classroom data refine and extend the results of these classic studies in important directions, and offer new evidence on the ability of competing models to explain price formation within and across trading periods in the continuous
double auction.

Thus the key results show some methodological novelty and are significant. The classroom experiments are mostly unpaid, which limits their validity in the eyes of economists, but the issue is addressed properly in the paper. Overall, then, I am satisfied with the validity of the data.

More specific comments follow.

1. One high level message is that market behavior does not vary much across the different subject pools considered, while the bargaining data does differ by subject pool in important dimensions. This result was suggested by Roth et al, and is now folk wisdom, but this study is the best empirical confirmation that I have seen of that dichotomy. The current version of the paper properly highlights this point.

2. The first new result to my knowledge is the clear documentation of a behavioral discontinuity in the Ultimatum Game: the acceptance rate of 0.5 offers is significantly higher than for 0.49 offers. Earlier studies had too few offers just below 0.5 to establish that interesting regularity. Again, the current version of the paper properly highlights this point.
We are pleased to see you share our view on this matter.

3. The reaction time results are interesting and revealing. Responders take longer on offers for which the shares of rejections and acceptances are more equal. This suggests that people take longer to make a decision when it is a close call. The reverse inference, that longer decision times indicate closer calls, may become one of the more important uses of reaction time data.
We are pleased to see you share our view on this matter.

4. To my thinking, the more important part of the paper concerns market behavior, for several reasons. First, it is the broadest confirmation to date of the existing consensus that market outcomes are similar across diverse subject pools.

5. Second, and more importantly, the data provide evidence on price formation within and across periods. I'll use the next several points to cover various aspects of the analysis.

6. The reported efficiencies of the CDA are among the lowest I have ever seen for such vanilla supply and demand. I suspect that this is due to the classroom setting with inexperienced subjects with little or no economic motivation. When I run CDA demos with students (usually on the first day of class), I often find that some shy students stay on the sidelines, and that some aggressive students make unfavorable trades, especially in the first few periods. In a more formal lab setting with salient pecuniary incentives, I see much less such behavior. The paper acknowledges these phenomena to some degree, but (as noted below in point 7) perhaps misses some of their consequences.

7. I went back and took a quick look at one of the articles from the 1990s. The main discrepancy I noticed from the current results on price formation was that the earlier studies found that price change autocorrelation moved (from around -0.5 in early periods with inexperienced subjects) towards zero with more experienced subjects and later periods. The current paper finds this autocorrelation to be around -0.5 and presents this as a major success for the ZI model. However, the ZI model assumes away trades that involve losses for either buyer or seller. Such trades, if occasional but at extreme prices (as suggested by the data in point 6) would have a major effect on estimated autocorrelations. E.g., suppose that most trades are in the vicinity of CE price 100, but (due to perhaps to inattention or boredom) a seller occasionally offers at p=20 or a buyer occasionally offers to buy at 200. Such aberrant traders would take a large paper loss (of no actual personal consequence in these classroom exercises) and would have a disproportionate effect of moving the autocorrelation towards -1.0.

Reply: Thank you for raising this excellent point. We have done a number of robustness checks per your suggestions and collected them in SOM B6. After filtering out all loss trades, the trade-to-trade price change autocorrelation only increases slightly from -0.457 to -0.427 so those trades were not driving the results.

8. My practical suggestion here is to run (and perhaps report in the Online Appendix) the autocorrelations again after filtering out any trades that involved a loss to either buyer or seller. Also, to avoid the "bid-ask bounce" effect which biases autocorrelations towards -

0.5, consider calculating the autocorrelation separately for the sequence of accepted bids and for the sequence of accepted asks. I conjecture that the resulting rhos will be noticeably closer to 0.

Reply: Thanks again for these great suggestions. We computed the trade-to-trade price change autocorrelation separately for accepted bids and accepted asks. The autocorrelation is -0.412 and -0.451 respectively which are again similar to the original pooled result.

9. Were there any other testable differences between the three models of price formation? E.g., the earlier literature looked at sequences of bids and asks between transactions, and those data might be available in the current data set. Also, do the available data support further tests of Easley and Ledyard's model of between-period adjustment?

****Reply: Thank you for the push to look more at these models for testable differences. The earlier literature assumes that traders are only allowed to modify the best bid and ask while the MobLab interface allows the top three bids and asks to be modified. Given this disparity, we have decided not to analyze the sequences. However, we have added the analysis of the source of inefficiency as the "Mutual Agreement" and "Against Nature" models predict that the source of inefficiency is from the least profitable trades not being executed (V-inefficiency) while the Zero Intelligence model predicts the source of inefficiency is the displacement of extra-marginal trades (EM-inefficiency). Our analysis finds evidence of both V-inefficiency and EM-inefficiency with the proportion being the latter decreasing as the market size becomes smaller. The Easley and Ledyard model also provides bounds on possible transaction prices. We find that only 14.5% of the transactions fail to belong to this range. These results are reported in the Buyer-Seller Double Auction section on pages 9 and 10 and the SOM B6.**

Reviewer #3 (experimental economics, ultimatum games)

Remarks to the Author:
The authors take advantage of the very large amount of data available from MobLab to study issues that can only be examined with very large amounts of experimental data. For markets, the authors find results consistent with models of strongly bounded rationality, although there is evidence of more rational behavior as well. Their ultimatum game data shows variation across nations and an odd break in data at the 50/50 split.

The paper is clearly written and the analysis is convincing. The authors are also honest about acknowledging the weaknesses of the data. Not all of the data involved use of incentives and individual level data about demographics isn't available. The large size of the dataset largely eliminates any concerns about power or p-hacking. The authors also do a good job explaining the advantage of having so many observations using exactly the same experimental setup.

Reading through the paper, I felt that too much weight was put on the quality of the dataset and not enough was put on the results. This dataset is only useful if it provide insights that would not be otherwise available. The result on zero intelligence models is interesting, but gets a little bit buried in the paper. I would like the paper to spend more time explaining how the predictions are derived and why these are important.

The ultimatum game data does not yield any one result that is as obviously interesting as the zero intelligence results, but the cross-country results could be important. How do these results compare to what was found in previous studies? Is there any overall pattern that organizes the cross country results? Is there any way of telling how sensitive the results are to the incentives being offered?

Reply: Your comment asks three important questions: Comparison to previous studies? Cross country results? Incentives?

On *"comparison to previous studies"*, the meta-analysis (Oosterbeek et al., 2004) cited in the text does show some country effects, but they are done by different investigators and typically confounded with design differences. No previous studies look at differences across types of classrooms (e.g., high school students versus advanced economics students).

On "any overall pattern that organizes the *cross country results"*? There are too few countries to speculate about this. If we had an interesting idea we would include it, but we simply do not. A clean cut at this question would have to standardize protocols and sample countries with particular theories in mind (e.g. the Cohn et al Science 2019 "lost wallet" study is a good example).

To look at how "*sensitive the results are"* to the incentives being offered, we searched exhaustively for syllabi of all the classes in our Ultimatum Game data set. We were able to find class syllabi for 58 sessions (out of 490 sessions). We categorize the incentive schemes to four types: no incentive, participation, course points and real money. We find that the patterns of behavior in the Ultimatum Game are similar across these incentive schemes. We discuss the analysis in more detail in the main text on page 16 and in SOM A6. The results are all very similar though there are too few with full performance incentives to be able to tell whether there is a jump in acceptance of exactly 50% offers.

In a revision, what I'd most like to see is more focus on the results. I only care about the dataset if it gives me valuable insights that are not available from other sources. I'd like to see the authors make a case that they are learning something important and could not have done so using more conventional data sources.

Reply: We agree and now paraphrase your point in our revised paper. We wrote on page 2: "The main reason to care about a new data set is that it generates valuable insights that are not available from other sources."

The main new insights are more robustness in double auctions than has ever been seen (since there is no meta-analysis of a large body of results, surprisingly) and novel results on trading dynamics. For ultimatum games, there is more cross-country data and a clear interesting effect of exactly-50% offers which has not been seen before.

Furthermore, keep in mind that NHB readers span a wide range of social sciences. Most readers are not economists. Many of these readers will be a bit familiar with ultimatum games, but most will probably not know about double-auction results at all. We suspect that, as experimental economists first were in the 1960s and 1970s, they will be surprised at how well the zero-parameter prediction of competitive equilibrium is able to explain prices and quantities.

Reviewer #4 (double auction and trust games)

Remarks to the Author:

This is a very nice meta-study of the Double Auction and Ultimatum Game. I mostly liked your narrative in the paper and the use of standard methods and models. I do have two major problems with the paper as it is.

The first is that MOBLAB experiments are used for instruction and parameters are chosen most often to confirm theoretical predictions. Figure 2 suggests only one set of DA parameters were used. Are these the default parameters used in MOBLAB? A lot is known about boundary experiments in the DA? are these never run in MOBLAB? A much more detailed description of the experimenter parameters is needed.

Reply: Table B.7 in the SOM now lists the top 10 market DA supply-demand configurations (the much more detailed description of parameters that you asked for). About 2000 of 6000 were the default configuration that is shown in Figure 2. However, most of the analyses used all the data. Unfortunately, the boundary experiments where CE convergence is less common are rarely used, probably for the reason you note— most instructors want to show their students that CE can work.

The text now says " There are many interesting supply-demand configurations, such as the ``swastika" design\supercite{smith1994economics} which previous experiments create highly unequal surplus distributions, and have been shown to not converge to CE as rapidly as the designs we report. Unfortunately, those were not used by instructors."

The second is that MOBLAB has many experimental designs that instructors can use. You have chosen the two designs that have already been replicated in hundreds of labs. Much more interesting would be experiments that have not been replicated as much. This would be much more informative. I would like to see an analysis of an experiment in MOBLAB where known theoretical predictions fail.

Note first, that the ultimatum game is one experiment in which known theoretical predictions (based on selfish subgame perfect equilibrium) fail. Many other regarding preference theories fail too because they do not allow a 'kink' at exactly equal offers.

Re: your second point, which the DA designs most instructors used have indeed been replicated, there is no publication that we could find which actually compares a large sample of data from studies with identical designs, or very similar designs, reporting all the statistics that we do. We think there is an

important distinction between something the profession takes as a stylized fact, and actually assembling a lot of evidence in one place with a wide variety of summary statistics.

With that said, further study of more rare experiments is certainly interesting. We hope that publication of these results would invite other economic scientists to use ata obtained from Moblab as well as other big data educational platforms that utilize experiments.

That said I found the paper interesting and suggest it be published in a more specialized journal such as the Journal of Experimental Economics.