Review of Gate 2: Experimental Design and Analysis Plan (EDAP) for Conservation Auction Behavior: Effects of Default Offers and Score Updating

October 20, 2020

Review by Sarah Jacobson

Thank you for inviting me to review this interesting proposal. I understand that I can remain anonymous, but I like the proposal so I am happy to identify myself.

To make sure I address the points requested in your instructions, I have copied your bullet-point list of items to address below, and I intersperse my comments on each point.

- **Experimental Design**
  - **Does the research propose appropriate methods to test a meaningful hypothesis?**

The hypotheses are meaningful.

The first set of hypotheses concerns whether the starting (default) values for bid down and cover practice affect the participant's eventual bid components. This is a reasonable question because we know that on complex decisions, even if they are highly consequential, people have a strong draw toward choosing the default option if it exists.

The second set of hypotheses concerns whether having bid score information update in real time as participants change their choices (as compared to having to go forward to the next screen to see the information, and back again to change the decision if desired) affects the eventual decision, and in particular whether people with different types of fields will respond differently to this more available information. This should reduce bidding errors. It's a little hard to envision what systematic effect that will have on behavior, but I think the researchers hypothesize that it will increase the correlation between value and bidding behavior.

The third set of hypotheses is not, I believe, per se interesting to the researchers, but can help interpret of all of the results. These hypotheses are about whether students and farmers behave the same way with regard to the preceding hypotheses. This is meaningful because if they are not different, then the data from student participants can be translated (with some caution) directly to predictions about farmer behavior; if they are different, then the sign and magnitude of the differences will be informative about the results we should expect to see of these treatments in a natural field setting. Since the researchers expect to be able to recruit 1000 farmers and run the same suite of treatments at the same scale with farmers as with students, there is less concern about needing to extrapolate the results from these students to farmers; rather, I suppose these students are useful to inform to what extent future studies can use student choices to understand farmer behavior in such a setting.

The research uses appropriate methods overall to test these hypotheses. The experiment structure and incentives for participants seem well designed to mimic the CRP, including using parameters that were derived from recent CRP signups. There were components I struggled to understand, though. For example, I didn't understand how the EBI score was calculated (to be fair, I'm confused about that in real life) so it was hard to see how much help the information updating was offering. Nor did I quite

understand how the EBI cutoff was to be determined in any particular session. I also couldn't figure out what constituted a session. I infer that all interactions will happen online, which is good because the pandemic may continue, but I don't know how many participants there are in a session, nor whether all participants in a session would be part of the same auction or whether they would be broken into auction groups. Similarly, I couldn't tell what participants knew about the number and composition of other participants in their auction. But none of these questions raise concerns for me; again, the design overall seems strong.

- **Evaluate the match between the proposed treatments and the research question.**

The control treatment provides no pre-filled in option in the bid entry boxes, and requires participants to go to the next screen to learn their EBI score, as is currently the case in CRP auctions.

The default treatment is different from the control in offering a pre-filled, "socially preferred" (low cost, high conservation benefit) bid. The difference between this and the control treatment is a good test of the hypothesis that a positive default can nudge participants toward making a bid closer to that socially preferred end of the spectrum, and therefore tests the research question. Conceptually, this treatment comprises both having a pre-filled option and having a nudge toward a high-scoring offer. The design does not separately test the effect of the default *level* versus the *existence* of a default. One could test both by calling the current T1 "T1high" and adding a T1low (pre-filled bid with high cost and low conservation benefit). This may be intellectually interesting, but the researchers may have chosen to omit it because they are not trying to prove out a new mechanism but rather trying to see whether this specific nudge would achieve the desired result in this situation.

The score updating treatment updates the offer score for the bid in real time as the participant changes the bid down and practice. This tests an alternative decision support display that is possible with modern software, allowing participants in actual CRP auctions to see the ramifications of their decisions more seamlessly. This is a good test of a practical solution.

I'm curious about why these two dimensions of treatment, the default and showing the offer scores in real time, are being tested together. There is no hypothesis about the interaction between the two, and I'm not sure I can see a strong theoretical reason for them to interact. Indeed, the researchers say they don't plan to estimate the interaction effect; why do an interaction treatment, then? I don't see a problem with doing it; it just could be better motivated.

- **Evaluate the external validity of the experiment.**

A potential external validity concern with this study is whether any effects observed in the lab will be predictive of responses to similar stimuli in the field. The problem is that CRP bidding is done by agents (farmers) with a strong incentive to make the bids that are best for them, broadly construed, where agents have experience with the auction setting and have access to extension resources and other sources of training and advice about how to approach bidding; thus it's not clear that biases and heuristics demonstrated by a sample of convenience will be demonstrated by these expert, experienced, highly incentivized agents.

The study design addresses this issue by conducting both a lab experiment with a sample of convenience, which is useful in getting as large a sample as possible at low cost, and a lab-in-field experiment with former participants in the CRP General Signup. Of course the incentives are less than those in an actual CRP signup auction, so behavior may not be identical to true field setting behavior, but it should be closer.

Beyond that issue, the experiment is designed to closely replicate features of the CRP General Signup, and therefore the lessons learned in the experiment should be easily translatable to the CRP.

- **Comment on the key features of the design such as the assignment of treatment, the statistical power, participant recruitment, and other features that you view as critical.**

Participants are assigned to treatment at an individual level. This is interesting, as it means that the participants in a given auction will be exposed to different stimuli. But as their underlying incentives are the same, this is not problematic; participants don't need to know that people they are bidding against are viewing different displays.

Recruiting participants at a university is very workable, especially when researchers have an established lab as these do; 1000 is a lot, but they know their subject pool and are confident they can get that number. Recruiting farmers is more difficult, but I have seen work from these scholars and others in which they have recruited farmers to participate in studies of this type, so I am confident they know what they need to do to get 1000 respondents. I appreciate their plan to recruit participants in waves to ensure they reach their target without going over budget given an unknown response rate.

The power analysis they present seems reasonable and relatively conservative to me.

I addressed other elements of the design above.

- Analysis Plan
  - Are the statistical tests for treatment effects appropriate?

The Tobit regressions of EBI score gain and bid down, and the ordered probit regression of cover choice, are appropriate empirical models for these outcomes. While we often use bivariate tests like ranksum for experiment data, in this case, the sample size is large enough that it is probably not necessary.

To analyze the use of the "back" button, depending on whether some subjects forgo this altogether or whether all use it and it's just a question of extent, I suppose they will use a probit or a Poisson or other count model.

Their justification of why they don't plan do to multiple hypothesis testing seems reasonable to me.

- Have the researchers left out any tests or analysis that should be conducted?

I'm curious what control variables the researchers expect to be important. I wonder if they plan to do a numeracy or deliberative thinking task, since the treatments focus on support for a complex decision.