How to run an experimental auction: A review of recent advances*

Maurizio Canavari†1, Andreas C. Drichoutis‡2, Jayson L. Lusk§3, and Rodolfo M. Nayga, Jr.¶4

[1]Alma Mater Studiorum-University of Bologna
[2]Agricultural University of Athens
[3]Purdue University
[4]University of Arkansas

First Draft: October 7, 2018
This Version: October 18, 2018

Abstract: In this paper, we review the recent advances in the literature of experimental auctions and provide practical advice and guidelines for researchers. We focus on issues related to randomization to treatment and causal identification of treatment effects, on design issues such as selection between different elicitation formats, multiple auction groups in a single session, and house money effects. We also discuss sample size issues related to recent trends about pre-registration and pre-analysis plans. We then present the pros and cons of moving auction studies from the lab to the field and review the recent literature on behavioral factors that have been identified as important for auction outcomes.

Keywords: auctions, randomization, sample size, pre-registration, field experiments, behavioral factors

JEL codes: C57, C90, D44

*We would like to thank Uri Simonsohn and Jay Corrigan for helpful suggestions.

†Associate Professor, Department of Agricultural and Food Science and Technology, Alma Mater-Studiorum-University of Bologna, Viale Fanin 50, 40126, Bologna, Italy, e-mail: maurizio.canavari@unibo.it.

‡Assistant Professor, Department of Agricultural Economics & Rural Development, Agricultural University of Athens, Iera Odos 75, 11855, Greece, e-mail: adrihout@aua.gr.

§Distinguished Professor and Head, Department of Agricultural Economics, Purdue University, 403 W. State St, West Lafayette, IN 47907-2056, USA, e-mail: jlusk@purdue.edu.

¶Distinguished Professor and Tyson Endowed Chair, Department of Agricultural Economics & Agribusiness, University of Arkansas, Fayetteville, AR 72701, USA, tel:+1-4795752299 e-mail: rnayga@uark.edu.

1 Introduction

In the more than ten years since the first thorough treatise of auctions as a methodological tool for value elicitation in applied economics and marketing research (Lusk and Shogren, 2007), the literature on experimental auctions has been accumulating at an increasing rate. For example, in the years 2016 and 2017, more than 100 papers per year were published on 'experimental auctions', representing roughly a 35% increase from an average of 80 papers/year for the years 2013 to 2015 (own calculations based on a Web of Science search for the terms 'experimental' + 'auctions'; see also Figure 1 for a time trend).

Experimental auctions have become a popular method for valuation research because they allow economists, psychologists, and marketers to determine the monetary value people place on non-market goods in order to carry out cost-benefit analysis, to determine the welfare effects of technological innovation or public policy, to forecast new product success, and to understand individual's behavior as citizens and consumers (Lusk and Shogren, 2007). Businesses are eager to develop an understanding of the factors affecting consumers' willingness-to-pay (WTP) for their products in the hope that this may lead to better product adoption and pricing decisions. One of the big advantages of auctions, often advertised by its proponents, is that (given incentive compatibility) auctions do not, in principle, suffer from the problem of stated preferences surveys because they are not hypothetical; i.e., they involve exchanging real money for real goods in an active market. In addition, in experimental auctions, the price paid is separate from what the winner(s) bid, so in theory they are incentive compatible mechanisms. Finally, in contrast to non-hypothetical choice experiments run with real products and money where one needs to estimate the WTP values using discrete choice models, one can directly obtain each respondent's WTP value in auctions from the bids. However, in practice there are a variety of factors that could potentially affect auction outcomes that deserve greater attention.

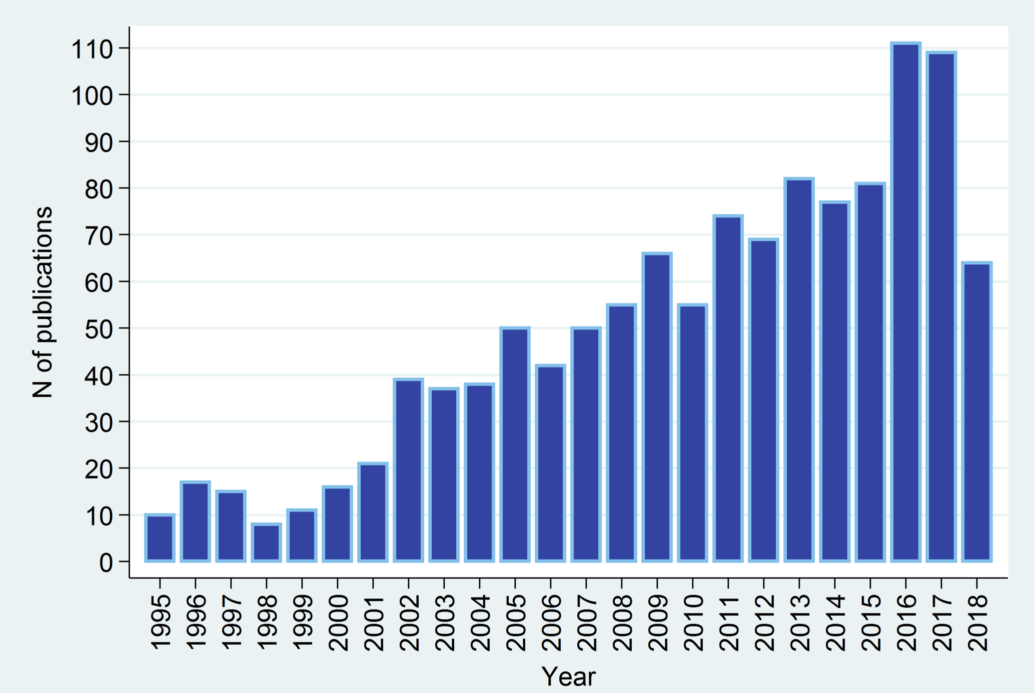
While Lusk and Shogren (2007) provide a thorough discussion of various issues involving experimental auctions (e.g., training and practice, endowment or full bidding approach, learning and affiliation etc.), there have been significant recent developments in the literature on many of these issues. For example, Lusk and Shogren (2007) when discussing the pros and cons of the endowment and the full bidding approach, they argue that if there are perfect field substitutes to products offered in the full bidding approach, then the bids for each of the products will be censored at the market price of the products and the differences in optimal bids might differ from the differences in values.[1] Since field substitutes have no effect on bids in the endowment approach, Lusk and Shogren (2007) recommend the endowment approach when outside options exist for the auction goods. Alfnes (2009) on the other hand, challenges this view and develops

[1] In the endowment approach, subjects are endowed with one product and are asked their WTP to exchange the endowed product with an upgraded product. In the full bidding approach, subjects bid on two or more products simultaneously.

a theoretical model that shows that if two alternatives that are offered in an auction differ in only one or two attributes, have the same set of field substitutes and are difficult to resell, the difference in optimal bids is equal to the difference in value.

Another example is the issue of eliciting valuations over multiple rounds in auctions and posting information (i.e., price from previous round) in between rounds. While Lusk and Shogren (2007) presented a balanced discussion on this issue of bid affiliation along with some empirical results indicating that posted prices have no effect on bids, their discussion did not settle the issue and instead led to an adversarial collaboration of researchers with different views on this issue. This adversarial collaboration resulted in the paper by Corrigan et al. (2012) which presented results from induced value experiments that show that posting prices between rounds creates larger deviations from induced values. In addition, they found that in an auction for lotteries, exposing subjects to price feedback makes them more likely to commit preference reversals. As a consequence, price posting between rounds has become much less common in recent literature.

Figure 1: Time trend of number of publications on the topic of 'experimental' + 'auction'



Source: Search for 'experimental' + 'auction' in Web of Science (August 19, 2018)

In addition to a glimpse of recent developments in the literature we offer above, in the rest of the paper we take a more systematic exploration of the practicalities of running an auction. Our general aim is to discuss issues that are important to consider in any experimental design as well as to discuss recent trends in the experimental auctions literature. Our literature review cannot be all inclusive since the volume of papers published since Lusk and Shogren's (2007) book is voluminous (take a look at Figure 1 again). Instead, we highlight papers that we believe are important from a methodological point of view; hence, we excluded many papers that just use experimental auctions as a tool for elicitation of values.

In the next section, we begin by discussing an issue that we believe is often ignored in studies that try to experimentally manipulate a factor in order to establish causality. This issue is randomization to treatment, which is a concern not just in auction studies but also to many experimental studies. In Section 3 we discuss more specific design issues that have not been given due attention such as a discussion on elicitation formats, the practice of forming multiple auction groups in a single session, house money effects in auctions etc. In Sections 4 and 5 we bring up the issues related to sample size and power as well as the need to focus not just on p-values but also on the magnitude of the estimates. We then discuss standard ways of analyzing auction data in Section 6 as well as an overview of recent advances. Then, we go over recent trends in relation to pre-registration and pre-analysis plans as a means to reduce p-hacking and increase replicability in the social sciences in Section 7. In Section 8 we discuss issues related to the conduct of auctions in the field vs. the lab. Our penultimate section (Section 9) is devoted in reviewing the literature on behavioral factors that appear to be important when considering what might affect auction outcomes. We then conclude in the final section.

2 Randomization to treatment in auctions

Many experimental auction studies are used as value elicitation vehicles to uncover consumers' WTP for novel goods and services and their attributes. As a value elicitation mechanism, these auctions do not normally apply the experimental method to estimate effect sizes from treatment and control groups since the aim is generally to construct a market that does not exist outside the auction environment. Controls are often then used in a multi-variate correlational research context in the form of independent variables to isolate the effect of these variables from other influences and to provide a causal interpretation. For example, one can regress bids on gender and then provide a potentially causal interpretation of gender effects on bidding behavior.

There are many other cases however where auction studies are conducted with the use of experimental designs that hold all other factors constant, so that the change in the outcome of interest can be associated with changes in the manipulated factor. This is casually referred to as the gold standard of experimentation for causal inference. For example, in the simplest of experimental designs, the single manipulated factor would be varied at two levels; e.g., providing information and not providing information. Causal interpretation of the manipulated factor (information in this particular example) is then based on the Neyman-Rubin model of causal inference due to Neyman (his original work appeared in Polish in a doctoral thesis submitted to the University of Warsaw; for excerpts reprinted in English and some commentary see Rubin, 1990; Speed, 1990; Splawa-Neyman et al., 1990) and Rubin (1974).

Auction studies that seek to estimate effect sizes from treatment and control groups often pay too little attention to the fact that causal interpretation based on the Neyman-Rubin model rests upon the validity of its assumptions. The Neyman-Rubin model explains causality through a framework of potential outcomes: each unit has two potential outcomes, one if the unit is treated and another one if it is untreated. A causal effect is then defined as the difference between the two potential outcomes; that is, the response of the same unit under a treatment and a control condition (the counterfactual).

In social science, however, we cannot observe both alternative conditions for the same unit because the unit changes irreversibly once it is exposed to a treatment. Note that in the Neyman-Rubin model, treatment effects from a within-subjects design do not have a causal interpretation and so there is a need to invoke additional assumptions (Holland, 1986; West and Thoemmes, 2010). In order to infer causality, one needs to compare the expected outcome of units that received different treatments in order to estimate the treatment effect. The key point is that by *randomly* assigning units to the treatments, the difference between the experimental and control units can be considered an unbiased estimate of the causal effect that the researcher is trying to isolate (Rubin, 1974). But why do we need to emphasize *random* assignment? Because only by random assignment will groups of units have the same expectation in the distribution of all covariates which are potentially relevant for the treatment outcome.[2] There is however an important caveat to remember: two randomly assigned groups will be comparable only with large enough sample sizes so that the sample average of individual characteristics becomes closer to the average of the population.

There is a practical implication coming out from the previous discussion given that running experimental auctions requires resources. Given a target sample size dictated by budget concerns (although there is an increasing demand for more sophisticated ways of determining sample size; see Section 4), simpler designs (e.g., two treatments), rather than complicated designs (e.g., 2 × 2 design), are more likely to achieve balance of characteristics that are potentially relevant for the treatment; i.e., to achieve randomization to treatment. The reason why randomization to treatment is important is because without it, we cannot be confident about the causal interpretation of the effect of the manipulated factor. For example, Briz et al. (2017) document a failure of randomization to treatment in experimental auctions by utilizing information from the practice auctions rounds. While practice auction rounds are normally

[2] It can be useful to view randomization through the lense of identification through instrumental variable (IV) regressions. A valid IV is one that is highly correlated with the explanatory variable of interest and can only affect the outcome variable through this explanatory variable of interest. A coin flip (or random draw) is a perfect IV; it completely determines assignment to the treatment or control but is not directly related to the outcome variable.

not reported or analyzed in auction studies, Briz et al. (2017) find a statistically significant treatment effect both in the practice as well as in real auction rounds. However, the treatment in their study was applied only after the practice auction rounds and hence the effect they find using the real auction data cannot be interpreted as causal but likely a result of some unbalance between the groups in some characteristics.

While it is quite a popular practice to use statistical tests to detect imbalance between groups in one or more characteristics, there is increasing discussion about how appropriate the use of such tests are. This is because any statistical test would test for the null $H_0 : \mu_A = \mu_B$ where μ_A and μ_B are the population means of two treatment groups. However, the researcher is interested in evaluating balance in the *sample*, not in the population where the samples come from. Thus, the issue of balance does not involve inference to populations (Ho et al., 2007; Imai et al., 2008). As far as the economics literature is concerned, the pitfalls of using balance tests is discussed in Deaton and Cartwright (2016).[3] Deaton and Cartwright (2016) advice that instead of reporting balance tests, researchers should report the standardized difference in means (Imbens and Rubin, 2016; Imbens and Wooldridge, 2009) calculated as $|\bar{x_1} - \bar{x_2}|/\sqrt{(s_1^2 + s_2^2)/2}$ for continuous variables and as $|\hat{p_1} - \hat{p_2}|/\sqrt{(p_1(1-p_1) + p_2(1-p_1))/2}$ for dichotomous variables with $\bar{x_j}$, $\hat{p_j}$ and s_j^2 ($j = 1, 2$) denoting the group means, prevalences and variances, respectively (Austin, 2009). The standardized difference is a scale-free measure and Cochran and Rubin's (1973) rule of thumb establishes a threshold of 0.25, below which the effect size of the difference is expected to be small.

Proper randomization to treatment does not always ensure complete balance. However, imbalance in a baseline variable is only potentially important if that variable is related to the outcome variable (Altman, 1985). Although baseline balance is not required to make valid inferences, the general advice is that even with randomization to treatment, observed covariates should be taken into account in the analysis (Senn, 1994, 2013).

In our view, randomization to treatment in auction studies is complicated by the fact that auctions generally require a group of subjects to gather in a single place at a specified time to perform an auction session. The literature has taken various approaches on this issue. In some studies, a whole session comprises a group over which an auction is performed (e.g., Lee and Fox, 2015), while in other studies, subjects in a given session are split into multiple auction groups (e.g., Drichoutis et al., 2017). In the latter case, it is possible to apply experimental treatments on a between-subject basis within the same session. If one considers time of the day and day of the week as additional confounds, then randomization to treatment will be harder to achieve given a certain sample size, especially when one adopts the session-group approach

[3] Hypothesis testing of imbalance is characterized as superfluous and misleading in the CONSORT (Consolidated Standards of Reporting Trials) statement endorsed by prominent medical journals (BMJ, Lancet etc.) (Moher et al., 2010, pp. 17).

rather than the multiple groups-session approach.

Randomization to treatment can also be affected even by very subtle things - so one needs to be very careful and cognizant of many issues. As an example, consider a case where subjects are seated in the lab in the order they arrive. It could as well be that if randomization is performed according to where subjects are seated (which could plausibly happen if ones uses zTree, for example, and connects computers in sequence to the server), then additional care has to be taken so that not all late-arrivals (students sometimes also arrive in groups) are allocated to the same auction group and potentially to the same treatment.

3 Experimental design issues

In this section we discuss a few issues that we believe can help in establishing good practice in the field.

3.1 Elicitation formats

One of the first design choices a researcher has to make is to decide on the mechanism that will be employed in the study. To get a rough sense of how popular some of these choices are, we used Google Scholar's search engine to tabulate the number of hits each auction mechanism has. Table 1 shows that the second price auction (SPA) clearly is the most popular mechanism in researchers' toolkit followed be the BDM mechanism. While the BDM mechanism is not an auction *per se*, it is often classified as such because it is seen as an alternative when one cannot easily put together a group of people (e.g., in field settings such as in supermarkets).

So why do some researchers choose other n$^{\text{th}}$ price auctions (NPAs) or the random NPA over the SPA? While we cannot definitively know the answer to this question in every single case, we believe that this choice comes as a response to a popular paper (Shogren et al., 2001) that showed that SPAs fall short in revealing preferences for subjects that are off-margin of the market clearing price (i.e., their value is not close to the 2$^{\text{nd}}$ highest price). In the random NPA, even off-margin bidders are not de-motivated to bid their true value because it is highly likely that their bid is close to the market clearing price. The disadvantage of a NPA is that it can logistically have a higher cost since the number of units sold in each auction increase proportionally with n. In addition, in the case of a *random* NPA one cannot predict how many units of the good will be needed or sold in the auction. Therefore, in cases when the good is hard to produce or is costly to produce (remember that in most cases auctions are used to elicit values for products that are not yet in the market), then it is likely that researchers will shift their preference to other mechanisms.

The BDM mechanism is a popular mechanism because of its ability to elicit valuations on an

Table 1: Number of Google Scholar search results citing different elicitation mechanisms

Mechanism	N of search results
Second (or 2nd) price auction	12,032
Third (or 3rd) price auction	154
Fourth (or 4th) price auction	84
Fifth (or 5th) price auction	91
(Random) nth price auction	432
BDM mechanism or BDM auction	1,513

Note: Table shows Google Scholar cumulative search results for the mechanism name as shown in the first column of the table and possible variants of the name. Results were retrieved on August 9, 2018 and were not checked if the respective study actually employed the mechanism or whether was just citing another study that employed the mechanism. The absolute numbers are, therefore, indicative and should only be interpreted with respect to the resulting ranking.

individual basis; i.e., you don't need a group of people. Therefore, it is normally favored outside of the lab where it is harder to recruit people to participate in an experiment at the same time. This is exactly the reason why the BDM mechanism is the favored mechanism in neuroimaging studies (see for example Kang and Camerer, 2013; Lehner et al., 2017; Linder et al., 2010; Tyson-Carr et al., 2018; Veling et al., 2017) where interaction between individuals while undertaking a brain scan is almost non-existent. Many neuroimaging studies cite Plassmann et al. (2007) as the earliest demonstration of the use of the BDM task while subjects take a brain scan although it was clearly preceded by Grether et al. (2007) and Rowe (2001).[4]

Although the BDM mechanism is the second most popular elicitation mechanism in the literature (see Table 1), there are a number of issues that researchers need to know and consider when using this mechanism. To begin with, Karni and Safra (1987) showed that the BDM mechanism is not incentive-compatible in valuing lotteries, even for rational agents. Even though this should not be a concern for experimental auction studies that seek to elicit valuations of products, Horowitz (2006) pointed out that the BDM mechanism may not be incentive-compatible even when the objects involve no uncertainty, as in the case of regular products. Although Horowitz (2006) point that this non-incentive-compatibility also holds for the Vickrey auction and general nth-price auctions, the literature has more often taken aim at the BDM, resulting in an accumulating body of research dealing with the pitfalls of using the BDM.

Banerji and Gupta (2014) provided theoretical and experimental results that confirm the role of expectations in the BDM mechanism. Specifically, they varied the support (i.e., the range) of

[4] The only neuroimaging study we are aware of that actually employed a two-person Vickrey auction is Delgado et al. (2008). In this study, subjects before entering the scanner met their opponent which would bid according to pre-defined strategies. The subject in the scanner would randomly receive one of four induced values and then could only select to bid one out of four options. In addition, in an interesting variant of the BDM mechanism, Lehner et al. (2017) use motor effort as the currency where subjects bid how many seconds they would be willing to apply 50% of their maximal grip force in order to receive the displayed reward of either money or food.

the randomly drawn bid for a chocolate and found a significant difference in valuations, a result which is in accordance with expectation-based reference points. Other relevant studies include Mazar et al. (2013) who tested the sensitivity of valuations to the underlying distribution in the BDM using travel mugs and Amazon vouchers; Urbancic (2011) using a within-subjects design and a gift certificate product redeemable for cookies; and Rosato and Tymula (2016) who used products with higher market values such as a backpack, an iPod Shuffle, and a pair of noise-canceling headphones. Cason and Plott (2014) provide evidence that the BDM mechanism is a problematic measurement tool because bids reflect mistakes rather than true preferences due to a failure of game form recognition; i.e., subjects behave as if they bid in a first price auction. More recently, Vassilopoulos et al. (2018) found that previous research findings that casted doubts on the incentive compatibility of the BDM mechanism were made on valid grounds. Specifically, they found that bids derived from the BDM mechanism are indeed dependent on the underlying distributions of the random competing bid (due to the expectations they generate) and on the anchoring of bids to the chosen price support.

But if the BDM mechanism is biased in all the ways described above, what is the alternative? Given that the main reason for using the BDM is that it avoids having to recruit many people to be at the same place at the same time to elicit their valuation, the next best solution could be to employ a SPA with just two subjects forming an auction group. Although this could be considered slightly more complicated than the BDM, it would still be possible to perform this in the field. For example, one could have two interviewers that will interview subjects almost simultaneously at two sides of a survey location. Instructions could explain to subjects that they are bidding against an unknown bidder on the other side of the location. Bids can then be easily compared to each other. This way, a SPA would be performed without having to resort to the biases associated with the BDM mechanism.

There is one potential caveat to our last suggestion. Subjects with altruistic incentives are likely to submit a zero bid in the hope that their bid would be chosen as the binding price and everyone else participating in the auction would get to buy the item for nothing. If such altruistic motives are prevalent, then small groups of auctions would enhance these motives. Whether this is a significant problem or not (i.e., how prevalent such a behavior could be in reality), is a topic worthy of further investigation.

3.2 Number of bidders in a group

An additional design choice which is related to our discussion in the previous subsection is the number of subjects in an auction group. Many studies in the literature have formed auction groups based on the number of subjects that showed up in a given session (e.g., Lee and Fox, 2015). If the number of subjects is not kept constant e.g., by turning away extra subjects, this will result in auction groups having different numbers of subjects. Given that aversion to turning away subjects should generally be correlated with the difficulty of recruiting subjects, then student subjects should be the easiest group of subjects to turn away. Turning away subjects is a very common practice in the experimental economics literature and this is exactly the reason why an experimenter needs to establish explicit show-up fees.

However, recent theoretical (Banerji and Gupta, 2014) and empirical studies (Rosato and Tymula, 2016) have shown that if subjects have reference dependent preferences, then the equilibrium bid is lower when the number of bidders is larger. Hence, one could eliminate a possible confound by keeping the number of bidders constant across auction groups. Given our discussion in the previous subsection about using a small number of subjects in each auction group (which can be as low as two subjects) that could be an alternative to the BDM, then one could also design a study using multiple auction groups in a given session (ideally subjects would be randomly matched into groups). This practice could be beneficial on two more fronts. One is that by having multiple groups in a given session and the fact that one can likely perform only one session at a time, perfect collinearity between session (a given session can confound time of the day effects and day of the week effects) and auction group is avoided. Furthermore, the general experimental economics literature often treats a group of subjects as one independent observation (e.g., Abbink and Hennig-Schmidt, 2006; Keser et al., 1998). Therefore, doing more auction groups in an experiment maximizes the number of independent observations. Another advantage of multiple auction groups is that it can reduce the risk of disruption of the experiment in case one of the participants decides to quit in the middle of the experiment; in this case, it would be possible to go forth with the auction, discarding only the auction group affected by the participant's defection.

3.3 House money and experimenter demand effects

Experimenters are sometimes rightfully concerned that when subjects receive an endowment of money, they might feel obliged to freely spend some of it in the experiment since they might consider this not their own money or they might feel obliged to reciprocate to the experimenter. This effect is called the 'house money' effect or sometimes called 'windfall money' (e.g., Corgnet et al., 2014; Jacquemet et al., 2009). Thus, observed bids may not reflect subjects' true valuation for the product but rather their need to reciprocate to the experimenter or their moral obligation to affirm that the product the experimenter is offering is of good value (i.e., an experimenter demand effect) (Zizzo, 2010).

A remedy could be to let subjects feel that they earned part of their endowment. For example, tasks from the experimental economics literature can be adopted (often called real effort tasks). One typical real effort task is the one from Abeler et al. (2011), also used in auctions studies in Drichoutis et al. (2017) and Kechagia and Drichoutis (2017). These real effort tasks do not normally require any prior knowledge, offer little learning possibility, and are simple enough to apply so that everybody can always complete the task successfully. Typically, in these tasks, subjects have to count and report the number of zeros shown in a matrix composed of zeros and ones. The difficulty of the task can be varied by varying the dimensions of the matrix. A 4×4 matrix for subjects from the general population and a 5×5 matrix for students can be employed, albeit this decision is rather *ad hoc*. This task can be repeated (the elements of the matrix change in each repetition but should be kept the same for all subjects at a given repetition) and subjects can earn a fixed payment of e.g., €0.5, every time they correctly solve the task within a given amount of time; e.g., 30 seconds. Since the task is purposefully easy, evidence from Drichoutis et al. (2017) and Kechagia and Drichoutis (2017) show that the vast majority of subjects solve this task correctly almost all of the time. It is crucial to make this task easy enough so that subjects would start off in the auction stage with approximately equal endowments, given that unequal endowments can confound bidding behavior. There are other real effort tasks that one could use but we are not aware of any rigorous assessment of the ability of different tasks to mitigate house money effects in the context of experimental auctions.

An alternative way is to let subjects bid with their own money. For example, one could provide gift vouchers as the participation fee but then make clear to subjects that they will have to pay for anything they bid at the auction. Davis et al. (2010) had a group of subjects physically receive a payment at the beginning of the session (to be considered their own money) while other subjects received it at the very end of the session with the rest of their earnings. Subjects that received money at the beginning of the session purchased information more frequently, which is consistent with increasing risk aversion. Rosenboim and Shavit (2012) prepaid one group of students two weeks before the experiment and found that this group put a greater effort to reduce their possible losses and that they also bid lower in a SPA. Zhang et al. (2017) used a delayed payment mechanism, where subjects at the day of the experiment had to pay with their own money, while they received their participation fees two weeks later. They found that the delayed payment mechanism reduced overbidding behavior especially for subjects with liquidity constraints; i.e., subjects that did not bring enough money to the session.

Taken together, the results from these studies seem to suggest that on the spot payments after a session induces more risk loving behavior (consistent with a house money effect) which results in overbidding. Therefore, a pre-paid mechanism (either two weeks before or in the form of a gift voucher that cannot be cashed) may counteract the house money effect and the resulting bias.

Overall, we believe that there is still a need for more research to examine the effect of different payment mechanisms in experimental auction settings. A comprehensive study that compares bidding behavior across various payments mechanisms would be of interest to the literature. In any case, this topic is strongly affected by regulations and ethical practices considered acceptable in the scientific community and by the non-trivial consideration that a subject should perceive that the reward obtained is worth the effort of participating in the study. This is especially challenging in field studies, since obtaining collaboration from operators in the field (for instance, a supermarket chain or a food specialty store hosting the data collection phase) is often conditional on the guarantee (or the expectation) that none of the participants will complain afterwards.

3.4 Number of repetitions and number of auctioned products

Generally, experimenters would prefer doing multiple rounds of an auction given previous evidence that it improves outcomes (Corrigan et al., 2012). However, the exact number of rounds to use in an auction is often a matter of trade-off between getting more observations and subjects spending more time in a session. For computerized experiments, this could be a trivial problem as the automated procedure allows the auction to roll faster than a paper and pencil auction experiment (although we don't see much of the latter anymore). The number of rounds could also be dictated by sample size calculations (see Equation 1 in Section 4) but normally a choice between for example an eight-round auction and a nine-round auction will not make a big difference in the resulting sample size.

As far as the number of products that one can simultaneously auction, there is normally a trade-off between increasing complexity versus eliciting information for more products as the number of auctioned products is increased. One also has to be careful because as the number of different products being auctioned goes up, the ability of subjects to differentiate between the different versions of the products might be decreased and confusion could arise. In addition, one has to be cautious about order effects when eliciting valuations for multiple products in multiple valuation tasks (Belton and Sugden, 2018).

Given that experimental auctions that use multiple products and multiple rounds generally employ the practice of randomly selecting one product and one round as binding, the consequence of having many products and many rounds is that the probability of any given round or product becoming binding is reduced. This could lead to subjects treating any particular round and product as a low-probability event and so the expected cost of misbehaving in any round or for any product can become relatively small (the cost of misbehaving was a heated discussion for first price auctions in Harrison (1989) and Harrison (1992); see also Lusk et al. (2007)).

3.5 Experimental instructions

We chose to place our last point about experimental instructions in this section because we believe that this is an integral part of a design. For experimental economics in general, instructions are very important because these help in explaining all aspects of the experiment to potentially unfamiliar subjects. Moreover, instructions can facilitate replicability of experiments as well as help identify small details that might explain some of the results obtained under a particular design.

In order to be able to evaluate a specific experimental auction design, one needs to take a look at the instructions that were provided to the subjects. Orally transmitted instructions to subjects without written transcripts would make it impossible for anyone to accurately evaluate or replicate a study, not to mention the possibility that the experimenter could introduce unknown confounds between sessions if there are improvisations or deviations from the script. Therefore, when written instructions are in place, it is also important for the experimenter to strictly follow the script.

One way to minimize or eliminate these confounds is by providing all instructions in electronic format (given that the experiment is computerized) and by creating interactive screens where subjects can familiarize themselves with the auction environment and answer practice questions. One should move on with the experimental auction only when all subjects have really understood the whole instructional set. That said, one should also be cognizant of the computer literacy of the subjects that participate in a given session. For example, while students are likely very good in taking instructions in electronic form and in interacting with a computer, this may not be true for subjects from the general population. While it is also important to encourage subjects to ask questions during a session, it is normally preferable that this be done in private so that the experimenter can first filter the question and answer she wants to provide to the group.

It goes without saying that experimental instructions should always be evaluated along with other methodological and statistical analysis standards. Editorial policies could enforce such submission of instructions by requiring this at the submission stage. From a reviewer's perspective, one could push for this practice to be uniformly applied in the field by refusing to review submissions without the experimental instructions. From the author's perspective, the fear of outright rejection could be enough to ensure that instructions are systematically and properly administered in experiments and then submitted to journals for evaluation.

4 Sample size and statistical testing issues

Scientific hypothesis testing relies on methods of statistical inference to empirically establish that an effect is not due to chance alone. This has been the gold standard of science ever since Ronald A. Fisher's era. A 'test of significance' (Fisher, 1925) of a treatment effect establishes that the effect is statistically significant when the test statistic allows us to reject the null hypothesis of no difference between two conditions based on a pre-specified low probability threshold.[5]

All statistical hypothesis tests have a probability of making one of two errors: an incorrect rejection of a true null hypothesis (type I error) representing a false positive; or a failure to reject a false null hypothesis (type II error) representing a false negative.[6] False positives have received a great deal of attention; academic journals are less likely to publish null results and p-value hacking makes false positives vastly more likely (Simmons et al., 2011).[7]

False positives may have more serious implications than false negatives by leading the research community into false avenues and wasting resources. The problem of false positives is further exacerbated by the fact that researchers may not only file-drawer entire studies but also file-drawer subsets of analyses that produce non-significant results (Simonsohn et al., 2014). In addition, researchers rarely take the extra step of replicating their original study (but see Kessler and Meier, 2014, for an exception).[8] Given the well known general lack of reproducibility of scientific studies in economics (Camerer et al., 2016) and psychology (Open Science Collaboration, 2015) respectively, there is growing concern over the credibility of claims of new discoveries based on statistically significant findings.[9]

There have been several calls for actions and proposed solutions for the seemingly high false positive rate. Nuzzo (2014) reports that p-values are widely misinterpreted as showing the exact chance of the result being a false alarm when such statements are really only qualified if the odds that a real effect is there are known in the first place. General rule of thumb conversions cited in Nuzzo (2014) indicate that "..... a p-value of 0.01 corresponds to a false-alarm probability of at least 11%, depending on the underlying probability that there is a true effect; a p-value of 0.05 raises that chance to at least 29%." Exact replications are therefore likely to uncover

[5] As a side note, according to Brodeur et al. (2016), Fisher supposedly decided to establish the 5% significance level since he was earning 5% of royalties for his publications.

[6] A type III error, typically not one that researchers often deal with, occurs when a researcher produces the right answer to the wrong question (Kimball, 1957). Kennedy (2002) warns that this is not to be confused with psychologists' type III error (Kaiser, 1960), which is concerned with concluding significance in the wrong direction.

[7] P-hacking refers to the practice of monitoring the data recording process or the outcomes of an experiment and choosing when to stop recording data, what variables to report, which comparisons to make, which observations to exclude, and which statistical methods to use in order to reach a p-value of 0.05. Brodeur et al. (2018) surveyed thousands of hypothesis tests reported in top economics journals in 2015 and show that selective publication and p-hacking is a substantial problem in research employing difference-in-difference methods and instrumental variables, while randomized control trials and regression discontinuity designs are less problematic.

[8] Clemens (2015) proposes the terms 'verification' and 'reproduction' to distinguish between replications and the terms 'reanalysis' and 'extension' to distinguish between robustness exercises.

[9] For psychological studies 36% of the replications yielded statistically significant findings while the mean effect size in the replications was approximately half the magnitude of the mean effect size of the original effects (Open Science Collaboration, 2015). For economic science studies, Camerer et al. (2016) found a significant effect in the same direction as in the original study for 11 replications (roughly 61%) while, on average, the replicated effect size was 66% of the original. More recently, Camerer et al. (2018) replicated 21 experimental studies in the social sciences published in Nature and Science between 2010 and 2015 and found a significant effect in the same direction as the original study for 13 (62%) studies while the effect size of the replications was on average about 50% of the original effect size.

false positives although the incentives for individual researchers to self-replicate one of their experiments are still currently weak.

Other researchers have taken aim at the p-value, calling for a change in the default p-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries (Benjamin et al., 2018). Another group of researchers responded that instead of adopting another arbitrary threshold, a better solution would be to make academics justify their use of specific p-values (Lakens et al., 2018). The criticism for p-values is not new. The epidemiologist Kenneth Rothman founded the journal *Epidemiology* in 1990 and as chief editor for a decade, he enforced the reporting of Confidence Intervals (CIs) instead of p-values. While his policy was successfully enforced, compliance was superficial as very few authors referred to CIs when discussing results (Fidler et al., 2004). More recently, journals like *Basic and Applied Social Psychology* and *Political Analysis* have moved one step further by banning p-values altogether (Gill, 2018; Trafimow and Marks, 2015) while at about the same time, the American Statistical Association issued a statement on the misuse of p-values and articulated principles of widespread consensus in the statistical community in order to improve conduct and interpretation of quantitative science (Wasserstein and Lazar, 2016).[10]

Recently, Simonsohn et al. (2014) introduced p-curve as a way to distinguish between selective reporting of non-existent effects and the truth.[11] This approach overcomes the limitations of previous approaches such as the 'funnel plots' method (Duval and Tweedie, 2000; Egger et al., 1997), the 'fail safe' method (Orwin, 1983; Rosenthal, 1979) and the 'excessive-significance' test (Ioannidis and Trikalinos, 2007)[12]. The p-curve tool requires a set of studies to be included in the analysis and as such, single-papers should contain multiple studies and at least one direct replication of one of the studies (Simonsohn et al., 2014). Given that self-replication is rare in the literature, it is quite hard to detect a false positive from single-paper studies.

Type II errors, on the other hand, have not been given similar attention. Zhang and Ortmann (2013) reviewed 95 papers published in *Experimental Economics* between 2010 and 2012 and

[10] Interestingly, Gigerenzer et al. (2004) mention that the *Journal of the Experimental Analysis of Behavior* and the *Journal of Mathematical Psychology* were launched as a way to escape *Journal of Experimental Psychology* editor's policy that made null hypothesis testing a necessary condition for the acceptance of papers and small p-values the hallmark of excellent experimentation.

[11] P-curve is the distribution of statistically significant p-values for a set of studies. Its shape is diagnostic of when one can rule out selective reporting as the sole explanation of a set of findings. P-curves that are left-skewed suggest the presence of intense p-hacking; i.e., researchers file-drawer the subsets of analyses that produce non-significant results.

[12] A 'funnel plot' is a scatterplot designed to check for the existence of publication bias by depicting a treatment effect against a measure of study size like total sample size, standard error of the treatment effect or the inverse variance of the treatment effect. The 'fail-safe' method consists of an algorithm in which an overall z-score is computed by summing individual z-scores and dividing by the square root of the number of scores. The 'fail-safe' is the number of studies needed to bring a significant overall p level up to some critical level like 0.05. The 'excessive-significance' test is an exploratory test for examining whether there is an excess of published statistically significant results as compared with what their true proportion should be in a body of evidence. See Simonsohn et al. (2014) for details on the limitations of these approaches.

found that only one article mentions statistical power and sample size issues. Replication studies (e.g., Maniadis et al., 2014) are particularly prone to false negatives because they are typically underpowered (Simonsohn et al., 2013).[13]

As far as experimental auction research is concerned, a priori sample size or power calculations are almost non-existent in agricultural economics journals. However, journals and editors outside the subdiscipline are now embracing the idea of including power calculations and we suspect that this would soon be discussed and considered as well in agricultural economics journals. While there are many statistical programs and packages that allow power analysis and/or calculation of optimal sample sizes (see for example Table 2 in Bellemare et al., 2016), researchers are better off getting their hands dirty. In auctions, sample size calculations are facilitated by the continuous nature of the observed variable (the bid) but could get slightly more complicated if the repeated nature of the auction setting needs to be taken into account. The reason we believe that sample size calculations should be an integral part of any experimental auction study is because researchers may end up with a result which is not statistically significant because the sample size was not large enough to detect a difference of practical significance. In addition, resources might be wasted by using a sample size that is much larger than is needed to detect a relevant difference.

In order to calculate an optimal sample size given a continuous outcome of interest (the bid), a dichotomous between-subjects treatment and a multiple-round design (i.e., subjects are asked to submit a bid in multiple rounds for the same product), we first need to assume appropriate values for the Type I and Type II errors. Following the standards in the literature, we can assume $\alpha = 0.05$ (Type I error) and $\beta = 0.20$ (Type II error). To compare the means from the two treatments, μ_0 and μ_1, with common variance of σ^2 in order to achieve a power of at least $1 - \beta$, given a number of repeated measurements M (i.e., auction rounds) as well as a value for the correlation ρ between observations, the per group/treatment minimum sample size is then given by (Diggle et al., 2002, p. 30; Liu and Wu, 2005;Kupper and Hafner, 1989):

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2(1 + (M-1)\rho)}{M(\frac{\mu_0 - \mu_1}{\sigma})^2} \tag{1}$$

Note that the formula simplifies to Equation 6 of List et al. (2011) in the case of single-round auctions by setting $M = 1$. To calculate a minimum sample size, one needs to feed the above formula with values for σ, the minimum meaningful difference $\mu_0 - \mu_1$ and ρ. The value of σ and ρ must not be calculated using the collected data (this is often called retrospective, post-hoc or observed power). This is because reporting posterior power (or posterior sample size) is

[13] For example, Simonsohn et al. (2013) argue that the null result obtained in the replication study of Maniadis et al. (2014), is just a noisy estimate and that the *relative* effect size is comparable to the original study of Ariely et al. (2003).

nothing more than reporting the p-value in a different way since power is a 1:1 function of the p-value (Gelman, 2018; Hoenig and Heisey, 2001). Drichoutis et al. (2015) and Briz et al. (2017) provide some examples on how one can use prior literature to find the relevant parameters for the sample size calculation. In this case, power is called prospective or *a priori* power (and similarly for sample size). Sample size calculations should be performed before data collection so that there is a clear stopping rule that defines the data-collection plan.

The minimum meaningful difference or expected effect size $d = \mu_0 - \mu_1$ can be derived from theoretical predictions or can be defined as the smallest effect size of interest. It can also be the product of a systematic literature review or it could be informed by auxiliary data, meta-analysis etc. (Gelman and Carlin, 2014). Gelman and Carlin (2014) also suggested using a broad range of effect sizes given that past estimates of effect sizes tend to be overestimates. This is because when the true effect is medium-sized, only small studies that (by chance) overestimate the magnitude of the effect will pass the threshold for discovery (Button et al., 2013).[14] Given a range of values for d, σ and ρ, one can find an interval of observations per treatment that are needed to detect a given effect size. This is a trivial calculation and a Stata code example is provided in the Electronic Supplementary Material. One can even experiment with the number of rounds at the experimental design stage and find an optimum number of rounds, given that more rounds are inversely related to the number of observations that are needed per treatment group.[15]

Sample size calculations can become more complicated however with more than one treatment level. One would first have to compare which contrasts are of interest. For example, consider a three treatment scenario where the control is compared with two treatments but the two treatments are never compared to each other. List et al. (2011) provide an example where the optimal allocation weighs more heavily to the control by allocating half of the sample to the control and one fourth to each of the other two treatments. Intuitively, the control is used in two contrasts while the other two treatments are used in just one comparison each; hence the

[14] In addition, the pressure of using the criterion of statistical significance may have led published research to systematically overestimate effect sizes (Lane and Dunlap, 1978) and report inflated effects (Fanelli and Ioannidis, 2013). Analyzing a large number of test statistics ($> 50,000$) published in the *American Economic Review*, the *Journal of Political Economy*, and the *Quarterly Journal of Economics* between 2005 and 2011, Brodeur et al. (2016) found a misallocation pattern in the distribution of the test statistics consistent with inflation bias. That is, researchers inflate the value of almost-rejected tests by choosing a slightly more 'significant' specification which amounts to 10% - 20% among the tests that are close to the significant threshold. They do not find that this problem arises in randomized control trials or laboratory experiments. As a side note, a consequence of the inflation bias is that if a replication study is powered based on the effect size of the original study, the power of the replicated study will be lower than intended (Simonsohn, 2015). Simonsohn (2015) suggests that the sample size of the replication study should be set to 2.5 times that of the original study (given an effect size of the original study that would give the study 33% power).

[15] This is easy to show mathematically. If one takes the partial derivative of n with respect to M in equation 1 we have: $\frac{\partial n}{\partial M} = \frac{2(z_{1-\alpha/2}+z_{1-\beta})^2(\rho-1)}{M^2(\frac{\mu_0-\mu_1}{\sigma})^2}$ and because $-1 < \rho < 1$ then $\frac{\partial n}{\partial M} < 0$ which is to say that n and M are inversely related.

allocation of observations in each treatment should not be equal.

Sample size calculations that have closed form expressions like the ones laid out above are typically used for simple statistical models. For more advanced statistical tests, estimation methods, and special design features, one would need to use simulation methods to approximate the power of complicated experimental designs. One recent contribution customized to the needs of experimental economists is the `powerBBK` package (Bellemare et al., 2016) implemented in Stata. The package allows the user to specify details about the experimental design e.g., the number of subjects, number of periods, within or between-subjects design, balance of the design, to specify individual heterogeneity by means of random-effects terms, accommodates non-linear models (i.e., logit, probit, tobit) etc.

Now let us take one step back. In the beginning of section 2 we discussed how experimental auctions are sometimes used as a value elicitation vehicle where the sole interest is in estimating the mean WTP, μ_{wtp}, of a random sample with variance σ^2 from an $N(\mu_{wtp}, \sigma^2)$ population but not in estimating a treatment effect. In this case, we can specify the maximum $100(1-\alpha)\%$ confidence interval width so that μ_{wtp} is estimated within a tolerance of $\pm A$ units. The minimum sample size n_m needed to achieve this precision is the smallest integer satisfying $n_m \geq [(\sigma/A)z_{1-\alpha/2}]^2$ (Kupper and Hafner, 1989).

5 Economic significance of estimates

One of the main caveats of p-values is that they do not provide a measure for the strength of an effect.[16] Furthermore, experiments with large sample sizes will be powerful enough to detect as statistically significant even small differences that maybe are not of practical or economic significance.

Standardized effect sizes can be an important complement to statistical significance testing. There are many effect size measures but one of the most popular is Cohen's (1988) d index, a pure number, free from measurement units, like many other effect size measures. The d index is the standardized mean difference of two means over the pooled standard deviation: $d = \frac{\bar{b}_1 - \bar{b}_2}{s}$ where $s^2 = \frac{\sum_{i=1}^{n_1}(b_{1i}-\bar{b}_1)^2 + \sum_{i=1}^{n_2}(b_{2i}-\bar{b}_2)^2}{n_1 - n_2 - 2} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 - n_2 - 2}$ is the common variance pooled over the variances $s_j^2 = \frac{\sum_{i=1}^{n_j}(b_{ji}-\bar{b}_j)^2}{n_j - 1}$ of the two groups for $j = 1, 2$. Given that the t-statistic to test whether the means are different is $t = \frac{\bar{b}_1 - \bar{b}_2}{s\sqrt{1/n_1 + 1/n_2}}$ we can then write $d = t\sqrt{1/n_1 + 1/n_2}$ (Cohen, 1988, pp.67). This last formula provides useful results because it links the power of

[16] There is a tendency to describe results that are near-threshold p-values as 'approaching' significance (Pritschet et al., 2016) which is consistent with treating significance as a continuum. This is a statistically flawed practice. In a popular blog-post, Hankins (2013) lists more than 500 linguistic terms that researchers use in order to report results that fail the significance test. In fact, there is a web application (called Signify: `http://perma.cc/MX9X-KA5Z`) which will let one type the p-value of their results and the application will come out with a label that makes that p-value sound significant.

the two-sample t-test with the difference between means d, the common standard deviations s and sample sizes n_1 and n_2. The inverse relationship between sample sizes and mean difference indicates that given a difference, larger samples will increase power. Or that given a total sample size, the standard error $s\sqrt{1/n_1 + 1/n_2}$ is minimized by having $n_1 = n_2$; that is, equally split the number of observations into the two treatments (see also Kenny, 1987, pp. 213-214).

As a crude guide, Cohen (1988) offers conventional operational definitions of 0.20, 0.50 and 0.80 for 'small', 'medium' and 'large' values of d, respectively. The values for these effects should be judged relative to the research field or to the specific content being employed in any given investigation. In Cohen's terminology, large or small effect sizes are not meant to classify treatment effects as 'important' or 'not important'. A 'small' effect size is to be interpreted as something that is really happening in the world but which can only be seen through careful study. A 'large' effect size is an effect which is big enough that can be spotted with a 'naked observational eye' (Cohen, 1988, pp. 13). Cohen (1988) notes that 'many effects sought in personality, social, and clinical-psychological research are likely to be small effects.' With respect to our previous discussion about how effect size and sample size are related and given a certain power level, this would imply that if one is to identify a psychological treatment effect in an auction setting, then one should aim for higher sample size because what she is really trying to find is likely small.

There are many other versions of standardized differences similar to Cohen's d. For example, Hedges's (1981) g applies a correction to the d index, Glass's Δ uses the standard deviation of the control group in the formula for d while Kline (2013) proposes reporting Glass's Δ using the standard deviation for each group. These are all trivial to calculate using today's software. For example, in Stata these can all be calculated using the `esize` command. This set of measures based on the differences of the means are often called the 'Difference' family or the 'd' family for short.

A second type of effect size measures, either called the 'Correlation' family or the 'r' family, quantifies the ratio of the variance attributable to an effect and is interpreted as the 'proportion of variance explained'. This family includes simple measures like the correlation coefficient r, the index q which is a measure of differences between correlation coefficients, the index eta-square η^2 which is analogous to the regression R^2 and the index ω^2 which is analogous to the adjusted R^2. These are all effect size measures easily computable with many software packages (for example with the `esize` command in Stata).

For effect sizes from more advanced econometric models, such as random effect models which are often employed for experimental auction data, the variances coming from different sources must be accounted for (Selya et al., 2012). Cohen's (1988) f^2, based on R^2 values of different versions of regression models, can circumvent the shortcomings of other standardized effect size measures. An application of this approach can be found in Briz et al. (2017). The Electronic

Supplementary Material provides an example using Stata code about how to calculate Cohen's (1988) f^2.

A final way to express a treatment effect in relative terms is to first calculate the predictions of the estimated model, then average out these predictions (which would correspond to an average predicted WTP) and find the ratio of the estimated coefficient of the treatment effect over the average prediction. This approach has been used by Kechagia and Drichoutis (2017).

6 How to analyze auction data

One of the distinct advantages of the auction approach, relative to discrete choice methods, is that the outcome of interest, WTP, is directly obtained. When combined with an experimental design where individuals are randomly allocated to control and treatment groups, a study's main hypotheses can be easily tested by comparing median, means, or other moments of the bid distribution across the groups. Moreover, because bids are a continuous measure of value, the distribution can be inspected without having to make distributional assumptions. As other studies have shown, bid distributions can be highly skewed or even bi-modal (e.g., Lusk et al., 2006).

Although auctions provide a direct estimate of the monetary value of a good, there is often interest in other measures of demand, such as market shares and elasticities. Auction data can also be used for these purposes without having to resort to econometric models or distributional assumptions. To illustrate, suppose individual i bid b_i^A for good A and b_i^B for good B. Which good would the individual be projected to choose if a retailer set prices p^A and p^B on goods A and B?

An individual would be projected to choose the good that provides the highest net value or consumer surplus, which is given by the difference in one's value for the good (i.e., the bid in an incentive compatible auction) and the price. Thus, A is predicted to be chosen over B if $b_i^A - p^A > b_i^B - p^B$ and $b_i^A - p^A > 0$. If $b_i^A - p^A < 0$ and $b_i^B - p^B < 0$, then the individual would be predicted to refrain from buying either A or B, as buying either of the goods would generate a loss.[17] In a sample of N consumers, the market share of A is simply a count of the number of individuals predicted to choose A divided by N. Own-price and cross-price arc elasticities can be determined by calculating how the market share of A changes when the assumed prices change. For a fuller treatment of this issue, including examples, see Lusk and Shogren (2007) or Lusk (2010).

Despite the fact that key insights from auction data often come from summary statistics,

[17] It is a straightforward matter to extend this logic to more than two goods, in which case the individual would be expected to choose the product that provided the highest net value. It is also possible to calculate shares if WTP premiums are elicited rather than full bids for each item (see Lusk, 2010).

there are common features of auction data that often prompt the need for econometric analysis. Most notably, auction bids are often censored from below at zero. It is also the case that auction bids for a conventional good can be censored from above at the field price of the good outside the experiment (Alfnes, 2009; Harrison, 2006). Although auction experiments can be constructed to allow negative bidding (Lee and Fox, 2015; Parkhurst et al., 2004), the practice is still uncommon. However, even in auctions that do not allow negative bids, it is possible to project the conditional mean bid if negative bids had been allowed through the straightforward application of the Tobit model. Likewise, the Tobit model can be used to estimate the conditional mean bid under the assumption that there was no censoring of bid at field prices.

While the Tobit model is widely used, our experience is that the interpretation of the model's coefficients are not well understood. The Tobit model draws a distinction between an uncensored latent or unobservable variable, b^*, (typically assumed to be Normally distributed) and the censored variable that is actually observed, b. In the case of censoring from below at zero, $b = b^*$ if $b > 0$, but $b = 0$ if $b \leq 0$. An important point is that the estimates from a Tobit model are the projected impacts on the mean of the b^*, i.e., the uncensored mean. The parameter from a simple model that includes only a constant term is the estimated mean of b^*, the uncensored distribution (i.e., the distribution that theoretically allows zero bids). As a result, when other explanatory variables are added to the model, the estimated coefficients relate to the marginal effects on the uncensored bid distribution. Overall, there are four values of potential interest in a Tobit model: a) marginal effects on the latent, uncensored variable, $\frac{\partial E[b^*|x]}{\partial x}$ (these are the raw coefficient estimates) b) on the observed, censored variable, $\frac{\partial E[b|x]}{\partial x}$ c) on positive bids, $\frac{\partial E[b|b>0,x]}{\partial x}$ and d) on the probability of being uncensored, $\frac{\partial Pr[b>0|x]}{\partial x}$ (see Drichoutis et al., 2017, for an application of the Tobit model with auction data).

But which marginal effects should one use? The answer depends on the question being asked. An example from outside the world of auctions might help provide some clarity. Imagine that a coach is interested in the marginal effect of halftime entertainment spending at a basketball game on the game's attendance. Data consist of many years of attendance records at games and spending at each halftime show. The distribution of attendance is censored from above at the stadium's capacity, and many observations are exactly equal to the capacity. The coefficient from the Tobit model associated with halftime spending is the estimate of the marginal effect of spending on the uncensored mean attendance $\frac{\partial E[b^*|x]}{\partial x}$, but as previously mentioned, it is also possible to estimate the marginal effect of spending on the attendance given that attendance is censored from above at capacity. Which estimate should be given to the coach? If there are no plans to expand the size of the stadium in the near future, then we should tell the coach the marginal effect on the censored distribution. If the coach is considering a stadium expansion before the next season, then we should tell him/her the marginal effect on the uncensored distribution. The decision of whether to expand the stadium might also be informed by the probability of attendance being censored at the capacity constraint.

A few additional comments about the Tobit model are in order. Just because bids might be censored at zero does not mean that a Tobit model is always required. The extent to which estimates from a Tobit model will diverge from ordinary least squares regression depends on the share of observations that are censored. Unless the share of observations is non-trivial (say, more than 5%), a Tobit model probably is not worth the trouble. It is also important to recognize that a Tobit model implicitly imposes the assumption that the effects of an explanatory variable are identical for censored and uncensored variables. One can relax this assumption by utilizing a double-hurdle model (Cragg, 1971). The double hurdle model is actually two models: 1) a probit model, where the dependent variable takes the value of 1 if the variable is censored and zero otherwise, and 2) a truncated regression model utilizing only those observations that are uncensored. It is sometimes the case that the double hurdle will yield insights that differ from the Tobit or ordinary least squares (e.g., Lusk and Fox, 2002). The double hurdle model allows one to specify a different set of independent variables that affect the decision to submit a positive bid and the second stage which is concerned with the level of the bid given a positive decision in the fist stage. In the special case where the set of independent variables is the same for both stages, then the Tobit model is nested within Cragg's double hurdle model (Burke, 2009).

Combined with our proposal of having smaller groups within a given session (see Section 3.2), and provided each subject submits bids in multiple rounds, this design creates a particular nesting with multiple levels. Bids from multiple rounds are nested within an individual, and the individual is nested within an auction group. This calls for the use of multilevel mixed-effects model to account for the lack of independence within these groups. An application can be found in Drichoutis et al. (2017).[18] Rather than explicitly modeling these random effects, it is also possible to produce clustered standard errors to account for such groupings. The extent to which one approach is preferred over the other depends on how comfortable one is in making parametric assumptions about the random effects.

Typical motivations for estimating regression models in experimental auction studies relate to the desire to control for censoring, as described above, or to control for potential differences in demographics across treatment groups. Randomization of participants to treatments or the use of within-subject designs lowers the need to worry about demographic controls; however, the general advice in Senn (1994, 2013) is that observed covariates should be taken into account in the analysis (see also discussion in Section 2). But there is another reason econometric models might be useful in analysis of auction data: consumer heterogeneity. Advances in

[18] In the latest versions of Stata one can estimate such a model via the `metobit` (`qsem`) command for version 15 (14).

discrete choice modeling have highlighted the importance and pervasiveness of heterogeneity in consumer preferences. As previously mentioned, auction studies are well suited to studying heterogeneity without econometric analysis. However, the latent class (also referred to as finite mixture) models or random coefficient models that have become common in the study of discrete choice data might also be useful in analysis of auction bids if there is reason to believe there might be heterogeneity in treatment effects.

Heterogeneity in treatment effects might arise in a variety of settings. For instance, many experimental economics studies have explored various price, information, or 'nudge' policies as they relate to healthfulness of food choice (e.g., Ellison et al., 2014; Muller et al., 2017). Imagine an experimental auction study where a control group bid on unlabeled items and a treatment group bid on the same items that now include red 'traffic light' nutritional label for unhealthy items. A simple regression model to test for the effect of the nutritional label is: $b^*_{ij} = \alpha_0 + \alpha_1 T + \beta Z_i + \epsilon_{ij}$, where b_{ij} is individual i's bid on product j, T is an indicator variable taking the value of 1 for the treatment with traffic light labels, Z is vector of demographic controls, ϵ is an error term, and α_0, α_1, and β are coefficients to be estimated. The primary interest is in the sign and significance of α_1. A reasonable hypothesis is that $\alpha_1 < 0$; inclusion of 'red' warning signs will reduce WTP for a food. However, there may be some people who will exhibit psychological reactance and respond to the policy in an unanticipated manner (Just and Hanks, 2015). Such a response, for example, can be observed if an individual does not like 'being told what to do' or interprets the policy as a threat to their autonomy or freedom. Thus, there may be some people for which we might expect $\alpha_1 > 0$.

The simple econometric model outlined above could be modified in several ways to identify heterogeneity in treatment effects. Perhaps the most straightforward approach is to include an interaction between T and some or all of the Z's: $b^*_{ij} = \alpha_0 + \alpha_1 T + \beta Z_i + \theta T Z_i + \epsilon_{ij}$. Now, the marginal effect of T on b^* is $\alpha_1 + \theta Z_i$. This approach is limited in the sense that heterogeneity in the treatment effect only arises through heterogeneity in observables, Z_i. One might hypothesize that, for example, men might be more likely to display reactance than females, but surely the effect is more complicated than its relationship to gender. One way to address this concern is to estimate a random coefficient model. In this case α_1 is replaced with the individual-specific parameter, α_{i1}, which is then assumed to be, for example, Normally distributed: $\alpha_{i1} \sim N(\overline{\alpha}_1, \sigma_{\alpha_1})$. This model allows for the estimation of the mean effect, $\overline{\alpha}_1$, but also allows for differential responses via the estimated standard deviation of the treatment effect, σ_{α_1}. A challenge with this approach is that it requires an assumption about the distribution of the treatment effect. In the case of the reactance example, it is not clear that one would expect the treatment effect to be normally distributed across people. An alternative approach is the latent class model. In this case, a number of classes, C, are specified and one estimates class-specific parameters, $\alpha_{0c} + \alpha_{1c} T + b_c Z_i$, along with the probability of respondents falling into each class.

Typically AIC or BIC measures are used to determine the number of classes, C. This approach can permit more distinct treatment heterogeneity, where for example, $\alpha_{1c=1}$ might be positive and $\alpha_{1c=2}$ might be negative, with the model revealing the share of the sample fitting each pattern.

There is much more that can be done with auction data to generate actionable insights. Lusk (2010) shows, for example, how auction bids can be used to identify consumer segments via cluster analysis or product groupings via factor analysis or multidimensional scaling.

7 Disclosure, pre-registration, pre-analysis and open data & materials

This section is motivated by the recent move of individuals, scientific societies and journals to embrace and promote more transparency in social science research. Here, we discuss issues that the various Agricultural Economics Associations and Journals in the field have not yet officially reacted to.[19] We believe that in the near future, these issues will become a higher priority and that most experimental research will eventually have to hold up to these standards. At the present time, however, there are only a handful of pre-registered studies that are related to experimental auctions and consumers' WTP (all of these are registered to AEA's CRT registry). However, we hope that our discussion here will stimulate more interest in the field of experimental auctions for agricultural economics research.

Given that experimental research is expanding and the advantages of experimentation are getting into the spotlight, along with the recent crisis in the replicability of experimental studies, there has been a call to set up practices that will promote transparency in social science research. Miguel et al. (2014) define three core practices for more transparency in social science research: disclosure, pre-analysis and pre-registration plans, and open-data/materials.

Disclosure is the systematic reporting of standards that researchers are (or should be) obliged to disclose such as the measures, manipulations, data exclusions, and the final sample size. Many prominent medical journals (BMJ, Lancet etc.) recommend or require that researchers adhere to the CONSORT (Moher et al., 2010). In the absence of an endorsement of similar standards by associations, Simmons et al. (2012) proposed a 21-word disclosure to accompany manuscripts regarding the authors' knowledge that they did not p-hack: 'We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study'. If needed, supplemental material can be used to support the disclosure.

Pre-registration involves specifying in detail, in an online repository, information such as

[19] But see Josephson and Michler (2018) for a discussion of ethical issues facing the profession and a few proposals to address these issues.

the number of subjects, the treatment and relevant stimuli, the outcome variable, prediction/hypothesis, the pre-analysis plan etc. that altogether constitute the plan for a study. The whole plan can be hosted in one of the available online repositories which can be accessed by interested parties (editors, reviewers, readers) and receives a time stamp. Although some repositories have time limited embargoes, the plan can be made public or not at the will of the researchers. For example, AEA's RCT registry can keep key information hidden until the time when the trial is completed. Some of these repositories are the Open Science Framework (https://osf.io/), AsPredicted (https://aspredicted.org/), American Economic Association's RCT registry (https://www.socialscienceregistry.org/), the Evidence in Governance and Politics registry (http://egap.org/content/registration) etc. Likely, registration for many economic experiments will be diverted to AEA's RCT registry since, as of January 2018, registration in the RCT registry has become mandatory for all submissions to AEA journals.

Some people share concerns about the time and effort costs involved with pre-registration. In addition, there are concerns as per if pre-registration can actually prevent deceiving practices. Simonsohn (2018) makes two arguments against this intuition. First, because people interpret ambiguity in self-serving ways, by reducing ambiguity through pre-registration will result in reduced self-serving biases. A smaller inclination to a self-serving bias will likely reduce the temptation to convince oneself after data collection that what worked was what was planned all along. Second, researchers may engage in deceiving practices either by omission or by commission. Likely, very few researchers would engage themselves in deception by commission and since pre-registration transforms a what would otherwise be a deceptive practice by omission to an explicit lie, we would expect a significant reduction of deceptive practices with widespread pre-registration.

So why would someone want to pre-register their research plan? Advocates claim this is the only way that criticism about data mining, p-hacking and other questionable research practices can be muted. Pre-registration creates a new step in the work-flow of research but it is seen as a good way to produce rigorous results that would allow a sharp distinction between confirmatory and exploratory analysis. To incentivize researchers to use pre-registration, many journals are now offering the option of registered reports. Registered reports involve peer reviewing a submitted research plan before data collection. If the research plan is given a positive evaluation, then the proposed paper will be given a conditional acceptance and a promise to publish it regardless of the outcome. Currently more than 100 journals use Registered Reports as a regular submission (the list is maintained by the Center for Open Science: https://perma.cc/KV4F-57ES). The vast majority of these journals are Psychology and Neuroscience journals; there is one political science journal (the *Journal of Experimental Political Science*) and one economics journal (the *Journal of Development Economics*).

We feel that the biggest drawback for researchers pre-registering their studies will be a commitment to a specific pre-analysis plan. Committing to a pre-analysis plan includes, among others, deciding on the precise definition of a primary outcome variable (or multiple co-primary outcomes) and of potential secondary outcomes, specific variable definitions, any inclusion or exclusion restrictions, statistical models, hypothesis testing methods (including correcting for multiple hypothesis testing if many primary outcomes are defined) and covariates (potentially including measures of standardized differences as described in Section 2), subgroup analysis etc.[20]

Olken (2015) summarizes the benefits and risks of committing to a pre-analysis plan. Researchers who put a lot of detail and effort in coming up with a pre-analysis plan could benefit from the careful thought processes required of going through their data analysis, and selecting which variables to collect and methods to apply for the analysis. This step could also involve researchers writing their statistical programs and run them on mock up data.[21] For the research community, it increases the confidence that the analysis did not just involve picking and reporting the most significant specification. While fully specifying all the analysis ex-ante could be considered an ambitious plan, this does not mean that additional analysis cannot be performed if not mentioned in the pre-analysis plan. One could for example still include results from analysis not included in the pre-analysis plan in the paper but these should be clearly indicated in the paper as not part of the pre-analysis plan. Coffman and Niederle (2015) offer arguments about te limited upside of pre-analysis plans and propose that economics move towards valuing replications and robustness checks of positive results instead.

To address the challenges of committing to a detailed pre-analysis plan, Anderson and Magruder (2017) and Fafchamps and Labonne (2016) revived an idea of testing the out-of sample performance of predictors in 'hold-out samples'.[22] Roughly speaking, the approach involves withholding a fraction of a sample, say half of it, and run an exploratory analysis on these data. One could then choose which hypotheses to test and the methods to test the hypotheses based on the exploratory analysis and then come up with a more well-informed pre-analysis plan using insights from the exploratory analysis. Once a specific pre-analysis plan has been decided and registered, the researcher can then use the other part of the sample to employ

[20] To pinpoint the pitfalls of subgroup analysis, Christensen and Miguel (2018) cite a story that first makes one laugh but then prompts deeper thinking. When a collaborative group of researchers were asked by journal editors to report subgroup analysis in a trial of aspirin and streptokinase use after heart attacks, the researchers found that this medication is beneficial, except for patients born under Libra and Gemini astrological signs, for whom there was a harmful effect.

[21] This is certainly possible, for example, if one runs a computerized experiment e.g., using zTree. Mock up data files can be generated before ever running an experiment and read into the statistical program of one's choice which could help on building the code for statistical analysis.

[22] This is an approach that has long been used in psychology and statistics (see citations in Anderson and Magruder, 2017) as well as to judge the predictive fit in the marketing literature (Erdem, 1996; Roy et al., 1996) and in the economics literature (Drichoutis and Lusk, 2014, 2016; Norwood et al., 2004) for model selection.

the pre-analysis plan. The split approach does come at a cost, however, since the split sample loses power relative to a full-sample pre-analysis plan on hypotheses which were anticipated.

The final point in Miguel et al. (2014) concerns open data and materials. We feel that agricultural economics journals have moved in the right direction over this and in line with data availability policies at top economics journals. For example, ERAE's guidelines to authors clearly states that 'The editors reserve the right to refuse to publish articles where the data, programs, etc. are not provided and where, in their view, there is no justifiable reason for not making them available.' The creation of journal archives has a long history dating back to Dewald et al. (1986) where by exploiting a change in the editorial policy of the *Journal of Money, Credit and Banking* that required authors to make the data available upon request, they found that the proportion of authors that submitted programs or data was significantly larger after the introduction of the policy. They then tried to replicate nine papers from authors that submitted their data to JMCB and found that only two of these could be replicated in their entirety. Dewald et al. (1986) suggested that journals require data and codes *at time of submission*.

In response to Dewald et al. (1986), the JMCB adopted a data and code archive policy. However, future replication attempts to the same journal (McCullough et al., 2006) were just marginally more successful: of the 186 empirical articles, only 69 had archive entries; 7 could not be replicated due to lack of software or the use of proprietary data; only 14 out of 62 articles could be replicated.

McCullough et al. (2008) examined compliance of depositing data in the journal's archive for journals that required this and found that the *Journal of Applied Econometrics* had a 99% compliance rate which they attributed to the fact that i) JAE had an editorial position for the archive manager and ii) that a paper is not published until the authors have made their data available. However, replication was extremely low which they attributed to the fact that no code was required to accompany the data files. In the meantime, things have not gotten better. Recently, Chang and Li (2017, 2018) were able to replicate only 33% of papers from 13 journals independently of the authors and 49% with help from the authors.

As far as experimental auctions research is concerned, we are not aware of any formal attempt yet to replicate results using data and codes from previously published papers. It is likely that a good percentage of this literature will follow the trends of the general economics literature. Independently of the current status of replicability of experimental auction results, there are things that we would endorse for the benefit of the profession and science in general. Major agricultural economics journals should make it mandatory to submit data and codes *at the time of submission*. Then an editor should be assigned as the manager of the archive. However, we also believe that it would be beneficial for overall transparency that the editor not only makes sure that data and codes are submitted but that any paper is accepted conditional on the managing editor or a third party being able to exactly replicate the results of the experimental paper using the data and codes provided by the authors during the submission (this is the verification process described in Clemens (2015) and Christensen and Miguel (2018)).

As far as availability of materials is concerned, as readers and reviewers of experimental auctions papers, we still often come across experimental auctions work that does not include experimental instructions. We believe this is a very crucial step that is often neglected in the review process, which could make a big difference when evaluating the merit of a paper. This is because even subtle words or the way phrases are expressed in the instructions could induce effects up and above what the authors believe their manipulation is inducing. We believe that one possible reason that instructions sometimes are not submitted with the paper is that there was never instructions distributed or shown to subjects; i.e., the experimenter orally explained the mechanics of the auctions and then let subjects bid on the products. As we explain in Section 3.5 this is not a practice we would endorse.

A final and perhaps subtle remark concerning transparency is that authors should make sure when using links pointing to websites that provide complementary information about their methods, analysis or any other material, that these links will outlive the paper. Normally, this is not the case because the internet changes rapidly. The solution is to make links permanent via services that archive websites e.g., http://archive.org/web and https://perma.cc/. The careful reader will notice that most of the links in this paper that point to online content, use one of these services. Another solution regarding materials associated with any research project is to post everything in an online repository like the Open Science Framework (https://osf.io/).

As a testament to the changes that other fields are undergoing, recently the journal *Cortex* (Chambers, 2018) has introduced the Transparency and Openness Promotion (TOP) guidelines (Nosek et al., 2015). The TOP Guidelines are a certification scheme in which journals and research organizations declare their level of adherence to a series of standards for enabling research transparency and reproducibility. These standards include, among others, availability of data (e.g., data must be posted to a repository), analysis code and digital research materials (code and materials must be posted to a repository; a Level 3 adherence standard would require that analysis is replicated independently before publication), preregistration of study procedures and analysis plans (e.g., authors declare if study has been pre-registered and provide access to reviewers), replication (e.g., the journal uses registered reports as a submission option) etc.

In addition, a number of journals have agreed that publishing peer review reports can be beneficial for the research community by increasing transparency of the assessment process (Available at https://perma.cc/Z7TM-G66C). These benefits might include reviewer and editorial accountability, training opportunities for educating students about the peer review process as well as a way to provide credit for peer review (since 2012, Publons https://publons.com provides a free service for academics to track, verify and showcase their peer review and editorial contributions for academic journals).

8 Field vs. lab

Harrison and List (2004) suggested a typology of economic experiments based on six classification criteria in terms of the nature of: subjects' pool, information, the good, the task/trading rules, and the stakes (amounts involved), plus the experimental environment. Harrison and List's (2004) typology distinguish 'conventional' lab experiments from 'artefactual', 'framed', and 'natural' field experiments, based on a growing role of a target population, context-relevant information, goods, and tasks, as well as investigations occurring in a natural environment where the subjects are not aware that they are being observed.

A number of studies have moved outside the lab setting to the location where consumers typically make their purchasing decisions (i.e, 'the field'). Gneezy (2016) call for an increase of the proportion of experimental studies based on field data in marketing research. Vecchio and Borrello (2018) maintain that many researchers who used experimental auctions in food consumer behavior studies think that more studies should be performed in real market environments.

The choice of whether to conduct research in the field or lab depends on numerous factors (Harrison and List, 2004), the importance of which may change depending on the specific purpose and audience of the study, while also keeping in mind that the methodological approach may also be affected by the experimental practices considered acceptable in a specific discipline (Croson, 2005).

The core trade-off between performing a lab or a field study that is often discussed is between control and realism or, to put in another way, between internal and external validity (Roe and Just, 2009). Lab settings allow researchers a greater degree of control but often under sterile environments. By contrast, field settings occur in a more natural setting that could include context-relevant information and cues, but are often harder to control. All in all, researchers must consider that conducting a study in the field, instead of in the lab, will necessarily introduce more noise and reduce the control over the experimental procedure, which could imply the need for more sophisticated models to consider control variables (Gneezy, 2016; Vecchio and Borrello, 2018).

Harrison and List (2004) maintain that lab and field studies should complement each other since they have different characteristics and given the possibility that what works in the lab does not necessarily works in the field and vice-versa. The key issue is that moving from a lab to the field can change bids, which is the most relevant information obtained in an experimental auction. Lusk and Fox (2003), for example, show that bids for a food product can increase when moving from a class-room lab setting to a bakery setting. The results are not necessarily surprising: people enter a field environment with the intention of making a purchase on the category of good in question. In addition, in the field many substitutes of the auctioned product can be readily available.

The move to field settings can have important implications on study findings related to effects and significance of factors and covariates. List (2004) showed that individuals with more experience in the field were less likely to suffer from the endowment effect; List (2003) showed that even experienced subjects, when put in a lab, appear to behave altruistically, but when moved back to a more natural environment behave more in accordance with self-interest; Dyer and Kagel (1996) showed a similar effect with regard to the winner's curse in common value auctions; Sousa and Munro (2012) confirmed this effect also in virtual online experiments.

In practical terms, field experiments can often dramatically reduce the cost of subject recruitment because the researcher travels to where the subjects are, unlike most lab experiments where subjects would travel to where the lab is. Incentives to participate in the field often include food products or coupons, which are an integral part of the design, like for instance in Lusk et al. (2001), Lusk et al. (2006), and Klain et al. (2014). However, until recently with the advent of mobile payment systems, such as Square, it was generally difficult to arrange payment mechanisms in field settings where people often did not have cash. Nevertheless, one can still encounter challenges in specific field settings like grocery stores or supermarkets if an agreement with store managers or company managers is required. For example, while traditional surveys are usually well accepted, an experiment involving sales for goods also sold in the store may be questioned or opposed by the managers who may also be worried about customer complaints that may arise from a situation that is not under their control. According to Gneezy (2016), collaborative experiments may require a lengthy process of reciprocal understanding between academic and non-academic partners, regarding the potential benefits and costs for the latter.

Another important aspect to consider when planning a field experiment are issues related to sampling methods and procedures. While a lab experiment could rely upon subjects randomly picked from a representative panel, it is usually much more difficult to select a random sample in the field, and so the chances of having a biased sample would generally be higher (Belot and James, 2014).

Some practical advice can be provided considering the discussion above:

(a) given the reduced control over the environment and unavoidable noise affecting a field experiment, it is highly desirable to reduce respondent's burden; therefore, it is advisable not to design long experiments and use very short instructions (1 page or less) for the participant. It is also recommended that researchers should devote a significant amount of time and effort in training data collectors.

(b) when choosing the auction mechanism, lean towards the simplest ones, such as the 2nd price auctions with very small groups (2-3 subjects).

(c) in case the experiment is aimed at providing industry-relevant information and statistical inference is among the expected outcomes, ensure an adequate sample size, use multiple locations in the field to get a representative sample of the target population and enforce measures to reduce selection bias to increase generalizability of results, according to the scope, purpose and audience of the study. If this is not feasible due to financial and time constraints, it is advisable to pair the auction study with a hypothetical (e.g., a choice experiment) survey (Lusk, 2010) and use the auction to calibrate the larger study results and correct for hypothetical bias (e.g., Alfnes and Rickertsen, 2007; Fox et al., 1998).

(d) consider the influence of the presence of perfect or partial substitutes in the locations and settings chosen for the field experiment, and control for these variables if possible.

(e) carefully plan the collaboration with the non-academic partners (e.g., store or restaurant managers), set up agreements clearly identifying advantages (such as purchase of the auctioned products and share of business-relevant study results), administrative burdens, costs and commitments.

(f) plan logistics issues very carefully, considering the trade-offs between on-site delivery and home delivery of the product: the former can rely upon the logistics, storage, and payment facilities of the non-academic partner in the field, while the latter can be implemented by collecting bids, payments, and delivery information on the spot and sending the purchased product to the respondents' address (this may need additional care in managing data because of privacy regulations).

9 Behavioral factors in auctions

A number of behavioral factors can influence bidding behavior in experimental auctions. In this section, we will discuss a number of studies that have directly examined some of these behavioral factors in relation to experimental auctions and their implications for the design of future studies. Since trying to review all the experimental auction behavioral studies is a daunting task, we will focus on a handful of contributions which directly test the effect of behavioral factors that have not been thoroughly examined in the past. So, while certainly also important, we will not cover issues related to the effect of information, labels, reference prices/products, and endowments.

9.1 Personality traits

Personality traits have entered economists' area of interest as important determinants for economic behavior. One of the most widely cited papers on the economics and psychology of personality traits (Borghans et al., 2008) makes a convincing case using evidence from the Perry preschool program. The program targeted disadvantaged African-American children in Michigan in the '70s and children were followed up to age 40. Apparently, the program was successful in changing *personality* and motivation of disadvantaged children which then resulted in measurable success on a variety of measures of socioeconomic achievement over the life cycles of participants (see Borghans et al., 2008, and citations therein).

Given the importance of personality in predicting outcomes and explaining variation in economically-relevant behaviors, it seems quite possible that personality traits could explain the differences in bidding behavior in experimental auctions. Grebitus et al. (2013) examined this issue in both hypothetical and non-hypothetical settings. Their results suggest that there is heterogeneity in valuation estimates across personality traits and that traits may partly explain the differences in behaviors or valuations from auction and choice experiments found in previous studies (Gracia et al., 2011; Lusk and Schroeder, 2006). Interestingly, their results also suggest that the effects of personality are stronger in non-hypothetical auctions than in hypothetical auctions. The implication of this result is that people will behave differently in real and hypothetical environments depending on their personality type, suggesting that personality traits may well explain a significant portion of hypothetical bias.

9.2 Cognitive Ability

It is well known that auction mechanisms may not always provide accurate valuation estimates, given behavioral anomalies on bidding behavior observed in lab experiments. For example, one consistent finding in experimental auction studies is that subjects tend to deviate from rational behavior and exhibit a pattern of overbidding in second price auctions (SPA) (Andreoni et al., 2007; Cooper and Fang, 2008; Drichoutis et al., 2015; Georganas et al., 2017; Kagel et al., 1987; Kagel and Levin, 1993). Kagel et al. (1987) inferred that subjects submit a higher bid in SPAs due to the impression that submitting higher bids improves the probability of winning with no real cost because the highest bidder pays the second highest bid. An alternative explanation was given by Morgan et al. (2003), who proposed that participants bid based on behavioral motives such as 'spite'. They suggested that subjects overbid in SPAs since the profit earned by a rival bidder could be reduced by a losing bidder's own bid. This overbidding behavior can be also explained by a 'joy of winning', in which subjects derive extra utility from winning the auction. Interestingly, Cooper and Fang (2008) found that small and medium overbids are consistent with the 'joy of winning' hypothesis, while large overbids are more consistent with the 'spite' hypothesis.

Understanding bid deviation in experimental auctions is important since it can potentially explain subjects irrational behavior, and it can also provide more clarity in determining when and how bids should be interpreted when trying to elicit homegrown valuations. Kagel et al. (1987) and Ausubel (2004) argued that the difficulty of understanding the SPA could lead to overbidding in the SPA as compared to the ascending-price English auction, even though both auction mechanisms are strategically equivalent. More recently, Li (2017) showed that the overbidding behavior in SPAs is due to the fact that it is not an obviously strategy-proof (OSP) mechanism, where an OSP mechanism is defined as one that a cognitively limited agent can recognize the weakly dominant strategy. This concept suggests that more cognitively able bidders will understand the strategic properties of an SPA better than low cognitive ability bidders. Lee et al. (2017) investigated this relationship directly by examining how individuals cognitive ability influences bid deviations in SPAs. They first measured subjects' cognitive abilities using a nonverbal Raven's Standard Progressive Matrices (RSPM) test and then classified subjects into two groups (i.e., a high cognitive ability group and low cognitive ability group) based on their RSPM test performance. Each group then participated in a series of induced value SPAs. Their results suggest that more cognitively able subjects behave in closer accordance with theory and that cognitive ability partially explains heterogeneity in bidding behavior. Their finding is important since it implies that experimental auction researchers must make sure that the auction mechanism used in the study is clearly understood, especially by low cognitive ability subjects.

9.3 Emotions

Roider and Schmitz (2012) examined how robust is the standard symmetric sealed-bid auctions model with risk-neutral players and private independent values. Specifically, they were interested in assessing subjects' bidding behavior when they anticipate the positive emotions of winning and the negative emotions of losing. They also investigated whether the introduction of anticipated emotions can shed light on various findings of bidding behavior in auctions with independent private values. Using a simple extension of the standard model of symmetric auctions —where bidders anticipate some (constant) positive emotions of winning and some (constant) negative emotions of losing — they showed that if bidders anticipate the joy of winning, bids will be larger than in the standard model in both first-price and second-price auctions. However, if bidders anticipate a disutility of losing, the implications depend on the auction format. In a SPA, bidders who still participate bid more when they anticipate negative emotions because they are more eager to avoid losing, and while bidders with very low valuations will not participate, all participating bidders overbid by the same amount due to the anticipated emotions. They also explained that in SPAs the joy of winning and the disutility of losing affect bids in the same way. In contrast, in first-price auctions, participating bidders with small valuations who anticipate that losing is painful, bid less than in the standard model. So for participating bidders with small valuations, a disutility of losing actually reduces bids, while bidders with high valuations bid more.

9.4 Mood

Mood states can influence behavior by influencing both the content and the process of cognition (Capra, 2004). Moods can also play an important role in the construction of preferences that, in turn, influence decision-making and judgment (Johnson et al., 2005; Lichtenstein and Slovic, 2006; Payne et al., 1999; Slovic, 1995). A few studies have examined the effect of moods in experimental auctions. Lerner et al. (2004) found that a negative mood state in the form of sadness (disgust) can increase (decrease) willingness to pay (WTP), while Capra et al. (2010) found only weak mood effects on WTP. Drichoutis et al. (2014) also explored how positive and negative mood states affect bidding behavior in experimental auctions to test the robustness of the findings of these two papers. Their study differs from the other two studies in that they focused not only on the effect of mood states on WTP but also on people's rationality, as represented by the rate of preference reversal for lotteries. They found that mood states can significantly affect the rate of preference reversal and bidding behavior in experimental auction valuation. Specifically, they showed that subjects under a positive mood state exhibit more rational behavior (i.e., fewer preference reversals) and provide lower bid values than others. Results from these studies suggest that researchers may need to take subjects' moods into account when conducting experimental auctions.

9.5 Other Regarding Preferences and Motives

The standard theory predicts that altruistic subjects underbid in the Vickrey auctions compared to the BDM, while spiteful subjects overbid in Vickrey auctions. Flynn et al. (2016) were not able to confirm these predictions, however. While they were able to observe aggregate underbidding in Vickrey auctions, their results were not driven by the choices of altruistic subjects.

9.6 Hormones

Behavioral economics has embraced the view that we can use the lens of biology to look at economic behavior. By now there is accumulating literature suggesting that gender differences in preferences and behavior can be attributed to hormonal differences between males and females.

Given that a number of studies have found that females tend to bid higher than males in auctions (Casari et al., 2007; Chen et al., 2009; Ham and Kagel, 2006; Pearson and Schipper, 2013), Chen et al. (2009) and Pearson and Schipper (2013) examined how the bidding and profits of females differ across the menstrual cycle. Menstrual cycle information can be used as a proxy of level of hormones in females that naturally fluctuate during the cycle. Chen et al. (2009) found that women bid higher than men in all phases of their menstrual cycle in a first-price auction but not in a second-price auction. Moreover, for first-price auctions they infer that higher bidding in the follicular phase and lower bidding in the luteal phase are driven entirely by oral hormonal contraceptives. Pearson and Schipper (2013) report that naturally cycling women bid significantly higher than men and earn significantly lower profits than men (in a first price auction) except during the midcycle (when fecundity is highest). They also found that women who use hormonal contraceptives bid significantly higher and earn substantially lower profits than men. This correlation they found between the use of hormonal contraceptives and bidding or profits, however, may be due to a selection effect or to hormones contained in contraceptives. All hormonal contraceptives contain synthetic versions of the sex hormone progesterone, and some also contain a version of estradiol. So, the evidence is far from conclusive.

Schipper (2015) conducted an auction experiment in which he collected salivary steroid hormones such as testosterone, estradiol, progesterone, and cortisol. He suggested that testosterone may affect bidding and profits via risk aversion. Basal testosterone has also been found to be positively correlated with 'aggression,' which may be another channel through which basal testosterone affects bidding in auctions. The results indicate that females bid significantly higher and earn significantly lower profits than males. Moreover, females who use hormonal contraceptives bid significantly higher and earn significantly lower profits. With respect to salivary basal hormones, Schipper (2015) found that bids are significantly positively correlated and profits are negatively correlated with basal salivary progesterone, but only in females who do not use hormonal contraceptives. In his study he did not find significant correlations between bidding or profits and salivary basal testosterone, estradiol, or cortisol.

9.7 Sensory cues

Many experimental auction studies conducted by agricultural and applied economists are focused on food products. The literature, however, has remained ambiguous as to whether sensory cues, such as taste or smell, needs to be measured in experimental auctions used to elicit consumers' WTP for food. While a number of studies have included sensory tests in their auctions, these studies did not completely isolate the role of taste in bidding behavior (e.g., Drichoutis et al., 2017; Feuz et al., 2004; Holmquist et al., 2012; Lusk et al., 2001; Platter et al., 2005; Umberger and Feuz, 2004; Umberger et al., 2002)

Typically in these studies, the subjects were asked to taste the food products and were then asked to bid on these products. So, the effect of taste was not directly tested. In other studies that included taste, a within-subjects design was utilized where subjects were progressively given more information about the foods in the auction and then eventually allowed to taste the food products prior to one of the later bidding rounds (Akaichi et al., 2017; Bi et al., 2012; Demont et al., 2013, 2012; Melton et al., 1996).

One study that directly tested the effect of taste in experimental auctions is Lewis et al. (2016b). They used a between-subjects design and were able to compare bids from an auction where subjects tasted the products and an auction where subjects did not taste the products. Their results suggest that taste influences bids, and they concluded that it would be valuable to include taste in auction designs when evaluating consumers' WTP for food products.

Studies that try to isolate the effect of sensory cues other than taste are even more rare. A recent study used controlled laboratory experiments to experimentally manipulate the ambient scent of the lab with a citrus fragrance (Kechagia and Drichoutis, 2017). The authors found that subjects that participated in a SPA auction inside the scented room were willing to pay up to 49% more than subjects who were not exposed to the scent.

In addition to the importance of sensory cues, there is also evidence that presence of the good to be auctioned matters. For example, Bushong et al. (2010) elicited valuations using the BDM mechanism under three different conditions: (a) text displays, (b) image displays, and (c) displays of the actual items. They found that subjects' bids were 40-61% larger in the real display than in the image and text displays. Their findings suggest the saliency of the 'tangibility' issue in experimental auctions. In particular, follow-up experiments in Bushong et al. (2010) suggest that the presence of real items trigger preprogrammed consummatory Pavlovian processes that promote behaviors that lead to contact with appetitive items when they are available.

9.8 Attention

Product evaluation can be influenced by the amount of attention paid to product stimuli, which can be linked to eye movement (see Orquin and Mueller Loose, 2013, for an overview on studies). When an individual looks at a stimulus, attention is paid to the stimulus (Wedel and Pieters, 2000). Lewis et al. (2016a) used eye tracking to measure how attention to brand, package attributes, and product information impacts consumer WTP for branded energy drinks. They found evidence that attention can explain the variation in consumers' WTP for branded energy drinks containing different sweeteners. In another study, Rihn and Yue (2016) examined the impact of extrinsic cues (specifically production method, origin, and nutrient content claim labels) on consumers' WTP for processed foods (apple juice and salad mix) using an experimental auction in combination with eye-tracking analysis. Their results suggest that consumer visual attention increases for important product attributes that positively or negatively impact their WTP bids.

10 Discussion and conclusions

This review provided a state of the art discussion of the best practices in the design and execution of experimental auctions, with the aim of improving cost-benefit and welfare analysis, theory testing, as well as marketing recommendations from their findings. Here, in the final section, we offer a brief list of recommendations distilled from our discussion above and then conclude with some suggested areas for future research.

Our first recommendation relates to power analysis. Sample size calculations should be performed at the design stage and these calculations should be used as stopping rules once the desired sample size has been achieved. Given finite resources, adopting this practice will likely lead to simpler designs (i.e., fewer treatments) in order to ensure sufficient number of subjects (and thus sufficient power) for the main queries of interest. It will also result in WTP estimates that are more precisely measured and treatment effects that have been more reliably identified. Another added benefit is that it should improve the credibility and publishability of null results.

Second, practice or training rounds in auctions should be more systematically analyzed when the aim of the study is to causally identify treatment effects. In order to causally identify a treatment effect, a sufficiently large sample size is needed to ensure that randomization to treatment was successful. One way to judge whether such randomization was successful is to analyze data from the training rounds as a form of a placebo test (Rosenbaum, 2002, p. 214) i.e., a test which examines the effect of the treatment on a variable known to be unaffected by the cause of interest. We believe that systematic adoption of this practice will lead to increased confidence in the underlying causal mechanisms.

Another important motivation for conducting practice and training rounds is that people often have misconceptions about the auction mechanisms, which can lead to non-truthful value revelation. Practice homegrown or practice induced value auctions should be always employed, unless inexperience with the mechanism is desirable. Quizzes and simple, easy to understand detailed instructions should be provided to subjects as well. All this material should then be made available to reviewers to allow them to directly assess the experimental design. This would also contribute to more transparency of research practices. The use of multiple-price lists (e.g., Andersen et al., 2006; Klain et al., 2014) or non-hypothetical choice experiments (Alfnes et al., 2006; Lusk and Schroeder, 2004) can be seen as an attempt to utilize more easily understood mechanisms, with the trade-off being less precise information about consumers' WTP than the auction method that have been the focus of this review.

When the experimental design includes multiple rounds of bidding, research results suggest that it is best to avoid providing price feedback between rounds since this could lead to a number of adverse effects on the bidding behavior. In addition, data on beliefs about prices of field substitutes should be collected when possible so that it can be incorporated in subsequent econometric analysis.

Field auction studies remain under-utilized and more of these studies should be conducted and compared with lab settings in order to highlight the role of the natural field context. One factor preventing the wider adoption of experimental auctions in field settings is the need to form auction groups. This is one reason why the BDM mechanism has been more popular in field settings than the Vickrey auction. However, given the limitations of the BDM mechanism discussed above, a 2nd price Vickrey auction could be conducted even with just two subjects in a group. Or, a group can be formed later, with payments and product delivery occurring at a later time. This was the approach used in a few previous studies (e.g., List and Lucking-Reiley, 2000). Research exploiting field-relevant variables in field settings to test hypotheses of interest is also lacking. For example, the natural variation of experience between dealers and non-dealers of sports-cards has been a key manipulating factor in early field valuation experiments (List and Lucking-Reiley, 2000; List and Shogren, 1998). A field-setting of particular interest to food and agricultural economists that remains largely unexploited is farmers markets (e.g., Toler et al., 2009).

Given the growth in the use of experimental auctions, one might suspect that all the low-hanging fruit (research-wise) has been picked. We are more optimistic about the possibilities for new discoveries. What remains to be done? Although mentioned in Lusk and Shogren (2007), there remains a dearth of studies showcasing the external validity of auction studies. To our knowledge, there has been no comparison of experimental auction behavior with scanner data, for example. There have been a number of such comparisons with other methods —like choice experiments (e.g., Brooks and Lusk, 2010; Chang et al., 2009; Lusk et al., 2006) —but not with auctions. That is, we need studies that compare auction generated data with real-world purchases. If auctions studies are shown to have good external validity, then this will boost the confidence of researchers in the experimental auctions field in promoting this value elicitation tool in academic and business circles.

There is also much to learn about how values in auctions are influenced by social networks. Demont et al. (2013) and Richards et al. (2014) both show that peoples values are significantly influenced by others' values. Understanding how beliefs and new information filter through such networks is key to understanding acceptance of technology, effects of media scares, or the success of advertising or new product introduction.

An avenue that has not been significantly explored yet in the experimental auctions literature is a road already taken by other subfields in economics; that is, the possibility of massively increasing sample sizes by doing more experiments online.[23] Auctions impose an additional challenge because of the simultaneity nature of the submission of bids. However, the interactive

[23] See also Katkar and Reiley (2007); Lucking-Reiley et al. (2007) for early studies on auctions using data from eBay and Augenblick (2016) for a more recent study focused on penny auctions.

nature of experiments is not totally impossible to overcome (e.g., Arechar et al., 2018). Furthermore, studies now consistently show that data obtained online (e.g., via Amazon's Mechanical Turk; but see Dreyfyss (2018) for a warning and some remedies) is not significantly harmed by the lack of control over the conditions under which the responses are recorded (Johnson and Ryan, 2018).

A concern for experimental auctions conducted for food items is the low-value of the good. Subjects with low induced values tend to bid further away from their induced value (e.g., Drichoutis et al., 2015) as compared to when high induced values are assigned to them, and in mechanisms like the second price auction, incentives for truthful value revelation are weaker for lower-value subjects (Lusk et al., 2007). This would imply that auctions with low value items are more likely to suffer from measurement error, which could be partly due to the lower cost of misbehaving for low value items. How can we increase the rather weak incentives for accurate preference revelation in experimental auctions? Cason and Plott (2014) show that although many subjects did not bid their induced value, they did state their correct valuation in a second round of bidding after they were exposed to their mistake by rereading the instructions and after receiving feedback. This is consistent with the findings in Malone and Lusk (2018) where in a discrete choice experiment they provide feedback to inattentive respondents which are subsequently given the opportunity to re-answer a 'trap question' that checks for attentiveness. In Malone and Lusk (2018) individuals who do not correctly revise their responses after missing a trap question have significantly different choice patterns than individuals who correctly answer the trap question. Therefore nudging individuals toward their true preference could be a way forward for induced value auctions. However, in homegrown value auctions we do not know what a true preference is. In this case, tools from the contingent valuation literature could be tested as for example, cheap talk scripts, budget constraint reminders and consequentiality scripts (see Drichoutis et al., 2017, for details on such scripts). Given that we do not know yet the effectiveness of these tools in experimental auctions, some proper evaluation of their effectiveness is another area of future research that would be worth exploring. There may also be alternative mechanisms that can further sharpen the gradient between participants bid-space and their payoff-space.

Experimental auctions allow us to elicit willingness-to-pay or other valuation measures which provide a mapping of preferences on the monetary space. A key assumption of classical economic analysis is that of stability of preferences. In economic analysis, individual preferences are considered to be stable over time. Andersen et al. (2008) argue that the assumption of stable preferences lies in the ability to assign causation between changing opportunity sets and choices in comparative statics exercises or, in Stigler and Becker's (1977) words, "no significant behavior has been illuminated by assumptions of differences in tastes". If preferences are volatile with respect to the passage of time, then Harrison et al. (2005) note that researchers and policymakers using out-of-sample predictions should worry about their conclusions. However, we know little about the individual and aggregate stability of WTP values over time (although see Lusk (2017) for comparison of repeated choice experiments over time or studies like Dillaway et al. (2011) or Shogren et al. (2000) that conducted experimental auctions with the same participants at different points in time). Prospective studies that would repeatedly elicit consumer valuations over time for a range of products using experimental auctions could provide some feedback (especially if compared with other elicitation mechanisms) about the appropriateness of auctions in value elicitation as a tool that satisfies the assumptions of economic theory.

Finally, the rise of the behavioral economics literature clearly shows that decision making errors and biases can be identified in experimental auction settings. The much more difficult issue is what we do with these biases. One stream of research has sought to develop methods or techniques to eliminate the biases, also referred to as 'debiasing' (e.g., Cherry et al., 2003; Kovalsky and Lusk, 2013; Shogren, 2006) with the presumption that more stable, well-informed, market-disciplined preferences are most suitable for cost-benefit analysis. However, many of the products we are interested in valuing are new or non-market goods (if they were traditional market goods, we could use conventional demand estimation methods applied to the revealed preference data). Understanding the process by which people learn and update their preferences for these novel products, even if they are unstable, is important. The findings of studies that examined behavioral biases have also undermined some of the conceptual foundations behind welfare economics (Just, 2017; Lusk, 2014), but have also raised interesting, researchable questions about how people might act on others' behavioral biases (e.g., Lusk et al., 2014) or respond to paternalism from others (Debnam, 2017; Just and Hanks, 2015).

References

Abbink, K. and H. Hennig-Schmidt (2006). Neutral versus loaded instructions in a bribery experiment. *Experimental Economics 9*(2), 103–121.

Abeler, J., A. Falk, L. Goette, and D. Huffman (2011). Reference points and effort provision. *American Economic Review 101*(2), 470–92.

Akaichi, F., J. R. M. Nayga, and L. L. Nalley (2017). Are there trade-offs in valuation with respect to greenhouse gas emissions, origin and food miles attributes? *European Review of Agricultural Economics 44*(1), 3–31.

Alfnes, F. (2009). Valuing product attributes in vickrey auctions when market substitutes are available. *European Review of Agricultural Economics 36*(2), 133–149.

Alfnes, F., A. G. Guttormsen, G. Steine, and K. Kolstad (2006). Consumers' willingness to pay for the color of salmon: A choice experiment with real economic incentives. *American Journal of Agricultural Economics 88*(4), 1050–1061. 10.1111/j.1467-8276.2006.00915.x.

Alfnes, F. and K. Rickertsen (2007). Extrapolating experimental-auction results using a stated choice survey. *European Review of Agricultural Economics 34*(3), 345–363.

Altman, D. G. (1985). Comparability of randomised groups. *Journal of the Royal Statistical Society. Series D (The Statistician) 34*(1), 125–136.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2006). Elicitation using multiple price list formats. *Experimental Economics 9*(4), 383–405.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2008). Lost in state space: Are preferences stable? *International Economic Review 49*(3), 1091–1112.

Anderson, M. L. and J. Magruder (2017). Split-sample strategies for avoiding false discoveries. *National Bureau of Economic Research Working Paper Series No. 23544*.

Andreoni, J., Y.-K. Che, and J. Kim (2007). Asymmetric information about rivals' types in standard auctions: An experiment. *Games and Economic Behavior 59*(2), 240–259.

Arechar, A. A., S. Gächter, and L. Molleman (2018). Conducting interactive experiments online. *Experimental Economics 21*(1), 99–131.

Ariely, D., G. Loewenstein, and D. Prelec (2003). "coherent arbitrariness": Stable demand curves without stable preferences. *Quarterly Journal of Economics 118*(1), 73–105. 10.1162/00335530360535153.

Augenblick, N. (2016). The sunk-cost fallacy in penny auctions. *The Review of Economic Studies 83*(1), 58–86. 10.1093/restud/rdv037.

Austin, P. C. (2009, November). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine 28*(25), 3083–3107.

Ausubel, L. M. (2004). An efficient ascending-bid auction for multiple objects. *American Economic Review 94*(5), 1452–1475.

Banerji, A. and N. Gupta (2014). Detection, identification, and estimation of loss aversion: Evidence from an auction experiment. *American Economic Journal: Microeconomics 6*(1), 91–133.

Bellemare, C., L. Bissonnette, and S. Kröger (2016, Nov). Simulating power of economic experiments: the powerbbk package. *Journal of the Economic Science Association 2*(2), 157–168.

Belot, M. and J. James (2014). A new perspective on the issue of selection bias in randomized controlled field experiments. *Economics Letters 124*(3), 326–328.

Belton, C. A. and R. Sugden (2018). Attention and novelty: An experimental investigation of order effects in multiple valuation tasks. *Journal of Economic Psychology 67*, 103–115.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E. J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson (2018). Redefine statistical significance. *Nature Human Behaviour 2*(1), 6–10.

Bi, X., L. House, Z. Gao, and F. Gmitter (2012). Sensory evaluation and experimental auctions: Measuring willingness to pay for specific sensory attributes. *American Journal of Agricultural Economics 94*(2), 562–568.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. t. Weel (2008). The economics and psychology of personality traits. *Journal of Human Resources 43*(4), 972–1059.

Briz, T., A. C. Drichoutis, and R. M. Nayga Jr (2017). Randomization to treatment failure in experimental auctions: The value of data from training rounds. *Journal of Behavioral and Experimental Economics 71*, 56–66.

Brodeur, A., N. Cook, and A. Heyes (2018). Methods matter: P-hacking and causal inference in economics. *IZA Discussion Paper No. 11796*.

Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics 8*(1), 1–32.

Brooks, K. and J. L. Lusk (2010). Stated and revealed preferences for organic and cloned milk: Combining choice experiment and scanner data. *American Journal of Agricultural Economics 92*(4), 1229–1241. 10.1093/ajae/aaq054.

Burke, W. J. (2009). Fitting and interpreting cragg's tobit alternative using stata. *Stata Journal 9*(4), 584–592.

Bushong, B., L. M. King, C. F. Camerer, and A. Rangel (2010). Pavlovian processes in consumer choice: The physical presence of a good increases willingness-to-pay. *American Economic Review 100*(4), 1556–71.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafö (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience 14*, 365.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu (2016). Evaluating replicability of laboratory experiments in economics. *Science 351*(6280), 1433–1436.

Camerer, C. F., A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikensten, L. Hummer, T. Imai, S. Isaksson, D. Manfredi, J. Rose, E.-J. Wagenmakers, and H. Wu (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour 2*(9), 637–644.

Capra, M. C. (2004). Mood-driven behavior in strategic interactions. *American Economic Review 94*(2), 367–372.

Capra, M. C., S. Meer, and K. Lanier (2010). The effects of induced mood on bidding in random nth-price auctions. *Journal of Economic Behavior & Organization 75*(2), 223–234.

Casari, M., J. C. Ham, and J. H. Kagel (2007). Selection bias, demographic effects, and ability effects in common value auction experiments. *American Economic Review 97*(4), 1278–1304.

Cason, T. N. and C. R. Plott (2014). Misconceptions and game form recognition: Challenges to theories of revealed preference and framing. *Journal of Political Economy 122*(6), 1235–1270.

Chambers, C. D. (2018). Introducing the transparency and openness promotion (top) guidelines and badges for open practices at cortex. *Cortex 106*, 316–318.

Chang, A. C. and P. Li (2017). A preanalysis plan to replicate sixty economics research papers that worked half of the time. *American Economic Review 107*(5), 60–64.

Chang, A. C. and P. Li (2018). Is economics research replicable? Sixty published papers from thirteen journals say "often not". *Critical Finance Review 7*.

Chang, J. B., J. L. Lusk, and F. B. Norwood (2009). How closely do hypothetical surveys and laboratory experiments predict field behavior? *American Journal of Agricultural Economics 91*(2), 518–534. 10.1111/j.1467-8276.2008.01242.x.

Chen, Y., P. Katuščák, and E. Ozdenoren (2009). Why can't a woman bid more like a man? *Games and Economic Behavior 77*(1), 181–213.

Cherry, T. L., T. D. Crocker, and J. F. Shogren (2003). Rationality spillovers. *Journal of Environmental Economics and Management 45*(1), 63–84.

Christensen, G. and E. Miguel (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature 56*(3), 920–80.

Clemens, M. A. (2015). The meaning of failed replications: A review and proposal. *Journal of Economic Surveys 31*(1), 326–342.

Cochran, W. G. and D. B. Rubin (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A 35*(4), 417–446.

Coffman, L. C. and M. Niederle (2015). Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives 29*(3), 81–98.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ, USA: Lawrence Erlbaum Associates.

Cooper, D. J. and H. Fang (2008). Understanding overbidding in second price auctions: An experimental study. *The Economic Journal 118*(532), 1572–1595.

Corgnet, B., R. Hernán-González, P. Kujal, and D. Porter (2014). The effect of earned versus house money on price bubble formation in experimental asset markets. *Review of Finance 19*(4), 1455–1488.

Corrigan, J. R., A. C. Drichoutis, J. L. Lusk, J. R.M. Nayga, and M. C. Rousu (2012). Repeated rounds with price feedback in experimental auction valuation: An adversarial collaboration. *American Journal of Agricultural Economics 94*(1), 97–115.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica 39*(5), 829–844.

Croson, R. (2005). The method of experimental economics. *International Negotiation 10*(1), 131–148.

Davis, L. R., B. P. Joyce, and M. R. Roelofs (2010). My money or yours: house money payment effects. *Experimental Economics 13*(2), 189–205.

Deaton, A. and N. Cartwright (2016). Understanding and misunderstanding randomized controlled trials. *National Bureau of Economic Research Working Paper No. 22595*.

Debnam, J. (2017). Selection effects and heterogeneous demand responses to the berkeley soda tax vote. *American Journal of Agricultural Economics 99*(5), 1172–1187. 10.1093/ajae/aax056.

Delgado, M. R., A. Schotter, E. Y. Ozbay, and E. A. Phelps (2008). Understanding overbidding: Using the neural circuitry of reward to design economic auctions. *Science 321*(5897), 1849–1852.

Demont, M., P. Rutsaert, M. Ndour, W. Verbeke, P. A. Seck, and E. Tollens (2013). Experimental auctions, collective induction and choice shift: willingness-to-pay for rice quality in Senegal. *European Review of Agricultural Economics 40*(2), 261–286.

Demont, M., E. Zossou, P. Rutsaert, M. Ndour, P. Van Mele, and W. Verbeke (2012). Consumer valuation of improved rice parboiling technologies in benin. *Food Quality and Preference 23*(1), 63–70.

Dewald, W. G., J. G. Thursby, and R. G. Anderson (1986). Replication in empirical economics: The journal of money, credit and banking project. *The American Economic Review 76*(4), 587–603.

Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data* (2nd ed.). New York, USA: Oxford University Press Inc.

Dillaway, R., K. D. Messer, J. C. Bernard, and H. M. Kaiser (2011). Do consumer responses to media food safety information last? *Applied Economic Perspectives and Policy 33*(3), 363–383. 10.1093/aepp/ppr019.

Dreyfyss, E. (2018, August 17). A bot panic hits Amazon's Mechanical Turk. Wired. Available at https://perma.cc/42U2-AQ58. Last accessed on September 9, 2018.

Drichoutis, A. C., S. Klonaris, and G. S. Papoutsi (2017). Do good things come in small packages? bottle size effects on willingness to pay for pomegranate wine and grape wine. *Journal of Wine Economics 12*(1), 84–104.

Drichoutis, A. C. and J. L. Lusk (2014). Judging statistical models of individual decision making under risk using in- and out-of-sample criteria. *PLoS ONE 9*(7), e102269.

Drichoutis, A. C. and J. L. Lusk (2016). What can multiple price lists really tell us about risk preferences? *Journal of Risk and Uncertainty 53*(2), 89–106.

Drichoutis, A. C., J. L. Lusk, and R. M. Nayga (2015). The veil of experimental currency units in second price auctions. *Journal of the Economic Science Association 1*(2), 182–196.

Drichoutis, A. C., A. Vassilopoulos, and J. L. Lusk (2014). Consumers' willingness to pay for agricultural products certified to ensure fair working conditions. Report to the John S. Latsis Public Benefit Foundation. Available at https://perma.cc/LFP7-XBJM. Last accessed on December 8, 2017.

Drichoutis, A. C., A. Vassilopoulos, J. L. Lusk, and J. R. M. Nayga (2017). Consumer preferences for fair labour certification. *European Review of Agricultural Economics 44*(3), 455–474.

Duval, S. and R. Tweedie (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics 56*(2), 455–463.

Dyer, D. and J. H. Kagel (1996). Bidding in common value auctions: How the commercial construction industry corrects for the winner's curse. *Management Science 42*(10), 1463–1475.

Egger, M., G. D. Smith, M. Schneider, and C. Minder (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ 315*(7109), 629–634.

Ellison, B., J. L. Lusk, and D. Davis (2014). The impact of restaurant calorie labels on food choice: Results from a field experiment. *Economic Inquiry 52*(2), 666–681.

Erdem, T. (1996). A dynamic analysis of market structure based on panel data. *Marketing Science 15*(4), 359–378.

Fafchamps, M. and J. Labonne (2016). Using split samples to improve inference about causal effects. *National Bureau of Economic Research Working Paper Series No. 21842.*

Fanelli, D. and J. P. A. Ioannidis (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences*.

Feuz, D. M., W. J. Umberger, C. R. Calkins, and B. Sitz (2004). U.S. consumers' willingness to pay for flavor and tenderness in steaks as determined with an experimental auction. *Journal of Agricultural and Resource Economics 29*(3), 501–516.

Fidler, F., N. Thomason, G. Cumming, S. Finch, and J. Leeman (2004). Editors can lead researchers to confidence intervals, but can't make them think:statistical reform lessons from medicine. *Psychological Science 15*(2), 119–126.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Flynn, N., C. Kah, and R. Kerschbamer (2016). Vickrey auction vs bdm: difference in bidding behaviour and the impact of other-regarding motives. *Journal of the Economic Science Association 2*(2), 101–108.

Fox, J. A., J. F. Shogren, D. J. Hayes, and J. B. Kliebenstein (1998). Cvm-x: Calibrating contingent values with experimental auction markets. *American Journal of Agricultural Economics 80*(3), 455–465.

Gelman, A. (2018). Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery In Press.*

Gelman, A. and J. Carlin (2014). Beyond power calculations:assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science 9*(6), 641–651.

Georganas, S., D. Levin, and P. McGee (2017). Optimistic irrationality and overbidding in private value auctions. *Experimental Economics 20*(4), 772–792.

Gigerenzer, G., S. Krauss, and O. Vitouch (2004). *The Null Ritual: What You Always Wanted to Know About Significance Testing but Were Afraid to Ask*, pp. 391–408. Th ousand Oaks, CA: SAGE Publications, Inc.

Gill, J. (2018). Comments from the new editor. *Political Analysis 26*(1), 1–2.

Gneezy, A. (2016). Field experimentation in marketing research. *Journal of Marketing Research 54*(1), 140–143.

Gracia, A., M. L. Loureiro, and J. R. M. Nayga (2011). Are valuations from nonhypothetical choice experiments different from those of experimental auctions? *American Journal of Agricultural Economics 93*(5), 1358–1373.

Grebitus, C., J. L. Lusk, and R. M. Nayga (2013). Explaining differences in real and hypothetical experimental auctions and choice experiments with personality. *Journal of Economic Psychology 36*(Supplement C), 11–26.

Grether, D. M., C. R. Plott, D. B. Rowe, M. Sereno, and J. M. Allman (2007). Mental processes and strategic equilibration: An fmri study of selling strategies in second price auctions. *Experimental Economics 10*(2), 105–122.

Ham, J. C. and J. H. Kagel (2006). Gender effects in private value auctions. *Economics Letters 92*(3), 375–382.

Hankins, M. (2013, April 21). Still not significant. *Probable Error.* Available at http://perma.cc/Z6B9-PHCV. Last accessed on August 6, 2018.

Harrison, G. W. (1989). Theory and misbehavior of first-price auctions. *The American Economic Review 79*(4), 749–762.

Harrison, G. W. (1992). Theory and misbehavior of first-price auctions: Reply. *The American Economic Review 82*(5), 1426–1443.

Harrison, G. W. (2006). Experimental evidence on alternative environmental valuation methods. *Environmental & Resource Economics 36*, 125–162.

Harrison, G. W., E. Johnson, M. M. McInnes, and E. E. Rutström (2005). Temporal stability of estimates of risk aversion. *Applied Financial Economics Letters 1*(1), 31–35.

Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic Literature 42*(4), 1009–1055.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics 6*(2), 107–128.

Ho, D. E., K. Imai, G. King, and E. A. Stuart (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis 15*(3), 199–236.

Hoenig, J. M. and D. M. Heisey (2001). The abuse of power. *The American Statistician 55*(1), 19–24.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Holmquist, C., J. McCluskey, and C. Ross (2012). Consumer preferences and willingness to pay for oak attributes in washington chardonnays. *American Journal of Agricultural Economics 94*(2), 556–561.

Horowitz, J. K. (2006). The becker-degroot-marschak mechanism is not necessarily incentive compatible, even for non-random goods. *Economics Letters 93*(1), 6–11.

Imai, K., G. King, and E. A. Stuart (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 171*(2), 481–502.

Imbens, G. W. and D. B. Rubin (2016). *Causal Inference for Statistics, Social, and Biomedical Sciences, An introduction.* Cambridge and New York: Cambridge University Press.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 5–86.

Ioannidis, J. P. A. and T. A. Trikalinos (2007). An exploratory test for an excess of significant findings. *Clinical Trials 4*(3), 245–253.

Jacquemet, N., R.-V. Joule, S. Luchini, and J. F. Shogren (2009). Earned wealth, engaged bidders? Evidence from a second-price auction. *Economics Letters 105*(1), 36–38.

Johnson, D. and J. Ryan (2018). Amazon mechanical turk workers can provide consistent and economically meaningful data. *Munich Personal RePEc Archive No. 88450*.

Johnson, E. J., M. Steffel, and D. G. Goldstein (2005). Making better decisions: From measuring to constructing preferences. *Health Psychology 24*(4), S17–S22.

Josephson, A. and J. D. Michler (2018). Viewpoint: Beasts of the field? ethics in agricultural and applied economics. *Food Policy 79*, 1–11.

Just, D. R. (2017). The behavioral welfare paradox: Practical, ethical and welfare implications of nudging. *Agricultural and Resource Economics Review 46*(1), 1–20.

Just, D. R. and A. S. Hanks (2015). The hidden cost of regulation: Emotional responses to command and control. *American Journal of Agricultural Economics 97*(5), 1385–1399. 10.1093/ajae/aav016.

Kagel, J. H., R. M. Harstad, and D. Levin (1987). Information impact and allocation rules in auctions with affiliated private values: A laboratory study. *Econometrica 55*(6), 1275–1304.

Kagel, J. H. and D. Levin (1993). Independent private value auctions: Bidder behaviour in first-, second- and third-price auctions with varying numbers of bidders. *The Economic Journal 103*(419), 868–879.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review 67*(3), 160–167.

Kang, M. and C. Camerer (2013). fmri evidence of a hot-cold empathy gap in hypothetical and real aversive choices. *Frontiers in Neuroscience 7*(104).

Karni, E. and Z. Safra (1987). "preference reversal" and the observability of preferences by experimental methods. *Econometrica 55*(3), 675–685.

Katkar, R. and D. H. Reiley (2007). Public versus secret reserve prices in ebay auctions: Results from a pokemon field experiment. *The B.E. Journal of Economic Analysis & Policy 5*(2).

Kechagia, V. and A. C. Drichoutis (2017). The effect of olfactory sensory cues on willingness to pay and choice under risk. *Journal of Behavioral and Experimental Economics 70*, 33–46.

Kennedy, P. E. (2002). Sinning in the basement: What are the rules? The ten commandments of applied econometrics. *Journal of Economic Surveys 16*(4), 569–589.

Kenny, D. A. (1987). *Chapter 13: The two-group design*, pp. 203–223. Little, Brown.

Keser, C., K.-M. Ehrhart, and S. K. Berninghaus (1998). Coordination and local interaction: experimental evidence. *Economics Letters 58*(3), 269–275.

Kessler, J. B. and S. Meier (2014). Learning from (failed) replications: Cognitive load manipulations and charitable giving. *Journal of Economic Behavior & Organization 102*(0), 10–13.

Kimball, A. W. (1957). Errors of the third kind in statistical consulting. *Journal of the American Statistical Association 52*(278), 133–142.

Klain, T. J., J. L. Lusk, G. T. Tonsor, and T. C. Schroeder (2014). An experimental approach to valuing information. *Agricultural Economics 45*(5), 635–648.

Kline, R. B. (2013). *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences.* American Psychological Association.

Kovalsky, K. L. and J. L. Lusk (2013). Do consumers really know how much they are willing to pay? *Journal of Consumer Affairs 47*(1), 98–127.

Kupper, L. L. and K. B. Hafner (1989). How appropriate are popular sample size formulas? *The American Statistician 43*(2), 101–105.

Lakens, D., F. G. Adolfi, C. J. Albers, F. Anvari, M. A. J. Apps, S. E. Argamon, T. Baguley, R. B. Becker, S. D. Benning, D. E. Bradford, E. M. Buchanan, A. R. Caldwell, B. Van Calster, R. Carlsson, S.-C. Chen, B. Chung, L. J. Colling, G. S. Collins, Z. Crook, E. S. Cross, S. Daniels, H. Danielsson, L. DeBruine, D. J. Dunleavy, B. D. Earp, M. I. Feist, J. D. Ferrell, J. G. Field, N. W. Fox, A. Friesen, C. Gomes, M. Gonzalez-Marquez, J. A. Grange, A. P. Grieve, R. Guggenberger, J. Grist, A.-L. van Harmelen, F. Hasselman, K. D. Hochard, M. R. Hoffarth, N. P. Holmes, M. Ingre, P. M. Isager, H. K. Isotalus, C. Johansson, K. Juszczyk, D. A. Kenny, A. A. Khalil, B. Konat, J. Lao, E. G. Larsen, G. M. A. Lodder, J. Lukavský, C. R. Madan, D. Manheim, S. R. Martin, A. E. Martin, D. G. Mayo, R. J. McCarthy, K. McConway, C. McFarland, A. Q. X. Nio, G. Nilsonne, C. L. de Oliveira, J.-J. O. de Xivry, S. Parsons, G. Pfuhl, K. A. Quinn, J. J. Sakon, S. A. Saribay, I. K. Schneider, M. Selvaraju, Z. Sjoerds, S. G. Smith, T. Smits, J. R. Spies, V. Sreekumar, C. N. Steltenpohl, N. Stenhouse, W. Świątkowski, M. A. Vadillo, M. A. L. M. Van Assen, M. N. Williams, S. E. Williams, D. R. Williams, T. Yarkoni, I. Ziano, and R. A. Zwaan (2018). Justify your alpha. *Nature Human Behaviour 2*(3), 168–171.

Lane, D. M. and W. P. Dunlap (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology 31*(2), 107–112.

Lee, J. Y. and J. A. Fox (2015). Bidding behavior in experimental auctions with positive and negative values. *Economics Letters 136*, 151–153.

Lee, J. Y., R. M. J. Nayga, C. Deck, and A. C. Drichoutis (2017). Cognitive ability and bidding behavior in second price auctions: An experimental study. *Munich Personal RePEc Archive No. 81495*.

Lehner, R., J. H. Balsters, A. Herger, T. A. Hare, and N. Wenderoth (2017). Monetary, food, and social rewards induce similar pavlovian-to-instrumental transfer effects. *Frontiers in Behavioral Neuroscience 10*(247).

Lerner, J., D. Small, and G. Loewenstein (2004). Heart strings and purse strings: Carryover effects of emotions on economic decisions. *Psychological Science 15*(5), 337–341.

Lewis, K. E., C. Grebitus, and R. M. Nayga (2016a). The impact of brand and attention on consumers' willingness to pay: Evidence from an eye tracking experiment. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie 64*(4), 753–777.

Lewis, K. E., C. Grebitus, and R. M. Nayga (2016b). The importance of taste in experimental auctions: consumers' valuation of calorie and sweetener labeling of soft drinks. *Agricultural Economics 47*(1), 47–57.

Li, S. (2017). Obviously strategy-proof mechanisms. *American Economic Review 107*(11), 3257–3287.

Lichtenstein, S. and P. Slovic (2006). *The construction of preference.* Cambridge, UK: Cambridge University Press.

Linder, N. S., G. Uhl, K. Fliessbach, P. Trautner, C. E. Elger, and B. Weber (2010). Organic labeling influences food valuation and choice. *NeuroImage 53*(1), 215–220.

List, J. A. (2003). Does market experience eliminate market anomalies?*. *The Quarterly Journal of Economics 118*(1), 41–71.

List, J. A. (2004). Neoclassical theory versus prospect theory: Evidence from the marketplace. *Econometrica 72*(2), 615–625.

List, J. A. and D. Lucking-Reiley (2000). Demand reduction in multiunit auctions: Evidence from a sportscard field experiment. *The American Economic Review 90*(4), 961–972.

List, J. A., S. Sadoff, and M. Wagner (2011). So you want to run an experiment, now what? some simple rules of thumb for optimal experimental design. *Experimental Economics 14*(4), 439.

List, J. A. and J. F. Shogren (1998). Calibration of the difference between actual and hypothetical valuations in a field experiment. *Journal of Economic Behavior & Organization 37*(2), 193–205.

Liu, H. and T. Wu (2005). Sample size calculation and power analysis of time-averaged difference. *Journal of Modern Applied Statistical Methods 4*(2), 434–445.

Lucking-Reiley, D., D. Bryan, N. Prasad, and D. Reeves (2007). Pennies from ebay: The determinants of price in online auctions*. *The Journal of Industrial Economics 55*(2), 223–233.

Lusk, J. (2010). Experimental auction markets for studying consumer preferences. In S. R. Jaeger and H. MacFie (Eds.), *Consumer-Driven Innovation in Food and Personal Care Products*, Woodhead Publishing Series in Food Science, Technology and Nutrition, pp. 332 – 357. Woodhead Publishing.

Lusk, J. L. (2014). Are you smart enough to know what to eat? a critique of behavioural economics as justification for regulation. *European Review of Agricultural Economics 41*(3), 355–373. 10.1093/erae/jbu019.

Lusk, J. L. (2017). Consumer research with big data: Applications from the food demand survey (foods). *American Journal of Agricultural Economics 99*(2), 303–320. 10.1093/ajae/aaw110.

Lusk, J. L., C. Alexander, and M. C. Rousu (2007). Designing experimental auctions for marketing research: The effect of values, distributions, and mechanisms on incentives for truthful bidding. *Review of Marketing Science 5*(1).

Lusk, J. L., M. S. Daniel, D. R. Mark, and C. L. Lusk (2001). Alternative calibration and auction institutions for predicting consumer willingness to pay for nongenetically modified corn chips. *Journal of Agricultural and Resource Economics 26*(1), 40–57.

Lusk, J. L. and J. A. Fox (2002). Consumer demand for mandatory labeling of beef from cattle administered growth hormones or fed genetically modified corn. *Journal of Agricultural and Applied Economics 34*(1), 27–38.

Lusk, J. L. and J. A. Fox (2003). Value elicitation in retail and laboratory environments. *Economics Letters 79*(1), 27–34.

Lusk, J. L., S. Marette, and F. B. Norwood (2014). The paternalist meets his match. *Applied Economic Perspectives and Policy 36*(1), 61–108. 10.1093/aepp/ppt031.

Lusk, J. L., F. B. Norwood, and J. R. Pruitt (2006). Consumer demand for a ban on antibiotic drug use in pork production. *American Journal of Agricultural Economics 88*(4), 1015–1033.

Lusk, J. L., J. R. Pruitt, and B. Norwood (2006). External validity of a framed field experiment. *Economics Letters 93*(2), 285–290.

Lusk, J. L. and T. C. Schroeder (2004). Are choice experiments incentive compatible? a test with quality differentiated beef steaks. *American Journal of Agricultural Economics 86*(2), 467–482. 10.1111/j.0092-5853.2004.00592.x.

Lusk, J. L. and T. C. Schroeder (2006). Auction bids and shopping choices. *Advances in Economic Analysis & Policy 6*(1).

Lusk, J. L. and J. F. Shogren (2007). *Experimental auctions, Methods and applications in economic and marketing research.* Cambridge, UK: Cambridge University Press.

Malone, T. and J. L. Lusk (2018). Releasing the trap: A method to reduce inattention bias in survey data with application to u.s. beer taxes. *Economic Inquiry (forthcoming).*

Maniadis, Z., F. Tufano, and J. A. List (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review 104*(1), 277–90.

Mazar, N., B. Koszegi, and D. Ariely (2013). True context-dependent preferences? the causes of market-dependent valuations. *Journal of Behavioral Decision Making*.

McCullough, B., K. A. McGeary, and T. D. Harrison (2008). Do economics journal archives promote replicable research? *Canadian Journal of Economics/Revue Canadienne d'Economique 41*(4), 1406–1420.

McCullough, B. D., K. A. McGeary, and T. D. Harrison (2006). Lessons from the jmcb archive. *Journal of Money, Credit and Banking 38*(4), 1093–1107.

Melton, B. E., W. E. Huffman, J. F. Shogren, and J. A. Fox (1996). Consumer preferences for fresh food items with multiple quality attributes: Evidence from an experimental auction of pork chops. *American Journal of Agricultural Economics 78*(4), 916–923.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, and M. Van der Laan (2014). Promoting transparency in social science research. *Science 343*(6166), 30–31.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gotzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ 340*.

Morgan, J., K. Steiglitz, and G. Reis (2003). The spite motive and equilibrium behavior in auctions. *Contributions in Economic Analysis & Policy 2*(1).

Muller, L., L. Anne, L. J. L., and R. Bernard (2017). Distributional impacts of fat taxes and thin subsidies. *The Economic Journal 127*(604), 2066–2092.

Norwood, B. F., M. C. Roberts, and J. L. Lusk (2004). Ranking crop yield models using out-of-sample likelihood functions. *American Journal of Agricultural Economics 86*(4), 1032–1043.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. L. Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni (2015). Promoting an open research culture. *Science 348*(6242), 1422–1425.

Nuzzo, R. (2014). Scientific method: Statistical errors - p values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature 506*, 150–152.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives 29*(3), 61–80.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science 349*(6251).

Orquin, J. L. and S. Mueller Loose (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica 144*(1), 190–206.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics 8*(2), 157–159.

Parkhurst, G. M., J. F. Shogren, and D. L. Dickinson (2004). Negative values in vickrey auctions. *American Journal of Agricultural Economics 86*(1), 222–235.

Payne, J. W., J. R. Bettman, and D. A. Schkade (1999). Measuring constructed preferences: Towards a building code. *Journal of Risk and Uncertainty 19*(1), 243–270.

Pearson, M. and B. C. Schipper (2013). Menstrual cycle and competitive bidding. *Games and Economic Behavior 78*(forthcoming), 1–20.

Plassmann, H., J. O'Doherty, and A. Rangel (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *The Journal of Neuroscience 27*(37), 9984–9988.

Platter, W. J., J. D. Tatum, K. E. Belk, S. R. Koontz, P. L. Chapman, and G. C. Smith (2005). Effects of marbling and shear force on consumers' willingness to pay for beef strip loin steaks. *Journal of Animal Science 83*(4), 890–899.

Pritschet, L., D. Powell, and Z. Horne (2016). Marginally significant effects as evidence for hypotheses:changing attitudes over four decades. *Psychological Science 27*(7), 1036–1042.

Richards, T. J., S. F. Hamilton, and W. J. Allender (2014). Social networks and new product choice. *American Journal of Agricultural Economics 96*(2), 489–516. 10.1093/ajae/aat116.

Rihn, A. L. and C. Yue (2016). Visual attention's influence on consumers' willingness-to-pay for processed food products. *Agribusiness 32*(3), 314–328.

Roe, B. E. and D. R. Just (2009). Internal and external validity in economics research: Tradeoffs between experiments, field experiments, natural experiments, and field data. *American Journal of Agricultural Economics 91*(5), 1266–1271.

Roider, A. and P. W. Schmitz (2012). Auctions with anticipated emotions: Overbidding, underbidding, and optimal reserve prices*. *The Scandinavian Journal of Economics 114*(3), 808–830.

Rosato, A. and A. A. Tymula (2016). Loss aversion and competition in Vickrey auctions: Money ain't no good. *The University of Sydney, Economics Working Paper Series 2016 - 09*.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.

Rosenboim, M. and T. Shavit (2012). Whose money is it anyway? using prepaid incentives in experimental economics to create a natural environment. *Experimental Economics 15*(1), 145–157.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin 86*(3), 638–641.

Rowe, D. B. (2001). Bayesian source separation for reference function determination in fmri. *Magnetic Resonance in Medicine 46*(2), 374–378.

Roy, R., P. K. Chintagunta, and S. Haldar (1996). A framework for investigating habits, "the hand of the past," and heterogeneity in dynamic brand choice. *Marketing Science 15*(3), 280–299.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Rubin, D. B. (1990). [On the application of probability theory to agricultural experiments. Essay on principles. Section 9.] Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science 5*(4), 472–480.

Schipper, B. C. (2015). Sex hormones and competitive bidding. *Management Science 61*(2), 249–486.

Selya, A. S., J. S. Rose, L. C. Dierker, D. Hedeker, and R. J. Mermelstein (2012). A practical guide to calculating Cohen's f^2, a measure of local effect size, from PROC MIXED. *Frontiers in Psychology 3*, 111.

Senn, S. (1994). Testing for baseline balance in clinical trials. *Statistics in Medicine 13*(17), 1715–1726.

Senn, S. (2013). Seven myths of randomisation in clinical trials. *Statistics in Medicine 32*(9), 1439–1450.

Shogren, J. F. (2006). Valuation in the lab. *Environmental & Resource Economics 34*(163-172).

Shogren, J. F., J. A. List, and D. J. Hayes (2000). Preference learning in consecutive experimental auctions. *American Journal of Agricultural Economics 82*(4), 1016–1021. 10.1111/0002-9092.00099.

Shogren, J. F., M. Margolis, C. Koo, and J. A. List (2001). A random nth-price auction. *Journal of Economic Behavior and Organization 46*(4), 409–421.

Simmons, J., L. Nelson, and U. Simonsohn (2012). A 21 word solution. *Dialogue: The Official Newsletter of the Society for Personality and Social Psychology 26*(2), 4–7.

Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science 22*(11), 1359–1366.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science 26*(5), 559–569.

Simonsohn, U. (2018, October 17). Re: Preregistration for lab experiment? Economic Science Association Experimental Methods Discussion Google group, Available at https://goo.gl/FQ5W2U.

Simonsohn, U., L. D. Nelson, and J. P. Simmons (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General 143*(2), 534–547.

Simonsohn, U., J. P. Simmons, and L. D. Nelson (2013). Anchoring is not a false-positive: Maniadis, tufano, and list's (2014) 'failure-to-replicate' is actually entirely consistent with the original. *Working paper, Available at http://dx.doi.org/10.2139/ssrn.2351926*.

Slovic, P. (1995). The construction of preference. *American Psychologist 50*(5), 364–371.

Sousa, Y. F. D. and A. Munro (2012). Truck, barter and exchange versus the endowment effect: Virtual field experiments in an online game environment. *Journal of Economic Psychology 33*(3), 482–493.

Speed, T. P. (1990). Introductory remarks on Neyman (1923). *Statistical Science 5*(4), 463–464.

Splawa-Neyman, J., D. M. Dabrowska, and T. P. Speed (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472.

Stigler, G. J. and G. S. Becker (1977). De Gustibus Non Est Disputandum. *The American Economic Review 67*(2), 76–90.

Toler, S., B. C. Briggeman, J. L. Lusk, and D. C. Adams (2009). Fairness, farmers markets, and local production. *American Journal of Agricultural Economics 91*(5), 1272–1278.

Trafimow, D. and M. Marks (2015). Editorial. *Basic and Applied Social Psychology 37*(1), 1–2.

Tyson-Carr, J., K. Kokmotou, V. Soto, S. Cook, N. Fallon, T. Giesbrecht, and A. Stancak (2018). Neural correlates of economic value and valuation context: an event-related potential study. *Journal of Neurophysiology 119*(5), 1924–1933.

Umberger, W. J. and D. M. Feuz (2004). The usefulness of experimental auctions in determining consumers' willingness-to-pay for quality-differentiated products. *Applied Economic Perspectives and Policy 26*(2), 170–185.

Umberger, W. J., D. M. Feuz, C. R. Calkins, and K. Killinger-Mann (2002). U.s. consumer preference and willingness-to-pay for domestic corn-fed beef versus international grass-fed beef measured through an experimental auction. *Agribusiness 18*(4), 491–504.

Urbancic, M. (2011). Testing distributional dependence in the becker-degroot-marschak mechanism. *Working paper, UC Berkeley*.

Vassilopoulos, A., A. C. Drichoutis, and R. M. Nayga Jr (2018). Loss aversion, expectations and anchoring in the bdm mechanism. *Munich Personal RePEc Archive No. 85635*.

Vecchio, R. and M. Borrello (2018). Measuring food preferences through experimental auctions: A review. *Food Research International*.

Veling, H., Z. Chen, M. C. Tombrock, I. A. M. Verpaalen, L. I. Schmitz, A. Dijksterhuis, and R. W. Holland (2017). Training impulsive choices for healthy and sustainable food. *Journal of Experimental Psychology: Applied 23*(2), 204–215.

Wasserstein, R. L. and N. A. Lazar (2016). The asa's statement on p-values: Context, process, and purpose. *The American Statistician 70*(2), 129–133.

Wedel, M. and R. Pieters (2000). Eye fixations on advertisements and memory for brands: A model and findings. *Marketing Science 19*(4), 297–312.

West, S. G. and F. Thoemmes (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods 15*(1), 18–37.

Zhang, L. and A. Ortmann (2013). Exploring the meaning of significance in experimental economics. *UNSW Australian School of Business Research Paper No. 2013-32*.

Zhang, Y. Y., R. M. Nayga, and D. P. T. Depositario (2017). Learning and the possibility of losing own money reduce overbidding: Delayed payment in experimental auctions. *Social Scence Research Network Available at http://dx.doi.org/10.2139/ssrn.2636040.*

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

Electronic Supplementary Material of

How to run an experimental auction: A review of recent advances

Sample size calculations: Stata code

```
*==============================================*
*==== Sample size calculations for          ===*
*==== two-sided two-sample tests of means   ===*
*==== in a repeated measures design         ===*
*==============================================*

local za=1.96          // Type I error (a=5%)
local zb=0.8416        // Type II error (b=20%)
local M=7              // N of repeated measurements e.g., rounds

display "N M rho sigma diff"
foreach diff of numlist 0.5(1)2  {      // define minimum detectable difference
foreach sigma of numlist 1(0.5)1 {      // define values for sigma
foreach rho of numlist 0.5 0.8   {      // define values for rho
display in yellow "N        M          rho      sigma    diff"
display in green round(2*( ((`za'+`zb')^2)*(`sigma'^2)*(1+(`M'-1)*`rho'))/(`M'*(`diff'^2)),1) ///
" `M' `rho' `sigma' `diff'"
}
}
}
```

Cohen's f^2: Stata code

```
*========================*
* Create mock up dataset *
*========================*
clear
version 14

set obs 300
gen id=_n
set seed 36774
gen bid1=round(runiform()*10,0.01)
gen bid2=bid1+round(runiform()*2,0.01)
gen bid3=bid2+round(runiform(),0.01)
gen bid4=bid1+round(runiform()*4,0.01)
gen bid5=round(runiform()*12,0.01)
gen age=18+round(runiform()*25,1)
gen treatment=round(runiform()*1,1)

reshape long bid, i(id) j(round) // Reshape data in long form

*=======================*
* Calculate Cohen's f^2 *
*=======================*
* Get R-sq abc
quietly xtmixed bid i.treatment c.age c.round || id: // Estimate the model
scalar N=e(N)                                        // Store number of observations
scalar v=N-e(df_m)-1                                 // Degrees of freedom
scalar u=1                                           // Degrees of freedom for the error variance
mltrsq                                               // This will calculate R-squared after two-level mixed models
                                                     // Need to install the mlt package if have not been installed before
scalar r1abc=e(sb_rsq_l1)                            // Snijders/Bosker R-squared Level 1
scalar r2abc=e(sb_rsq_l2)                            // Snijders/Bosker R-squared Level 2

* Get R-sq ab
quietly xtmixed bid i.treatment c.age || id:
mltrsq
scalar r1ab=e(sb_rsq_l1)
scalar r2ab=e(sb_rsq_l2)
* Get R-sq First estimate R2a, then calculate f2 for each variable
 * for the treatment variable
   quietly xtmixed bid i.treatment || id:
   mltrsq
   scalar r1a=e(sb_rsq_l1)
   scalar r2a=e(sb_rsq_l2)
   do CI.do
 * for the age variable
```

```
   quietly xtmixed bid c.age || id:
   mltrsq
   scalar r1a=e(sb_rsq_l1)
   scalar r2a=e(sb_rsq_l2)
   do CI.do
```

The code below should be placed in a separate do file and saved with the name CI.do.

```
*==============================================*
*===== Calculate confidence intervals =========*
*==  (put this code in a separate file )   ==*
*== (and save it as CI.do)   ==*
*==============================================*
foreach x of numlist 1/2 {               // Loop over the Snijders/Bosker Level-1 and Level-2 R-squared

scalar f2`x'=(r`x'ab - r`x'a)/(1 - r`x'abc) // This is the f-squared

* First set the confidence level
scalar conf=0.95                        // e.g., 0.95 if 95% or 0.9 if 90%
scalar ulim=1-(1-conf)/2
scalar llim=(1-conf)/2
scalar fval=(f2`x'*v)/u                // This is the F-statistic: f^2*v/u
scalar lcp=0                           // Lower centrality parameter, lamda (note that lambda ranges from 0 to 10000)
scalar ucp=50                          // Upper limit of centrality parameter

scalar probl=nF(u,v,lcp,fval)          // Initial probability given an lcp=0
if probl<=ulim {
scalar lcp=0
}
else if probl>ulim{
while probl>ulim {
scalar lcp=lcp+0.0001                  // Iteratively try lcp values with a step of 0.0001
scalar probl=nF(u,v,lcp,fval)          // this produces the probability of F stats being lower
                                       // than fval given a non-centrality paramert lcp
}
}
scalar probu=nF(u,v,ucp,fval)          // Initial probability given an ucp=10000, lamda
while probu<=llim {
scalar ucp=ucp-0.0001                  // Iteratively try ucp values with a step of 0.0001
scalar probu=nF(u,v,ucp,fval)
}

scalar lf2`x'=lcp/(u+v+1)
scalar uf2`x'=ucp/(u+v+1)

* Output the results
di in green "LEVEL-" `x' " R-SQUARED: "  "lcp= " in yellow lcp in green " ucp= " ///
in yellow ucp in green " Prob for lower cp= " in yellow probl in green " Prob for upper cp=" ///
in yellow probu in green " Cohen's f2= " in yellow f2`x' in green " Lower CI=" in yellow lf2`x' ///
in green " Upper CI=" in yellow uf2`x'
}
```