Randomization to treatment failure in experimental auctions: The value of data from training rounds

Briz, Teresa and Drichoutis, Andreas C. and Nayga, Rodolfo M.

Universidad Politecnica de Madrid, Agricultural University of Athens, University of Arkansas

2 July 2014

Randomization to treatment failure in experimental auctions: The value of data from training rounds *

Teresa Briz[†1], Andreas C. Drichoutis[‡2], and Rodolfo M. Nayga, Jr.[§3,4,5,6]

[1]Universidad Politécnica de Madrid
[2]Agricultural University of Athens
[3]University of Arkansas
[4]Norwegian Institue of Bioeconomy Research
[5]Korea University
[6]The National Bureau of Economic Research

Abstract: In the experimental auctions literature, it is common practice to train subjects, who are often unfamiliar with the auction procedure, by conducting a few training (often hypothetical) auctions. Data from these practice auctions are rarely reported in scientific papers. We argue that valuable information can be garnered by looking at data coming from the training rounds of experimental auctions. As a case study, we use data from an experiment that seeks to elaborate on the mediating role of mood states on projection bias. Following a mood induction procedure, subjects are found to bid more under negative mood (as compared to positive mood) for products that are delivered in the future but bid less under negative mood for products that are delivered in present time. We show that if we had neglected insights gained from the training auction data, we would not have been able to detect a failure of randomization to treatment that rendered us biased estimates of the true causal effects due to unobserved heterogeneity.

Keywords: projection bias, mood induction, experimental auctions, randomization to treatment.

[*] We gratefully acknowledge the critical reviews by multiple referees on previous versions of our paper. Two reviewers of this journal made numerous suggestions that have improved and focused the article in substantial ways. The introduction and discussion sections of the paper draw from several excellent ideas provided by the referees.

[†] Departamento de Economía, E.T.S.Ingenieros Agrónomos, Universidad Politécnica de Madrid, Madrid, Spain, e-mail: teresa.briz@upm.es.

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

[§] Professor and Tyson Endowed Chair, Department of Agricultural Economics & Agribusiness, Division of Agriculture, University of Arkansas, Fayetteville, AR 72701, USA, and Adjunct Professor, Norwegian Institute of Bioeconomy Research, Korea University, and NBER, e-mail: rnayga@uark.edu.

JEL Classification Numbers: C12, C90.

1 Introduction

The experimental turn that followed mathematical formalization, economic modeling, and the emergence of econometrics, has been increasingly recognized as an equally important fourth transformative power in terms of its impact on economics (Svorenčík, 2015).[1]

In economic analysis, the '*ceteris paribus*' phrase is used whenever we want to isolate an effect from other influences, that is, whenever we want to give a causal interpretation of an effect. Experiments are being referred to as the golden standard for causal inference because through proper experimental designs, they allow holding all other factors constant, so that the change in the outcome of interest can be associated with changes in the manipulated factor. The essential features of experimental design —one of the fundamental aspects of the methods used in experimental economics —are often referred to as "…control and comparison, with *randomization* [emphasis added] an essential part of control" (Siegel, 1964). Randomization was popularized in Fisher's (1925) *Statistical methods for research workers*, but dates back even earlier to more than a century ago (Peirce and Jastrow, 1885). The statistical description and interpretation of a randomized experiment is now commonly called 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 Speed, 1990; Splawa-Neyman et al., 1990; Rubin, 1990) and Rubin (1974).

Experimental economics bases causal interpretation on the Neyman-Rubin model and the validity of its assumptions. The Neyman-Rubin model approach to causality, is explained 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 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). The problem, however, is that we cannot observe the same unit under both the control and the treatment condition. This

[1] This particular opinion expressed in the introduction of Svorenčík's (2015) thesis (his monograph provides the first account of the history of experimental economics), echoes the view of several experimental economists. For example, Plott (1991) describes the experimental turn as a revolution that would transform economics into an experimental science. Similarly, Guala (2010) describes the experimental turn as a methodological revolution towards a 'tool-based' science "…where models, statistics, and mathematics played the role both of instruments and, crucially, of objects of investigation" in which experiments subsequently entered the economists' toolkit.

would require to observe the outcome for the same unit in both alternative conditions. Therefore, one of the potential outcomes will always be missing which is widely known as the 'fundamental problem of causal inference'. In social sciences, unlike natural sciences, the unit changes irreversibly once it is exposed to a treatment and although '. . . we may have the same unit measured on both treatments in two trials [. . .] we cannot be certain that the unit's responses would be identical at both times' due to carryover or time trend effects (Rubin, 1974).

Recognizing the inability of the researcher to observe the counterfactual (if it was possible this would be the ideal experiment), the Neyman-Rubin model infers causality by comparing the expected outcome of units that received different treatments in order to estimate the treatment effect. By randomly assigning the treatments, the difference between the experimental and control units is an unbiased estimate of the causal effect that the researcher is trying to isolate (Rubin, 1974).[2] The comparison between two groups of units is valid, however, only if the two groups have the same expectation in the distribution of all covariates which are potentially relevant for the treatment outcome. The law of large numbers (LLN) ensures that by enlarging the sample size, two randomly assigned groups will indeed be comparable, since the sample average of individual characteristics in larger samples tends to become closer to the average of the population. However, the need of having a large sample so that randomization can reduce the differences between the treatment and control groups is something that experimental economists do not generally pay particular attention to. In fact, a statistically significant effect coming from a relatively small sample could be valued more than a statistically significant treatment effect coming from larger samples given that collecting more data could lead to any consistent statistical test rejecting the null as the number of observations goes to infinity.

Although by the LLN the more observations collected, the more likely it is that randomization will work, a practical decision has to be made as to how many observations would be needed for it to be considered a large enough sample. This is because studies with large samples may not be cost effective and are generally more difficult to carry out. Sample size calculations are increasingly being carried out in experimental studies (e.g., see Drichoutis et al., 2015, as a showcase in experimental auctions) but these are likely carried out as a

[2] Thus, one can think of the Neyman-Rubin model as an approximation to the ideal experiment. As one reviewer notes, the Neyman-Rubin model stipulates that we cannot obtain unconfounded responses under both treatment and control for the same individual; which is to say that a treatment effect in a within-subjects design does not have a causal interpretation under the Neyman-Rubin model. West and Thoemmes (2010) nicely summarize the additional assumptions needed in the context of the Neyman-Rubin model, given randomization, to provide a sufficient basis for an unbiased estimate of the magnitude of the causal effect. Holland (1986) discusses two additional approximations to the ideal design, namely the within-subjects design and unit homogeneity, and the assumptions behind these approximations (see also footnote 5 in West and Thoemmes, 2010).

protection against false negatives; that is, experimenters want to make sure they collect as many observations as needed to detect a given effect size with a pre-specified power; no less, no more than needed[3].

In this study we show that, in the context of experimental auctions, useful information can be conveyed by looking at the data coming from the training or practice rounds. Experimental auctions have become a popular tool for applied economists to elicit people's willingness to pay (WTP) values due to their demand revealing properties. These auctions are considered demand revealing because of the (theoretically) incentive compatible nature of the auction mechanisms. However, experimental auctions are often unfamiliar to subjects. Consequently, most practitioners agree that employing a training phase prior to the actual valuation task is essential for subjects to abandon market-like heuristics such as "buying low" or for demonstrating the incentive compatibility of the auction. This advice is echoed in Lusk and Shogren's (2007, pp. 62) experimental auctions book which largely reflects the established procedure of conducting experimental auctions.[4]

While preceding an actual auction with practice rounds is common, bids from practice rounds are rarely recorded. Corrigan et al. (2014) note that a well-known economist said this about his work on experimental auctions, "I pulled up three old data sets associated with various published papers. Alas, it seems I did not enter the practice round data (normally with candy bars) for any of them." Indeed, it is uncommon in this literature to pay particular attention to practice/training rounds. Only a handful of papers have done so. For example,

[3] All statistical hypotheses 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. 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). Type II errors, on the other hand, have not been given similar attention. In an attempt to quantify type I errors, Brodeur et al. (2016) analyzed 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 and 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 marginally significant. With respect to type II errors, Zhang and Ortmann (2013) reviewed 95 papers published in *Experimental Economics* between 2010 and 2012 and found that only one article mentioned statistical power and sample size issues.

[4] As one reviewer notes, training rounds may endanger causal interpretation for the same reasons that a within-subjects design does not establish causality under the Neyman-Rubin model. The potential downside of training rounds is that training may have some undesirable effect on subjects because by the time subjects are exposed to the treatment, they may have already changed their behavior due to the training. In the context of experimental auctions, this could be undesirable if, for example, training induces anchoring. On the other hand, training could lead to a desirable change in behavior due to learning; for example, to abandon the 'buy low' heuristic. We deliberately abstain from drawing general conclusions about the value of training in experiments outside the realm of experimental auctions. While a recent study by Roux and Thöni (2015) contributes to the debate about whether control questions have carry-over effects on oligopoly market behavior, there are wider issues here that have not garnered much attention in the literature that warrant further exploration.

although Drichoutis et al. (2011) did not specifically analyze the bid data from the training auctions, they found that subjects with extensive training gave significantly higher bids in the real auctions that followed the practice rounds than minimally trained subjects. In another study, Corrigan et al. (2014) examined the relationship between practice and real bids from two auction experiments where participants bid on homegrown-value goods. They found a positive correlation between practice and real bids but that this was mitigated by repetition. In addition, the reporting and analysis of practice rounds (in the context of a BDM mechanism) was a key issue of contention between Plott and Zeiler (2005) and Isoni et al. (2010)[5].

In this paper we use data from our study on projection bias in the context of experimental auctions[6]. We opted to look at the mediating effect of induced mood states on subjects' WTP for some products at three different delivery dates[7]: i) in present time (right after

[5] Plott and Zeiler (2005) offered a different interpretation than what has been frequently reported in the literature of the disparity between the willingness to accept (WTA) and willingness to pay (WTP) measures by showing that "...observed gaps are symptomatic of subjects' misconceptions about the nature of the experimental task" and that "the differences reported in the literature reflect differences in experimental controls for misconceptions as opposed to differences in the nature of the commodity". Subsequently, Isoni et al. (2010) noted that although Plott and Zeiler's (2005) experiments involved 14 tasks eliciting WTP and WTA for lotteries and one eliciting WTP or WTA for a mug, Plott and Zeiler's (2005) paper reports only the results from the task involving mugs. The exclusion of lottery data was justified on the grounds that these tasks were used only for training subjects. Isoni et al. (2010) applied Plott and Zeiler's (2005) elicitation procedure to both mugs and lotteries and found no significant WTP-WTA gap for mugs. However, they observed a significant and persistent gap for lotteries of the same kind found in Plott and Zeiler's (2005) unreported lottery data. Plott and Zeiler (2011) then argued that properties of the goods, such as the randomness associated with lotteries, might carry an inherent possibility of misconceptions, which might lead to gaps like those observed in Plott and Zeiler (2005) and Isoni et al. (2010).

[6] Loewenstein et al. (2003) coined the term "projection bias" to describe the general bias that has been documented in relation to the prediction of future tastes and particularly to describe people's assumption that tastes or preferences will remain the same over time. Quoting from Loewenstein et al. (2003), projection bias is "...a general bias in the prediction of future tastes: people tend to understand qualitatively the directions in which their tastes will change, but systematically underestimate the magnitudes of these changes. Hence, they tend to exaggerate the degree to which their future tastes will resemble their current tastes."

Experimental auctions have increasingly been used as a method to reveal subjects' preferences for market and non-market goods, where preferences are defined in monetary terms by values. As such, auctions are value elicitation methods that put people in an active (constructed) market environment. Valuations are elicited for use in cost-benefit analysis and to estimate welfare effects. Recently, Briz et al. (2015) studied the role of projection bias in experimental auctions by examining the bidding behavior of hungry and non-hungry subjects. They found that the difference in bids between a hot state (hunger) and a cold state (satiation) almost doubled when subjects had to predict their future tastes. Subjects who had to predict their future willingness to pay from their current tastes, tended to over-predict their hunger and under-predict satiation.

[7] The interplay of projection bias and mood is implied by affective forecasting, a process that describes the prediction of one's affect in the future (Wilson and Gilbert, 2003). Affective forecasting errors are closely related to several cognitive biases; one of these being projection bias. Projection bias can arise when '...people attempt to come up with an unbiased estimate of what their affective state will be in the future, but their assessment is contaminated by unique influences on their current well-known affective state' (Wilson and Gilbert, 2003). That is, projection bias occurs because individuals mispredict the present and future states of affective forecasting that differ in terms of physical arousal. The reason why this is of interest to economists is because projection bias implies a violation of utility maximization since the utility derived from future consumption is influenced by the affective state at the moment of choice (Kahneman and Thaler,

the auction), ii) one week later and before typical lunch time, and iii) one week later and after typical lunch time. We used two different types of products, a ham-cheese sandwich for which craving at lunch time is relevant and a ballpoint pen for which craving during lunch hours should not be relevant. Our results show that mood states actually mediate the effect of projection bias on subjects' WTP. However, a more careful look at the data coming from the training rounds indicates the presence of a similar effect for a set of (different) products used in the training rounds. Given that mood was induced only after the training rounds, we conclude that the effect we observe is not due to treatment assignment; that is, the treatment manipulation does not cause the difference we observe, which renders the received estimates as biased estimates of the true causal effects due to unobserved heterogeneity. Therefore, if data from the training rounds have not been analyzed, our study would have falsely attributed the difference in bids to the treatments.

Our results are worrying and troublesome because experimental economists do not often worry about causality since subjects are randomized into treatment and control groups. A treatment is then administered and any observed difference between the responses from those groups typically receives a causal treatment-response interpretation. Our per treatment sample size is typical for auction experiments (see Lee et al., 2011; Lusk et al., 2001; Neugebauer and Perote, 2007; Olivola and Wang, 2015, for just a few out of numerous examples in the literature). Moreover, we show that sample size calculations bound detectable effect sizes in the range of the actual effects we observe. Hence, without examining data from the training rounds, it would have been impossible to suspect failure of randomization. The problem we identify is as serious as the problem from false positives: failure of randomization can lead

2006). In this respect, Mehra and Sah (2002) note the relation between mood and projection bias: they study fluctuations in individuals' subjective parameters (the discount factor and the level of risk-aversion) due to mood and how subjects project their current mood into what their future sequence of moods will be. Patrick et al. (2005) provided some preliminary evidence of the interaction of mood and projection bias by showing that mood has an impact on the evaluation of future events but only when the future event is valence neutral and not affectively valenced. A few other studies have used cloudcover, as a proxy for mood, to study the effect of mood at the time of making a decision on future outcomes. Simonsohn (2010) analysed the enrollment decisions of prospective students who visited a US university in relation to cloudcover on the day of the visit and found evidence consistent with projection bias (i.e., cloudiness biased upwards the estimated future utility of attending that university). He explains this finding in terms of induced sadder mood due to the weather, which made belonging to an academically challenging institution more appealing and therefore influenced college decisions through memory. More recently, Madeira (2015) studied mental health care decision making and showed that cloud coverage increased the probability that a patient filled an antidepressant prescription on appointment day with the physician, which implies that projection bias interacts with mood in influencing the antidepressant treatment decision. Finally, Busse et al. (2015) find that the choice to purchase a convertible or a four-wheel-drive is highly dependent on the weather at the time of purchase in a way that can be explained by projection bias. All in all, there has been a strand of the literature that implies that mood at the time when making a decision for a future state might have a significant impact on this decision. This would be consistent with the interaction of mood and projection bias and has been the main force behind the empirical endeavor of this research project.

the research community into false avenues and wasted resources[8]. The next section describes our experimental design and methods. Section 3 describes our results which showcase our conclusion of failure of randomization to treatment. We conclude in the last section.

2 Methods

The experiment was carried out in February and March, 2013 at the Universidad Politécnica de Madrid. Announcements on the website of the University had been made a few weeks before the scheduled sessions. Students also received bulk announcement emails, sent to their university email accounts. The announcement did not provide specific details about the experiment, just general information that the experiment was about a study on consumer behavior. Participants responded with their weekday preferences and were then randomly assigned to one of the sessions. For each session, exactly eight participants were assigned with appropriate over-recruitment to account for no-shows.

The experiment consisted of 6 treatments in 24 different sessions (4 sessions/treatment). In all, 192 students participated in the experiment (8 subjects/session) and each subject only took part in one session. The experiment involved a 2 (mood inducement)×3 (time of delivery of the product) between-subjects experimental design. With respect to mood inducement (described momentarily), half of the subjects were induced to a positive mood state and the other half were induced to a negative mood state. We also varied the time of delivery of the auctioned products to test the effect of projection bias. In one third of the sessions (8 sessions) the product was given right after the auction (control treatment); in another third of the sessions (8 sessions), the product was given one week later at 1 pm (which is typically considered "before lunch" in the Spanish culture and according to students' habits); and in the other third of the sessions (8 sessions), the product was given one week later at 3 pm (which is typically considered an "after lunch" time). We auctioned simultaneously one food and one non-food product to test the effect of craving on subjects' WTP. We would expect a priori that delivering the product before or after lunch time is relevant for the food item but is not relevant for the non-food item. For the "before lunch" and "after lunch" future delivery treatments, the highest bidders were given an exchangeable coupon and were told that a fresh sandwich bought on the delivery date would be available in the exact same place of the auction site one week later. The place where the auctions took place is one of the main classroom buildings in the center of campus, just a few meters from the head office building. In the control sessions, subjects were given the product right

[8] The problem of false positives is further exacerbated by the fact that researchers 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 (for an exception see Kessler and Meier, 2014).

after the auction, which is the standard procedure for auctions. The experimental design is depicted in Table 1.

Table 1: Experimental design

		Mood	
		Negative	Positive
Time of delivery	Future at 1 pm	4 sessions × 8 subjects	4 sessions × 8 subjects
	Future at 3 pm	4 sessions × 8 subjects	4 sessions × 8 subjects
	Present	4 sessions × 8 subjects	4 sessions × 8 subjects

To achieve some variation over hunger levels, we varied the time of the sessions. Half of the sessions were conducted at 12 pm and half of the sessions were conducted at 1 pm. We only allowed a one hour difference between scheduled sessions to minimize potential "time of the day" effects on bidding behavior (Demont et al., 2012, 2013; Hoffman et al., 1993; Menkhaus et al., 1992; Morawetz et al., 2011). Since we could not fully control subjects' eating behavior before participating in the auction, we also asked subjects to self-report their hunger level.

When subjects arrived at the lab they were randomly seated and were assigned a six-digit participant number. They were told that all their answers were confidential, that answers would only be used for this specific study and that they would be given €10 at the end of the session for their participation. We elicited subjects' WTP using a 2$^{\text{nd}}$ price auction (Vickrey, 1961). The experimenter (one of the authors conducted all sessions) carefully explained the auction mechanism by means of numerical examples that were also projected in a screen. Subjects then participated in hypothetical practice auctions to familiarize themselves with the procedure. In this training phase, subjects bid separately for a USB pendrive and a mug in three repeated rounds. The importance to bid their true value for the goods during the auction was emphasized to the subjects. At the end of the third round, one of the rounds and one of the products were chosen randomly as binding. Although subjects knew this (training) procedure was hypothetical and thus the binding round was not binding at all, we used this language to mimic as closely as possible the auction procedure of the real rounds. No information was posted between rounds (Corrigan et al., 2012).

Next, subjects were induced in either a positive mood state or a negative mood state. To induce subjects into different moods, we exposed them to picture stimuli. The stimuli consisted of 40 color pictures representing either pleasant or unpleasant scenes. Half of the subjects were induced to a positive mood state (exposed to pleasant pictures) and the other half were induced to a negative mood state (exposed to unpleasant pictures). Pictures were selected from the International Affective Picture System (IAPS) (Lang et al., 2008).[9] Each

[9] The library numbers for IAPS pictures used in this study for positive mood inducement are: 1340,

one of the pictures was shown for 6 seconds with a 10 second gap in between, in order to let participants rate their emotional experience on a 5-point Likert scale anchored by a 'smiley' face and a 'frowned' face (see experimental instructions in Appendix A).

To quantify the mood induction effect, we used the Positive and Negative Affect Scale (PANAS), developed by Watson et al. (1988). This scale consists of 20 items using 5-point scales (1 = very slightly/not at all to 5 = extremely). The scale is sub-divided in two 10-item scales for positive affect (PA) and negative affect (NA). The terms comprising each sub-scale for negative affect are: afraid, scared, nervous, jittery, irritable, hostile, guilty, ashamed, upset, distressed; while the terms for positive affect are: active, alert, attentive, determined, enthusiastic, excited, inspired, interested, proud and strong. We used the Spanish version of the PANAS scale validated by Robles and Páez (2003).

Following mood induction, we carried out the real (non-hypothetical) auction. The auctioned products were a ballpoint pen and a ham-and-cheese sandwich. Before the auction, subjects were able to examine visually the auctioned products: a standard black ballpoint pen, and a non-branded ham-and-cheese packed sandwich. We explained that the auction procedure was the same as the one used for the practice auction. Depending on treatment assignment, subjects were told that they would get the product i) right after the auction, ii) one week later at 1 pm, iii) one week later at 3 pm. In all treatments (including the future delivery treatments), the highest bidder had to pay the 2$^{\text{nd}}$ highest price for the product that they won right after the auction. Subjects were informed about this policy before they started bidding. Subjects bid in three repeated rounds with no information being posted in between rounds. At the end of the third round, one round and one product were chosen as binding.

At the end of the experiment, subjects signed a participation sheet and were given €10 (minus the 2$^{\text{nd}}$ highest price, if the person was the highest bidder in a binding round/product). In the present delivery treatment, subjects were also given the product they paid for, while in the future delivery treatments subjects were given a coupon to be redeemed a week later.

1440, 1441, 1463, 1630, 1659, 1999, 2035, 2071, 2158, 2224, 2314, 2352, 2391, 2501, 2550, 2791, 4628, 5831, 8496; for negative mood inducement are: 1019, 2053, 2205, 2375, 2455, 2456, 2688, 2700, 2703, 3350, 6212, 6520, 8485, 9040, 9075, 9254, 9332, 9341, 9410, 9560. The pictures in the positive mood treatment generally depict happy moments in life, happy people, (cute) animals and their interaction with people and cartoon characters. The pictures in the negative mood treatment depict illness, grief, sorrow, death, starvation, accidents and pollution.

3 Results

3.1 Picture stimuli

We first explore whether subjects rated the emotional experience of the picture stimuli consistently with *a priori* expectations. Subjects in the positive mood treatment were exposed to a series of 20 pleasant pictures while subjects in the negative mood treatment were exposed to a series of 20 unpleasant pictures. They were then asked to indicate how each picture made them feel on a 5-point Likert scale anchored by a 'smiley' face and a 'frowned' face. We then summed their responses over the 20 pictures.

Figure 1 depicts kernel density estimators of picture evaluation scores by treatment (present delivery, future delivery at 1 pm, future delivery at 3 pm). Vertical lines depict median values. It is evident that pleasant pictures were scored lower while unpleasant pictures were scored higher across all treatments. Note that the smiley scale was reversed, so that a lower score indicates feelings associated with a 'smiley' when watching the picture while a higher score indicates feelings associated with a 'frown'. A t-test of whether pleasant and unpleasant picture evaluation scores are significantly different in each treatment, highly reject the null (p-value<0.001). Similar results are obtained using non-parametric Wilcoxon-Mann-Whitney tests.

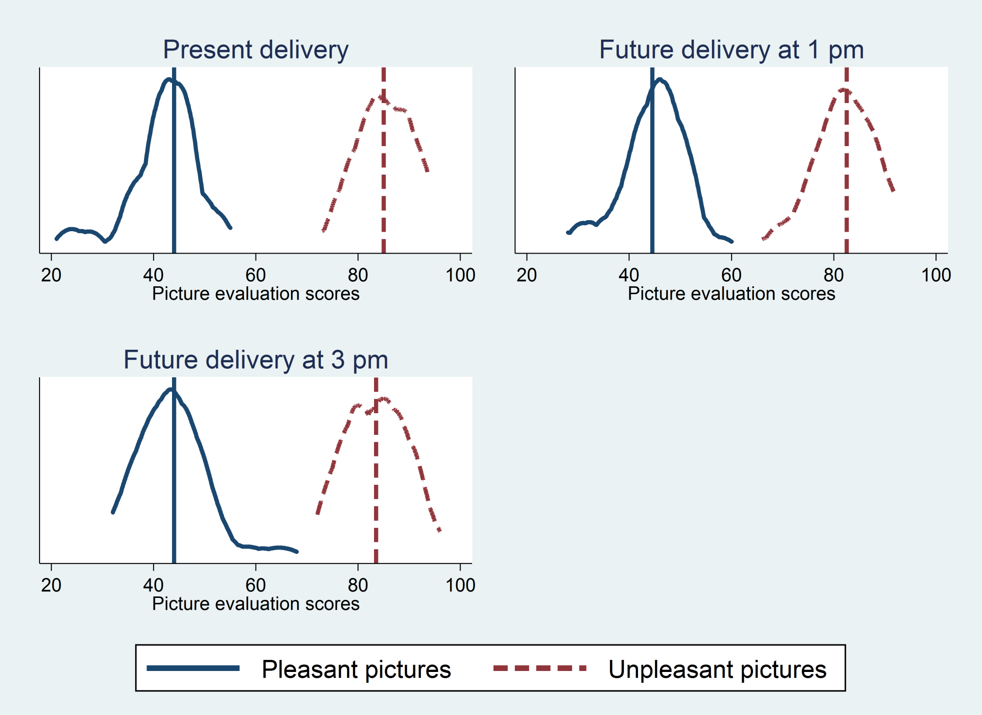


Figure 1: Kernel density estimators of picture evaluation scores by treatment

In addition, pleasant and unpleasant pictures were evaluated similarly across treatments.

An ANOVA test indicates that evaluation scores for pleasant pictures do not differ across treatments (F-statistic=1.44, p-value=0.241). The same goes for unpleasant picture evaluation scores (F-statistic=2.41, p-value=0.096). We obtain similar results using the non-parametric Kruskal-Wallis test for pleasant pictures ($\chi^2 = 2.509$, p-value=0.285) and for unpleasant pictures ($\chi^2 = 3.997$, p-value=0.135), respectively. Figure B.1 in Appendix B depicts kernel density estimates of pleasant picture evaluation scores and unpleasant picture evaluation scores by treatment. The graphs confirm results from the statistical tests above. In addition, Kolmogorov-Smirnov tests indicate that we cannot reject the null of equality of distributions of the picture evaluation scores shown in Figure B.1. All in all, results from the tests above indicate that the set of pleasant pictures corresponded to a pleasant emotional experience while the set of unpleasant pictures corresponded to an unpleasant emotional experience.

3.2 Picture stimuli and induced mood

Once we established that subjects perceived pleasant (unpleasant) picture stimuli as pleasant (unpleasant), the next important question is whether the picture stimuli were adequate in inducing positive and negative mood states. For this purpose, we summed the individual items of the PANAS scale to form the two sub-scales of positive and negative affect.

Figure 2 displays kernel density estimates for the positive and negative affect scales. The top panel corresponds to subjects in the positive mood treatment (these are the subjects that were exposed to pleasant picture stimuli) while the lower panel corresponds to subjects in the negative mood treatment (these are the subjects that were exposed to unpleasant picture stimuli). Vertical lines depict median values. In the positive mood treatment, the distribution for the positive affect scale is more to the right (indicating higher positive affect) while the distribution for the negative affect scale is more to the left (indicating lower negative affect). Thus, it appears that the pleasant picture stimuli are consistent with a gap in affect scores in the expected direction. Let us now contrast the top panel of Figure 2 with the bottom panel in the same figure. Compared with the positive mood treatment, positive affect is more to the left (implying lower levels of positive affect), while negative affect is more to the right (implying higher levels of negative affect). Thus, when comparing the positive mood and the negative mood treatments, distributions of positive and negative affect are in the expected directions.

Visual differences in Figure 2 are also confirmed by statistical analysis. A t-test of whether positive affect scores is different than negative affect scores, highly rejects the null in the positive mood inducement (p-value<0.001) as well as in the negative mood inducement (p-

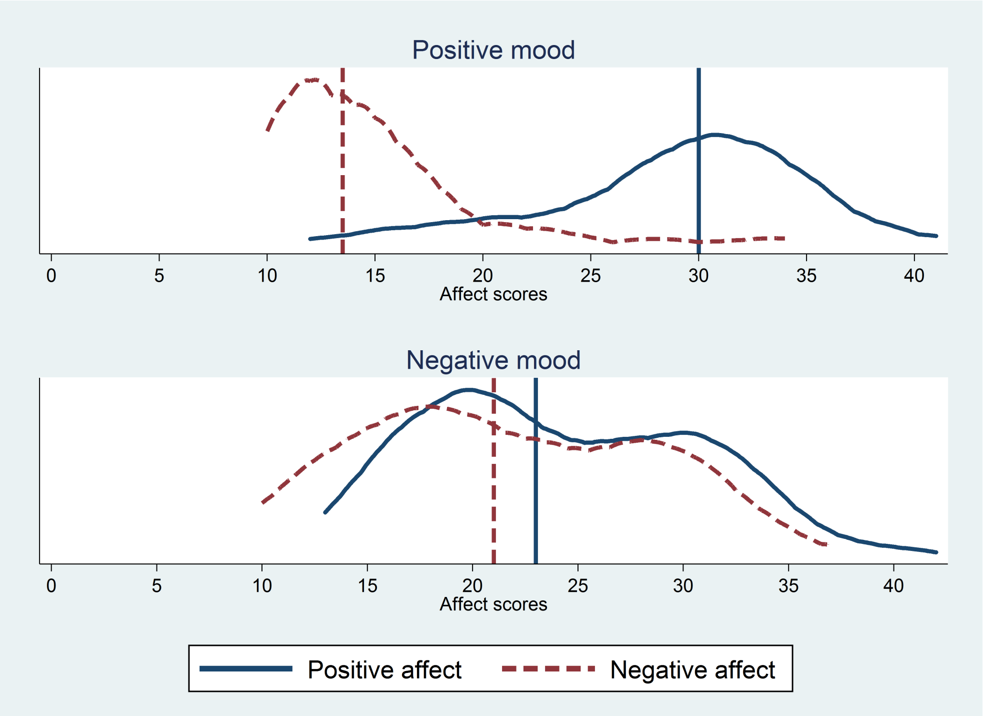


Figure 2: Kernel density estimators of positive and negative affect by mood induction treatment

value=0.021). Results from Wilcoxon-Mann-Whitney tests agree with previous conclusions. Kolmogorov-Smirnov tests on whether the distributions of positive affect and negative affect are equal, highly reject the null in all cases. Additionally, one may wonder if induced positive affect levels are similar in the positive mood treatment and in the negative mood treatment. The evidence is clear: the t-tests, Wilcoxon-Mann-Whitney tests and Kolmogorov-Smirnov tests all indicate that the positive affect in the positive mood treatment is statistically different (and higher) than in the negative mood treatment. The same set of tests show that the negative affect in the positive mood treatment is statistically different (and lower) than in the negative mood treatment.

As one reviewer noted, given that our design did not include a neutral mood treatment, all we can claim is that the degree of positive and negative affect is different across positive and negative mood treatments but not how a particular treatment changed mood and in what direction. However, all that matters is that the positive and the negative mood treatments exhibit affect levels in the expected directions and are different from each other. It is also important to note that the result pertaining to the positive and negative affect levels being close in the negative mood treatment, is not surprising given that positive and negative affect are aspects of mood that co-exist.

3.3 Bidding behavior

Table 2 shows the mean, standard deviation and median values of bids (pooled over the three rounds) per product and treatment. The top two panels show the descriptive statistics for the sandwich and pen products. With respect to the sandwich product, we can see that subjects under negative mood bid more in both the future delivery treatments while they bid less in the present delivery treatment (as compared to positive mood).

Table 2: Descriptive statistics of bids per treatment and product

Product	Treatment	Positive mood			Negative mood		
		Mean	SD	Median	Mean	SD	Median
Sandwich	Present delivery	1.52	0.77	1.50	1.27	0.66	1.20
	Future delivery at 1 pm	1.35	0.67	1.23	1.71	1.15	1.50
	Future delivery at 3 pm	0.98	0.84	0.83	1.17	0.85	1.00
Pen	Present delivery	1.05	0.73	1.00	0.82	0.47	0.75
	Future delivery at 1 pm	0.82	0.58	0.70	1.25	0.97	1.00
	Future delivery at 3 pm	0.75	0.76	0.50	0.74	0.81	0.50
USB pendrive	Present delivery	4.04	2.24	3.00	4.52	2.86	4.00
	Future delivery at 1 pm	5.07	3.16	5.00	7.74	5.11	7.00
	Future delivery at 3 pm	5.08	3.99	4.00	5.20	3.78	4.00
Mug	Present delivery	2.84	1.72	2.50	2.67	2.09	2.50
	Future delivery at 1 pm	2.78	1.73	2.25	4.94	3.89	3.75
	Future delivery at 3 pm	2.73	2.30	2.00	3.45	2.34	3.00

Notes: SD stands for standard deviation.

Although the analysis up to now is purely descriptive, a picture starts to emerge. It looks as if mood states interact with time of delivery of the product which would imply that some form of interaction between projection bias and mood is at work. Table 3 shows the results from statistical tests of whether bids (per product and per time of delivery treatment) differ between positive and negative mood. Three sets of tests are presented: a t-test, a Wilcoxon-Mann-Whitney test (WMW), and a Kolmogorov-Smirnov (KS) equality of distributions test[10]. The top two panels of Table 3 show that the difference in bids shown in Table 2 between the positive and negative mood treatments are statistically significant for the present and future delivery at 1 pm treatments. This result holds for both the sandwich and ballpoint pen under all three sets of tests. The statistical tests are not very clear for the sandwich product in the case of the future delivery at 3 pm treatment since the WMW test rejects the null only at the 10% level and fails to reject the null with a t-test and KS

[10] The t-test, the Wilcoxon-Mann-Whitney and the Kolmogorov-Smirnov tests are meant to be complementary to each other. The t-test performs a test on the equality of means. The WMW test, tests the hypothesis that samples are from populations with the same distribution and the KS test, tests the equality of distributions. The KS test is sensitive to any differences in the two distributions like differences in shape, spread or median while the WMW test is mostly sensitive to changes in the median.

test. Thus, for the future delivery at 1 pm we can conclude that there is a significant effect showing up already in the training rounds.

Table 3: Tests of differences in bids between the mood treatments

			Wilcoxon-	Kolmogorov-
Product	Treatment	t-test	Mann-Whitney	Smirnov
			test	test
		t	z	D
Sandwich	Present delivery	-2.428**†	-2.025**	0.167
	Future delivery at 1 pm	2.605**†	1.776*	0.208**†
	Future delivery at 3 pm	1.549	1.959*	0.156
Pen	Present delivery	-2.646***†	-1.536	0.240***‡
	Future delivery at 1 pm	3.719***‡	2.794***‡	0.292***‡
	Future delivery at 3 pm	-0.116	-0.327	0.063
USB pendrive	Present delivery	1.300	1.057	0.135
	Future delivery at 1 pm	4.357***‡	4.004***‡	0.365***‡
	Future delivery at 3 pm	0.216	0.586	0.073
Mug	Present delivery	-0.628	-1.180	0.115
	Future delivery at 1 pm	4.975***‡	4.090***‡	0.260***‡
	Future delivery at 3 pm	2.150**†	2.596***†	0.240***‡

Note: * p-value< 0.1, ** p-value< 0.05, *** p-value< 0.01. † indicates cases for which up to four different multiple test procedures indicate non-credibility of the null hypothesis. ‡ indicates cases for which nine to eleven different multiple test procedures indicate non-credibility of the null hypothesis.

Given that in Table 3 we test a family of hypotheses for each test statistic, a multiple comparisons problem arises. It is well known that as the number of comparisons increases, the probability of an incorrect rejection of a true null hypothesis increases. If all the tests are mutually independent, then the probability of at least one true null hypothesis being rejected would equal to $1-(1-\alpha)^N$ where N is the number of hypotheses being tested. For each test statistic, we perform 12 comparisons (3 treatments×4 products) so that for $\alpha = 5\%$ significance level, the chance of rejecting at least one true hypothesis would be 45.96%. To account for the multiple comparisons problem, we utilized multiple test procedures in order to calculate a corrected overall critical p-value such that an individual null hypothesis is considered to be acceptable only if its corresponding p-value is greater than the corrected overall critical p-value. Several multiple test procedures for defining an upper confidence bound for the set of null hypotheses that are true can be used (Newson and the ALSPAC Study Team, 2003, offer an exposition and Stata implementation for eleven different methods). Table 3 indicates all the cases for which the null hypothesis is not credible, that is, their p-values are lower than the corrected overall critical p-values. Given that we used several methods to calculate corrected p-values, the table indicates with a dagger (†) cases for which up to four different multiple test methods indicate non-credibility of the null hypothesis and a double dagger (‡) to indicate cases for which nine to eleven different multiple test methods indicate non-credibility of the null hypothesis[11]. Thus, visual inspection of Table 3 shows that most effects remain intact when accounting for multiple hypothesis testing.

At this point, we posit that most empirical researchers would be looking for narrative arguments to support the claim that there is some form of interaction between mood and projection bias as well as to explain how the small difference between the pen and sandwich results (in the future delivery at 3 pm treatment) could be attributed to craving. After all, the statistical tests look fairly convincing.

The bottom two panels in Table 2 show the descriptive statistics of bids from the practice rounds for the USB pendrive and mug. Although these are data that researchers usually discard and (almost) never formally analyze, it is interesting to note that for both products subjects bid more in the negative mood treatments (as compared to the positive mood treatments) and future delivery treatments. Note that although we present the descriptive statistics per treatment for the two practice products, the treatments are irrelevant for the practice rounds. For one, in the practice rounds the language in the instructions did not mention any future delivery of the products nor were subjects made aware that there was a future delivery treatment in the later part of the experiment. Second, mood induction had not even been applied in the practice rounds nor were subjects made aware that they were going to be exposed to a series of images as part of a mood induction procedure. The bottom two panels in Table 3 present the statistical tests for the treatment effects shown in Table 2. These roughly support the discussion above.

Conceptually, the problem with the results described above is that if a treatment effect is evident before the treatment is even applied, then the effect is likely not due to the treatment but rather an indication of randomization failure. Hence, we argue that practice rounds in experimental auctions may convey useful information and so data from practice rounds should always be analyzed before interpreting the treatment effect as causal. Even though one could argue that it is not necessary to analyze practice rounds separately since one could as well perform randomization/balance tests (i.e., checking whether a set of observed covariates is statistically different between the treatment and control) as a test for randomization failure, it is questionable whether such randomization tests are really meaningful.[12,13] Ho

[11] We utilized the set of multiple test methods listed in Stata's `multproc` routine. More details can be provided by the authors upon request.

[12] As a reviewer noted, an obvious advantage of analyzing the training round data is that as long as personal characteristics do not fully determine behavior, we can obtain additional information by looking at actual training round bidding behavior.

[13] Such a test can be reported separately for each variable (preferably after correcting for multiple hypothesis testing) or as a joint test (given that a large volume of individual tests increases the probability of finding at least one being significant). With our data, we fail to reject the null of no difference for

et al. (2007) categorize such tests under the term 'the Balance Test Fallacy'. Their main point is that '. . . balance is a characteristic of the observed sample, not some hypothetical population' and that 'the idea that hypotheses tests are useful for checking balance is therefore incorrect.' Mutz and Pemantle (2015) note that randomization is a process rather than an outcome and that balance tests are neither necessary to detect randomization problems nor do such tests indicate sufficient evidence of a randomization problem. They conclude that the process of randomization is either done correctly or not; there is no middle ground.[14]

Naturally, we can control for observable characteristics in a regression context. Table 4

gender ($\chi^2 = 5.19$, p-value=0.393), income level ($\chi^2 = 17.42$, p-value=0.294), hunger level (Kruskal-Wallis $\chi^2 = 8.35$, p-value=0.138), liking of ham-cheese sandwiches ($\chi^2 = 24.88$, p-value=0.206) and whether subjects had brought lunch with them the day of the experiment ($\chi^2 = 0.62$, p-value=0.987) but reject the null of no difference between treatments for a few observable characteristics such as age (Kruskal-Wallis $\chi^2 = 26.72$, p-value< 0.001) and education level (Fisher's exact p-value=0.001). A multiple hypothesis test procedure, similar to what we followed in Table 3, indicates that for age and education we can reject the null of no credibility of the null hypothesis.

[14] As a reviewer has further pushed us in discovering, we found that balancing tests are an issue also widely discussed in Randomized Control Trials (RCTs) in the medical literature. For example, in the CONSORT (Consolidated Standards of Reporting Trials) statement, endorsed by prominent medical journals (BMJ, Lancet etc.), hypothesis testing of imbalance is characterized as superfluous and misleading (Moher et al., 2010, pp. 17). In the economics literature, Deaton and Cartwright (2016) discuss the pitfalls of balance tests along the lines we present below.

In statistical notation, one would state the null hypothesis of such tests as $H_0 : \mu_A = \mu_B$ 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 is entirely for the sample and involves no inference to populations (Ho et al., 2007; Imai et al., 2008). By design, however, the purpose of hypotheses tests is to use sample information to make an inference about a population parameter. By these metrics, a balance test is a logical contradiction.

Altman and Doré (1990) note that if randomization has been done fairly (that is, the researcher has not cheated the randomization process in order to favor a treatment; as a side note, Senn (1994) demonstrates how one could allocate units in a way which favors one treatment over another but which does not lead to statistically significant imbalance), the null hypothesis is by definition true; so if one insists on using some statistical test to test for imbalance, we would expect 5% of such comparisons to be significant at the 5% level. Therefore, one test in twenty should give a significant result just by chance when $\alpha = 5\%$, so when ten variables are compared (a typical number of covariates in many experimental studies), it is very likely that none is significant. One should also keep in mind that randomization does not ensure balance but rather produces random deviations of relative size inversely proportional to the square root of the sample size (Mutz and Pemantle, 2015). When sample sizes increase, the expected random deviations between the two groups are reduced, which reflects the greater expected balance between the groups (Senn, 2013).

Significance tests assess the probability that observed baseline differences could have occurred by chance when, given randomization, we know that any differences are caused by chance (Altman, 1985; Moher et al., 2010). Imai et al. (2008) advance the view that researchers should check for balance but not with hypotheses tests. They support the use of quantile-quantile plots that directly compare the empirical distribution of two variables while Deaton and Cartwright (2016) suggest the use of a distance measure such as the normalized difference in means (Imbens and Wooldridge, 2009, equation 3). Imbalance in a baseline variable is only potentially important, if that variable is related to the outcome variable (Altman, 1985). Although there is no requirement to have baseline balance for valid inferences, the general advice is that even with randomization to treatment, observed covariates should be taken into account (Senn, 1994, 2013). This advice goes against the popular practice of not controlling for observable characteristics after the researcher has failed to reject the null based on balance tests and suggests that results based on unconditional tests should be taken with a grain of salt.

Table 4: Random effects regressions (with demographics)

	(1)		(2)		(3)		(4)	
	Sandwich		Pen		USB		Mug	
Constant	-0.178	(0.595)	0.788	(0.518)	7.953***	(2.526)	2.711	(1.752)
Positive mood	0.327*	(0.190)	0.252	(0.180)	-1.282	(0.878)	-0.064	(0.609)
Future at 1 pm	0.620***	(0.187)	0.450***	(0.174)	3.191***	(0.848)	2.230***	(0.588)
Future at 3 pm	0.098	(0.185)	-0.036	(0.176)	0.249	(0.857)	0.674	(0.595)
Positive mood× Future at 1 pm	-0.873***	(0.270)	-0.686***	(0.253)	-1.463	(1.232)	-2.365***	(0.855)
Positive mood× Future at 3 pm	-0.565**	(0.265)	-0.219	(0.251)	1.189	(1.224)	-0.847	(0.849)
Round 2	-0.017	(0.028)	0.052*	(0.028)	-0.133	(0.134)	0.074	(0.078)
Round 3	0.021	(0.028)	0.064**	(0.028)	0.077	(0.134)	0.173**	(0.078)
Hunger$_4$	-0.249	(0.152)						
Hunger$_3$	0.051	(0.249)						
Hunger$_2$	0.369	(0.246)						
Hunger$_1$	-0.419	(0.263)						
Age	0.049**	(0.019)	0.005	(0.018)	-0.081	(0.088)	0.025	(0.061)
Undergrad student	0.113	(0.149)	0.129	(0.138)	-1.530**	(0.673)	-0.352	(0.467)
Male	-0.212*	(0.111)	-0.186*	(0.104)	-0.502	(0.505)	-0.054	(0.350)
Income$_2$	-0.296*	(0.179)	-0.200	(0.169)	-0.665	(0.822)	-0.403	(0.570)
Income$_3$	-0.104	(0.183)	-0.133	(0.173)	0.732	(0.842)	-0.456	(0.584)
Income$_4$	-0.236	(0.201)	-0.200	(0.185)	0.250	(0.904)	0.337	(0.627)
Did not bring lunch	-0.120	(0.115)						
Like sandwich$_2$	0.610***	(0.197)						
Like sandwich$_3$	0.840***	(0.189)						
Like sandwich$_4$	0.785***	(0.196)						
Like sandwich$_5$	0.885***	(0.213)						
σ_u	0.694***	(0.037)	0.672***	(0.036)	3.274***	(0.176)	2.289***	(0.121)
σ_ϵ	0.278***	(0.010)	0.269***	(0.010)	1.311***	(0.047)	0.764***	(0.028)
N	576		576		576		576	
Log-likelihood	-366.297		-347.810		-1259.616		-981.978	
AIC	782.594		727.620		2551.232		1995.955	
BIC	891.497		797.318		2620.930		2065.653	

Standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01

shows the results from random effects regressions separately for each product. The regression model for the sandwich product controls for extra factors that are relevant to the sandwich product such as hunger level, liking of cheese-ham sandwich in general and whether the respondent brought lunch with him/her the day of the experiment. What is clear in Table 4 is that there are significant treatment effects consistent with our discussion above which, however, are also evident for the training products for which the treatments were not even applied. Note that observable characteristics do not seem to play a significant role since regression results from models with just the treatment variable produce almost identical results (see Table 5)[15].

Given the importance of practice/training bids, it may be preferable then to examine these bids in conjunction with the main auction bids. This is not likely to be a problem in terms of increasing the length of the papers since results from practice auction bids can be part of an appendix and most journals now offer this option to their prospective authors. In addition, examining pre-treatment outcomes and the characteristics across treatments is now a standard practice in field experiments or quasi-experimental work. Nevertheless, we posit that this is still considered 'not standard' for researchers doing laboratory experiments.

One of the main reasons for the use of laboratory experiments is that it allows us to randomize treatments. Moreover, the combination of having a high degree of control and greater ability to achieve homogeneity amongst the population studied allows us to have greater confidence on the unbiasedness of the treatment estimates. Therefore, using training rounds to detect unobserved heterogeneity may be considered just a second best, since it would make more sense to increase the number of observations per treatment so that the law of large numbers ensures that a randomized treatment is really orthogonal to potential outcomes. As far as we know, however, there is no rule of thumb on the determination of the optimal number of observations per treatment where the law of large numbers kicks in. The closest to this concept is sample size calculations that establish conditions under which it is possible to detect a certain (desirable) effect size with a specified power. As shown in Appendix C, our per treatment sample size is adequately safe for detecting any difference between treatments larger than 0.5 but also smaller differences under certain conditions. So the question "how many observations per treatment are required for the law of large numbers to kick in?" does not have a clear answer. As such, it is difficult to give

[15] Although we focus on statistical significant results here, it might be important to consider effect sizes as well. In Appendix D we present standardized effect sizes as complements to statistical significance testing. In addition, one can interpret coefficient estimates in terms of predicted WTP in order to get a sense of proportions. The average predicted WTP for Table 4 is €1.34 for the sandwich, €0.91 for the pen, €5.27 for the USB and €3.24 for the mug. To illustrate our point, take the coefficient estimate for 'Future at 1 pm' variable which happens to be statistically significant for all models listed in Table 4. The estimated coefficient would correspond to a 46.33% (=0.62/1.34), 49.67% (=0.45/0.91), 60.51% (=3.19/5.27) and 68.91% =(2.23/3.24) change, respectively, of the average predicted WTP, which are substantial effects.

Table 5: Random effects regressions

	(1)		(2)		(3)		(4)	
	Sandwich		Pen		USB		Mug	
Constant	1.264***	(0.143)	0.781***	(0.125)	4.542***	(0.616)	2.585***	(0.420)
Positive mood	0.251	(0.201)	0.234	(0.175)	-0.481	(0.865)	0.174	(0.590)
Future at 1 pm	0.441**	(0.201)	0.432**	(0.175)	3.216***	(0.865)	2.276***	(0.590)
Future at 3 pm	-0.097	(0.201)	-0.081	(0.175)	0.674	(0.865)	0.784	(0.590)
Positive mood× Future at 1 pm	-0.605**	(0.284)	-0.663***	(0.247)	-2.192*	(1.223)	-2.334***	(0.835)
Positive mood× Future at 3 pm	-0.439	(0.284)	-0.221	(0.247)	0.360	(1.223)	-0.893	(0.835)
Round 2	-0.017	(0.028)	0.052*	(0.028)	-0.133	(0.134)	0.074	(0.078)
Round 3	0.021	(0.028)	0.064**	(0.028)	0.077	(0.134)	0.173**	(0.078)
σ_u	0.786***	(0.042)	0.682***	(0.037)	3.374***	(0.181)	2.320***	(0.123)
σ_ϵ	0.278***	(0.010)	0.269***	(0.010)	1.311***	(0.047)	0.764***	(0.028)
N	576		576		576		576	
Log-likelihood	-389.226		-350.619		-1265.110		-984.437	
AIC	798.453		721.237		2550.221		1988.875	
BIC	842.014		764.798		2593.782		2032.436	

Standard errors in parentheses. * p<0.1, ** p<0.05 *** p<0.01

practical advice to experimenters on when it is safe to stop collecting data to ensure the orthogonality of the treatment to potential outcomes. Therefore, using training rounds to detect unobserved heterogeneity remains a practical and viable alternative in the context of experimental auctions.

4 Conclusion

Our results show the relevance of analyzing all available experimental data from experimental auctions before making claims about the significance of treatment effects. In the context of experimental auctions, we argue that data from training rounds could contain valuable information that researchers most often dismiss.

Using a case study on the interplay of mood and projection bias in experimental auctions, we demonstrate that data analysis from training rounds can help experimenters assess if there is a failure of randomization to treatment and hence not fall prey to attributing a non-causal effect as causal. When we analyze the data from the real auction, we detect significant treatment effects that show that projection bias is mediated by mood. However, we found that similar effects are present in the training rounds and products. Given that treatments were applied only after the training rounds, we conclude that the effect we observe is not due to assignment to treatment. The fundamental methodological problem we identify is not limited to this experiment or auction experiments in general. It should be relevant for all economics experiments that try to identify treatment effects.

As a further step in this research agenda, it would be interesting to collect auction datasets from published papers for which data from the training rounds have been recorded and are available. An analysis of data coming from the training rounds alongside the actual auction results reported in the published papers would be a first step in redirecting the research community from false avenues, if failures of randomization to treatment have indeed been prevalent. Additionally, we only used one type of auction, a 2$^{\text{nd}}$ price auction, for our experiment which by design may not produce strong enough incentives for all bidders to reveal their true preferences. If inference mistakes interact with incentivization, then it might be interesting in future studies to also examine and compare results from different types of auctions that have been designed to engage off-margin bidders in larger group sizes (like the random n$^{\text{th}}$ price auction; see for example Shogren et al., 2001).

For future auction experiments, we suggest that journal editors and reviewers ask authors to report data and analysis from the training phase of experimental auctions. If authors find significant treatment effects from their experiments, then these effects should not be present in the training part of the auction, conditional on the treatments not taking place during the training part of the auction study. All in all, we cannot stress enough the importance of employing larger sample sizes in experiments to ensure success of randomization to treatment. Since there might be a worry that this would lead to almost anything passing for statistically significant, we should also stress the importance of complementing any analysis with effect size calculations that would help in judging the economic significance of the treatment effects (see for example Appendix D).

References

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

Altman, D. G. and C. J. Doré (1990). Randomisation and baseline comparisons in clinical trials. *The Lancet 335*(8682), 149–153.

Briz, T., A. C. Drichoutis, and L. House (2015). Examining projection bias in experimental auctions: The role of hunger and immediate gratification. *Agricultural and Food Economics 3*(22).

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.

Busse, M. R., D. G. Pope, J. C. Pope, and J. Silva-Risso (2015). The psychological effect of weather on car purchases. *The Quarterly Journal of Economics 130*(1), 371–414.

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

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.

Corrigan, J. R., M. C. Rousu, and D. P. T. Depositario (2014). Do practice rounds affect experimental auction results? *Economics Letters 123*(1), 42–44.

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

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.

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.

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., J. Rodolfo M. Nayga, and P. Lazaridis (2011). The role of training in experimental auctions. *American Journal of Agricultural Economics 93*(2), 521–527.

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

Guala, F. (2010). experimental economics, history of. In S. N. Durlauf and L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd ed.)., pp. 99–106. London: Palgrave Macmillan UK.

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.

Hoffman, E., D. J. Menkhaus, D. Chakravarti, R. A. Field, and G. D. Whipple (1993). Using laboratory experimental auctions in marketing research: A case study of new packaging for fresh beef. *Marketing Science 12*(3), 318–338.

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

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

Isoni, A., G. Loomes, and R. Sugden (2010). The willingness to pay-willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations: Comment. *American Economic Review 101*(2), 991–1011.

Kahneman, D. and R. H. Thaler (2006). Anomalies: Utility maximization and experienced utility. *Journal of Economic Perspectives 20*(1), 221–234.

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.

Lang, P. J., M. M. Bradley, and B. N. Cuthbert (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical report A-8, University of Florida, Gainesville, FL.

Lee, J. Y., D. B. Han, R. M. Nayga, and S. S. Lim (2011). Valuing traceability of imported beef in Korea: an experimental auction approach*. *Australian Journal of Agricultural and Resource Economics 55*(3), 360–373.

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.

Loewenstein, G., T. O'Donoghue, and M. Rabin (2003). Projection bias in predicting future utility. *The Quarterly Journal of Economics 118*(4), 1209–1248.

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

Madeira, T. (2015). Weather, mood, and use of antidepressants: The role of projection bias in mental health care decisions. *Working paper*.

Mehra, R. and R. Sah (2002). Mood fluctuations, projection bias, and volatility of equity prices. *Journal of Economic Dynamics and Control 26*(5), 869–887.

Menkhaus, D. J., G. W. Borden, G. D. Whipple, E. Hoffman, and R. A. Field (1992). An empirical application of laboratory experimental auctions in marketing research. *Journal of Agricultural and Resource Economics 17*(1), 44–55.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gtzsche, 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*.

Morawetz, U. B., H. De Groote, and S. C. Kimenju (2011). Improving the use of experimental auctions in Africa: Theory and evidence. *Journal of Agricultural and Resource Economics 36*(2), 263–279.

Mutz, D. C. and R. Pemantle (2015). Standards for experimental research: Encouraging a better understanding of experimental methods. *Journal of Experimental Political Science 2*(2), 192–215.

Neugebauer, T. and J. Perote (2007). Bidding as if risk neutral in experimental first price auctions without information feedback. *Experimental Economics 11*(2), 190–202.

Newson, R. and the ALSPAC Study Team (2003). Multiple-test procedures and smile plots. *Stata Journal 3*(2), 109–132.

Olivola, C. Y. and S. W. Wang (2015). Patience auctions: the impact of time vs. money bidding on elicited discount rates. *Experimental Economics*, 1–22.

Patrick, V. M., A. Fedorikhin, and D. MacInnis (2005). The future is colored pink or blue: The effect of mood on affective forecasting. In G. Menon and A. R. Rao (Eds.), *NA* -

Advances in Consumer Research, Volume 32, pp. 339–340. Duluth, MN, USA: Association for Consumer Research.

Peirce, C. S. and J. Jastrow (1885). On small differences in sensation. *Memoirs of the National Academy of Sciences 3*, 73–83.

Plott, C. R. (1991). Will economics become an experimental science? *Southern Economic Journal 57*(4), 901–919.

Plott, C. R. and K. Zeiler (2005). The willingness to pay-willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations. *The American Economic Review 95*(3), 530–545.

Plott, C. R. and K. Zeiler (2011). The willingness to pay-willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations: Reply. *American Economic Review 101*(2), 1012–28.

Robles, R. and F. Páez (2003). Estudio sobre la traducción al español y las propiedades psicométricas de las escalas de afecto positivo y negativo (PANAS). *Salud Mental 26*(1), 69–75.

Roux, C. and C. Thöni (2015). Do control questions influence behavior in experiments? *Experimental Economics 18*(2), 185–194.

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.

Selya, A. S., J. S. Rose, L. C. Dierker, D. Hedeker, and R. J. Mermelstein (2012). A practical guide to calculating Cohens 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., M. Margolis, C. Koo, and J. A. List (2001). A random nth-price auction. *Journal of Economic Behavior and Organization 46*(4), 409–421.

Siegel, A. E. (1964). Sidney Siegel: a memoir. In S. Messick and A. H. Brayfield (Eds.), *Decision and Choice. Contributions of Sidney Siegel*. New York: McGraw-Hill Book Co.

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. (2010). Weather to go to college. *The Economic Journal 120*(543), 270–280.

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.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement 61*(4), 605–632.

Snijders, T. A. and R. J. Bosker (1999). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage publications Ltd.

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.

Svorenčík, A. (2015). *The experimental turn in Economics: A history of Experimental Economics*. University of Utrecht, Utrecht school of economics, Dissertation series #29.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance 16*(1), 8–37.

Watson, D., L. A. Clark, and A. Tellegen (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology 54*(6), 1063–1070.

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

Wilson, T. D. and D. T. Gilbert (2003). *Affective Forecasting*, Volume Volume 35, pp. 345–411. Academic Press.

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

A Appendix: Experimental Instructions

[This is an English translation of the original instructions written in Spanish]
[Text in brackets was not shown to subjects]

Welcome announcement
Thank you for agreeing to participate in this survey. The survey concerns the economics of decision making.

In a short while, we will conduct a series of experimental auctions with known products.

You have been randomly assigned a participant number for this entire session. All information collected is strictly confidential and will only be used for this specific project.

To thank you for your participation, you are going to be given 10€ at the end of the session. At the auction, you will have the opportunity to bid on and get the product, which will be given to the highest bidders (according to the rules I will describe momentarily).

If you have any questions, you may ask an assistant or the moderator. Do not communicate with other participants of this session.

In this experiment we explore several topics; therefore you will participate in several phases.

Page **1** out of **8**

------Page break------

The 2nd Price Vickrey Auction

In the tasks to follow you will participate in a type of auction known as a 2nd price auction. The 2nd price auction has 5 basic steps:

Step 1: We'll describe to you the product to be auctioned.

Step 2: Each one of you, will submit a bid for buying the product.

Step 3: The monitor will collect the bid sheets and rank all bids from highest to lowest.

Step 4: The person that submits a bid higher than the 2nd highest price buys the product **but will pay the price of the 2nd highest bidder**. If your bid is not higher than the second highest bid then you don't purchase the good.

Consider this numerical example:

Suppose 8 people bid in an auction in order to buy a USB memory stick (16GB). Each bidder submits a bid separately. The submitted bids are given in the table below:

Person	**Bid**
1	12
2	15
3	20
4	18
5	30
6	25
7	35
8	0

Page **2** out of **8**

------Page break------

After ranking bids from highest to lowest, we have:

Person	**Bid**
7	35
5	**30**
6	25
3	20
4	18
2	15
1	12
8	0

Persons 7 purchases one unit of the good because s/he bid higher than any other person but only pays 30 (second highest bid). All the other participants in the auction pay nothing and do not receive a memory stick.

In this auction, the best strategy is to bid exactly what the item is worth to you. Consider the following: if you bid less than what the object is worth to you, then you may not buy the product and miss a good opportunity for buying something at a price you were actually willing to pay.

Conversely, if you bid more than what the object is worth to you, you may end up having to pay a price higher than what you really wanted to. Thus, your best strategy is to bid exactly what the object is worth to you. The tasks you will do today are not hypothetical and have real monetary consequences.

Do you have any questions?

Page **3** out of **8**

------Page break------

Training auction [Hypothetical: USB memory stick]

We will now do a training task. This task is designed to allow you to familiarize yourself with the 2nd price auction. We will repeat this auction for three rounds. We will then select one round as binding by having one of you selecting a number from 1 to 3 from an urn. The numbers correspond to rounds, so if s/he picks number 1 then round 1 is binding, if s/he picks number 2 then round 2 is binding etc. [It was emphasized to subjects that 'binding' in the hypothetical context of the practice auction was only meant to simulate the actual auction to follow]

In this auction we will auction a memory stick. Take a look at this picture.

[Experimenter shows picture of usb memory stick in the screen]

You have all been provided with yellow slips, wherein you will write down and record your bid.

Please, take the yellow slip number 1 and write down the maximum you are willing to pay to purchase this usb stick.

After you've finished writing your bids, the monitor will go around the room and collect the bid sheets. I will then rank bids from highest to lowest, determine the 2nd highest price and the person with bid above the 2nd highest price. In private, at the front of the room, bids will be ranked from lowest to highest.

The bid is private information and should not be shared with anybody else. Please be quiet while the auction is carried out.

[Once the first round is finished, second round starts]

Now, please, take the yellow slip number 2 and write down the maximum you are willing to pay to purchase this usb stick.

[Same procedure is followed. Once the second round is finished, third round starts]

Now, please, take the yellow slip number 3 and write down the maximum you are willing to pay to purchase this usb stick.

[Experimenter collects bid sheets]

[The experimenter asks one person to draw a number from an urn; Number determines binding round]

[IDs of highest bidder and 2nd highest price for the binding round are determined and announced]

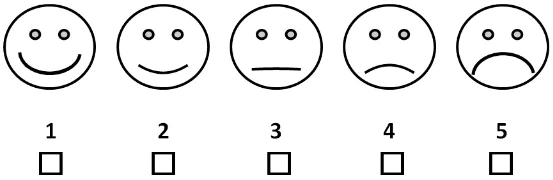
------Page break------

Picture evaluation phase [**Mood inducement phase**]

In this phase we will show you a sequence of 20 pictures.

You will see a slide with the number of the picture for 2 seconds, after it, each picture will be shown on the screen for 6 seconds, and then you have 10 seconds to describe "how you felt while watching the picture." Please, look at the pictures carefully and keep quiet.

To describe how you felt, we have provided a scale with faces that you should use for ranking.



Please, make sure the number of the picture matches the number of scale used and tick the appropriate box.

We remind you to please remain silent during the whole session.

Page **5** out of **8**

------Page break------

Feelings evaluation [**Mood measurement**]

This question consists of a number of words and phrases that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word.

Indicate to what extent you feel like this right now. Use the following scale to record your answers:
[Original word in Spanish is provided in parenthesis]

very slightly/not at all	a little	moderately	quite a bit	extremely
1	2	3	4	5

Page **6** out of **8**

------Page break------

___ Interested (Motivado)
___ Upset (Molesto a disgusto)
___ Excited (Emocionado)
___ Distressed (De malas)
___ Strong (Firme)
___ Guilty (Culpable)
___ Scared (Temeroso)
___ Hostile (Agresivo)
___ Enthusiastic (Entusiasmado)
___ Proud (Estar orgulloso)
___ Irritable (Irritable)
___ Alert (Alerta)
___ Ashamed (Avergonzado)
___ Inspired (Inspirado)
___ Nervous (Nervioso)
___ Determined (Decidido)
___ Attentive (Estar atento)
___ Jittery (Inquieto)
___ Active (Activo)
___ Afraid (Inseguro)

The real product auctions

[**Present auction treatment**]

We will now auction a pen and a non-branded ham and cheese sandwich. The sandwiches are kept in a refrigerator and have been bought this morning. We will follow the same procedure as we did for the memory stick. We will repeat this auction for three rounds. We will then select one round as binding by having one of you selecting a number from 1 to 3 from an urn. The numbers correspond to rounds, so if s/he picks number 1 then round 1 is binding, if s/he picks number 2 then round 2 is binding etc. Finally, we will select one product as binding. We will select either 1 or 2 from an urn, being 1 the pen and 2 the ham and cheese sandwich. Please, pass the sandwiches and the pens around.

Please, take the pink slip number 1 and write down the maximum you are willing to pay to purchase the sandwich and the pen, respectively.

After you've finished writing your bids, the monitor will go around the room and collect the bid sheets. The monitor will then rank bids from highest to lowest, determine the 2nd highest price and the person with bids above the 2nd highest price for each product, respectively.

The bid is private information and should not be shared with anybody else. Please be quiet while the auction is carried out.

[Once the first round is finished, second round starts]

Now, please, take the pink slip number 2 and write down the maximum you are willing to pay to purchase the sandwich and the pen, respectively.

[Same procedure is followed. Once the second round is finished, third round starts]

Now, please, take the pink slip number 3 and write down the maximum you are willing to pay to purchase the sandwich and the pen, respectively.

[Experimenter collects bid sheets]

[The experimenter asks one person to draw a number from an urn; Number determines binding round. The experimenter asks one person to draw a number from an urn. The number determines the binding product.]

[ID of highest bidder and 2nd highest price for the binding product and round are determined and announced.]

[**Future auction treatment**]

We will now auction a pen and a non-branded ham and cheese sandwich. We will follow the same procedure as we did for the memory stick. We will repeat this auction for three rounds. We will then select one round as binding by having one of you selecting a number from 1 to 3 from an urn. The numbers correspond to rounds, so if s/he picks number 1 then round 1 is binding, if s/he picks number 2 then round 2 is binding etc. Finally, we will select one product as binding. We will select either 1 or 2 from an urn, being 1 the pen and 2 the ham and cheese sandwich. Please, pass the sandwiches and the pens around.

[**Before lunch treatment**]

The selected product will be given at 1 pm, in this exact place in one week from today. In case the binding product is a sandwich, we will have available fresh sandwiches made on the day of delivery. To ensure you get your pen or sandwich next week, you will be given a coupon to be redeemed, and my own professional card, in case you have any problem showing up. You can stop by my office to get the product in a weeks' time, if for some reason you are not able to pick it up from this room this time next week.

[**After lunch treatment**]

The selected product will be given at 3 pm, in this exact place in one week from today. In case the binding product is a sandwich, we will have available fresh sandwiches made on the day of delivery. To ensure you get your pen or sandwich next week, you will be given a coupon to be redeemed, and my own professional card, in case you have any problem showing up. You can stop by my office to get the product in a weeks' time, if for some reason you are not able to pick it up from this room this time next week.

[Common text for before and after lunch treatments]

However, you will have to pay for the product today when the session is finished. The market price for the pen/sandwich determined from this auction (2nd highest price) will be deducted from the participation fees of the highest bidders.

Are there any questions?

Please, take the pink slip number 1 and write down the maximum you are willing to pay to purchase the sandwich and the pen, respectively.

After you've finished writing your bids, the monitor will go around the room and collect the bid sheets. The monitor will then rank bids from highest to lowest, determine the 2nd highest price and the person with bid above the 2nd highest price for each product, respectively.

The bid is private information and should not be shared with anybody else. Please be quiet while the auction is carried out.

[Once the first round is finished, second round starts]

Now, please, take the pink slip number 2 and write down the maximum you are willing to pay to purchase the sandwich and the pen, respectively.

[Same procedure is followed. Once the second round is finished, third round starts]

Now, please, take the pink slip number 3 and write down the maximum you are willing to pay to purchase the sandwich and the pen, respectively.

[Experimenter collects bid sheets]

[The experimenter asks one person to draw a number from an urn; Number determines binding round. The experimenter asks one person to draw a number from an urn. The number determines the binding product.]

[ID of highest bidder and 2nd highest price for the binding product and round are determined and announced.]

Page **7** out of **8**

------Page break------

Final questionnaire

The final task involves filling out a questionnaire. I will distribute the questionnaire in a minute. Please make sure your ID number is on the top left corner and raise your hands when you are done.

Thank you very much for your participation!

Page **8** out of **8**

------Page break------

B Appendix: Additional figures

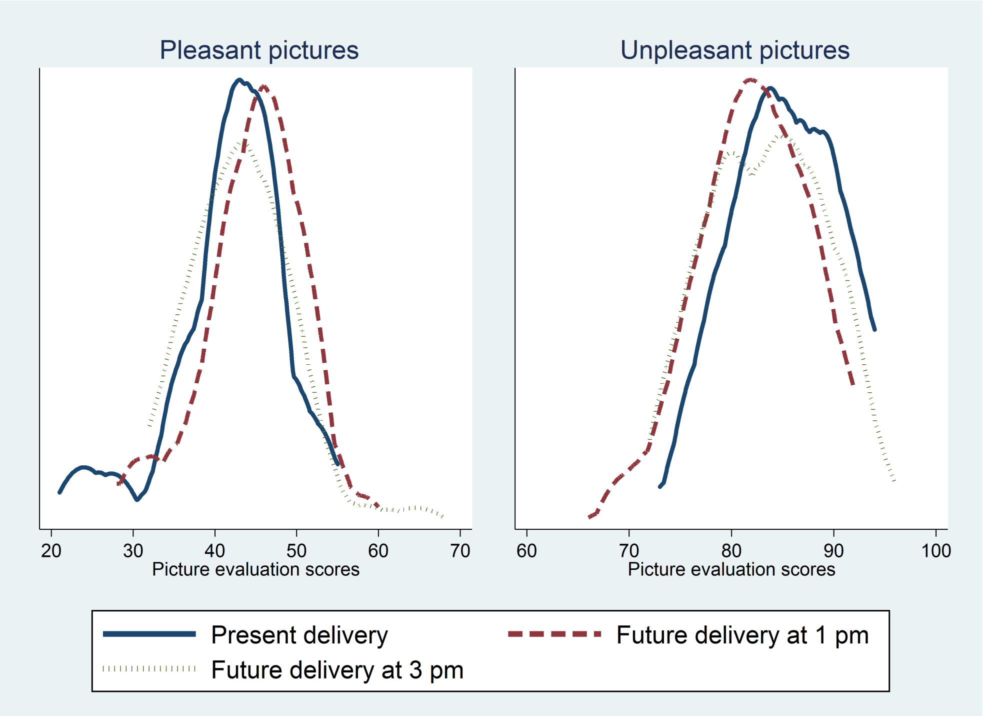


Figure B.1: Kernel density estimators of picture evaluation scores by treatment

C Appendix: Sample size calculations

Sample size calculations allow experimenters to establish conditions under which it is possible to detect a certain (desirable) effect size with a specified power (Type II error). Assuming $\alpha = 0.05$ (Type I error) and $\beta = 0.20$ (Type II error), the per treatment minimum sample size required to compare two means μ_0 and μ_1, with common variance of σ^2 in order to achieve a power of at least $1-\beta$ is given by (Diggle et al., 2002, p. 30; Liu and Wu, 2005):

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

The formula above takes into account the fact that subjects submit multiple bids in our experiment, by including the number of repeated measurements M (i.e., auction rounds in our case) as well as a value for the correlation ρ between observations on the same subject. To calculate a minimum sample size one needs to feed the above formula with values for σ, M, ρ and the minimum meaningful difference $\mu_0 - \mu_1$.[16]

To specify the necessary parameters for the above formulas, we used data from an older 2$^{\text{nd}}$ price experiment for sandwich products conducted also with a Spanish student sample (Briz et al., 2015). For this particular study, we calculated a value of $\sigma = 0.55$ and an intraclass correlation coefficients of $\rho = 0.83$.

Consequently, we calculated the minimum sample size for a range of values of σ between 0.4 and 0.7 with a step of 0.1 and for three values of ρ (0.7, 0.8 and 0.9). As the minimum meaningful difference, $d = \mu_0 - \mu_1$, we used values of 0.2 to 0.5 with steps of 0.05 which matches well with the range of treatment effects reported in Tables 4 and 5.

Table C.1 shows sample size calculations for different values of σ, ρ and d. It is obvious that the lower the minimum meaningful difference d, the higher the standard deviation σ and the higher the correlation ρ, a larger sample size is needed to detect the desired effect size with 80% power. Table C.1 highlights in gray all cases for which the per treatment sample size of our experiment (32 subjects) is lower or equal than the corresponding calculation i.e., cases for which the sample size of the present study can detect as statistically significant the minimum meaningful difference with 80% power. For the range of values of σ and ρ considered here, our study cannot detect differences lower or equal to 0.2 with 80% power but can safely detect any difference larger or equal to 0.5.

[16] The reader will note that in the special case where the experiment does not have multiple measurements per subject, that is $M = 1$ and $\rho = 0$, the formula reduces to $n = \frac{2(z_{1-\alpha/2}+z_{1-\beta})^2}{(\frac{\mu_0-\mu_1}{\sigma})^2}$.

Table C.1: Sample size calculations for different values of σ, ρ and d

		$\sigma = 0.4$	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.7$
$d = 0.2$	$\rho = 0.7$	50	78	113	154
	$\rho = 0.8$	54	85	122	167
	$\rho = 0.9$	59	92	132	179
$d = 0.25$	$\rho = 0.7$	32	50	72	98
	$\rho = 0.8$	35	54	78	107
	$\rho = 0.9$	38	59	84	115
$d = 0.3$	$\rho = 0.7$	22	35	50	68
	$\rho = 0.8$	24	38	54	74
	$\rho = 0.9$	26	41	59	80
$d = 0.35$	$\rho = 0.7$	16	26	37	50
	$\rho = 0.8$	18	28	40	54
	$\rho = 0.9$	19	30	43	59
$d = 0.4$	$\rho = 0.7$	13	20	28	38
	$\rho = 0.8$	14	21	31	42
	$\rho = 0.9$	15	23	33	45
$d = 0.45$	$\rho = 0.7$	10	16	22	30
	$\rho = 0.8$	11	17	24	33
	$\rho = 0.9$	12	18	26	35
$d = 0.5$	$\rho = 0.7$	8	13	18	25
	$\rho = 0.8$	9	14	20	27
	$\rho = 0.9$	9	15	21	29

D Appendix: Cohen's f^2

Standardized effect sizes can be important complements to statistical significance testing. Although there is currently a menagerie of measures of effect sizes, these are often inappropriate for standardizing effects coming from more advanced econometric models, such as random effect models employed in this paper, where variance accounted by 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 by employing a signal-to-noise ratio in the form of an F test:

$$F(u,v) = \frac{PV_S/u}{PV_E/v} \tag{A.2}$$

where PV_S is the source of the dependent's variable Y variance accounted from some source, PV_E is proportion of error variance, u is the number of independent variables (IVs) for the source i.e., the degrees of freedom for the numerator and v is the degrees of freedom for the error variance. Cohen (1988) defines different formulas for F depending on how the source and error are defined. The most general case, from which other cases can be derived as special cases, assumes that the source of variance in Y comes from three different sources: a set of variables $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ made up of k, m and n variables respectively. Using similar notation to Cohen (1988), define the proportion of variance accounted for by a set of factors $\mathbf{A}$ as $R^2_{Y\cdot\mathbf{A}}$ and the proportion of variance accounted for by $\mathbf{A}$ and $\mathbf{B}$ as $R^2_{Y\cdot\mathbf{A},\mathbf{B}}$, then the proportion of Y variance accounted for by $\mathbf{B}$, *over and above* what is accounted for by $\mathbf{A}$ is denoted as $R^2_{Y\cdot(\mathbf{B}\cdot\mathbf{A})} = R^2_{Y\cdot\mathbf{A},\mathbf{B}} - R^2_{Y\cdot\mathbf{A}}$. Given the additional set of variables $\mathbf{C}$, the error variance proportion is given by $1 - R^2_{Y\cdot\mathbf{A},\mathbf{B},\mathbf{C}}$.

The F statistic in equation A.2 can be written as $F(u,v) = f^2\frac{v}{u}$ where:

$$f^2 = \frac{PV_S}{PV_E} = \frac{R^2_{Y\cdot(\mathbf{B}\cdot\mathbf{A})}}{1 - R^2_{Y\cdot\mathbf{A},\mathbf{B},\mathbf{C}}} = \frac{R^2_{Y\cdot\mathbf{A},\mathbf{B}} - R^2_{Y\cdot\mathbf{A}}}{1 - R^2_{Y\cdot\mathbf{A},\mathbf{B},\mathbf{C}}} \tag{A.3}$$

As a crude guide, Cohen (1988) offers conventional operational definitions of 0.02, 0.15 and 0.35 for 'small', 'medium' and 'large' values of f^2, respectively. These terms should not be taken literally since the 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.'

$F(u,v)$ has a non-central F distribution with non-centrality parameter equal to $\lambda = f^2(u+v+1)$ (Cohen, 1988, pp. 414). In order to construct two-sided $100(1-\alpha)\%$ confidence intervals (CIs) for λ given the sample $F(u,v)$, one needs to find the non-centrality parameters λ_{lower} and λ_{upper} that correspond to (Smithson, 2001):

$$Pr(u,v,F,\lambda_{lower}) = 1 - \frac{a}{2} \tag{A.4}$$

$$Pr(u, v, F, \lambda_{upper}) = \frac{a}{2} \tag{A.5}$$

The recovered non-centrality parameters[17] can be transformed back to the f^2 scale as (Cohen, 1988, pp. 430):

$$f^2_{lower} = \frac{\lambda_{lower}}{u + v + 1} \tag{A.6}$$

$$f^2_{upper} = \frac{\lambda_{upper}}{u + v + 1} \tag{A.7}$$

Table D.2: Cohen's f^2 and 95% Confidence intervals

	Based on SB's R_1^2			Based on SB's R_2^2		
Variable	f^2	95% CI		f^2	95% CI	
Positive Mood	0.008	0.000	0.029	0.008	0.000	0.030
Future at 1 pm	0.023	0.005	0.055	0.025	0.006	0.058
Future at 3 pm	0.001	0.000	0.013	0.001	0.000	0.014
Positive mood× Future at 1 pm	0.022	0.004	0.053	0.024	0.005	0.056
Positive mood× Future at 3 pm	0.012	0.001	0.036	0.012	0.001	0.038

Notes: Columns 2-4 show Cohen's f^2 and respective 95% confidence intervals based on Snijders and Bosker's (1999) level-1 R_1^2. Columns 5-7 are based on Snijders and Bosker's (1999) level-2 R_2^2.

[17] For example, in Stata these can be calculated using the function `nF(df1,df2,np,f)`.