One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects

BY ZACHARIAS MANIADIS, FABIO TUFANO AND JOHN A. LIST

Web Appendix

I. Complementary Evidence

Several complementary experiments on anchoring effects reveal insights similar to our experiment, pointing to relatively low effects. Fortunately, while we were executing our original replication study reported in the paper, other researchers have explored the robustness of the original Ariely, Loewenstein and Prelec (2003) (ALP) results. In this informal review we include all studies that we are aware of, which satisfy the following criteria: they are published after ALP, they concern either standard consumption goods or hedonic experiences very similar to the ones used by ALP, and their structure allows direct comparisons to their study.[1]

We summarize them in Table 1. We use effect sizes in percentage terms for their simplicity in interpreting. In a more formal meta-analytic treatment standardized effects such as Hedges' g should be used (see Cooper, Hedges and Valentine, 2009). In the first two rows, we present the results of a clever study by Simonson and Drolet (2004), who performed a series of experiments with purely hypothetical decisions. In their first treatment they elicited the WTP and WTA for four products: a toaster, a cordless phone, a backpack, and a headphone radio. In the table we report the empirical results (the average of the median WTP and WTA for the four products) of their Study 1, which is directly comparable with ALP.[2] As can be seen, this study found moderately large anchoring effects for WTP, but no effects for WTA.[3]

---

[1] For instance, the anchoring question should be similar the analogous question in ALP. For the case of common market goods, this means that it generates a unique, unambiguous, and clearly uninformative anchor.

[2] Simonson and Drolet (2004) had an additional treatment where they asked subjects to explain how they would form their reservation prices, and other treatments where they asked subjects to imagine that they have already made the choice to sell the item, etc. On average, these treatments found small anchoring effects, but we do not consider them directly comparable to ALP.

[3] In a private communication, Aimee Drolet kindly noted that the follow-up studies to Study 1 and parts of their Study 2 included conditions which were the same as SD1 and SD2. Regarding Study 2, the relevant treatments reveal an effect size of 52% for WTP (N=29) and of -1% for WTA (N=83). The data we received from Drolet regarding the relevant follow-up studies to Study 1 reveal that for WTP the effect size was 65% (N=75) and for WTA it was -14% (N=26). Therefore, these sessions gave very similar results to SD1 and SD2. Drolet also noted that the results from their paper might not be fully comparable with ALP for several reasons. In particular, they did not use the BDM mechanism or any other monetary incentive, they collected the data in large survey runs, they did not show the physical products (but pictures), and the goods were 'utilitarian', rather than 'hedonic'. She also highlighted that, in consumer research, multiple replications - such as those reported in this footnote - serve as a means

TABLE 1—THE ANCHORING EFFECTS IN RECENT SIMILAR STUDIES

Authors*	Type of Study	Anchor		Results		Effect	N
		Low	High	Low	High	(%)	
SD1	WTA, goods	0-49	50-99	35	32.5	-7	53
SD2	WTP, goods	0-49	50-99	25	38	52	73
FLM1	WTP, goods	0-49	50-99	12.43	12.76	3	78
FLM2	WTA, goods	0-49	50-99	21.65	20.93	-3	79
FLM3	WTA, goods	0-49	50-99	17.88	20.5	15	79
BEJS	WTP, goods	0-49	50-99	50	73	45	116
ALL	WTA, goods	0-4.9	5.0-9.9	4.46	4.99	12	121
Tufano	WTA, liquid	0.05	1.25	1.33	1.39	4	116
MTL	WTA, sounds	0.10	0.50	0.196	0.252	28	76

*SD: Simonson and Drolet (2004); FLM: Fudenberg, Levine, and Maniadis (2012); BEJS: Bergman et al. (2010); ALL: Alevy, Laundry, and List (2011); Tufano: Tufano (2010); MTL: Our paper.

TABLE 2—THE ANCHORING EFFECTS IN ALP

Number of study	Type of study	Anchor		Results		Effect	N
		Low	High	Low	High	(%)	
1	WTP, goods	0-49	50-99	14.237	25.017	76	55
2	WTA, sounds	0.10	0.50	0.398	0.596	50	88
3	WTA, sounds	0-4.9	5.0-9.9	3.550	5.760	62	90
4	WTA, sounds	0.10	1.00	0.430	1.300	202	53
5	WTA, sounds	0.10	0.90	0.335	0.728	117	44

Notes: The amounts in the "Anchor" columns denote the size (or range) of the anchor price in the low and high treatment, in each study. In the "Results" columns, the amounts represent the average WTP or WTA (depending on the study) in each of the two treatments. "Effect" denotes the effect size, or the percentage change in the average outcome due to the treatment. In the last column, "N" denotes the total sample size of the specific 'high anchor' and 'low anchor' conditions. Study 2 had 132 subjects randomly assigned to six conditions, four of which we are reporting here. In expectation, these four conditions had 88 subjects, and we use this estimate. Study 5 involves multiple anchors: a different one in each round. Thus, we report the results from the first round, where subjects have been exposed to a unique anchor, which is the case which is comparable with all the other studies reported here.

In rows 3-6 we report the results of Fudenberg, Levine, and Maniadis (2012) and Bergman et al. (2010), who attempted to replicate Experiment 1 of ALP in two careful sets of experiments. They used similar goods to ALP and a similar incentive structure.[4] Fudenberg, Levine, and Maniadis (2012) elicited both WTP and WTA for

of reducing the noise in this style of survey evidence. Despite these reservations, we included the study in the analysis as we had done from the beginning in order to restrict our own "degrees of freedom" (see 'research bias' section of our paper).

[4] It should be noted that exact replication is difficult, since Experiment 1 of ALP was conducted after a class, which introduced relevant concepts for the experiment, and it is not clear what was explained during the class. In two of their three treatments, Fudenberg, Levine, and Maniadis (2012) adjusted for this by explaining why the Becker-Degroot-Marschak (Becker, Degroot,

the goods, and Bergman et al. (2010) elicited WTP only. Fudenberg, Levine, and Maniadis (2012) found very weak or no effects in all their treatments, while Bergman et al. (2012) found positive, but weaker effects than ALP.

Likewise, Alevy, Laundry, and List (2011) performed a field experiment, where subjects were recruited at a market for sport memorabilia. They used a jar of salted peanuts, which is a standard consumer product, and a protocol similar to ALP. Their results (row 7) suggest that consumers in markets do not have the anchoring tendencies observed in ALP.[5] Finally, Tufano (2010) explored the robustness of anchoring for hedonic experiences, rather than goods, by using a binary-choice elicitation procedure. In particular, he examined whether WTA to drink a bad-tasting liquid (similar to the one used by ALP, in a treatment not reported in the paper) was sensitive to an anchor. As the results in row 8 show, the anchoring manipulation had no effect.

In summary, most studies seem to find smaller percentage effects than ALP, and some find no effects. (For comparison, Table 2 reproduces the effects of ALP). The picture that emerges from the empirical evidence is that anchoring effects in economic valuations are real, since the effects are typically positive, but the magnitude of the effect should still be considered an open question. The claim that traditional economic models need radical revision seems premature based on the existing evidence alone. We would welcome more research, and believe that a formal meta-analysis would be very useful in further clarifying the issue.

II. Test of Homogeneity of Effect Sizes

Unfortunately the data of the ALP study are no longer available. Given the available information, we were still able to test for the homogeneity of the standardized effect sizes (Hedges' g) of the two studies (ALP and ours), focusing on the difference between the 10-cent anchor condition versus the 50-cent anchor condition. In particular, given the F-test reported by ALP and our data, we were able

and Marschak, 1964) mechanism is incentive compatible. In their third treatment, Fudenberg, Levine and Maniadis (2012) did not provide this explanation. Bergman et al. (2010) did not provide any additional explanations of the Becker-Degroot-Marschak mechanism. Moreover, Fudenberg, Levine, and Maniadis (2012) could not use bottles of wine, and they used an academic agenda/planner and a financial calculator instead.

[5] Alevy, Laundry and List (2011) also performed the same treatment for a pack of sports cards, for which no clear market price exists. We believe that this type of good goes beyond the spirit of "common market good" and we did not include this result in the table (for completeness, they found an effect size of -8%).

to test for the homogeneity of effect sizes, by calculating the statistic Q as follows (see Card, 2011, p. 185).

Let $w_i = 1/SE^2$ be the weight of study i (with SE being the 'Standard Error' of the effect size estimate for study i), ES_i be the effect size estimate (i.e., Hedges' g) for study i and $\overline{ES}$ be the weighted mean effect size. The significance test for homogeneity is based on the variable Q, which is given as follows:

$$Q = \sum_{i=1}^{2}[w_i\ (ES_i - \overline{ES})^2] = \sum_{i=1}^{2}\left[w_i\ \left(ES_i - \frac{\sum_{i=1}^{2} w_i ES_i}{\sum_{i=1}^{2} w_i}\right)^2\right]$$

$$= \sum_{i=1}^{2}(w_i ES_i^2) - \frac{(\sum_{i=1}^{2} w_i ES_i)^2}{\sum_{i=1}^{2} w_i} =$$

$$= 18.62594 - \frac{23.44237}{38.61957} = 4.396243,$$

Note that for ALP we obtained $ES = 2\sqrt{F_{(1,86)}/(n_1 + n_2)} = 0.935$ and $SE = 2\sqrt{[(n_1 + n_2)/(n_1 n_2)] + [ES^2/2(n_1+n_2)]}$ with n_1=44 and n_2=44 being the respective sample sizes of the low and high anchor treatments. For obtaining ALP's ES we used the formulas described in Cooper, Hedges and Valentine (2009) and Card (2011) for retrieving effect sizes for meta-analysis. For our study, we obtained ES and SE directly from the raw data (Hedges' $g = 0.258$). The Q statistic is distributed as χ^2 with degrees of freedom $df = k - 1$, where k is the number of studies considered. Our Q statistic is such that the hypothesis of homogeneous treatment effects is rejected at the 5-percent level ($p = 0.036$).

We need to note that ALP had 132 subjects in six conditions (i.e., 3 anchor conditions × 2 sequence conditions). By pooling together the two sequence conditions, the expected sample size in each relevant anchor condition was 44 subjects. Note that, after carefully following the description of ALP, it seems to us that the degrees of freedom in their F-test were something similar to 86 rather than 126, so there might be a typo in the reported analysis of ALP (p.82). So, in our analysis we have assumed their sample size was 44 in each condition, and our results should be viewed with extra caution.[6]

[6] We thank Dan Ariely for his kind effort to explain. However, in our understanding, we were still unable to reach a definitive interpretation of the F-test ALP used.

Moreover, it is important to emphasize that we do not suggest in the paper that a single failed replication provides conclusive condemning evidence against the original study. On the contrary, we hope that the insights of our theoretical framework will shift the focus away from statistical significance as the sole criterion. We also hope to encourage the practice of analyzing the totality of available evidence, in the spirit of meta-analysis that emphasizes treatment effects rather than p-values.

REFERENCES

Alevy, Jonathan, Craig Laundry, and John A. List. 2011. "Field Experiments of Anchoring of Economic Valuations". University of Alaska Working Paper 2011-12.

Ariely, Dan, George Loewenstein, and Drazen Prelec. 2003. "'Coherent Arbitrariness': Stable Demand Curves Without Stable Preferences." *Quarterly Journal of Economics* 118(1): 73-105.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak.1964. "Measuring Utility by a Single-Response Sequential Method". *Behavioral Science* 9(3): 226-232.

Bergman, Oscar, Tore Ellingsen, Magnus Johannesson, and Cicek Svensson. 2010. "Anchoring and Cognitive Ability." *Economics Letters* 107(1): 66-68.

Card, Noel A. 2011. *Applied Meta-Analysis for Social Science Research*. New York: Guilford Press.

Cooper, Harris, Larry V. Hedges, and Jeffrey C. Valentine. 2009. *Handbook of Research Synthesis and Meta-Analysis*. New York: Russell Sage Foundation.

Fudenberg, Drew, David K. Levine, and Zacharias Maniadis. 2012. "On the Robustness of Anchoring Effects in WTA and WTP Experiments." *AEJ Microeconomics* 4(2): 131-141.

Simonson, Itamar, and Aimee L. Drolet. 2004. "Anchoring Effects on Consumers' Willingness-to Pay and Willingness-to-Accept." *Journal of Consumer Research* 31(4): 681-690.

Tufano, Fabio. 2010. "Are "True" Preferences Revealed in Repeated Markets? An Experimental Demonstration of Context-Dependent Valuations." *Experimental Economics* 13(1): 1-13.