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

BY ZACHARIAS MANIADIS, FABIO TUFANO, AND JOHN A. LIST[‡]

Abstract: Some researchers have argued that anchoring in economic valuations casts doubt on the assumption of consistent and stable preferences. We present new evidence that explores the strength of certain anchoring results. We then present a theoretical framework that provides insights into why we should be cautious of initial empirical findings in general. The model importantly highlights that the rate of false positives depends not only on the observed significance level, but also on statistical power, research priors, and the number of scholars exploring the question. Importantly, a few independent replications dramatically increase the chances that the original finding is true.

JEL classification codes: D12 • C91

[†] This paper has been published in *The American Economic Review 2014, 104(1): 277–290. G*o to http://dx.doi.org/10.1257/aer.104.1.277 to visit the article page as well as for additional materials.

[‡] Maniadis: Economics Division, School of Social Sciences, University of Southampton, Southampton SO17 1BJ, UK (e-mail: Z.Maniadis@soton.ac.uk); Tufano: CeDEx, School of Economics, University of Nottingham, University Park, Nottingham NG7 2RD, UK (e-mail: fabio.tufano@nottingham.ac.uk); List: Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637 (e-mail: jlist@uchicago.edu). Part of the research was done when Maniadis was at Bocconi University and Tufano was at University of Milano-Bicocca. A previous, longer version of this paper was titled: "One Swallow Doesn't Make a Summer: How Economists (Mis-) Use Experimental Methods and Their Results." We are grateful to David K. Levine for his comments and invaluable support. We are also indebted to Colin F. Camerer, Robin Cubitt, Aimee Drolet, Jacob Goeree, Charles Plott, Uri Simonsohn, Roberto Weber, and three anonymous referees for their very detailed comments. We thank Dan Ariely, George Loewenstein, and Drazen Prelec for their comments and for providing us with useful experimental materials and data from Experiment 1 (but unfortunately the material—with the exception of the noise samples—and data from Experiment 2 are no longer available). We also thank the BLESS lab at the University of Bologna for their hospitality. Maniadis thanks Fabio Maccheroni for financial support. The authors obtained the informed consent of participants, who were volunteers recruited according to the rules of the University of Bologna. IRB approval for this type of nonmedical studies is not required at the University of Bologna, and there is no equivalent body devoted to approve, monitor, and review economic experimentation with human subjects. The authors have no financial or other material interests related to this research to disclose.

I. Introduction

Much of modern economics is predicated on the notion of durable and meaningful consumer preferences. However, in an influential study, Ariely, Loewenstein, and Prelec (2003)—henceforth, ALP—report that people's preferences are characterized by a very large degree of arbitrariness. In particular, they provide evidence that subjects' preferences for an array of goods and hedonic experiences are strongly affected by normatively irrelevant cues, namely, anchors.[1] Importantly, the ALP results suggest that arbitrariness of preferences is extremely strong, even in situations where we would expect traditional economic theory to have descriptive power.

Summing up the implications of their results, ALP argue (p. 102) that: "These results challenge the central premise of welfare economics that choices reveal true preferences (...). It is hard to make sense of our results without drawing a distinction between 'revealed' and 'true' preferences." The broader literature has followed, as these results have been received as strong evidence against traditional normative economics. If economic preferences are unstable and subject to the vagaries of the environment, then even the simplest choices may not be traced back to any optimization principles. In this case, a reevaluation of the fundamental building blocks of utility theory is warranted.

Our study begins by revisiting the seminal ALP results. In doing so, we present new experimental evidence of anchoring. We find a standardized effect size less than a third of what ALP found, and a descriptive percentage effect size about half of ALP's (the effect sizes are roughly zero before excluding outliers following ALP's methodology); and across several outcome measures we find no significant differences between treatments. Our data thus point to more modest effects than reported in the original study, although an important caveat is that the data we focus on here comes from one small sized experiment.

More importantly, however, one must recognize that many novel and surprising experimental results might not be robust—not because of falsification or something egregious, but merely due to the mechanics of the problem. Our main objective is to illustrate this basic point by means of a theoretical framework that provides insights into the mechanics of proper inference. The model highlights that we should be cautious when interpreting new experimental findings. For example, we show that the common benchmark of simply evaluating *p*-values when determining whether a result is a true association is flawed.

The common reliance on statistical significance as the sole criterion leads to an excessive number of false positives. In this sense, our theoretical model suggests that many surprising new empirical results are likely not recovering true associations. Our framework highlights that, at least in

[1] Decades of controlled experiments have also provided evidence that, in certain contexts, people might not always have predefined preferences, but "construct" them, when facing a choice problem (e.g., Lichtenstein and Slovic 2006, for a summary of the accumulated evidence).

principle, the decision about whether to call a finding noteworthy, or deserving of great attention, should be based on the estimated probability that the finding represents a true association, which follows directly from the observed *p*-value, the power of the design, the prior probability of the hypothesis, and the tolerance for false positives.

Beyond providing a paradigm in which to view new empirical results, our model indicates that we need a new approach for deciding which findings to highlight among the set of results from an empirical exercise. Since new and surprising results many times spur research meant to extend the original analysis, publishing false positives may have a costly effect in terms of misallocated resources. For the economics profession, the stakes are important because after the publication of such results it is possible to make the logical error that if the conventional economic model is rejected, then theory based on psychology is necessarily correct. In this way, entirely new research efforts may commence based on false insights. We should note that we use anchoring research as a lens to understand the problem, and we do not consider our evidence alone conclusive on the import of anchoring.

II. The ALP Investigation and Our New Evidence

Consider how a typical "anchoring" experiment is conducted. Subjects enter the experimental laboratory and, before starting the experimental task, they are exposed to a salient, but irrelevant, number. For example, a subject is asked to take the last two digits of her social security number and to turn those numbers into a dollar value (i.e., if your numbers are 12 then you provide a value of $12). Then, she is asked whether she would buy a certain item for the dollar value thus formed. Subsequently, the subject is asked the maximum amount of money she would pay for a certain item, commonly called "willingness to pay" (WTP).[2]

A. *The Original Investigation*

ALP's first experiment was conducted in a classroom, and the items involved were six common market goods: a cordless trackball, a cordless keyboard, a bottle of average wine, a bottle of rare wine, a design book, and a pack of Belgian chocolates. The other four experiments were conducted in a laboratory, and the relevant items were different durations of a high-pitched noise, which subjects heard through their headphones.

[2] Similarly, when the decision involves selling, rather than buying an item, or when the item is a "bad," willingness to accept (WTA) is elicited.

For illustration and quantitative comparison purposes, we summarize ALP's results in Table 1.[3] In Table 1 all numbers in the "Anchor" and "Results" columns are denominated in US dollars. We present ALP's Experiments 1–5, in the order in which they were presented in their paper, in rows 1 to 5. As can be seen, the smallest percentage effects are in the order of 50 percent, and the largest are approximately 200 percent.

ALP interpret these data as importantly refuting the foundations of economics. The literature has broadly concurred:[4] Fehr and Hoff (2011) interpret the ALP results as "striking" evidence that preferences are reference-dependent and suggest that a person might have multiple preference orderings, depending on the "social identity" invoked at the moment of a choice. For Kahneman and Sugden (2005), these results seem to imply that "individuals can be unsure what their preferences 'really' are" (p. 167), and they argue that stated WTP and WTA should not be used in policy valuation. For Beshears et al. (2008), the ALP results reflect the powerful influence of third-party manipulation on consumer choices, which casts doubt on whether these choices represent "normative preferences." Likewise, Bernheim and Rangel (2007, 2009) use the results as motivating evidence for proposing modifications of the traditional, revealed preference–based welfare analysis.

B. The Replication Study

As Levitt and List (2009) discuss, there are at least three levels at which replication can take place. The first of these entails reanalyzing the original data generated by an experiment in order to corroborate the results. A second conception of replication refers to implementing an experiment under a similar protocol to the original experiment to verify whether similar findings can be obtained using different subjects. The third notion of replication (the most general one) pertains to the employment of a new research design with the purpose of testing the hypotheses of the first study.[5] Our primary focus here and in the theoretical model below is on the second notion of replication. Yet our fundamental points apply equally to the third replication concept. As Table 1 shows, ALP's main body of evidence concerns the "annoying sounds" treatment.[6]

As ALP argue, these hedonic goods are particularly appropriate for testing economic valuation, since they involve a very simple experience, a sample of which can be readily provided without satiating the subjects. Moreover, a market price does not exist, and neither do outside-the-lab substitutes. For

[3] In Experiments 1 and 3 the anchor was a random price between \$0 and \$99 (or between \$0 and \$9.9). We follow the natural convention of considering as a "low" anchor one that belongs to the lower half of this support. Moreover, in Experiment 4, the authors report the mean WTA for each of the three durations of the sound, and we have taken the average of the three means.

[4] All of the following authors base their arguments on a large set of evidence, and not only on ALP's results. However, as we shall argue, the ALP results are singularly important because they provide extremely strong evidence in favor of arbitrary preferences, in some of the most favorable environments conceivable for traditional economic theory.

[5] This taxonomy is in agreement with the one proposed by Cartwright (1991). Hunter (2001) also defines threelevels of replication, but the first level he suggests concerns the exact repetition of the original study in all dimensions, rather than the routine checking of the original study. His other two levels are largely equivalent with ours.

[6] ALP's cleverly designed study included six experiments. The five experiments that we are presenting in Table 1, and an additional one that did not involve WTA or WTP in monetary units and, therefore, is not comparable with the other studies. This experiment showed that anchoring matters even when people express their preferences directly in terms of substituting one hedonic experience for another, and not only when they are substituting one hedonic experience for money.

these reasons, it is possible that the large effects found in ALP, for the simple hedonic experiences, represent the "true" effects of anchoring, net of all possible distorting factors.

TABLE 1—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	132
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 sample size of the given study. 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 order to examine this hypothesis, we replicated Experiment 2 of ALP as closely as possible. Our experiment took place at the BLESS (Bologna Laboratory for Experiments in Social Science) lab of the University of Bologna (Italy). It consisted of six experimental sessions programmed and conducted in a computerized environment using z-Tree (Fischbacher 2007). A total of 116 subjects, recruited and randomly invited through ORSEE, the Online Recruitment System for Economic Experiments (Greiner 2004), attended our experiment. Participants were students of the University of Bologna drawn from a range of academic disciplines.[7]

In each session, subjects entered the lab and were asked to put on their headphones, and to keep them on for the duration of the experiment. Then, they listened to a sample of 30 seconds of the annoying sound, which was the same as ALP's sound. The anchoring question followed: each subject was asked whether she/he would be willing to repeat the same experience for a given amount of money (the anchor).[8] Subsequently, subjects participated in nine experimental rounds. In each round, subjects were asked to state the minimum amount of money (i.e., WTA) for which they would be willing to hear the same sound, with certain duration.

[7] Eighteen participants (i.e., 15.52 percent of our sample) had attended at least one different experiment before taking part in our own (14 out of those 18 subjects had participated in only one experiment). Our sample featured a narrow majority of men (56.90 percent), while consisting almost exclusively of Italian nationals (95.69 percent). Subjects received a show-up fee of 5 euros, plus their earnings from the experiment. Average payoffs per subject were equal to 7.65 euros, including the show-up fee.

[8] As in ALP, in our experiment the anchoring question was not incentivized.

Then, the Becker-DeGroot-Marschak (1964) mechanism was implemented to ensure incentive compatibility.[9] As in ALP, we varied the anchoring manipulation and the sequence in which the sound durations appeared. The anchoring manipulation involved an anchor either of 10 cents, of 50 cents, or no anchor at all. We crossed these three treatments with two sequences: an increasing sequence and a decreasing one.[10] In the increasing (decreasing) sequence, the first round had a sound of 10 (60) seconds, the second of 30 (30) seconds, and the third of 60 (10) seconds. This triplet was repeated three times, for a total of nine rounds.

The results are summarized in Figure 1.[11] For the increasing-sequence (decreasing-sequence) treatment, the average stated WTA in the 10-cent anchor condition, the no-anchor condition, and the 50-cent anchor condition was 23.05 (16.50), 25.16 (20.37), and 28.79 (21.61), respectively. For both sequences pooled, the average stated WTA was equal to 19.60, 22.76, and 25.20, respectively. Using the pooled data from the two sequences, and comparing only the 10-cent and the 50-cent anchor conditions, we find that our descriptive percentage effect size[12] is equal to 28.57 percent, about half of what ALP found.[13] In terms of the corresponding standardized effect size, we find that Hedges' g = 0.258, with confidence intervals (−0.19, 0.71). The p-value of the two-sided t-test for differences in the average WTA of each subject, across the 10-cent and 50-cent anchor treatments, was equal to 0.253.[14] Moreover, the average payoffs per subject did not differ significantly in the 10-cent anchor condition (2.70 euros) from the 50-cent anchor condition (2.62 euros) [p = 0.746, two-sided t-test]. A similar nonsignificant difference appears in the average number of listened sounds per subject [6.750 in the 10-cent treatment versus 6.275 in the 50-cent treatment, p = 0.392, two-sided t-test]. Our experiment thus points to considerably lower point estimates than the original study.[15]

[9] In particular, in each round a random price was drawn by the computer. The number was drawn from a triangular distribution with mode zero, and maximum 100 cents. Subjects were shown a picture of this distribution. If their stated WTA was lower than this price, they would receive the computer's random price and listen to the sound. If their stated WTA exceeded this price, they would neither receive any money, nor listen to the sound. Subjects were told that this process ensured that it was in their best interest to state their true minimum for listening to the sound.

[10] All six experimental sessions had 20 subjects, except the 10-cent anchor condition in the increasing-sequence treatment and the 10-cent anchor condition in the decreasing-sequence treatment which had 17 and 19 participants, respectively (due to subjects that did not show up).

[11] The scales of the axes were chosen purposely to increase visual comparability with ALP's Figure I.

[12] Effect size expressed as a percentage change in the average outcome due to the treatment is an intuitive, easily accessible measure. Moreover, it can be calculated for ALP's study while not relying on the accuracy of the reported statistical tests. We refer to the percentage effect size for descriptive purposes only.

[13] If our interpretation of the F-test reported in the ALP study is correct, the analogous effect size for their condition was 0.935, with confidence intervals (0.50, 1.38). So there seems to be significant overlap in the confidence intervals. For tentative evidence that our 10-cent versus 50-cent treatment effects are meaningfully different from ALP, see the online Appendix, Section II.

[14] With respect to the third treatment, the p-value of the two-sided t-test for the 10-cent condition versus the no-anchor condition is equal to 0.431. The p-value of the two-sided t-test for the 50-cent condition versus the noanchor condition is equal to 0.610. It is important to underline the fact that although we use t-tests for comparability with ALP, using the nonparametric Wilcoxon rank-sum test gives the same results. In particular, no difference is significant even at the 10 percent level. Moreover, it should be emphasized that following ALP, for the purpose of statistical analysis we truncated responses greater than 100 cents to 101 cents. Using the non-truncated data, the average WTA in the 10-cent anchor condition is greater than in the 50-cent anchor condition (27.92 versus 27.81).

[15] This statement demands important qualifications, however. If one considers simply the difference of means across our study and ALP, different inference can emerge. This is because our confidence interval for the difference of means across treatments (10-cent anchor versus 50-cent anchor) is (−4.085, 15.288). While this interval does not include the point estimate reported in ALP (19.78), we believe that their confidence interval has significant overlap with ours. If one takes their reported F-tests as given in their paper, then the ALP confidence interval for the difference of means is (10.85, 28.7). A potential caveat, however, is that this interval is generated using a test reported as F (1,126), which cannot be possible since they, in expectation, only had 88 subjects in these two treatments. Without having their original data (which are lost), it is difficult to make valid statistical comparisons without knowledge of important factors (such as whether there is dependence in the data, how many observations were actually in each treatment, etc.). We conclude, however, that it must be the case that the confidence intervals for the difference of means across treatments between the two studies overlap, as the confidence intervals for Hedges' g probably also do. Thus, we urge caveat lector when making inference solely across this one experiment and ALP. This is why we

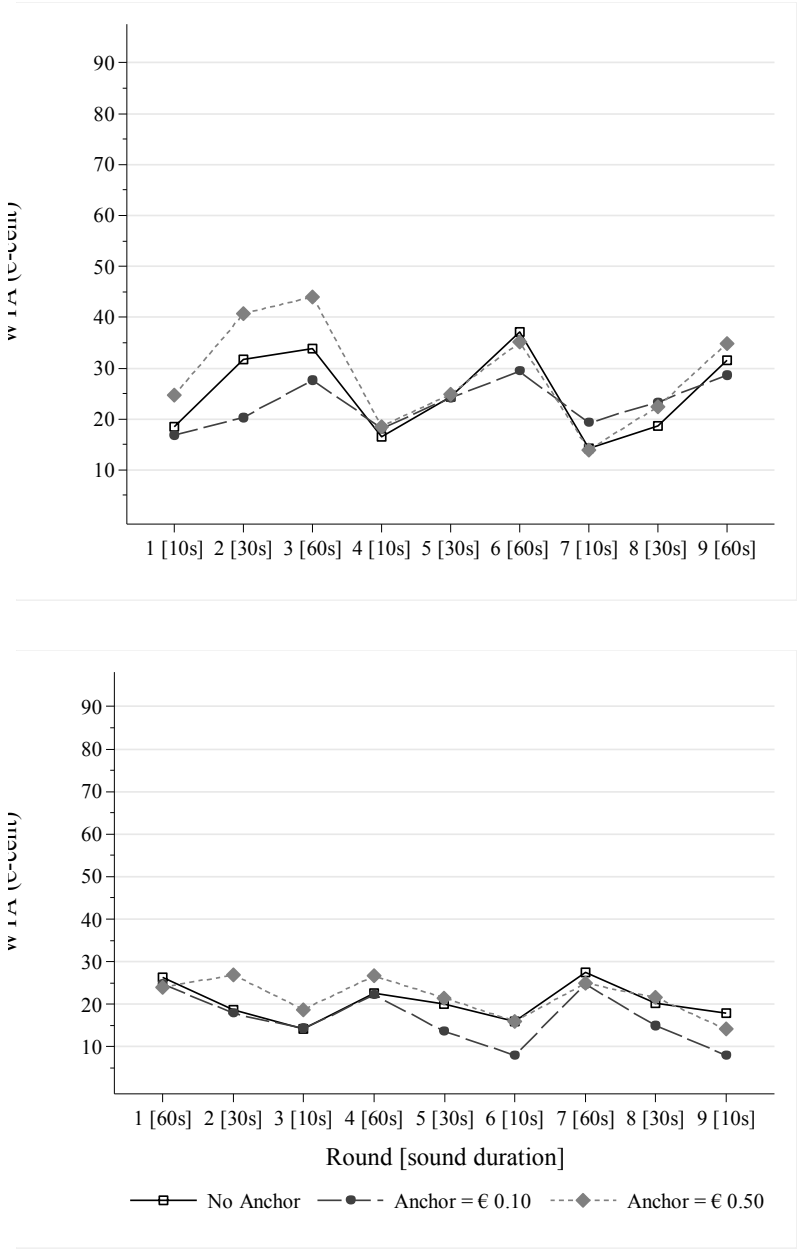

FIGURE 1. RESULTS FROM THE INCREASING (*top panel*) AND DECREASING (*bottom panel*) SEQUENCE

There are several factors that might account for the fact that our experiment found different results than ALP. Our subjects were students enrolled at the University of Bologna, while ALP had MIT undergraduates. There was more than a decade of difference between the two studies. Moreover, the z-Tree experimental interface is different than the one ALP used (which unfortunately was not available to us). Furthermore, it is possible that our results stem from a very unlucky draw and are not representative of the true underlying phenomenon. Finally, and most importantly, we should stress that our data are from one small scale experiment.

consider our experiment as only one small piece of evidence and consider data from multiple studies below to motivate our more general contribution, the inferential framework.

All these are important considerations, but by combining our work with other complementary evidence that found (if any) much smaller effects than ALP (e.g., Fudenberg, Levine, and Maniadis 2012; Alevy, Landry, and List 2011; Bergman et al. 2010; Tufano 2010; Simonson and Drolet 2004—see the online Appendix for details), we also believe that the ALP point estimates potentially overestimate anchoring effects.[16] In summary, the picture that emerges from the totality of the empirical evidence is that anchoring effects in economic valuations are real, since the effects are typically positive, but the magnitude of the effect in economically relevant environments remains an open question. The claim that traditional economic models need radical revision seems premature based on this evidence alone. We would welcome more research.[17]

III. The Importance of Replication

Why do we believe that replication exercises like the one reported here are important? Viewed through the lens of a simple theoretical framework, we show that by their very nature, studies that report strong and highly surprising findings are most likely not revealing true associations—not due to researcher malfeasance, rather because of the underlying mechanics of the methods. Such a model, which we describe now, also provides insights into factors that exacerbate or attenuate such effects.[18]

A. The Basic Framework

Building on a formal methodology developed in the health sciences literature (Wacholder et al. 2004; Ioannidis 2005; Moonesinghe, Khoury, and Janssens 2007), we let n be the number of scientific associations that are being examined in a specific research field. Let π represent the fraction of these that are true associations. We use α as the typical significance level in the field (usually $\alpha = 0.05$) and $1 - \beta$ denotes the typical power of an experimental design in this field.[19] We are interested

[16] Our approach is to consider treatment effects as our focus, not statistical significance. This is in the spirit of meta-analysis, and in the online Appendix we present an informal review of our study and other relevant studies that performed experiments very close to that of ALP. Dan Ariely and George Loewenstein noted that anchoring effects have been found also for works of art, housing prices, and judicial compensation decisions (Beggs and Graddy 2009; Simonsohn and Loewenstein 2006; Sunstein et al. 2002). These results are of independent importance, and they concern highly complex goods, for which standard utility theory might arguably not have high explanatory power. We still believe that quantifying the average anchoring effects in the class of experiments that closely follow ALP, using the totality of the relevant evidence, is important for determining the economic significance of anchoring. In fact, we agree that if the ALP treatment effects are representative of the average effects for simple consumer goods and hedonic experiences, a radical reevaluation of consumer theory might be in order.

[17] Interestingly, we seem to get both a lower percentage anchoring effect and a lower percentage effect of increasing the sound from 10 to 60 seconds (83.6 percent in our study, 133.7 percent in ALP). The size of the "anchoring effect" relative to the "duration effect" is not very different across the two studies. Moreover, we find some suggestive evidence of repetition, or learning, effects. These also deserve further scrutiny.

[18] An astute reviewer raised other issues that should be of concern to experimentalists, including (i) representativeness of the population (most studies use undergraduate students as subjects and generalize to the population of interest) and (ii) multiple testing. The interested reader should see Levitt and List (2007) for a recent discussion of (i) and Romano and Wolf (2005) and Romano, Shaikh, and Wolf (2008) for good discussions of the latter.

[19] For simplicity, we assume that the research practices in the field are relatively homogeneous and, therefore, the choices of sample size can be captured by this single level of power. It is straightforward to notice that the arguments of our model can be made on the basis of a single study (so n plays only an expositional role). Therefore, this assumption is innocuous.

in the probability that a declaration of a research finding, made upon reaching statistical significance, is true. We denote this as the Post-Study Probability (*PSP*).

This probability depends on the mechanics of statistical inference and can be found as follows: of the n associations, $\pi \cdot n$ associations will be true, and $(1-\pi) \cdot n$ will be false. Among the true ones, $(1-\beta) \cdot \pi \cdot n$ will be declared true, and among the false associations, $\alpha(1-\pi) \cdot n$ will be declared true even though they are false (i.e., they are false positives). The *PSP* is equal to the number of true associations which are declared true divided by the number of all associations which are declared true:

$$(1) \qquad PSP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}.$$

Of first note is that the *PSP* from equation (1) is increasing in the prior π: the higher the priors about the existence of a phenomenon the weaker the evidence that is needed to substantiate it. With respect to the role of sample size, the derivative of *PSP* from equation (1) with respect to power $(1-\beta)$ is positive. This means that *PSP* is a positive function of sample size via the power of the experimental study.[20]

B. Researchers' Competition

Now assume that there are k independent researchers working simultaneously on each of n associations in a specific field.[21] Let each researcher's study have the same power $(1-\beta)$. The probability that at least one of the k researchers will declare a true association as true is $(1-\beta^k)$. Likewise, the probability that a false relationship is declared true by at least one of k researchers is $1-(1-\alpha)^k$. Accordingly, out of the $\pi \cdot n$ true relationships, $(1-\beta^k) \cdot \pi \cdot n$ will be declared true. And, of the $(1-\pi) \cdot n$ false relationships, $[1-(1-\alpha)^k]\,(1-\pi) \cdot n$ will be declared (mistakenly) true, or will be false positives. Hence, the *PSP* in the presence of competition by independent researchers (PSP^{Comp}) is equal to

$$(2) \qquad PSP^{Comp} = \frac{(1-\beta^k)\pi}{(1-\beta^k)\pi + [1-(1-\alpha)^k](1-\pi)} \quad .$$

This is decreasing in k as long as $(1-\beta)>\alpha$. Since power is typically greater than the significance level, equation (2) reveals that as the number of investigators examining a typical phenomenon increases (competition intensifies), the probability that an *initial* declared research finding is true decreases.[22]

[20] For more details about this relationship, see Maniadis, Tufano, and List (2013).
[21] Note that we use the term "researcher" referring indifferently to both a single researcher and a research team.
[22] Of course, it is also the case that competition will tend to increase the number of replications, so its effect in the medium term could be to increase the average reliability of research findings—under the assumption of no editorial reluctance in publishing replication studies. Here we focus on proper interpretation of initial findings.

C. Research Bias

The aforementioned analysis implicitly assumed that the degrees of freedom in doing research do not play a role. As the recent paper by Simmons, Nelson, and Simonsohn (2011) emphasizes, there is by now a large literature that shows the existence of self-serving biases in interpreting ambiguous evidence and reaching defensible conclusions that satisfy the research objectives (Babcock and Loewenstein 1997; Dawson, Gilovich, and Regan 2002).[23]

We define the "bias" u as "the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced" (Ioannidis 2005, p. 697).[24] In particular, the parameter u denotes the fraction of all cases where a positive research finding has been declared because of the bias, although it should not have been declared. Recall that if n relationships are tested in the field, $\pi \cdot n$ will be true and $(1-\pi) \cdot n$ will be false. Of the true relationships, $(1-\beta) \cdot \pi \cdot n$ will be declared true, and $\beta \cdot \pi \cdot n$ will be declared false due to pure noise.

When in addition to noise there is research bias, a fraction u of the latter will be declared true because of the bias, so that $(1-\beta) \cdot \pi \cdot n + u \cdot \beta \cdot \pi \cdot n$ will be declared true. Using analogous reasoning, one can verify that out of the false associations, $\alpha \cdot (1-\pi) \cdot n + u \cdot (1-\alpha) \cdot (1-\pi) \cdot n$ will be declared true. In this case, *PSP* with bias (PSP^{Bias}) is equal to

$$PSP^{Bias} = \frac{(1-\beta)\pi + \beta\pi u}{(1-\beta)\pi + \beta\pi u + [\alpha + (1-\alpha)u](1-\pi)} \quad . \tag{3}$$

The derivative of PSP^{Bias} with respect to u is negative when $\pi (1-\pi) [\alpha+\beta-1]$ is smaller than zero which implies that $\alpha<1-\beta$, which, as we argued, is typically true.[25] Therefore, as equation (3) shows, an increase in research bias decreases the probability that a published research finding corresponds to the truth.[26]

[23] For a related discussion, we direct the interested reader to Dufwenberg (forthcoming), who discusses the biased nature of the publication process toward accepting studies that report surprising results and proposes an innovative solution.

[24] Notice that the bias, as defined above, differs from chance variability, which can lead to false positive findings, even though a study was correctly conducted in each and every of its aspects. Also note that our concept of bias here does not refer to biased beliefs regarding the experimental results (perhaps due to ideology). The bias that we refer to is purely behavioral and concerns how the study is conducted. It differs from the bias in beliefs in nontrivial ways: for example, it could depend—perhaps subconsciously—on the incentives for publication, while the bias in beliefs should not depend on it. It is worth exploring how the other type of bias (in beliefs) operates: for example, it is possible that having opposing biases in beliefs could be quite helpful, in the sense that they increase replications that falsify initial findings. It might also lead to nontrivial results if a refereeing process with biased beliefs is introduced. However, these analyses fall outside the scope of this short article.

[25] A reviewer correctly noticed that it is possible that the bias could operate in an asymmetric manner. Our general result—that *PSP* declines in the presence of the bias—is robust even to the case where the bias operates much more on the nonsignificant true associations than on the nonsignificant false ones. However, the decline in the *PSP* would not be as large as in the symmetric case, depending on the level of the asymmetry.

[26] In principle, the behavioral bias could exist in both directions, in the sense that a researcher might prefer to fail to get an effect, so that the net effect of behavioral biases might not be obvious. However, we believe that complicating the model in order to capture opposing biases would not change the direction of the effect. The reason is that there is a fundamental asymmetry in exploratory research: it is much more common that an experimentalist would like to discover an effect, rather than to reveal its absence. Second and third generation studies, of course, do not have the same properties.

D. Discussion

How can our model help us to determine whether some given initial findings are indeed true effects? For this, we need to specify a prior π, the power of the design $(1-\beta)$, the number of independent researchers k, and then use equation (2) above to calculate the *PSP*. Since it is difficult to pinpoint these variables exactly, in a thought experiment we consider various combinations of the variables to provide meaningful ranges.[27]

TABLE 2—THE PSP ESTIMATES AS A FUNCTION OF PRIOR PROBABILITY(π), POWER, AND COMPETITION (k) OF THE STUDY

	Power=0.80				Power=0.50				Power=0.20			
π	k=1	k=5	k=15	k=50	k=1	k=5	k=15	k=50	k=1	k=5	k=15	k=50
	PSP											
0.01	**0.14**	**0.04**	**0.02**	**0.01**	**0.09**	**0.04**	**0.02**	**0.01**	**0.04**	**0.03**	**0.02**	**0.01**
0.02	**0.25**	**0.08**	**0.04**	**0.02**	**0.17**	**0.08**	**0.04**	**0.02**	**0.08**	**0.06**	**0.04**	**0.02**
0.05	**0.46**	**0.19**	**0.09**	**0.05**	**0.34**	**0.18**	**0.09**	**0.05**	**0.17**	**0.14**	**0.09**	**0.05**
0.10	0.64	**0.33**	**0.17**	**0.11**	0.53	**0.32**	**0.17**	**0.11**	**0.31**	**0.25**	**0.17**	**0.11**
0.20	0.80	0.52	**0.32**	**0.21**	0.71	0.52	**0.32**	**0.21**	0.50	**0.43**	**0.31**	**0.21**
0.35	0.90	0.70	0.50	**0.37**	0.84	0.70	0.50	**0.37**	0.68	0.62	**0.49**	**0.37**
0.55	0.95	0.84	0.69	0.57	0.92	0.84	0.69	0.57	0.83	0.78	0.69	0.57

TABLE 3—THE PSP ESTIMATES AS A FUNCTION OF PRIOR PROBABILITY (π), POWER AND BIAS (u) OF THE STUDY

	Power=0.80				Power=0.50				Power=0.20			
π	u=0	u=0.10	u=0.25	u=0.50	u=0	u=0.10	u=0.25	u=0.50	u=0	u=0.10	u=0.25	u=0.50
	PSP											
0.01	**0.14**	**0.05**	**0.03**	**0.02**	**0.09**	**0.04**	**0.02**	**0.01**	**0.04**	**0.02**	**0.01**	**0.01**
0.02	**0.25**	**0.10**	**0.06**	**0.03**	**0.17**	**0.07**	**0.04**	**0.03**	**0.08**	**0.04**	**0.03**	**0.02**
0.05	**0.46**	**0.23**	**0.13**	**0.08**	**0.34**	**0.17**	**0.10**	**0.07**	**0.17**	**0.09**	**0.07**	**0.06**
0.10	0.64	**0.39**	**0.25**	**0.16**	0.53	**0.30**	**0.19**	**0.14**	**0.31**	**0.18**	**0.13**	**0.11**
0.20	0.80	0.59	**0.43**	**0.30**	0.71	**0.49**	**0.35**	**0.26**	0.50	**0.33**	**0.26**	**0.22**
0.35	0.90	0.75	0.61	**0.48**	0.84	0.67	0.54	**0.43**	0.68	0.51	**0.43**	**0.38**
0.55	0.95	0.87	0.78	0.68	0.92	0.83	0.73	0.64	0.83	0.70	0.63	0.58

Table 2 presents such combinations, and the resulting *PSP*.[28] Similarly, in Table 3 we specify u, the research specific bias, leaving aside the number of competing researchers, and use equation (3) to calculate the relevant *PSP*. Tables 2 and 3 convey a strong message: we should be very careful not to make strong inference from a first, surprising research finding. Tables 2 and 3 also indicate that it is not unlikely that the *PSP* after the initial study is less than *0.5*, as several plausible parameter combinations yield this result (presented by bold fonts). In addition, one important feature left on the sidelines in this analysis is the possible interaction between competition and bias that may lead to

[27] Notice that the combinations of variables (prior, design power, etc.) that we consider should be understood as related to novel and surprising findings, hence, for instance, we do not consider large priors. We welcome empirical studies for estimating the values of those variables in our discipline.

[28] Note that for all our tables we assume that $\alpha=0.05$.

interpreting the above estimates as an upper bound. Summing up, there are several factors that might lead to false positives, and many of them stem from the incentives of the current academic system (see Oswald 2007; Glaeser 2008; Young, Ioannidis, and Al-Ubaydli 2008).

We believe that the best solution to the inference problem is replications. Our framework suggests that a little replication can lead to far reaching benefits. To illustrate, we consider several specifications of our model, each with a different number of competing researchers k (see also Maniadis, Tufano, and List 2013; Moonesinghe, Khoury, and Janssens 2007). Then, we calculate the probability that anywhere from zero (only the original study) to three replication studies (so a total of four studies) find a significant result, given that the relationship is true and given that it is false. Then, we derive the *PSP* the usual way, as the fraction of the true associations declared true over all associations declared true, for each level of replication. The results reported in Table 4 (for the case of *k=10*) show that with just two independent replications of the initial finding, the improvement in *PSP* is dramatic. Indeed, for studies that report "surprising" results—those that have low π values—the *PSP* increases more than threefold upon a couple of replications.

TABLE 4—THE PSP ESTIMATES AS A FUNCTION OF PRIOR PROBABILITY (π), POWER AND NUMBER OF REPLICATIONS (i)

	Power = 0.80				Power = 0.50				Power = 0.20			
π	i=0	i=1	i=2	i=3	i=0	i=1	i=2	i=3	i=0	i=1	i=2	i=3
	PSP											
0.01	**0.02**	**0.10**	**0.47**	0.91	**0.02**	**0.10**	**0.45**	0.89	**0.02**	**0.07**	**0.22**	0.54
0.02	**0.05**	**0.19**	0.64	0.95	**0.05**	**0.19**	0.63	0.94	**0.04**	**0.13**	**0.36**	0.71
0.05	**0.12**	**0.38**	0.82	0.98	**0.12**	**0.38**	0.81	0.98	**0.10**	**0.28**	0.60	0.86
0.10	**0.22**	0.56	0.91	0.99	**0.22**	0.56	0.90	0.99	**0.20**	**0.45**	0.76	0.93
0.20	**0.38**	0.74	0.96	1.00	**0.38**	0.74	0.95	1.00	**0.36**	0.64	0.88	0.97
0.35	0.57	0.86	0.98	1.00	0.57	0.86	0.98	1.00	0.55	0.80	0.94	0.98
0.55	0.75	0.93	0.99	1.00	0.75	0.93	0.99	1.00	0.73	0.90	0.97	0.99

IV. Conclusions

Economists and policymakers rely on utilitarian analysis as a crucial step in guiding policymaking. Within the United States alone, every economically significant[29] proposed rulemaking must undergo a formal benefit cost analysis. In many cases, the economic analysis critically relies on empirical measures derived from experimental or survey methods. The stakes are heightened further when such empirical methods are also used to test the foundations of theoretical models; in those cases where extant theory is rejected, profound paradigmatic changes can ensue. In both instances, if the original empirical findings are untrue, the social cost can be quite high.

[29] Based on the historical standard introduced as part of President Reagan's Executive Order 12291 and maintained currently under Executive Order 13563, an "economically significant" policy is that which has an annual effect on the US economy of $100 million or more, or adversely impacts the economy, a sector of the economy, productivity, public health, the environment, or a host of other relevant facets of the US economy.

Such considerations are especially important in light of recent findings due to ALP. In this study we provide new experimental evidence of anchoring in economic valuations. We proceed to provide a theoretical basis showing why replication is so important for science. Combined, our theory and empirical work highlights that we should be cautious when interpreting new empirical findings. For example, in the model we show that the common benchmark of simply evaluating *p*-values when determining whether a result is a true association is flawed. Two other considerations – the statistical power of the test and the fraction of tested hypotheses that are true associations – are key factors to consider when making appropriate inference. The common reliance on statistical significance as the sole criterion leads to an excessive number of false positives. The problem is exacerbated as journals make "surprise" or counterintuitive results necessary for publication. But by their very nature such studies are most likely not revealing true associations – not because of researcher malfeasance, merely because of the underlying mechanics of the methods.

While this message is pessimistic, there is good news: our analysis shows that a few independent replications dramatically increase the chances that the original finding is true. As Fisher (1935) emphasized, a cornerstone of the experimental science is replication. Inference from empirical exercises could be advanced considerably if scholars begin to adopt concrete requirements to enhance the replicability of results, as for instance starting to actively encourage replications within a given study.[30] We trust that our own estimates of anchoring, which are from a small sample and have wide confidence intervals, will stimulate replication. In fact, as the saying goes "one swallow doesn't make a summer".[31]

REFERENCES

Alevy, Jonathan, Craig Landry, and John A. List. 2011. "Field Experiments of Anchoring of Economic Valuations." University of Alaska Working Paper 2011–02.

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

Babcock, Linda, and George Loewenstein. 1997. "Explaining Bargaining Impasse: The Role of Self-Serving Biases." *Journal of Economic Perspectives* 11 (1): 109–26.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak. 1964. "Measuring Utility by a Single-Response Sequential Method." *Behavioral Science* 9 (3): 226–32.

Beggs, Alan, and Kathryn Graddy. 2009. "Anchoring Effects: Evidence from Art Auctions." *American Economic Review* 99 (3): 1027–39.

[30] We do not regard internal replication as a sufficient requirement to establish the robustness of the original results. However, we envisage it as a first step in the right direction (see also Simmons, Nelson, and Simonsohn 2011).

[31] As an instance of why we should be cautious, our experiment does not allow us to declare a failed replication in the strong sense suggested by Simonsohn (2013), who proposed that a replication should use a sample size two and a half times the original.

Bergman, Oscar, Tore Ellingsen, Magnus Johannesson, and Cicek Svensson. 2010. "Anchoring and Cognitive Ability." *Economics Letters* 107 (1): 66–8.

Bernheim, B. Douglas, and Antonio Rangel. 2007. "Toward Choice-Theoretic Foundations for Behavioral Welfare Economics." *American Economic Review* 97 (2): 464–70.

Bernheim, B. Douglas, and Antonio Rangel. 2009. "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics." *Quarterly Journal of Economics* 124 (1): 51–104.

Beshears, John, James J. Choi, David Laibson, and Brigitte C. Madrian. 2008. "How Are Preferences Revealed?" *Journal of Public Economics* 92 (8–9): 1787–94.

Cartwright, Nancy. 1991. "Replicability, Reproducibility, and Robustness: Comments." *History of Political Economy* 23 (1): 143–55.

Dawson, Erica, Thomas Gilovich, and Dennis Regan. 2002. "Motivated Reasoning and Performance on the Wason Selection Task." *Personality and Social Psychology* 28 (10): 1379–87.

Dufwenberg, Martin. Forthcoming. "Maxims for Experimenters." In *Methods of Modern Experimental Economics*, edited by G. Frechette and A. Schotter. New York: Oxford University Press.

Fehr, Ernst, and Karla Hoff. 2011. "Introduction: Tastes, Castes and Culture: The Influence of Society on Preferences." *Economic Journal* 121 (556): F396–412.

Fisher, Ronald A. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10 (2): 171–78.

Fudenberg, Drew, David K. Levine, and Zacharias Maniadis. 2012. "On the Robustness of Anchoring Effects in WTP and WTA Experiments." *American Economic Journal: Microeconomics* 4 (2): 131–45.

Glaeser, Edward L. 2008. "Researcher Incentives and Empirical Methods." In *The Foundations of Positive and Normative Economics*, edited by Andrew Caplin and Andrew Schotter, 300–19. New York: Oxford University Press.

Greiner, Ben. 2004. "An Online Recruitment System for Economic Experiments." In *Forschung und Wissenschaftliches Rechnen 2003* (*GWDG Bericht 63*), edited by Kurt Kremer and Volker Macho, 79–93. Göttingen: Gesellschaft für Wissenschaftliche Datenverarbeitung.

Hunter, John E. 2001. "The Desperate Need for Replications." *Journal of Consumer Research* 28 (1): 149–58.

Ioannidis, John. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): 1418–22.

Kahneman, Daniel, and Robert Sugden. 2005. "Experienced Utility as a Standard of Policy Evaluation." *Environmental and Resource Economics* 32 (1): 161–81.

Levitt, Steven D., and John A. List. 2007. “Viewpoint: On the Generalizability of Lab Behaviour to the Field.” *Canadian Journal of Economics* 40 (2): 347–70.

Levitt, Steven D., and John A. List. 2009. “Field Experiments in Economics: The Past, the Present, and the Future.” *European Economic Review* 53 (1): 1–18.

Lichtenstein, Sarah, and Paul Slovic. 2006. *The Construction of Preference*. New York: Cambridge University Press.

Maniadis, Zacharias, Fabio Tufano, and John A. List. 2013. “The Beauty of Replication: How Economists Should Use Experimental Methods and Their Results.” Unpublished.

Maniadis, Zacharias, Fabio Tufano, and John A. List. 2014. “One Swallow Doesn’t Make a Summer: New Evidence on Anchoring Effects: Dataset.” *American Economic Review*. http://dx.doi.org/10.1257/aer.104.1.277.

Moonesinghe, Ramal, Muin J. Khoury, and Cecile J.W. Janssens. 2007. “Most Published Research Findings Are False-But a Little Replication Goes a Long Way.” *PLoS Medicine* 4 (2): 218–21.

Oswald, Andrew J. 2007. “An Examination of the Reliability of Prestigious Scholarly Journals: Evidence and Implications for Decision-Makers.” *Economica* 74 (293): 21–31.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf. 2008. “Formalized Data Snooping Based on Generalized Error Rates.” *Econometric Theory* 24 (2): 404–47.

Romano, Joseph P., and Michael Wolf. 2005. “Stepwise Multiple Testing as Formalized Data Snooping.” *Econometrica* 73 (4): 1237–82.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology: Undisclosed

Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22 (11): 1359–66.

Simonsohn, Uri. 2013. “Evaluating Replication Results.” SSRN Working Paper 2259879.

Simonsohn, Uri, and George Loewenstein. 2006. “Mistake #37: The Effect of Previously Encountered Prices on Current Housing Demand.” *Economic Journal* 116 (508): 175–99.

Simonson, Itamar, and Aimee Drolet. 2004. “Anchoring Effects on Consumers’ Willingness-to-Pay and Willingness-to-Accept.” *Journal of Consumer Research* 31 (3): 681–90.

Sunstein, Cass R., Daniel Kahneman, David Schkade, and Ilana Ritov. 2002. “Predictably Incoherent Judgments.” *Stanford Law Review* 54 (6): 1153–1215.

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

Wacholder, Sholom, Stephen Chanock, Montserrat Garcia-Closas, Laure El-ghormli, and Nathaniel Rothman. 2004. “Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies.” *Journal of the National Cancer Institute* 96 (6): 434–42.

Young, Neal, John Ioannidis, and Omar Al-Ubaydli. 2008. “Why Current Publication Practices May Distort Science.” *PLoS Medicine* 5 (10): e201.