Non-Naïve Participants Can Reduce Effect Sizes

In press at Psychological Science

Jesse Chandler[1], Gabriele Paolacci[2], Eyal Peer[3], Pam Mueller[4], and Kate Ratliff[5]

[1]Institute for Social Research, University of Michigan; [2]Erasmus University Rotterdam;

[3]Bar-Ilan University; [4]Princeton University; and [5]University of Florida

Corresponding Author:

Jesse Chandler, University of Michigan, Institute for Social Research, 426 Thompson St.,

Ann Arbor, MI 48105

E-mail: jjchandl@umich.edu

Abstract

Although researchers often assume their participants are naïve to experimental materials, this is not always the case. We investigated how prior exposure to an experiment affects subsequent experimental results. Participants in this study completed the same set of 12 experimental tasks at two points in time, first as a part of the “Many Labs” replication project (Klein et al., 2014) and again days, a week or a month later. Effect sizes were markedly lower in the second wave. The reduction was most pronounced when participants were assigned to a different condition in the second wave. We discuss the methodological implications of these findings.

Keywords

non-naivety, repeated participation, effect sizes, judgment and decision-making, panel conditioning, research methods, open data, open materials

Non-Naïve Participants Can Reduce Effect Sizes

When researchers conduct a study, they often assume that participants are naive to the research materials, either because the pool of participants is large (e.g., in the case of Internet samples) or because prior exposure to research is limited (e.g., in the case of first year college students). However, people may belong to a participant pool for several years, and small numbers of “professional” survey-takers tend to dominate responses (Chandler, Mueller, & Paolacci, 2014; Goyder, 1986; Hillygus, Jackson, & Young, 2014). Undergraduate subject pools experience faster turnover than others (1–4 years) but are also likely to be used by researchers with greater overlap of interests (e.g., researchers in the same lab). Compounding these issues, people may belong to multiple participant pools, increasing exposure to experimental materials (Hillygus et al., 2014), or may gain knowledge of research materials through college courses, other pool members or media coverage.

Prior exposure to experimental materials can change how participants behave in subsequent related studies through various pathways. For example, performance can be improved through practice (e.g., Chandler et al., 2014), beliefs can change through additional cognitive elaboration (Sherman, 1980; Sturgis, Allum & Brunton-Smith, 2009), and motivation to perform well can be deterred by boredom or increased by a desire to please the experimenter. Statistically, this can affect both group means (e.g., uniform improvements can lead to a ceiling effect that compresses scores) and standard deviations (e.g., inattention caused by boredom can increase within-condition variance), resulting in changes in observed effect sizes. In practice, the effects of prior participation on experimental data have only been identified in a narrow range of circumstances, and typically only in domains where a hypothesis about the purpose of the experiment can be used to respond in a manner that makes participants look more favorable (for a review, see Weber & Cook, 1972).

The effect of repeated participation may be particularly apparent when participants are exposed to different experimental conditions, as illustrated though comparisons of within- and between-subjects experiments. Exposure to earlier conditions informs responses to subsequent conditions, leading to different results than logically equivalent between-subjects designs (Charness, Gneezy, & Kuhn, 2012; Greenwald, 1976). Depending on the available information, observed effects can be inflated (Fox & Tversky, 1995), attenuated (Hershey & Schoemaker, 1980), or reversed (Birnbaum, 1999).

Information contained within psychological experiments is of little relevance to research participants and should be forgotten quickly. Recently, however, researchers have noted a correlation between responses to psychological measures and indirect measures of prior participation in similar experiments such as memory of prior participation (Greenwald & Nosek, 2001), the chronological order of studies themselves (Rand et al., 2014), measures of the total number of completed experiments (Chandler et al., 2014) or naturally varying levels of experience with a task (Mason, Suri, & Watts, 2014). These findings suggest that exposure to research materials can influence effect sizes more generally, but this possibility has not been directly tested. To address this gap, we examine how prior exposure to study materials affects responses.

Method

Design and procedure

To test the effects of non-naivety on commonly-used psychological measures and potential moderators of these effects we conducted a two-stage study on Amazon Mechanical Turk (MTurk), an online crowdsourcing service frequently used for experimental research (for a review see Paolacci & Chandler, 2014). All participants completed a set of two-condition experiments in Wave 1 (W1; previously reported in Klein et al., 2014). Participants were then invited to participate in a study in Wave 2 that included the same tasks(W2). Sample size for W1 was predetermined for an existing study. Sample size for W2 consisted of all participants from W1 who responded to the survey invitation. For each experiment, assignment to experimental condition was randomized at the individual task level, and thus participants completed each task in either the same condition as in W1 or in the alternative condition.

Further, we explored whether the effect of prior exposure is more pronounced under conditions when it should be particularly easy to remember previous materials. The visual similarity of the experiments was manipulated by randomly assigning participants to complete the experiments on either the same online platform or a different, visually distinct platform. The duration between the completion of W1 and W2 was also manipulated by randomly assigning participants to be re-contacted a few days, about a week, or about a month later. This produced a 3 (Time Delay: a few days (DL), about a week (W), about a month (M), between participants) × 2 (Visual Similarity: same, different, between participants) × 2 (Condition: same, different, between participants) design. This study was approved by the University of Michigan’s IRB.

Wave 1.

One thousand adults were recruited from MTurk to participate in a "decision making and attitudes survey" as a part of a larger research study (Klein et al., 2014). Participants were restricted to those classified as U.S. residents who had completed at least fifty HITs (prior tasks) with an approval ratio of at least 95%. Participants were paid $0.70 and were told that the experiment would take approximately 13 min to complete. Participants completed 14 between-subjects experimental tasks and an Implicit Association Task. After finishing all tasks, participants(described below) before completing additional measures and demographic information (for full details see Klein et al., 2014). Participants were thanked for their time but were not debriefed.

Wave 2.

A unique qualification was created on MTurk for each delay period, and qualifications were randomly assigned to all W1 participants who responded from U.S. IP addresses at W1 (*N* = 950; see Chandler et al., 2014 for technical details). This qualification was required to participate in W2, ensuring that i.) only W1 participants could complete the W2 measures, and ii.) W1 participants completed the W2 HIT after the time delay to which they were randomly assigned. Workers who completed this HIT were paid $1 for their time. Participants completed the experimental tasks a second time, before providing demographic information and indicating for each experiment whether they remembered participating in it in the past.

Once a Wave 2 HIT was created, the eligible participants received an e-mail sent via the MTurk APIand told "As a result of your participation in a prior study, you have qualified to complete another set of tasks. Based on worker comments, this study is shorter than the first one you participated in, and the pay rate is also higher. This study will likely take about 10 minutes. Thank you for your past participation and we hope that you are willing to participate again!”

Participants.

Six hundred eighty-seven participants participated in both W1 and W2. Response rates were approximately 72% across all time-delay conditions. To eliminate the possibility that differences in effect size between wavesbetween was a result of attrition, participants were included in the analysis if they completed all experimental tasks s in both W1 and W2 from a U.S. IP address and submitted a payment request (N = 638; 55% women; M_{age} = 36, SD = 12.8, range 18–75; 83% White). In cases where the participant was recorded as attempting the study more than once, only the first attempt was used unless the participant saw none of the tasks in this attempt (i.e., read the consent form only in their first attempt).

Experimental tasks

Participants completed the following tasks in both Wave 1 and Wave 2. For brevity, an IAT task conducted in Wave 1 was not included in Wave 2 materials. Two additional tasks (not reported here) did not detect significant between condition differences at Wave 1 and were dropped from subsequent analyses (because our purpose was to examine changes in bona fide effects; see Klein et al., 2014, for more details about the experimental tasks).

Sunk costs.

Replicating Oppenheimer, Meyvis, and Davidenko (2009), participants were asked to “Imagine that your favorite football team is playing an important game. You have a ticket to the game that you [have paid handsomely for] [have received for free from a friend]. However, on the day of the game, it happens to be freezing cold. What do you do?” Participants rated their likelihood of attending the game on a 9-point scale (1 = *Definitely stay at home*, 9 = *Definitely go to the game*).

Gain vs. loss framing for combating disease.

Replicating Tversky and Kahneman (1981), participants were asked to imagine that the U.S. is preparing for the outbreak of an disease and select from two courses of action: Program A, under which [200 people will be saved] [400 people will die] or Program B, under which there is a 1/3 probability that 600 people will be saved [no people will die] and 2/3 probability that no people will be saved [600 people will die].

Anchoring and adjustment.

In a partial replication of Jacowitz and Kahneman (1995), participants made four quantitative estimates (the distance from New York to San Francisco, the population of Chicago, the height of Mount Everest, and how many babies are born per day in the United States) after being told that the target is greater than or less than a specified value.

Retrospective gambler’s fallacy.

Replicating Oppenheimer and Monin (2009), participants were asked to imagine seeing a man rolling dice. The man rolls 3 sixes, [2 sixes and a three]. Participants then estimated how many times the man had rolled the dice before they had entered the room to watch him.

Low-vs.-high category scales.

Replicating Schwarz, Hippler, Deutsch, and Strack (1985) participants were asked to estimate how much TV they watch daily on either a low-frequency Likert-type scale (Up to ½ hour a day, 1/2 hour to 1 hour a day, 1 to 1 ½ hours a day,1.5–2 hr, 2 to 2 ½ hours a day, More than 2 ½ hours a day) or a high-frequency scale (Up to 2 ½ hours a day, 2 ½ hours to 3 hours a day, 3 to 3 ½ hours a day, 4 to 4 ½ hours a day, More than 4 ½ hours a day).

Norm of reciprocity.

Conceptually replicating Hyman and Sheatsley (1950), participants were asked (yes/no) in a randomly assigned order whether i.) North Korea should allow American reporters in and allow them to report the news back to American papers and ii.) whether America should allow North Korean reporters into the United States and allow them to report back to their papers.

Allowed/Forbidden.

Replicating Rugg (1941), participants were asked to indicate (yes/no) whether (1) The United States should allow speeches against democracy, or (2) The United States should forbid speeches against democracy.

Quote Attribution.

In a conceptual replication of Lorge and Curtis (1936), participants were presented with a quote ("I have sworn to only live free, even if I find bitter the taste of death") attributed to either George Washington or Osama Bin Laden and indicated the extent to which they agree with the quote (1 = *Strongly agree*, 9 = *Strongly disagree*).

Imagined contact.

Closely replicating Husnu and Crisp (2010), participants were asked to either imagine and describe (for one minute) meeting a Muslim stranger for the first time or, in the control condition, imagine walking in the outdoors, and describe it (for one minute).

Participants then responded to four measures indicating their willingness to interact with Muslims.

Results

Non-naivety reduced observed effect sizes

Table 1 shows effect sizes in W1 and W2. For the anchoring effect, results outside of the specified upper and lower anchors were dropped from analysis and treated as missing values. The rightmost column contains the interaction between Condition and Wave from a within-subjects GZLM analysis with Condition, Wave and their interaction as factors that functions as a within-subjects significance test of the overall decline. Correlations across time periods for each experiment and within W2 across experiments can be found in the supplementary materials.

[Insert Table 1 about here.]

As can be seen, only one effect size increased from W1 to W2 (the low vs. high scales task), while 11 of the 12 observed effects exhibited statistically significant declines, $p < .01$ for an exact sign test. Declines were statistically significant in 5 of 12 measurements using traditional omnibus tests of the interaction between Condition and Wave. The largest effects in W1 tended to show correspondingly larger decline effects, suggesting that a floor effect might have limited our ability to observe significant results for phenomena with smaller effect sizes. Although not statistically significant, declines in smaller effects may have greater practical implications for researchers who evaluate the truth of observed effects through standard frequentist tests of statistical significance.

Meta-regression (Lipsey & Wilson, 2001) was used to illustrate the average effect of non-naivety on participant responses. A meta-analytic approach was selected because the dependent measures of interest consist of a mixture of categorical and continuous dependent measures and thus cannot be integrated into a single analysis until transformed into a common metric. This approach was preferred over simple averaging because it assigns greater weight to effects with smaller standard errors.

To prevent the four measures of anchoring from exerting an undue influence, a single effect size was estimated for them by taking the simple average of effect sizes and standard error across the four anchoring experiments. Another concern was that increased variance in Wave 2 responses may have resulted from the varied treatment effects of the different conditions in Wave 2. To address this, , effect sizes for W1 and W2 were estimated for each of the nine effects of interest under each of the twelve possible combinations of Wave 2 treatment level, yielding 216 effect size estimates that along with the overall effects observed at W1 were regressed on dummy codes for the different experimental paradigms and Wave.[1] Overall, effect sizes declined from T1 (weighted *d* = .82) to T2 (weighted *d* = .63) by *d* = .19, a drop of about 25%.

Moderators of the decline of effect sizes

Additional analyses were conducted to examine the effect of the variables of theoretical interest and their higher order interactions on individual experiments. GZLM was used for all effects except the anchoring measures, which were examined together in a single linear mixed model to account for their dependence. W2 Condition, whether the participant was in the same condition across both waves (Same Condition, dummy coded), Platform, Duration, and their higher order interactions were included in all models. Due to issues with model fit, the analysis of Allowed/Forbidden had to be simplified by eliminating Duration.[2] As can be seen in Table 2, the effect of being in different conditions in each wave (represented by the interaction between Condition and Same Condition) was significant for Anchoring, Quote Attribution, and Sunk Costs, suggesting that attenuation of these effects was driven by information gained from exposure to both experimental conditions.

[Insert Table 2 about here.]

To illustrate how various treatment conditions affected the decline of effect sizes in aggregate, dummy variables for task (accounting for differences in attenuation across experimental paradigms), and the three theoretically relevant factors of Same Condition, Platform, Delay, and their higher order interactions were regressed on W2 effect size estimates using meta-regression (Lipsey & Wilson, 2001). As can be seen in Table 3, the effect of condition was pronounced relative to the effects of other variables and their interactions. The decline in effect sizes from Wave 1 to Wave 2 Wave 1 and Wave 2 effect sizes for responses to different conditions and the same conditions was examined separately for experiments in which participants were assigned to different conditions. The effect of wave was substantial for participants assigned to different conditions, d = 0.23, but was observed even among participants assigned to the same condition, d = 0.14. In addition to condition, the delay between Wave 1 and Wave 2 also played a role, with mean effect sizes returning toward the magnitude observed in Wave 1 as time progressed. Survey platform and the higher higher-order interactions between theoretically relevant variables had minimal influence on effect size (Table 3). Mean Wave 2 effect sizes across conditions and the grand mean of Wave 1 effect sizes are displayed in Figure 1.

[Insert Table 3 and Figure 1 about here.]

Memory for prior participation

A related question is whether people report participating in a prior task, and whether this self-reported memory mediates the observed reduction of effect sizes. If the reduction of effect sizes is related to the probability that a participant reports their prior participation, researchers could simply exclude participants who report having participated in similar experiments. Memory for participation in prior studies was high, but far from unanimous, ranging between 35% and 80% of participants for each task. In addition to reporting whether they had previously participated in the relevant experimental tasks, participants were also asked if they had ever participated in a plausible but likely never posted experimental task in which they sorted images of dinosaur fossils. It is unlikely that memory for participating in previous studies reflects acquiescence bias, as the proportion of participants who reported participating in sorting dinosaur fossils was close to zero.

To examine whether participants were more likely to recognize an identical (as opposed to similar) experimental task, e ratio of experiments that participants remembered versus did not remember was calculated for both experiments to which they were assigned to the same condition at W1 and at W2 and for experiments to which they were assigned to different conditions at W1 and W2. A 2 (Same Condition: within participants) × 2 (Platform: between participants) × 3 (Time, between participants) mixed model ANOVA revealed that participants tended to remember fewer experiments as time passed, $F(2, 626) = 16.57$, $p < .001$, $\eta_p^2 = .05$. There was no main effect of Same Condition or Platform and no higher order interactions between variables, all *F*s < 1. The amount of time between W1 and W2 was thus the only reliable predictor of self-reported memory for prior participation.

Reporting prior participation does not predict effect size attenuation

To test whether memory for prior participation was related to the attenuation of effect sizes, GZLM was used for all effects except the anchoring measures, which were examined together in a single linear mixed model to account for their dependence. Dummy variables for task (accounting for differences in attenuation across experimental paradigms), and the two theoretically relevant factors of Same Condition, memory for prior participation (Memory) and their interaction were regressed on W2 effect size estimates. Due to issues with model fit, the analysis of Allowed/Forbidden had to be simplified by eliminating the three-way interaction. The effect of remembering prior participation (represented by the interaction between Condition and Memory) was significant only for Anchoring, $F(2, 2181.6) = 5.79$, $p < .01$. For all other experiments these two way and three way interactions were not significant, *p*s > .22.

To examine the overall effect of self-reported memory, W2 effect sizes were calculated for participants who did vs. did not remember each experiment, yielding a total of eighteen effect size measures. Effect sizes were regressed on dummy variables for the different experimental paradigms, and a dummy variable indicating whether the effect size represented those who did or did not report remembering the prior experiment, a dummy variable indicating whether participants were assigned to the same or a different experimental condition and their interaction using meta-regression (Lipsey & Wilson, 2001). Neither the effect of Memory, $\beta = 0.12$ nor the interaction between Memory and Same Condition, $\beta = 0.11$ were significant at the $p < .05$ level, even under the (possibly liberal) assumption that the responses provided by participants were statistically independent from each other. Moreover, both were substantially smaller than the main effect of being assigned to the Same Condition, β = 0.23. These findings suggest that self-reported memory for prior participation is at best a poor indicator of which participants will display attenuated effect sizes due to prior participation.

Discussion

Pior exposure to research materials can reduce the effect size of true research findings. Effect sizes decreased by about 25%, when they were replicated on the same sample. This decline can have a surprisingly large effect on experimental power, even when non-naïve participants are a fraction of the total sample. To illustrate, research has suggested that 10% of all respondents on MTurk are responsible for 40% of all experimental responses (Chandler et al., 2014) and could thus be considered "non-naïve." In a subsequent study, these highly productive workers composed 25% of a large sample of workers (Chandler et al., 2014). A 25% reduction among a quarter of the sample would reduce an average sized behavioral science effect of *d* = .43 (Richard, Bond, & Stokes-Zoota, 2003) to *d* = .40 (ignoring further declines due to increased within-group variance created by a mix of naïve and non-naïve participants). For a two-condition experiment with 80% power, this would require increasing the sample size from N = 172 to N = 200 to compensate (~15%; Faul, Erdfelder, Buchner, & Lang, 2009).

Effect sizes observed at W2 were generally smaller, but this difference was not uniformly statistically significant. Larger effect sizes were more likely to demonstrate a significant reduction, perhaps because observable between-conditions variance is constrained for true effects that are closer to zero. Additionally, some effects may be more susceptible to attenuation effects. For example, it is easy to imagine that for the anchoring questions, memory of prior anchoring values may inform numerical estimates.

Effect sizes were particularly attenuated when participants were exposed to alternative conditions of an experiment, suggesting that this additional contextual information undermines effect sizes, analogous to participating once in a within-participant experiment (Greenwald, 1976). However, effects were also attenuated among participants exposed to the same condition twice. While there is no direct evidence of a mechanism underlying this decline, one possible explanation is that asking questions multiple times may lead to elaboration (Sherman, 1980; Sturgis et al., 2009). Elaboration could reduce observed effect sizes if it undermines intuitive responses to one or both conditions, or brings to mind idiosyncratic information that increases within-condition variance. If so, changes in effect sizes should be especially pronounced towards hypothetical or unfamiliar situations (Hirt & Sherman, 1982).

Self-reported participation is an imperfect measure of prior participation. It does not identify all prior participants, or even those who demonstrated a particularly large behavioral effect of prior participation. While this may be surprising, it is not inconsistent with the hypothesis that the additional information gained from exposure to both conditions explains the attenuation effect: forgetting of the source from which information is learned occurs more quickly or separately from forgetting the information itself (Johnston, Hashtroudi, & Lindsay, 1993). This finding suggests that researchers cannot simply ask participants to identify whether they have participated before, and underscore the importance of directly monitoring prior participation. When this is not possible, researchers should strive to design procedures and stimuli that differ from those known to the tested population (Chandler et al., 2014; Rand, 2014), or increase their sample size to offset the anticipated decrease in power.

Author Contributions

J. Chandler developed the study concept. All authors contributed to the study design. Testing and data collection were performed by J. Chandler and P. Mueller. J. Chandler and E. Peer analyzed the data. J. Chandler, G. Paolacci and E. Peer wrote the manuscript, and all authors provided critical revisions. All authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Supplemental Material

Additional supporting information can be found at http://pss.sagepub.com/content/by/supplemental-data

Open Practices

All data and materials have been made publicly available via Open Science Framework and can be accessed at https://osf.io/4vx3z/. The complete Open Practices Disclosure for this article can be found at http://pss.sagepub.com/content/by/supplemental-data. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ and http://pss.sagepub.com/content/25/1/3.full.

Notes

1. We thank Jelte Wicherts for this insight.

2. Specifically, a quasicomplete separation (lack of variance of dependent measures within a specific combination of predictors) prevented estimation of parameters for the full model.

References

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods*, *4,* 243–249.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46,* 112–130.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81,* 1–8.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149–1160. doi:10.3758/BRM.41.4.1149

Fox, C., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics, 110,* 585–603.

Goyder, J. (1986). Surveys on surveys: Limits and potentialities. *Public Opinion Quarterly, 50,* 27–41. doi:10.1086/268957

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin, 83,* 314–320.

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift für Experimentelle Psychologie, 48*(2), 85–93.

Hershey, J. C., & Schoemaker, P. J. H. (1980). Prospect theory's reflection hypothesis: A critical examination. *Organizational Behavior and Human Performance, 25,* 395–418.

Hillygus, D. S., Jackson, N., & Young, M. (2014). Professional respondents in nonprobability online panels. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 219–237). West Sussex, England: Wiley.

Hirt, E. R., & Sherman, S. J. (1985). The role of prior knowledge in explaining hypothetical events. *Journal of Experimental Social Psychology, 21,* 519–543.

Husnu, S., & Crisp, R. J. (2010). Elaboration enhances the imagined contact effect. *Journal of Experimental Social Psychology, 46,* 943–950.

Hyman, H. H., & Sheatsley, P. B. (1950). The current status of American public opinion. In J. C. Payne (Ed.), *The teaching of contemporary affairs* (pp. 11–34). New York, NY: National Council for the Social Studies.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin, 21,* 1161–1166.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114,* 3–28.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152.

Lipsey, M. W., & Wilson, D. B. (Eds.). (2001). *Applied Social Research Methods Series: Vol. 49. Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lorge, I., & Curtis, C. C. (1936). Prestige, suggestion, and attitudes. *The Journal of Social Psychology, 7,* 386–402.

Mason, W., Suri, S., & Watts, D. J. (2014). Long-run learning in games of cooperation. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation* (pp. 821–838). New York, NY: Association for Computing Machinery.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867–872.

Oppenheimer, D. M., & Monin, B. (2009). The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making, 4,* 326–334.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23,* 184–188.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications, 5,* Article 3677. Retrieved from http://www.nature.com/ncomms/2014/140422/ncomms4677/full/ncomms4677.html

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331-363.

Rugg, D. (1941). Experiments in wording questions: II. *Public Opinion Quarterly, 5,* 91–92.

Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly, 49,* 388–395.

Sherman, S. J. (1980). On the self-erasing nature of errors of prediction. *Journal of Personality and Social Psychology, 39,* 211-221.

Sturgis, P., Allum, N., & Brunton-Smith, I. (2009). Attitudes over time: The psychology of panel conditioning. *Methodology of longitudinal surveys*, 113-126.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458.

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin, 77*, 273-295.

Table 1.

Experimental Effects of Condition and Wave

	Wave 1		Wave 2		Condition × Wave interaction (Wald)
Task	Comparison of conditions	Effect size (Cohen's *d*)	Comparison of conditions	Effect size (Cohen's *d*)	
Allow/forbid	$\chi^2(1, N = 638) = 24.61$***	0.40	$\chi^2(1, N = 638) = 7.32$**	0.22	4.35*
Anchoring: births	*t*(489) = 25.83***	2.11	*t*(557) = 20.65***	1.65	6.40**
Anchoring: Mount Everest	*t*(608) = 31.39***	2.55	*t*(559) = 20.83***	1.79	37.63***
Anchoring: Chicago	*t*(595) = 18.06***	1.46	*t*(545) = 12.62**	1.02	$3.28^{\dagger}$
Anchoring: NY to SF	*t*(552) = 12.78***	1.04	*t*(602) = 10.73***	0.86	1.45
Gain vs. loss framing	$\chi^2(1, N = 638) = 56.15$***	0.62	$\chi^2(1, N = 638) = 21.55$***	0.37	$2.88^{\dagger}$
Imagined contact	*t*(636) = 4.38***	0.35	*t*(636) = $1.79^{\dagger}$	0.14	6.07*

Low vs. high scales	χ^2(1, *N* = 638) = 27.09***	0.42	χ^2(1, *N* = 638) 47.24***	0.57	0.66
Norm of reciprocity	χ^2(1, *N* = 637) = 8.50**	0.23	χ^2(1, *N* = 636) = 0.90	0.08	0.20
Retrospective gambler’s fallacy	*t*(429) = 5.43***	0.43	*t*(558) = 2.90***	0.23	0.08
Quote attribution	*t*(633) = 6.13***	0.48	*t*(636) = 3.56***	0.28	4.67*
Sunk costs	*t*(612) = 5.30***	0.42	*t*(636) = 2.36*	0.19	0.56

Note: Positive *t* values indicate that the effect was in the theoretically predicted direction. Reported *t* values and degrees of freedom assume equality of variance unless Levine’s test indicated that this assumption was violated. Fractional degrees of freedom were rounded down. The chi-square tests of the interaction between condition and wave used a generalized estimating equation treating Wave 1 and Wave 2 data as nonindependent.

†*p* < .10. **p* < .05. ***p* < .01. ****p* < .001.

Table 2.

Effects of Condition, Platform, and Delay in Each Task

Variable	Allow/forbid	Anchoring	Gain vs. loss framing	Imagined contact	Low vs. high scales	Norm of reciprocity	Retrospective gambler’s fallacy	Quote attribution	Sunk costs
Condition	7.26*	981.39***	20.10*	3.73	44.61***	1.12	8.22**	14.89***	5.61*
Same condition	0.08	7.30**	1.282	0.58	0.00	2.02	2.90†	2.71	0.96
Platform	0.21	0.08	0.017	0.63	1.53	1.85	1.09	8.93**	0.00
Delay	—	0.51	0.012	0.43	1.45	3.36	0.25	1.24	1.09
Condition × Same Condition	0.09	30.07***	0.24	0.01	0.03	0.01	1.95	4.47*	4.72*
Condition × Platform	1.1	0.223	0.14	2.05	1.11	0.15	0.55	0.78	3.54†

Condition × Delay	—	1.82	0.37	1.70	0.23	1.50	1.50	1.66	4.30
Same Condition × Platform	0.09	0.21	0.76	1.24	2.44	0.04	0.01	0.84	0.83
Same Condition × Delay	—	0.11	1.27	2.80	1.79	3.22	3.16	2.24	3.67
Platform × Delay	—	0.63	0.45	2.75	1.42	0.56	1.47	2.22	1.94
Condition × Same Condition × Platform	0.24	0.01	0.14	2.75†	0.01	5.37*	0.09	2.11	2.01

Condition × Same Condition × Delay	—	2.70	1.78	2.26	0.90	1.47	1.16	2.02	1.87
Condition × Platform × Delay	—	4.88**	0.11	$5.01^{\dagger}$	3.40	0.23	0.28	6.57*	1.14
Same Condition × Platform × Delay	—	0.82	0.51	2.54	2.82	3.22	1.22	0.54	0.94
Condition × Same Condition ×	—	1.56	1.52	$4.95^{\dagger}$	0.18	1.34	0.48	2.21	1.08

Platform × Delay

Note: All values were obtained from Wald chi-square tests of significance, except for the values for anchoring, which are the results of an analysis of variance.

$^{\dagger}p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

Table 3.

Results From the Meta-Regression of Wave 2 Responses

Variable	*b*	β
Anchoring	1.17	0.74
Gain vs. loss framing	0.17	0.12
Imagined contact	–0.04	–0.03
Low vs. high scales	0.37	0.25
Norm of reciprocity	–0.24	–0.17
Quote attribution	0.11	0.08
Retrospective gambler’s fallacy	0.12	0.09
Sunk costs	0.00	0.00
Platform	0.03	0.06
Delay	0.03	0.06
Same condition	0.08	0.17
Platform × Wave	0.02	0.03
Same Condition × Wave	–0.05	–0.08
Same Condition × Platform	–0.01	–0.02
Same Condition × Platform × Delay	–0.01	0.02

Note: Experimental paradigms are represented as dummy variables with allow/forbid serving as the reference group.

Fig. 1.

Average effect size (Cohen's *d*) at Wave 1 and Wave 2. Wave 2 results are shown for participants in the same and different condition as in Wave 1 and for those who used the same and different platform as in Wave 1, separately for the three delay periods.