Investigating variation in replicability: A “Many Labs” Replication Project

Richard A. Klein and Kate A. Ratliff

University of Florida

Brian A. Nosek

University of Virginia

Authors’ Note: The ultimate project team will include more than 20 co-authors. The project team will be assembled and confirmed upon registration of the project.

Abstract

Although replication is a central tenant of science, direct replications are rare in the field of psychology. This pre-registered replication will bolster understanding of replication in psychology by examining how effects replicated directly vary across samples and situations. The study will test 12 effects in a single experimental package across numerous labs. Further, by comparing the conditions under which the study was administered (e.g., participant compensation, geographic location of lab) we can test whether any predict replication success. At one extreme, sample and situational characteristics might have little effect on the tested effects – variation is a function of reliability of the estimates. At the other extreme, effects might be highly contextualized – replicating only with sample and situational characteristics that are highly consistent with the original circumstances. More likely, the reality is somewhere in between – and some effects will be more invariant across sample and setting than others. These results will provide an initial basis for identifying correlates of replicability across a variety of effects.

Investigating variation in replicability: A “Many Labs” Replication Project

Replication is a central tenant of science that helps ensure the veracity of empirical findings. In order to be confident in the validity of an empirical result, others should be able to reproduce the conditions of the original demonstration and obtain the same effect (Open Science Collaboration, 2012). Replication in psychology may be particularly difficult because human behavior is simultaneously influenced by many variables inside and outside the person. An effect observed in one cultural context may not be observed in another (Henrich, Heine, & Norenzayan, 2010). Subtle changes to the situational context may neutralize, or even reverse, an otherwise robust phenomenon. And, individual differences among humans may moderate any particular result.

This malleability of human thought and behavior to suit the context is a feature, not a bug. That said, the challenges this malleability creates for predicting human behavior do not excuse psychological science from having replicability as a core tenet of its practice. The special issue of *Social Psychology* to which we are submitting this proposal aims to increase the valuation of replication, and perhaps make some progress in identifying the conditions under which particularly important effects are obtained. The purpose of the present project is to complement specific replication studies with a study of the variation in replicability across numerous effects and laboratory settings.

In this “Many Labs Project” we propose to replicate twelve documented effects in as many independent laboratories as possible. Of the effects we propose to include, some are known to be highly replicable; for others, ease of replication is unknown. Some likely depend on social context or participant sample, whereas others are not expected to vary as much. Some of the experiments are older and highly cited; others are new and intriguing. We will bundle the selected studies together into a brief, easy-to-administer experiment that will be delivered through a single on-line infrastructure (<http://projectimplicit.net/>) to all of the participating labs.

There are many factors that can influence the replicability of an effect: power, sample characteristics, setting characteristics, and procedural variations. The present design standardizes procedural characteristics in order to examine effects of sample and setting on replicability. There is much discussion, but little systematic evaluation, about the extent to which psychological effects are highly dependent on sample and setting. Features of the laboratories and the participant samples will be coded as potential predictors of replicability. At one extreme, sample and situational characteristics might have little effect on the tested effects – variation is a function of reliability of the estimates. At the other extreme, effects might be highly contextualized – replicating only with sample and situational characteristics that are highly consistent with the original circumstances. Surely, the reality is somewhere in between – and some effects will be more invariant across sample and setting than others. The present sample of 12 psychological effects is not random nor is it representative of all possible effects. As such, the contribution of the present investigation is to establish a paradigm and demonstration evidence of replicability across samples and settings. Ideally, the results will stimulate theoretical developments about the conditions under which replication will be robust to the inevitable variation in circumstances of data collection.

Further, it will offer an opportunity to consider the interactive roles of direct replication and generalizability. Every direct replication has an infinite number of differences from an original demonstration. As such successful replication both verifies the original effect and extends its generality. As a consequence, direct replication validates and theoretically extends (or constrains) a phenomenon.

**Method**

*Sampling Plan*

To be included in the study, each participating laboratory commits to collecting data from *at least* 80 participants. Requiring too few participants would make the replicability estimates highly unstable in each laboratory; requiring too many participants would enable only a few laboratories to participate. A minimum of 80 was chosen as a balance between these conflicting goals. It allows for reasonable statistical power for each individual data collection while requiring a sampling plan that enables many labs to get involved in the project (see section on confirmatory data analysis for power analyses). We will encourage all laboratories to make their sample sizes as large as possible to provide maximally stable estimates from each site.

Further, we will administer the study in two virtual laboratory environments where large samples are accessible – Project Implicit and Amazon’s Mechanical Turk. Project Implicit (http://implicit.harvard.edu) is a non-profit organization and international collaboration focusing on implicit social cognition - thoughts and feelings outside of conscious awareness and control. Project Implicit operates a virtual laboratory to administer behavioral research on the Internet (see Nosek, 2005, for more information). More than 10,000 volunteers complete studies at Project Implicit every week, with more than 1 million study sessions completed per year. Participants learn about the site through media coverage, links from other sites, search engines, word of mouth, or class assignment. During their initial visit to the site, participants who register to participate in research complete a demographics questionnaire. During each subsequent visit, registered participants are randomly assigned to an eligible study from a pool of all active studies in the system. Visitors can complete as many studies as they would like, but those randomly assigned to a particular study are not assigned to it again on future visits.

The other virtual environment we will use is Amazon’s Mechanical Turk (MTurk; [www.MTurk.com](http://www.mturk.com)). MTurk is an online marketplace for getting individuals to complete tasks (see Buhrmester, Kwang, & Gosling, 2011, for more information). After registering and providing bank account information, “workers” can choose which tasks they wish to complete. MTurk participants are typically paid between 10 and 50 cents for participation in studies lasting 5 to 20 minutes. We will recruit at least 1000 participants from both Project Implicit and MTurk (see section on confirmatory data analysis for power analysis).

*Anticipated Sample Characteristics*

This data collection will include many undergraduate samples from multiple countries, an online sample with MTurk participants, and an online sample with Project Implicit volunteers. We cannot readily estimate the sample characteristics of the undergraduate samples from participating laboratories; however, the researchers who have expressed interest in the project so far come from Germany, Italy, Spain, Brazil, The Czech Republic, The Netherlands, Turkey, and all over the United States. Each participating university team will commit to (1) administering the protocol as specified, (2) collecting at least 80 participants (and as many as is feasible), (3) video recording a simulation of the setting and procedure of study administration, and (4) completing a short questionnaire about the features of recruiting, sample, undergraduate population, and variations from the standard protocol.

We can anticipate MTurk and Project Implicit characteristics based on past research. Paolacci and colleagues (2010) found that a large sample of MTurk workers consisted of 65% women, the mean age was 36 years, workers came from 66 countries (47% United States, 35% India, 18% other), 66.7% earn below $60,000 per year, and the workers tended to have attained higher education levels than the general population (though these numbers were not reported). Ratliff and Nosek (2010) found that 75% of Project Implicit visitors completed the study they began. The sample of participants who complete the study consisted of 65% women, and the mean age was 30.1 years. The race of the participants was as follows: 6% were Asian, 7% were African–American, 3% were Hispanic, 76% were white, 4% were multiracial or ‘other’ and 4% did not report their ethnicity. Nosek and Smyth (2011) found that the average age of their Project Implicit sample was 27 years, with half older than age 22; 65% were women; Whites were 76%, Asians 6%, Blacks 5%, Hispanics 4%, and 9% were another ethnicity or race; 77% were U.S. citizens, 4% Canadian, 4% were from the United Kingdom, and 15% from other countries; and 13% had a high school degree or less education, 51% had some college/university experience, 19% had a BA/BS or equivalent degree, and 17% reported an advanced degree.

*Recruitment and Participant Remuneration*

On February 21st we posted about the Many Labs Project in the online forum of the Open Science Collaboration (see Appendix A). In the following week, nearly 20 researchers from all over the world expressed interest in taking part in the initiative. This provides an initial base of likely participation. Once the design is finalized and registered, we will expand the invitation with a public announcement. It is likely that a second call for collaborators, combined with outreach on the Society for Personality and Social Psychology listserv, will yield a relatively large number of total contributors. Participating researchers will be credited as co-authors on the final publication (if accepted).

Recruitment of samples and compensation to participants will be determined by each participating lab individually, leading to a mix of participants receiving course credit, payment, or no compensation. If there is an effect of compensation it should be detected when comparing results across labs.

Project Implicit participants will be volunteers. The traffic flow at Project Implicit is ~10,000 participants per week across all studies and web sites. Participants that register to complete studies in the Project Implicit participant pool (<http://implicit.harvard.edu/>) will be eligible to complete the study and will be randomly assigned to complete this study from the pool of available studies (see Nosek, 2005 for more information). Once assigned to the study, participants will never again be assigned to it.

MTurk participants will receive 40 cents for their participation. We will post a job (called a “Human Intelligence Task” or “HIT”) titled “Decision-Making Study” that includes no restrictions on who can serve as a “worker”. Workers who access the site will then have the opportunity to choose our study from the list of tasks available (able to be sorted according to criteria such as payment amount and amount of time needed for completion).

*Materials*

Twelve two-condition “mini-experiments” have been chosen for inclusion in the experiment packet. We have piloted this type of "dense effect" design in our own labs and, although we did not test all of the experiments we propose here, we have confirmed that similar effects can indeed be elicited with this type of experimental design. The studies were selected based on the following criteria:

1. *Suitability for online presentation.* Our primary concern is to give each study a “fair” replication that is true to the original design. All of the included studies were able to be adapted to an online format while staying true to the original demonstration of the effect. By administering the study through a web browser, we are able to control exactly how the studies themselves are presented at every site to ensure consistency of procedure. This allows us to focus on other contextual variables, and alleviates some concerns about the quality of the replication procedure itself varying between labs.
2. *Length of study.* In order to administer a number of studies in one experimental session, and to facilitate recruitment of collaborators, we selected studies that could be administered quickly. The entire study package will take less than 15 minutes to complete, on average.
3. *Diversity of effects.* Within the limitations listed above, we sought to include a wide variety of studies. We have included different types of effects (e.g., order effects, priming effects, wording effects), from various time periods (from 1936 to 2012), and with differing levels of prominence (from only a handful of citations to several thousand citations). By doing so, we hope to have a wide sample of studies on which to examine the variation between labs.

*Included studies*

Below we provide a brief description of each experiment, the relevant results (if available), and why we chose to include each particular experiment.

1. *Sunk costs (Oppenheimer et al., 2009).* Sunk costs are those that have already been incurred and cannot be recovered (Knox & Inkster, 1968). Oppenheimer et al. (2009) adapted a task from Thaler (1985), and asked participants 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*). The authors reported that, consistent with prior research on sunk costs (Thaler, 1985), participants were slightly more likely to go to the game if they had paid for the ticket (*M* = 7.46) than if the ticket had been free (*M* = 6.93), *F*(1, 211) = 2.74, *p* = .1, partial *η2*=.01. One caveat is that the alpha level used for these significance tests should be noted, as the difference is not significant at a conventional alpha of .05. In an unpublished replication conduct in the laboratory at the University of Virginia (N = 180), we found an effect size of 0.67 for this same procedure. This task was chosen as a demonstration of a classic line of research, as well as being a task that can be precisely replicated as it was originally presented.
2. *Gain versus loss framing for combating disease (Tversky & Kahneman, 1981)*. In this study Tversky and Kahneman demonstrated how the decisions people make often depend on how the choices are presented. Their 1981 study examined how changing the framing of a question from a negative result to a positive result influenced how much the participants indicated a willingness to “gamble” to receive a better outcome. 307 students at Stanford University and the University of British Columbia were randomly presented one of the two following conditions in a classroom setting: (1) “Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the program are as follows: If Program A is adopted, 200 people will be saved. If Program B is adopted, there is a 1/3 probability that 600 people will be saved and 2/3 probability that no people will be saved. Which of the two programs would you favor?” (2) “Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the program are as follows: If Program C is adopted, 400 people will die. If Program D is adopted, there is a 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. Which of the two programs would you favor?” In the original study, in condition 1, Program A was adopted by 72 percent of participants, while 28 percent favored Program B. In condition 2 this effect was reversed, as Program C was adopted by only 22 percent while 78 percent selected Program D. This represents an asymmetry in risk aversion as questions are posed in a positive or negative frame. This study is representative of some of the most well-known findings in psychology involving framing effects. It has spurred much research (8,753 citations in Google Scholar) and is highly replicable. Like the other effects we included, there could be differences across contexts, but the present research literature emphasizes the persistence of these heuristic biases across samples and settings.
3. *Anchoring (Jacowitz & Kahneman, 1995).* A substantial literature investigates the phenomenon of anchoring – presenting people with a number prior to making a quantitative judgment can influence the subsequent judgment. Jacowitz and Kahneman (1995) presented a number of scenarios in which participants estimated size or distance after first receiving an anchor that was clearly too large or too small. This effect is very robust and has been cited 291 times. Participants in the original study were 103 students at the University of California, Berkeley, who participated for course credit. Participants were asked 3 questions about each of 15 different topics, in which they estimated some quantity about the topic. First, they indicated if the quantity was greater or less than an anchor value. Second, they estimated the quantity. Lastly, they indicated their confidence in the estimate. Some of the 15 topics included things like the distance from San Francisco to New York City, the population of Chicago, the height of Mount Everest, and how many babies are born per day in the United States. The authors found that the estimates of the quantity were biased in the direction of the anchor that was presented. For the purposes of the replication we will shorten this experiment to just ask one question each about four different topics. The questions will be as follows, and participants will respond in an open-response format:
4. Distance from San Francisco to New York City
   1. The distance from San Francisco to New York City is longer than 1,500 miles. How far do you think it is?
   2. The distance from San Francisco to New York City is shorter than 6,000 miles. How far do you think it is?
5. Population of Chicago
   1. The population of Chicago is more than 200,000. What do you think the population of Chicago is?
   2. The population of Chicago is less than 5,000,000. What do you think the population of Chicago is?
6. Height of Mount Everest
   1. Mount Everest is taller than 2,000 ft. How tall do you think Mount Everest is?
   2. Mount Everest is shorter than 45,500 ft. How tall do you think Mount Everest is?
7. Babies Born per day in the United States
   1. More than 100 babies are born per day in the United States. How many babies do you think are born in the United States each day?
   2. Less than 50,000 babies are born per day in the United States. How many babies do you think are born in the United States each day?

For our replication attempt, we will be primarily concerned with the difference in estimates between the low-anchor group and the high-anchor group. We included his experiment because anchoring may play a substantial role in many areas of life where people are making estimates and are presented with an initial suggestion. The widely applicable findings from the initial finding make it a strong candidate, and this is also a well-known effect that should replicate well. This may also reveal some unexpected results if the impact of anchoring differs between locations/settings, or different areas differ in the degree to which they are affected by low vs. high anchors.

1. *Retrospective gambler’s fallacy (Oppenheimer & Monin, 2009).* Oppenheimer and Monin (2009) investigated whether the rarity of an independent, chance observation influenced what participants thought had occurred before that event. The following situation was posed to 80 Stanford University undergraduates as part of a packet of unrelated one-page questionnaires, in a pencil and paper format: “Participants were asked to imagine that they are in a casino and happen to pass a man rolling dice. In one version of the survey, participants were told they witnessed three dice being rolled which all came up 6’s. In a second condition, they witnessed three dice being rolled, two of which came up 6’s and one of which came up a 3. A final condition told participants that they witnessed the rolling of two dice, both of which came up 6’s.” (p. 328). Participants in all conditions were then asked to estimate, in an open-ended format, how many times the man had rolled the dice before they had entered the room to watch him. The authors reported, “Participants believed that a sequence of die rolls was more than three times as long when a set of three 6’s were observed (*M* = 34.2) than when there were only two 6’s (*M* = 10.6), which in turn was believed to be longer than the representatively random sequence of two 6’s and a 3 (*M* = 3.2). The differences between groups was reliable, omnibus *F*(2, 77) = 4.8, *p* < .05, Cohen’s *f* = .18. Pairwise comparisons showed that all differences between conditions were reliable as well, *t*(47, 48, 57) = 1.94, 2.32, 2.65, *p* < .05, Cohen’s *d =* .56, .67, .69).” (pp. 328-329). We included this experiment because this finding represents a new branch in a line of well-established research in the gambler’s fallacy (Tune, 1964; Darke & Freedman, 1997). It is a relatively untested new result (cited 10 times) providing some diversity compared to the tests presented thus far.
2. *Low-vs.-high category scales (Schwarz et al., 1985).* Schwarz and colleagues (1985) demonstrated that simple shifts in response options can alter participants’ reports of their behavior. The research suggests that people infer from response options what is low, normal, and high amounts of a behavior, and self-assess accordingly. In the classic demonstration, 132 German participants completed an experiment in a pencil-and-paper format. Participants were asked how much TV they watch daily on either a high-frequency or low-frequency scale: (1) Low-frequency condition: Up to a half hour a day, Half an hour to one hour a day, One to one and a half hours a day, One and a half to two hours a day, Two to two and a half hours a day, More than two and a half hours a day, (2) High-frequency condition: Up to two and a half hours a day, Two and a half to three hours a day, Three to three and a half hours a day, Three and a half to four hours a day, Four to four and a half hours a day, More than four and a half hours a day. The results showed 16.2 percent of the respondents who were presented the low category range reported watching TV for more than two and a half hours, while 37.5 percent of the respondents presented the high category range did so, *X2* (1) = 7.7, *p* < .005. This method elegantly demonstrates an important phenomenon and is exceptionally simple to administer. Also, with the widespread use of scales as a response format not only in psychology but in many fields and mainstream data collection, it is important to know how these effects may vary across contexts.
3. *Norm of reciprocity (Hyman and Sheatsley, 1950).* People tend to hold more favorable views of behaviors of members of ingroups than for members of outgroups. However, when confronted with a decision about allowing or denying the same behavior to an ingroup and outgroup, people may feel an obligation to reciprocity, or consistency in their evaluation of the behaviors (Hyman & Sheatsley, 1950). In this classic article, the authors asked two questions in counterbalanced order to American participants. One question asked whether communist countries should allow American reporters in and allow them to report the news back to American papers, while the other asked whether America should allow communist reporters into the United States and allow them to report back to their papers. Participants showed much more support for allowing communist reporters into America when that question was asked after the question about allowing American reporters into the communist countries (73 percent in favor) than when it was asked first (36 percent support). This examination of social evaluation adds diversity to the results we are replicating. Additionally, this finding has a strong possibility of revealing different results depending on where it is administered (due to differences in political ideologies), which may spur future research itself.
4. *Allowed/Forbidden (Rugg, 1941)*. How a question is worded can have a pronounced influence on how respondents answer. In an early demonstration of these effects, Rugg (1941) found that respondents were less likely to support forbidding speeches against democracy than they were to indicate those speeches should be “not allowed”. In the study, American respondents were asked one of two questions on a pencil and paper survey: (1) The United States should allow speeches against democracy, or (2) The United States should forbid speeches against democracy. Participants responded by indicating “Yes” or “No”. It was found that 62 percent of participants indicated “No” when asked if speeches against democracy should be allowed, but only 46 percent indicated “Yes” when asked if these speeches should be forbidden. This finding has spurred much future research into why this effect occurs, and is noted as one of the largest wording effects found (see Schuman and Presser, 1981, for a review). This finding also has widespread applicability as many policy decisions can be worded in either frame. For this reason, it may prove valuable to know how such basic wording effects vary across contexts and cultures.
5. *Quote Attribution (Lorge & Curtis, 1936)*. The source of information has a great impact on how that information is perceived and evaluated. Lorge and Curtis (1936) examined how an identical quote would be perceived if it was attributed to a famous liked or disliked individual. The authors recruited 99 unemployed, educated adults to volunteer for the study. Participants were given a varied list of quotations and asked to rate how much they agree with each (1 = completely, 2 = agree but not as much as those given a 1, 5 = neither agree nor disagree, 8 = disagree but not as much as those given a 9, 9 = strongly disagree). Several of the quotations were included twice. The quotation of interest was, “I hold it that a little rebellion, now and then, is a good thing, and as necessary in the political world as storms are in the physical world.” One time that the quote was presented it was attributed to Thomas Jefferson, a liked individual, and the other time it was presented it was attributed to Vladmir Lenin, a disliked individual. It was found that changes in the ratings of the quotations corresponded with participants liking for the presumed speakers (reported in Moskowitz, 2004). The ratings of the individual quotations are not presented in the original paper. The replication of this study will be conceptual, using a different quotation and authors. Participants in the replication will read the following quote: “I have sworn to only live free, even if I find bitter the taste of death” which is a loose modification of actual quotations by both George Washington and Osama Bin Laden. Using this quotation allows us to avoid deception in our research. This study was included as another opportunity for potential ideological differences to be revealed across testing sites, and for the influence of these differences to be measured.
6. *Flag Priming (Carter et al., 2011; Study 2)*. Priming effects occur when a presented stimulus unconsciously influences later responses. The American flag is a powerful symbol in American culture, and Carter et al. (2011) examined how exposure to the flag may influence voting intentions. In a pilot study, the authors found that participants associated the flag more with the Republican Party than the Democratic Party, and hypothesized that this association may bias flag-primed participants towards the Republican Party. They found evidence supporting this hypothesis, as brief exposures to the American flag indeed led to a conservative shift in U.S. participants. 70 participants completed the experiment in exchange for either $5 or extra credit in a psychology class. To manipulate the priming of an American flag or a control prime, participants were presented with four photographs and asked to estimate the time of day at which each photo was taken. For those in the flag-prime condition, the American flag was in the background of two of these photos, while in the control condition the same photos were presented but the flags were edited out. Following this priming manipulation, participants completed an 8-item questionnaire assessing views toward various political issues (e.g., abortion, gun control, affirmative action, etc.). Participants indicated their agreement with each position on 7-point scales from “1 – Completely Disagree” to “7 – Completely Agree”. The results supported their hypothesis, as participants in the flag-primed condition indicated significantly more conservative positions (*M* = 3.10) than those in the control condition (*M* = 2.65), *t*(64) = -2.04, *p* < .05, Cohen’s *d* = .51. The inclusion of this experiment provides a test for how priming effects generalize across settings. This particular paradigm was selected because it is important to understand how voting intentions and public opinion may be swayed by the location of polling sites or the makeup of ballots.
7. *Currency priming and system justification (Caruso et al., 2012)*. Money is another powerful symbol that most of us encounter every day, and that may impact us in ways we do not realize. Caruso et al. (2012) hypothesized that making money salient would induce increased support for free-market systems, and they provide evidence that merely exposing participants to money increases their endorsement of the current social system. 30 participants from a university subject pool completed the study online for a chance to win $25. The background screen was manipulated between subjects. In one condition the background showed a faint picture of U.S. $100 bills; in the other condition the background was a blurred, unidentifiable version of the same picture. After this screen, participants indicated their age, gender, and ethnicity. Next, participants completed an 8-question “system justification scale” (Kay & Jost, 2003) on which they responded to each item on a nine-point scale ranging from strongly disagree to strongly agree: In general, you find society to be fair; In general, the American political system operates as it should; American society needs to be radically restructured; The United States is the best country in the world to live in; Most policies serve the greater good; Everyone has a fair shot at wealth and happiness; Our society is getting worse every year (reverse-scored); Society is set up so that people usually get what they deserve. Participants in the money-prime condition scored higher on the system justification scale than those in the control condition, *t*(28) = 2.12, *p* = .043, *d* = 0.80. The inclusion of this experiment provides another test for how priming effects replicate across settings, and may reveal differences between replications indicating that money may have different associations depending on context.
8. *Imagined contact (Husnu & Crisp, 2010, Study 1).* Contact theory proposes that interacting with members of different racial and ethnic groups can reduce prejudice towards those groups when the interaction is positive (Allport, 1954). This effect has been shown even when participants merely imagine contact with outgroup members (Turner et al., 2007). Husnu and Crisp (2010) demonstrated this finding in a study with 33 British non-Muslim undergraduate students at the University of Kent. Participants were assigned to either an imagined contact group, or a control group: “I would like you to take a minute to imagine yourself meeting a British Muslim stranger for the first time. During the conversation imagine you find out some interesting and unexpected things about the stranger.” Or “I would like you to take a minute to imagine you are walking in the outdoors. Try to imagine aspects of the scene about you (e.g., is it a beach, a forest, are there trees, hills, what's on the horizon).” Participants were given one minute to imagine the scene, and then were instructed to “describe as many aspects of the scenario you just imagined as possible” for an additional minute on a response sheet. Participants then responded to the following questions as dependent measures: (1) How much do you intend to interact with British Muslims in the future? (1 = not at all, to 9 = very much), (2) How much time do you think you might spend learning about Islam in the future? (1 = none at all, to 9 = a lot of time), (3) How important do you think interacting with British Muslims is? (1 = not at all important, to 9 = highly important), and (4) How willing would you be to attend a mosque gathering to learn more about Islamic beliefs and practices? (1 = not at all willing, to 9 = very willing). The mean of these four items was taken as a composite index of intentions to engage in future contact (α=.82). Participants in the “imagined contact” group scored significantly higher (*M* = 5.93, *SD* = 1.67) on the composite measure of intentions to engage in future actual contact than participants in the control group (*M* = 4.69, *SD* = 1.26), *t*(31) = -2.39, *p* = .023, *d* = .86. This result represents a line of relatively new research that may have implications for diversity training, while adding diversity to our set of studies itself by being a very different paradigm from most others we included. Understanding how well these effects replicate across settings and populations may provide important insights as to how promising those prospects for future research are, and whether these “imagined contact” instructions work across settings.
9. *Sex differences in implicit math attitudes and relations with self-reported attitudes (Nosek, Banaji, & Greenwald, 2002).* Women are much less likely than men to pursue education and careers in science, technology, engineering, and math (STEM) fields. Nosek, Banaji, and Greenwald (2002) investigated this issue in two studies of Yale undergraduates. Participants completed four Implicit Association Tests (IATs) in random order. In this replication we will focus on the math attitudes IAT (associations between math/arts and good/bad). Self-reported math attitudes were measured with a combination of: (1) feeling thermometers (preference ratings based on a 0–100 scale from *cold/unfavorable* to *warm/favorable*) to assess participants’ feelings of warmth toward math and arts, and (2) six semantic differential scales (reported as five scales) measuring attitudes toward math and arts. Dichotomous pairs of adjectives anchored each end of a 7-point scale (from -3 to 3): *good–bad*, *happy–sad*, *delightful–disgusting*, *beautiful–ugly*, *approach–avoid*, and *unafraid– afraid*. Strong correspondence between the two self-report measures (*r* = .76) justified standardizing and combining them. They key findings to be replicated are that: (a) women’s implicit math were significantly more negative than men’s (Cohen’s *d* = 1.01 in Study 1 and 0.90 in Study 2), and (b) across studies, implicit math attitudes were significantly correlated with explicit math attitudes (*r* = .42). This article has been cited 373 times. It provides the possibility of a number of different important discoveries, such as the IAT working differently across different populations/settings, finding different trends in different populations, or finding that implicit and explicit attitudes in this content area vary depending on where the data is gathered (see Nosek & Smyth, 2011). Further, there is evidence that the gender gap in math performance - and perhaps evaluation - is declining and varies across contexts (Nosek et al., 2009). This suggests that the effect could be weaker now than in the past, and may show substantial variation across measurement contexts.

*Known Differences from Original Studies*

Below we outline the study-specific difference between our proposed replication studies and the original versions. After that we discuss more general differences.

1. *Sunk costs (Oppenheimer et al., 2009).* The replication version of this study will include one change from the original version – administration of this task will be adapted from pencil and paper to a computer format.
2. *Gain versus loss framing for combating disease (Tversky & Kahneman, 1981)*. The replication version of this study will include two changes from the original version. First, administration of this task will be adapted from pencil and paper to a computer format. Second, the wording of the question will be adjusted to allow for international replication. Specifically, “U.S.” will be replaced with “the country”, and the word “Asian” will be omitted from “an unusual Asian disease”.
3. *Anchoring (Jacowitz & Kahneman, 1995).* The replication version of this study will include two changes from the original version. First, the experiment will be shortened from the original version, using only four topics instead of fifteen, and only asking one question about each instead of three. Second, the administration of this task will be adapted from pencil and paper to a computer format.
4. *Retrospective gambler’s fallacy (Oppenheimer & Monin, 2009).* The replication version of this study will definitely include two changes and will possibly include more changes from the original version. The exact wording for this item is not in the published manuscript. We will contact the authors and attempt to retrieve the precise question wording which will be presented exactly. Failing that, we will reconstruct the wording from the information given in the article, and provide the authors with a link to the completed study to get them to approve that it closely follows the original implementation. Second, the condition where the man rolls two 6’s will be removed, so there will be only two conditions. Third, administration of this task will be adapted from pencil and paper to a computer format.
5. *Low-vs.-high category scales (Schwarz et al., 1985).* The replication version of this study will have two minor changes from the original. First, the materials will be translated from German to English. Second, it will be adapted from pencil and paper to a computer format.
6. *Norm of reciprocity (Hyman and Sheatsley, 1950).* The replication version of this study will include two changes from the original version. First, administration of this task will be adapted from pencil and paper to a computer format. Second, we will change the question slightly to ensure the “other country” is a suitable, modern target. The two questions will be worded as follows: (1) “In your opinion, should our government let reporters from North Korea come here and send back to their media the news as they see it?”, (2) “In your opinion, should the government of North Korea let our reporters go there and send back to our media the news as they see it?”. For international replication, the target country will be adapted as needed (determined by the researcher heading the replication) to ensure the target country remains suitable for the purposes of the study.
7. *Allowed/Forbidden (Rugg, 1941)*. The replication version of this study will include two changes from the original version. First, administration of this task will be adapted from pencil and paper to a computer format. Second, to facilitate international replication, the words “The United States” will be replaced with the name of the country in which it is being administered.
8. *Quote Attribution (Lorge & Curtis, 1936)*. This is the study that will differ most between original and replication. First, administration of this task will be adapted from pencil and paper to a computer format. Second, the replication of this study will be conceptual, using a different quotation and authors in order to avoid deception (as described in the Method section). Note that we will use “George Washington” as the positive target for all samples, not just the United States samples. Third, this will be a between-subjects manipulation rather than a within-subjects manipulation; participants will see and rate only one version of the quotation.
9. *Flag Priming (Carter et al., 2011; Study 2)*. The replication version of this task will be practically identical for U.S. participants. For international replications, the survey questions will be adapted on an individual basis to ensure they are appropriate for the political climate of the target country (as judged by the researcher heading the replication). However, the flag prime will always consist of an American flag. This is primarily because the original priming materials (images with a flag or without) make it impractical to substitute the flags of other countries. In addition, because the American flag is an international symbol it is plausible that it will still exert an effect on political attitudes in other countries. For this effect, we will only treat replications conducted in the United States as true replication attempts in terms of verifying the original effect.

Additionally, the original authors have identified two potential moderators to this effect. To investigate these potential moderators, we will include three additional items at the end of the replication study: (1) How much do you identify with being American? (1, not at all – 11, very much), (2) To what extent do you think the typical American is a Republican or Democrat? (1, Democrat – 7, Republican), (3) To what extent do you think the typical American is conservative or liberal? (1, Liberal – 7, Conservative). These items are also presented in Appendix D. By including these items at the end of the study the direct replication will remain unchanged, but we will be able to analyze these potential moderators in a follow up analysis.

1. *Currency priming and system justification (Caruso et al., 2012)*. This experiment will be presented exactly as in the original, although it will be modified as needed for international replications. For international replications, the money prime will be changed to a faint image of that country’s national currency, and the control prime will be a blurred (unidentifiable) version of that image. The system-justification questions will be adapted to reflect the name of the relevant country.
2. *Imagined contact (Husnu & Crisp, 2010, Study 1).* The replication version of this study will include two changes from the original version. First, administration of this task will be adapted from pencil and paper to a computer format. Second, the word “British” will be removed from all references to “British Muslims” to permit international replications.
3. *Sex differences in implicit math attitudes and relations with self-reported attitudes (Nosek, Banaji, & Greenwald, 2002).* The materials used in the replication will be identical to the original study; however, due to time constraints, we will shorten the experiment to only administer one IAT and set of explicit items.

There will be two additional differences from original studies that will be true of all of the replication versions:

1. Materials will be translated into the dominant language of the data collection setting.
2. Researchers interested in participating in this project were told that they could add this study to the beginning, middle, or end of another lab session. That combined with the reality of presenting twelve separate studies means that each study will be run in the context of every other study rather than being a fully-independent experiment. In addition, because these studies are all packaged together and administered at the same time, it is possible some order effects may appear. No tasks were included that were anticipated to have a lasting effect that would bleed into the next task. Even so, the order of the tasks will be randomized so each individual study will be first at some point, allowing for empirical evaluation of order effects in the aggregate sample.

*Procedure*

The package of studies will be replicated across many different environments, although always on computers. This includes purely online implementation with no face-to-face contact at all (e.g., Amazon Mechanical Turk; Project Implicit, university participant pool online options) as well as in-lab studies with experimenters present during the entire experiment. Experimenters in laboratory studies will not be aware of participant condition for each task. Additionally, experimenters will not interact with participants during the data collection portion unless participants have questions.

Participants will first complete a demographic questionnaire. Next all participants will complete an instructional manipulation check (Oppenheimer et al., 2009).In the instructional manipulation check, participants are presented with a question that contains an instruction to ignore the question asked and instead to click the header at the top of the screen. Participants who click the title at the top of the screen are considered to pass the Instructional Manipulation Check (IMC). Oppenheimer et al. (2009) found that 54% of participants passed their version of the IMC. We chose to include the IMC because it may offer a quick way for experimenters to increase the power of their experiments. Including it in this replication allows us to test whether it provides a benefit across a variety of studies and contexts, particularly whether contexts with high IMC failure rates show less replicability of findings than those with low failure rates.

Following the IMC, participants will be randomly presented with one condition of each of the 12 experiments, in randomized order. They will go through the tasks at their own pace and will be debriefed when they are finished.

Contextual factors will vary somewhat between individual labs conducting the replication, but these variations will be recorded and analyzed along with the final data. Each participating lab will be required to complete a post-study questionnaire in which they describe important features of their replication setting (see Appendix B).

*Confirmatory Analysis Plan*

Researchers will be informed of the study-completion deadline before committing to the project. Data from those researchers who have committed to the project but who fail to contribute 80 participants by the study deadline will not be included in the analysis.

Each of the twelve studies included in the experimental package has two conditions. All but one of the studies involves a simple comparison between the two conditions (a mean difference, frequency difference, or, in one case, a correlation difference). In the next section we report for each study our estimated statistical power given the original effect size and 40 or 500 participants per condition. This gives the minimum and maximum expected power for each individual test. Note that the total power, aggregated across all replication attempts, will be substantially higher than any of these estimates. We also report the statistical analysis that we will conduct for each replication.

1. *Sunk costs (Oppenheimer et al., 2009).* In the Oppenheimer et al. (2009) demonstration of sunk costs, participants were slightly more likely to go to the game if they had paid for the ticket (*M* = 7.46, SD not reported) than if the ticket had been free (*M* = 6.93, SD not reported), *F*(1, 211) = 2.74, *p* = .1, partial *η2*=.01. To test for replication we will conduct an independent samples *t*-test with condition (paid ticket, free ticket) as the independent variable and likelihood of going to the game as the dependent variable. Based on the effect size (Cohen’s *d* = 0.23) from the original study, 40 participants per condition will give us approximately 28% statistical power to find an effect; 500 participants per condition will give us approximately than 98% statistical power to detect an effect. This is likely an underestimate of the actual power because the Oppenheimer et al. (2009) demonstration is unusually low for this paradigm. As mentioned previously, in an unpublished replication conduct in the laboratory at the University of Virginia (N = 180), we found an effect size of 0.67 for this same procedure. This results in power estimates of approximately 91% statistical power to detect an effect with 40 participants per condition and greater than 99% statistical power to detect an effect with 500 participants per condition. No participant’s data will be excluded from analysis.
2. *Gain versus loss framing for combating disease (Tversky & Kahneman, 1981)*. In the original study, 72% of participants in Condition 1 chose to adopt Program A (saving 200 people) and 28% chose Program B (a 1/3 probability that 600 people will be saved). In Condition 2 this effect was reversed, as Program C (400 people die) was adopted by only 22 percent while 78 percent selected Program D (1/3 probability that no one will die). Although formal analyses were not reported in the original paper, our chi-square analysis reveals, *X2*(1) = 16.19, *p* < .0001, *r* = .32. To test for replication we will conduct a chi-square analysis on program choice with condition (gain-frame or loss-frame) as a between-subjects variable. Based on the effect size (*r* = 0.32) from the original study, 40 participants per condition will give us approximately 77% statistical power to detect an effect; 500 participants per condition will give us greater than 99% power to detect an effect. No participant’s data will be excluded from analysis.
3. *Anchoring (Jacowitz & Kahneman, 1995).* In the original study, a point-biserial correlation was computed between subjects’ estimates and the anchor they had seen. The mean point-biserial correlation over the 15 topics was .42. Although formal significance testing of this statistic was not reported in the original paper, a one-sample *t*-test reveals this effect to be significant, *t*(103) = 4.651, *p* < .0001. To test for replication we will analyze each of the four questions separately by conducting independent samples *t*-tests on the difference between estimates from the high and low anchor conditions. Based on the point-biserial correlation from the original study of .42, 40 participants per condition will give us approximately 98% statistical power to detect an effect. 500 participants per condition will give us greater than 99% chance to detect an effect. This study will also have two exclusion rules for open response data: (1) responses must be interpretable as a number, and (2) all responses outside of the anchor end points will be converted to missing data (e.g., responses greater than 5 million or less than 200,000 for the Chicago population item).
4. *Retrospective gambler’s fallacy (Oppenheimer & Monin, 2009).* In the original study, participants believed that a sequence of previous die rolls was more than three times as long when a set of three 6’s was observed (*M* = 34.2) than when there were two 6’s and a 3 (*M* = 3.2), *t*(57) = 2.65, *p* < .05, Cohen’s *d =* 0.69). To test for replication we will use an independent samples *t*-test with condition (three 6’s, two 6’s one 3) as the independent variable and expected length of sequence as the dependent variable. Based on the effect size (Cohen’s *d* = 0.69) from the original study, 40 participants per condition will give us approximately 70% statistical power to detect an effect; 500 participants per condition will give us greater than 99% statistical power to detect an effect. Responses greater than three standard deviations from the mean will be excluded from the analysis.
5. *Low-vs.-high category scales (Schwarz et al., 1985).* In the original study, 16.2 percent of the respondents who were presented the low category range reported watching TV for more than two and a half hours, while 37.5 percent of the respondents presented the high category range did so, *X2*(1)= 7.7, *p* < .01. Although the original authors did not report an effect size, we calculated it as *r* = .24. To test for replication we will conduct a chi-square analysis on likelihood of reporting watching 2.5 hours of television per day, with condition (high-frequency, low-frequency) as a between-subjects variable. Based on the effect size (*r* = 0.24) from the original study, 40 participants per condition will give us approximately 57% statistical power to find an effect; 500 participants per condition will give us greater than 99% power to detect an effect. No participant’s data will be excluded from analysis.
6. *Norm of reciprocity (Hyman and Sheatsley, 1950).* In the original study, 73% of participants supported allowing communist reporters into America when it was asked after a question about allowing American reporters into the communist country, but only 36% supported allowing communist reporters into America when the question about it was asked first. To test for replication we will conduct a chi-square analysis on statement-agreement with condition (US-first, North Korea-first) as a between-subjects variable. Because sample size is not reported in the original paper, it is not possible to calculate an effect size from the original data. However, it is possible to determine that, for this type of analysis, 40 participants per condition will give us approximately 15% power to detect a small effect (*r* = .10), 77% power to detect a medium effect (*r* = .30), and 99% power to detect a large effect (*r* = .50). 500 participants per condition will give us greater than 99% power to detect even a small effect. No participant’s data will be excluded from analysis.
7. *Allowed/Forbidden (Rugg, 1941)*. In the original study, 62% of participants indicated “No” when asked if speeches against democracy should be allowed, but only 46 percent indicated “Yes” when asked if these speeches should be forbidden. To test for replication we will conduct a chi-square analysis on policy choice with condition (allowed, forbidden) as a between-subjects variable. Because sample size is not reported in the original paper, it is not possible to calculate an effect size from the original data. However, it is possible to determine that, for this type of analysis, 40 participants per condition will give us approximately 15% power to detect a small effect (*r* =.10) 77% power to detect a medium effect (*r* = .30), and 99% power to detect a large effect (*r* = .50). 500 participants per condition will give us greater than 99% power to detect even a small effect. No participant’s data will be excluded from analysis.
8. *Quote Attribution (Lorge & Curtis, 1936)*. In the original study, participants indicated more agreement with a quotation from a liked individual than a disliked individual (exact numbers unknown; see method). To test for replication we will conduct an independent samples *t*-test with condition (liked, disliked speaker) as the independent variable and agreement with the quote as the dependent variable. Because sample size is not reported in the original paper, it is not possible to calculate an effect size from the original data. However, it is possible to determine that, for this type of analysis, 40 participants per condition will give us approximately 11% power to detect a small effect (Cohen’s *d* = 0.10), 38% power to detect a medium effect (Cohen’s *d* = 0.30), and 72% power to detect a large effect (Cohen’s *d* = 0.50). 500 participants per condition will give us approximately 47% statistical power to detect a small effect and greater than 99% power to detect a medium or large effect. No participant’s data will be excluded from analysis.
9. *Flag Priming (Carter et al., 2011; Study 2)*. In the original study, attitudes of participants in the flag-prime condition (*M* = 3.10) were significantly closer to the Republican end of the scale than were attitudes of participants in the control condition (*M* = 2.65), *t*(64) = -2.04, *p* < .05. Based on the effect size (Cohen’s *d* = 0.51) from the original study, 40 participants per condition will give us approximately 73% power to detect an effect. 500 participants per condition will give us greater than 99% statistical power to detect an effect. To ensure participants are examining the photos, those participants who do not provide a “time of day this photo was taken” estimate for all four priming photos will be excluded from analysis.
10. *Currency priming and system justification (Caruso et al., 2012)*. In the original study, participants in the money-prime condition (*M* = 4.96, *SD* = 1.27) scored higher on the system justification scale than those in the control condition (*M* = 3.99, *SD* = 1.19), *t*(28) = 2.12, *p* = .043, *d* = 0.80. To test for replication we will conduct an independent samples *t*-test on system justification with condition (money-prime, control) as a between-subjects variable. Based on the effect size (Cohen’s *d* = 0.80) from the original study, 40 participants per condition will give us approximately 97% statistical power to detect an effect. 500 participants per condition will give us greater than 99% statistical power to detect an effect. Data will be analyzed only from those participants who respond to at least six of the eight system justifications items; otherwise, no participant’s data will be excluded from analysis.
11. *Imagined contact (Husnu & Crisp, 2010, Study 1).* In the original study, participants in the “imagined contact” group scored significantly higher (*M* = 5.93, *SD* = 1.67) on the composite measure of intentions to engage in future actual contact than participants in the control group (*M* = 4.69, *SD* = 1.26), *t*(31) = -2.39, *p* = .023, *d* = .86. To test for replication we will conduct an independent samples *t*-test on contact intention with condition (imagined-contact, control) as the independent variable. Based on the effect size (Cohen’s *d* = 0.86) from the original study, 40 participants per condition will give us approximately 98% statistical power to find an effect; 500 participants per condition will give us greater than 99% statistical power to detect an effect. No participant’s data will be excluded from analysis.
12. *Sex differences in implicit math attitudes and relations with self-reported attitudes (Nosek, Greenwald, & Banaji, 2002).* In the original study, women’s implicit math attitudes were significantly more negative than men’s (Cohen’s *d* = 1.01 in Study 1 and 0.90 in Study 2), and (b) across studies, implicit math attitudes were significantly correlated with explicit math attitudes (*r* = .42). To test for replication of (a) we will conduct an independent samples *t*-test on implicit math attitudes with participant sex (men, women) as the independent variable. Based on the combined effect size (Cohen’s *d* = 0.96) from the original study, 40 or 500 participants per condition will give us greater than 99% power to detect an effect. To test for replication of (b) we will compute a correlation coefficient for the relationship between implicit and explicit math attitudes. Based on the effect size (*r* = .42) from the original study, 80 participants will give us approximately 98% power to detect an effect; 500 participants per condition will give us greater than 99% power to detect an effect. For all analyses, data will be excluded from those participants who do not respond to a particular item for both math and arts, and who do not respond to at least 6 of the eight total explicit attitude items. Data from the IAT will be analyzed using the *D* algorithm (Greenwald, Nosek, & Banaji, 2003) with the following features: response latencies < 400ms and >10,000ms will be removed, and trial latencies will be calculated from the beginning of the trial until the time of a correct response. We will exclude data from those participants who have greater than 40% errors on a single block or greater than 30% errors overall.

The specific analysis plan for each individual replication study is outlined above; however, there are multiple ways we can consider the data from each replication in the aggregate. “Successful” replication can be evidenced with many metrics: (1) The number of significant replications of each experiment, (2) the number of significant replications of each experiment with only the two high-powered samples, (3) the overall replicability when aggregated across all replications, and (4) the number of significant replications of each experiment (overall successful replications, successful replications with the two high-powered samples, and replications in the aggregate) when including only those participants who passed the Instructional Manipulation Check.

“Vote counting” methods of replication are easy but weak means of understanding an effect. With this design there is substantial opportunity for meta-analytic data aggregation in order to obtain (a) a highly precise estimate of the effect size with narrow confidence intervals, and (b) evaluation of whether heterogeneity of the effect across sample/setting exceeds variation expected by chance. The latter is particularly important for evaluating the effect of sample and setting variation on replicability.

For each of these studies we will report the following across all replication attempts: (a) descriptive statistics (mean, standard deviation, median, mode, range, etc.) for each condition across all replication attempts, (b) average effect size, (c) the range of effect sizes, and (d) the number of times the difference between Condition 1 and Condition 2 achieved statistical significant (*p* < .05, one-tailed test). We will also report the range of p-values. We will also report each effect for those who did and did not pass the Instructional Manipulation Check.

The prior analyses emphasized null-hypothesis significance testing for the evaluation of replication. However, treating a *p* = .06 as a failure to replicate and *p* = .04 as a success is not sufficient for evidencing actual replicability. As such the aggregate analysis, and the comparative analysis across laboratories will focus on effect size and confidence intervals. The real purpose of replication is to increase the precision of the estimated effect size. The value of the present design is that there will be an aggregate result with a relatively precise estimate, and variation in individual results - each with their own precision.

We do not have strong a priori hypotheses about the variation in effect sizes that we will observe across samples and settings. Therefore, comparative analyses will be reported as exploratory. For a list of contextual differences we will collect from researchers see Appendix B. For a list of the sample characteristics we will examine, see Appendix C. We will primarily use multilevel modelling analysis to examine the preceding items as predictors of replication. A detailed analysis plan will be registered separately, prior to beginning the actual analysis.

Through these analyses, we will provide speculative interpretations for why some effects showed little variation across samples and settings and others showed extensive variation. Ultimately, this part of the analysis will provide theory generation opportunity for the conditions under which psychological effects will show strong sensitivity to sample and context.

*Provisions for quality control*

All replications will be conducted via links provided to the individual labs by the core team of researchers. Thus, there will be no differences in the administration of the study materials themselves. In addition, all data will be automatically recorded into a single database, and data analysis will be conducted by the core team. All labs will fill out a survey to report the conditions the study was run under to ensure all aspects of the running were up to ethical and procedural standards. To ensure quality, each lab will be expected to videotape one sample experimental session and to submit the videotape along with their questionnaire. This videotaping will increase the accountability of each lab, and also allow for retrospective coding of differences between labs that may not have been included in the “lab conditions” questionnaire (see Appendix B).

*Implications*

An increased emphasis on replication will bolster the quality of research and accuracy of published findings in psychology. Replication enables verification and provides an opportunity to examine generalizability. The present investigation will assess the replicability of 12 effects across a number of settings and participant samples. This method will produce precise estimates of the effect sizes, and also provide evidence of boundary conditions or moderators of the effects. More broadly, this project will demonstrate how to investigate the contextual and sample specific factors that can affect replicability. Understanding such factors will improve theoretical understanding of the phenomena.

References

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?. *Perspectives on Psychological Science*, *6*(1), 3-5.

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychological science*, *22*(8), 1011-1018.

Caruso, E. M., Vohs, K. D., Baxter, B., & Waytz, A. (2012). Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General.*

Darke, P. R., & Freedman, J. L. (1997). Lucky events and beliefs in luck: Paradoxical effects on confidence and risk-taking. *Personality and Social Psychology Bulletin*, *23*, 378–388.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85(2)*, 197-216.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29-29.

Husnu, S., & Crisp, R. J. (2010). Elaboration enhances the imagined contact effect. *Journal of Experimental Social Psychology*, *46*(6), 943-950.

Hyman, H. H., & Sheatsley, P. B. (1950). The current status of American public opinion. In *The Teaching of Contemporary Affairs*, pp. 11-34. New York: National Council of Social Studies.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161-1166.

Kay, A.C., & Jost, J.T. (2003). Complementary justice: Effects of "poor but happy" and "poor but honest" stereotype exemplars on system justification and implicit activation of the justice motive. *Journal of Personality and Social Psychology, 85(5),* 823-837.

Knox, R. E., & Inkster, J. A. (1968). Postdecision dissonance at post time. *Journal of personality and social psychology*, *8*(4p1), 319.

Lorge, I., & Curtiss, C. C. (1936). Prestige, suggestion, and attitudes. *The Journal of Social Psychology*, *7*(4), 386-402.

Moskowitz, G. B. (2004). *Social cognition: Understanding self and others*. Guilford Press.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = Male, Me = Female, therefore Math ^= Me. *Journal of Personality and Social Psychology, 83(1)*, 44-59.

Nosek, B. A., & Smyth, F. L. (2011). Implicit social cognitions predict sex differences in math engagement and achievement. *American Educational Research Journal*, *48*(5), 1125-1156.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., et al. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences, USA, 106*, 10593–10597.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7,* 657-660.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 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*(5), 326-334.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411-419.

Ratliff, K. A., & Nosek, B. A. (2010). Creating distinct implicit and explicit attitudes with an illusory correlation paradigm. *Journal of Experimental Social Psychology, 46,*721-728.

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

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.

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*(3), 388-395.

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing science*, *4*(3), 199-214.

Tune, G. S. (1964). Response preferences: A review of some relevant literature. *Psychological Bulletin*, *61*, 286–302.

Turner, R. N., Crisp, R. J., & Lambert, E. (2007). Imagining intergroup contact can improve intergroup attitudes. *Group Processes and Intergroup Relations, 10*, 427−441.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458.

**Appendix A**

Text of recruitment message posted to the Open Science Framework Google Group (<https://groups.google.com/forum/?hl=en&fromgroups=#!forum/openscienceframework>).

Subject: Collaborators Needed for "Many Labs" Replication Project

Dear Colleagues,

We are currently seeking collaborators for a wide-scale replication project to be submitted for consideration in a special issue of *Social Psychology* “Replications of Important Results in Social Psychology.”

Thegoal of the project is to take a small set of documented effects in social psychology that are extremely easy to administer. We will perform replications in as many independent labs as possible. Some of the effects are known to be highly replicable, for others there is less knowledge about replicability. Also, some are thought to depend on the social context or sample, whereas others may not be. This will allow the results to be compared across various locations, participant populations, and lab set-ups, in order to learn more about the role these factors play in replicability.

We have identified an initial list of about 12 effects that take just seconds or a few minutes each to administer. All effects can be automated and will be run through a single experiment script (using the web-based Project Implicit infrastructure). This way, all replication teams can run the identical study script in their laboratories with their own samples, and it will only take a few clicks to launch.

Participants will be able to complete the study in 10-15 minutes. This way, the study could be administered as an independent session or in “extra time” after another data collection. We are also happy to work with you to resolve any issues that may impede your ability to administer the experiment.

With efforts such as the Reproducibility Project already underway, understanding the situational variables that influence replication takes on an important role. In addition to aiding in the advancement of this research, all participating researchers will be credited as co-authors on the final publication.

If you are interested in joining the project or would like to get further information or ask questions, please contact the investigators at [manylabsproject@gmail.com](mailto:manylabsproject@gmail.com).

Sincerely,

The Many Labs Team

Richard A. Klein, University of Florida

Dr. Kate A. Ratliff, University of Florida

Dr. Brian A. Nosek, University of Virginia

**Appendix B**

List of contextual information to be collected from researchers.

1. Features of the experimenter(s) – Gender, ethnicity.
2. How many participants were administered the survey at once?
3. Were participants separated or not?
4. Did participants participate in another study prior to being administered the replication? If so, please indicate what that study was.
5. How were participants recruited?
6. How were participants compensated (paid/volunteer/course credit)?
7. Was the study presented in the original language?

**Appendix C**

List of demographic questions to collect from participants:

1. Gender
2. Age
3. Race and ethnicity
4. Political ideology
5. Country of citizenship
6. Native language
7. Major (if any)
8. Prior exposure to included experiments (we will provide a list with brief descriptions, and they will mark if they have seen or learned about it previously).

**Appendix D**

Additional items for flag priming study (to be administered at the end of the study package).

1. How much do you identify with being American? (1, not at all – 11, very much)
2. To what extent do you think the typical American is a Republican or Democrat? (1, Democrat – 7, Republican)
3. To what extent do you think the typical American is conservative or liberal? (1, Liberal – 7, Conservative)