Investigating Variation in Replicability

A "Many Labs" Replication Project

Richard A. Klein,[1] Kate A. Ratliff,[1] Michelangelo Vianello,[2] Reginald B. Adams Jr.,[3] Štěpán Bahník,[4] Michael J. Bernstein,[5] Konrad Bocian,[6] Mark J. Brandt,[7] Beach Brooks,[1] Claudia Chloe Brumbaugh,[8] Zeynep Cemalcilar,[9] Jesse Chandler,[10,36] Winnee Cheong,[11] William E. Davis,[12] Thierry Devos,[13] Matthew Eisner,[10] Natalia Frankowska,[6] David Furrow,[15] Elisa Maria Galliani,[2] Fred Hasselman,[16,37] Joshua A. Hicks,[12] James F. Hovermale,[17] S. Jane Hunt,[18] Jeffrey R. Huntsinger,[19] Hans IJzerman,[7] Melissa-Sue John,[20] Jennifer A. Joy-Gaba,[17] Heather Barry Kappes,[21] Lacy E. Krueger,[18] Jaime Kurtz,[22] Carmel A. Levitan,[23] Robyn K. Mallett,[19] Wendy L. Morris,[24] Anthony J. Nelson,[3] Jason A. Nier,[25] Grant Packard,[26] Ronaldo Pilati,[27] Abraham M. Rutchick,[28] Kathleen Schmidt,[29] Jeanine L. Skorinko,[20] Robert Smith,[14] Troy G. Steiner,[3] Justin Storbeck,[8] Lyn M. Van Swol,[30] Donna Thompson,[15] A. E. van 't Veer,[7] Leigh Ann Vaughn,[31] Marek Vranka,[32] Aaron L. Wichman,[33] Julie A. Woodzicka,[34] and Brian A. Nosek[29,35]

[1]University of Florida, Gainesville, FL, USA, [2]University of Padua, Italy, [3]The Pennsylvania State University, University Park, PA, USA, [4]University of Würzburg, Germany, [5]Pennsylvania State University Abington, PA, USA, [6]University of Social Sciences and Humanities Campus Sopot, Poland, [7]Tilburg University, The Netherlands, [8]City University of New York, USA, [9]Koç University, Istanbul, Turkey, [10]University of Michigan, Ann Arbor, MI, USA, [11]HELP University, Kuala Lumpur, Malaysia, [12]Texas A&M University, College Station, TX, USA, [13]San Diego State University, CA, USA, [14]Ohio State University, Columbus, OH, USA, [15]Mount Saint Vincent University, Nova Scotia, Canada, [16]Radboud University Nijmegen, The Netherlands, [17]Virginia Commonwealth University, Richmond, VA, USA, [18]Texas A&M University-Commerce, TX, USA, [19]Loyola University Chicago, IL, USA, [20]Worcester Polytechnic Institute, MA, USA, [21]London School of Economics and Political Science, London, UK, [22]James Madison University, Harrisonburg, VA, USA, [23]Occidental College, Los Angeles, CA, USA, [24]McDaniel College, Westminster, MD, USA, [25]Connecticut College, New London, CT, USA, [26]Wilfrid Laurier University, Waterloo, ON, Canada, [27]University of Brasilia, DF, Brazil, [28]California State University, Northridge, CA, USA, [29]University of Virginia, Charlottesville, VA, USA, [30]University of Wisconsin-Madison, WI, USA, [31]Ithaca College, NY, USA, [32]Charles University, Prague, Czech Republic, [33]Western Kentucky University, Bowling Green, KY, USA, [34]Washington and Lee University, Lexington, VA, USA, [35]Center for Open Science, Charlottesville, VA, USA, [36]PRIME Research, Ann Arbor, MI, USA, [37]University Nijmegen, The Netherlands

Abstract. Although replication is a central tenet of science, direct replications are rare in psychology. This research tested variation in the replicability of 13 classic and contemporary effects across 36 independent samples totaling 6,344 participants. In the aggregate, 10 effects replicated consistently. One effect – imagined contact reducing prejudice – showed weak support for replicability. And two effects – flag priming influencing conservatism and currency priming influencing system justification – did not replicate. We compared whether the conditions such as lab versus online or US versus international sample predicted effect magnitudes. By and large they did not. The results of this small sample of effects suggest that replicability is more dependent on the effect itself than on the sample and setting used to investigate the effect.

Keywords: replication, reproducibility, generalizability, cross-cultural, variation

Replication is a central tenet of science; its purpose is to confirm the accuracy of empirical findings, clarify the conditions under which an effect can be observed, and estimate the true effect size (Brandt et al., 2013; Open Science Collaboration, 2012, 2014). Successful replication of an experiment requires the recreation of the essential conditions of the initial experiment. This is often easier said than done. There may be an enormous number of variables

Social Psychology 2014; Vol. 45(3):142–152
DOI: 10.1027/1864-9335/a000178

influencing experimental results, and yet only a few tested. In the behavioral sciences, many effects have been observed in one cultural context, but not observed in others. Likewise, individuals within the same society, or even the same individual at different times (Bodenhausen, 1990), may differ in ways that moderate any particular result.

Direct replication is infrequent, resulting in a published literature that sustains spurious findings (Ioannidis, 2005) and a lack of identification of the eliciting conditions for an effect. While there are good epistemological reasons for assuming that observed phenomena generalize across individuals and contexts in the absence of contrary evidence, the failure to directly replicate findings is problematic for theoretical and practical reasons. Failure to identify moderators and boundary conditions of an effect may result in overly broad generalizations of true effects across situations (Cesario, 2014) or across individuals (Henrich, Heine, & Norenzayan, 2010). Similarly, overgeneralization may lead observations made under laboratory observations to be inappropriately extended to ecological contexts that differ in important ways (Henry, MacLeod, Phillips, & Crawford, 2004). Practically, attempts to closely replicate research findings can reveal important differences in what is considered a direct replication (Schmidt, 2009), thus leading to refinements of the initial theory (e.g., Aronson, 1992; Greenwald, Pratkanis, Leippe, & Baumgardner, 1986). Close replication can also lead to the clarification of tacit methodological knowledge that is necessary to elicit the effect of interest (Collins, 1974).

Overview of the Present Research

Little attempt has been made to assess the variation in replicability of findings across samples and research contexts. This project examines the variation in replicability of 13 classic and contemporary psychological effects across 36 samples and settings. Some of the selected effects are known to be highly replicable; for others, replicability is unknown. Some may depend on social context or participant sample, others may not. We bundled the selected studies together into a brief, easy-to-administer experiment that was delivered to each participating sample through a single infrastructure (http://projectimplicit.net/).

There are many factors that can influence the replicability of an effect such as sample, setting, statistical power, and procedural variations. The present design standardizes procedural characteristics and ensures appropriate statistical power in order to examine the effects of sample and setting on replicability. At one extreme, sample and situational characteristics might have little effect on the tested effects – variation in effect magnitudes may not exceed expected random error. At the other extreme, effects might be highly contextualized – for example, replicating only with sample and situational characteristics that are highly consistent with the original circumstances. The primary contribution of this investigation is to establish a paradigm for testing replicability across samples and settings and provide a rich data set that allows the determinants of replicability to be explored. A secondary purpose is to demonstrate support for replicability for the 13 chosen effects. 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.

Method

Researcher Recruitment and Data Collection Sites

Project leads posted a call for collaborators to the online forum of the Open Science Collaboration on February 21, 2013 and to the SPSP Discussion List on July 13, 2013. Other colleagues were contacted personally. For inclusion, each replication team had to: (1) follow local ethical procedures, (2) administer the protocol as specified, (3) collect data from at least 80 participants,[1] (4) post a video simulation of the setting and administration procedure, and (5) document key features of recruiting, sample, and any changes to the standard protocol. In total, there were 36 samples and settings that collected data from a total of 6,344 participants (27 data collections in a laboratory and 9 conducted online; 25 from the US, 11 from other countries; see Table 1 for a brief description of sites and for a full descriptions of sites, site characteristics, and participant characteristics by site).

Selection of Replication Studies

Twelve studies producing 13 effects were chosen based on the following criteria:

1. *Suitability for online presentation.* Our primary concern was to give each study a ''fair'' replication that was true to the original design. By administering the study through a web browser, we were able to ensure procedural consistency across sites.
2. *Length of study.* We selected studies that could be administered quickly so that we could examine many of them in a single study session.
3. *Simple design.* With the exception of one correlational study, we selected studies that featured a simple, two-condition design.

[1] One sample fell short of this requirement ($N = 79$) but was still included in the analysis. All sites were encouraged to collect as many participants as possible beyond the required 80, but the decision to end data collection was determined independently by each site. Researchers had no access to the data prior to completing data collection.

Table 1. Data collection sites

Site identifier	Location	*N*	Online (O) or laboratory (L)	US or international (I)
Abington	Penn State Abington, Abington, PA	84	L	US
Brasilia	University of Brasilia, Brasilia, Brazil	120	L	I
Charles	Charles University, Prague, Czech Republic	84	L	I
Conncoll	Connecticut College, New London, CT	95	L	US
CSUN	California State University, Northridge, LA, CA	96	O	US
Help	HELP University, Malaysia	102	L	I
Ithaca	Ithaca College, Ithaca, NY	90	L	US
JMU	James Madison University, Harrisonburg, VA	174	O	US
KU	Koç University, Istanbul, Turkey	113	O	I
Laurier	Wilfrid Laurier University, Waterloo, Ontario, Canada	112	L	I
LSE	London School of Economics and Political Science, London, UK	277	L	I
Luc	Loyola University Chicago, Chicago, IL	146	L	US
McDaniel	McDaniel College, Westminster, MD	98	O	US
MSVU	Mount Saint Vincent University, Halifax, Nova Scotia, Canada	85	L	I
MTURK	Amazon Mechanical Turk (US workers only)	1,000	O	US
OSU	Ohio State University, Columbus, OH	107	L	US
Oxy	Occidental College, LA, CA	123	L	US
PI	Project Implicit Volunteers (US citizens/residents only)	1,329	O	US
PSU	Penn State University, University Park, PA	95	L	US
QCCUNY	Queens College, City University of New York, NY	103	L	US
QCCUNY2	Queens College, City University of New York, NY	86	L	US
SDSU	SDSU, San Diego, CA	162	L	US
SWPS	University of Social Sciences and Humanities Campus Sopot, Sopot, Poland	79	L	I
SWPSON	Volunteers visiting www.badania.net	169	O	I
TAMU	Texas A&M University, College Station, TX	187	L	US
TAMUC	Texas A&M University-Commerce, Commerce, TX	87	L	US
TAMUON	Texas A&M University, College Station, TX (Online participants)	225	O	US
Tilburg	Tilburg University, Tilburg, Netherlands	80	L	I
UFL	University of Florida, Gainesville, FL	127	L	US
UNIPD	University of Padua, Padua, Italy	144	O	I
UVA	University of Virginia, Charlottesville, VA	81	L	US
VCU	VCU, Richmond, VA	108	L	US
Wisc	University of Wisconsin-Madison, Madison, WI	96	L	US
WKU	Western Kentucky University, Bowling Green, KY	103	L	US
WL	Washington & Lee University, Lexington, VA	90	L	US
WPI	Worcester Polytechnic Institute, Worcester, MA	87	L	US

4. *Diversity of effects.* We sought to diversify the sample of effects by topic, time period of original investigation, and differing levels of certainty and existing impact. Justification for study inclusion is described in the registered proposal (http://osf.io/project/aBEsQ/).

The Replication Studies

All replication studies were translated into the dominant language of the country of data collection (N = 7 languages total; 3/6 translations from English were back-translated). Next, we provide a brief description of each experiment, original finding, and known differences between original and replication studies. Most original studies were conducted with paper and pencil, all replications were conducted via computer. Exact wording for each study, including a link to the study, can be found in the supplementary materials. The relevant findings from the original studies can be found in the original proposal.

1. *Sunk costs (Oppenheimer, Meyvis, & Davidenko, 2009).* Sunk costs are those that have already been incurred and cannot be recovered (Knox & Inkster, 1968). Oppenheimer et al. (2009; adapted from Thaler, 1985) asked participants to imagine that they have tickets to see their favorite football team play an important game, but that it is freezing cold on the day of the game. Participants rated their likelihood of attending the game on a 9-point scale (1 = *definitely stay at home*, 9 = *definitely go to the game*). Participants were marginally more likely to go to the game if they had paid for the ticket than if the ticket had been free.

2. *Gain versus loss framing (Tversky & Kahneman, 1981).* The original research showed that changing the focus from losses to gains decreases participants' willingness to take risks – that is, gamble to get a better outcome rather than take a guaranteed result. Participants imagined that the US was preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Participants were then asked to select a course of action to combat the disease from logically identical sets of alternatives framed in terms of gains as follows: Program A will save 200 people (400 people will die), or Program B which has a 1/3 probability that 600 people will be saved (nobody will die) and 2/3 probability that no people will be saved (600 people will die). In the ''gain'' framing condition, participants are more likely to adopt Program A, while this effect reverses in the loss framing condition. The replication replaced the phrase ''the United States'' with the country of data collection, and the word ''Asian'' was omitted from ''an unusual Asian disease.''
3. *Anchoring (Jacowitz & Kahneman, 1995).* Jacowitz and Kahneman (1995) presented a number of scenarios in which participants estimated size or distance after first receiving a number that was clearly too large or too small. In the original study, participants answered 3 questions about each of 15 topics for which they estimated a quantity. First, they indicated if the quantity was greater or less than an anchor value. Second, they estimated the quantity. Third, they indicated their confidence in their estimate. The original number served as an anchor, biasing estimates to be closer to it. For the purposes of the replication we provided anchoring information before asking just for the estimated quantity for four of the topics from the original study – distance from San Francisco to New York City, population of Chicago, height of Mt. Everest, and babies born per day in the US for countries that use the metric system, we converted anchors to metric units and rounded them.
4. *Retrospective gambler's fallacy (Oppenheimer & Monin, 2009).* Oppenheimer and Monin (2009) investigated whether the rarity of an independent, chance observation influenced beliefs about what occurred before that event. Participants imagined that they saw a man rolling dice in a casino. In one condition, participants imagined witnessing three dice being rolled and all came up 6's. In a second condition two came up 6's and one came up 3. In a third condition, two dice were rolled and both came up 6's. All participants then estimated, in an open-ended format, how many times the man had rolled the dice before they entered the room to watch him. Participants estimated that the man rolled dice more times when they had seen him roll three 6's than when they had seen him roll two 6's or two 6's and a 3. For the replication, the condition in which the man rolls two 6's was removed leaving two conditions.
5. *Low-versus-high category scales (Schwarz, Hippler, Deutsch, & Strack, 1985).* Schwarz and colleagues (1985) demonstrated that people infer from response options what are low and high frequencies of a behavior, and self-assess accordingly. In the original demonstration, participants were asked how much TV they watch daily on a low-frequency scale ranging from ''up to half an hour'' to ''more than two and a half hours,'' or a high-frequency scale ranging from ''up to two and a half hours'' to ''more than four and a half hours.'' In the low-frequency condition, fewer participants reported watching TV for more than two and a half hours than in the high-frequency condition.
6. *Norm of reciprocity (Hyman & Sheatsley, 1950).* 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 the original study, American participants answered two questions: whether communist countries should allow American reporters in and allow them to report the news back to American papers and whether America should allow communist reporters into the United States and allow them to report back to their papers. Participants reported more support for allowing communist reporters into America when that question was asked after the question about allowing American reporters into the communist countries. In the replication, we changed the question slightly to ensure the ''other country'' was a suitable, modern target (North Korea). For international replication, the target country was determined by the researcher heading that replication to ensure suitability (see supplementary materials).
7. *Allowed/Forbidden (Rugg, 1941).* Question phrasing can influence responses. Rugg (1941) found that respondents were less likely to endorse forbidding speeches against democracy than they were to not endorse allowing speeches against democracy. Respondents in the United States were asked, in one condition, if the US should allow speeches against democracy or, in another condition, whether the US should forbid speeches against democracy. Sixty-two percent of participants indicated ''No'' when asked if speeches against democracy should be allowed, but only 46% indicated ''Yes'' when asked if these speeches should be forbidden. In the replication, the words ''The United States'' were replaced with the name of the country the study was administered in.
8. *Quote Attribution (Lorge & Curtiss, 1936).* The source of information has a great impact on how that information is perceived and evaluated. Lorge and Curtiss

(1936) examined how an identical quote would be perceived if it was attributed to a liked or disliked individual. Participants were asked to rate their agreement with a list of quotations. 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.'' In one condition the quote was attributed to Thomas Jefferson, a liked individual, and in the other it was attributed to Vladimir Lenin, a disliked individual. More agreement was observed when the quote was attributed to Jefferson than Lenin (reported in Moskowitz, 2004). In the replication, we used a quote attributed to either George Washington (liked individual) or Osama Bin Laden (disliked individual).

9. *Flag Priming (Carter, Ferguson, & Hassin, 2011; Study 2).* The American flag is a powerful symbol in American culture. Carter et al. (2011) examined how subtle exposure to the flag may increase conservatism among US participants. Participants were presented with four photos and asked to estimate the time of day at which they were taken. In the flag-prime condition, the American flag appeared in two of these photos. In the control condition, the same photos were presented without flags. Following the manipulation, participants completed an 8-item questionnaire assessing views toward various political issues (e.g., abortion, gun control, affirmative action). Participants in the flag-primed condition indicated significantly more conservative positions than those in the control condition. The priming stimuli used to replicate this finding were obtained from the authors and identical to those used in the original study. Because it was impractical to edit the images with unique national flags, the American flag was always used as a prime. As a consequence, the replications in the United States were the only ones considered as direct replications. For international replications, the survey questions were adapted slightly to ensure they were appropriate for the political climate of the country, as judged by the researcher heading that particular replication (see supplementary materials). Further, the original authors suggested possible moderators that they have considered since publication of the original study. We included three items at the very end of the replication study to test these moderators: (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*).
10. *Currency priming (Caruso, Vohs, Baxter, & Waytz, 2013).* Money is a powerful symbol. Caruso et al. (2013) provide evidence that merely exposing participants to money increases their endorsement of the current social system. Participants were first presented with demographic questions, with the background of the page manipulated between subjects. In one condition the background showed a faint picture of US$100 bills; in the other condition the background was a blurred, unidentifiable version of the same picture. Next, participants completed an 8-question ''system justification scale'' (Kay & Jost, 2003). Participants in the money-prime condition scored higher on the system justification scale than those in the control condition. The authors provided the original materials allowing us to construct a near identical replication for US participants. However, the stimuli were modified for international replications in two ways: First, the US dollar was usually replaced with the relevant country's currency (see supplementary materials); Second, the system justification questions were adapted to reflect the name of the relevant country.
11. *Imagined contact (Husnu & Crisp, 2010; Study 1).* Recent evidence suggests that merely imagining contact with members of ethnic outgroups is sufficient to reduce prejudice toward those groups (Turner, Crisp, & Lambert, 2007). In Husnu and Crisp (2010), British non-Muslim participants were assigned to either imagine interacting with a British Muslim stranger or to imagine that they were walking outdoors (control condition). Participants imagined the scene for one minute, and then described their thoughts for an additional minute before indicating their interest and willingness to interact with British Muslims on a four-item scale. Participants in the ''imagined contact'' group had significantly higher contact intentions than participants in the control group. In the replication, the word ''British'' was removed from all references to ''British Muslims.'' Additionally, for the predominately Muslim sample from Turkey the items were adapted so Christians were the outgroup target.
12. *Sex differences in implicit math attitudes (Nosek, Banaji, & Greenwald, 2002).* As a possible account for the sex gap in participation in science and math, Nosek and colleagues (2002) found that women had more negative implicit attitudes toward math compared to arts than men did in two studies of Yale undergraduates. Participants completed four Implicit Association Tests (IATs) in random order, one of which measured associations of math and arts with positivity and negativity. The replication simplified the design for length to be just a single IAT.
13. *Implicit math attitudes relations with self-reported attitudes (Nosek et al., 2002).* In the same study as Effect 12, self-reported math attitudes were measured with a composite of feeling thermometers and semantic differential ratings, and the composite was positively related with the implicit measure. The replication used a subset of the explicit items (see supplementary materials).

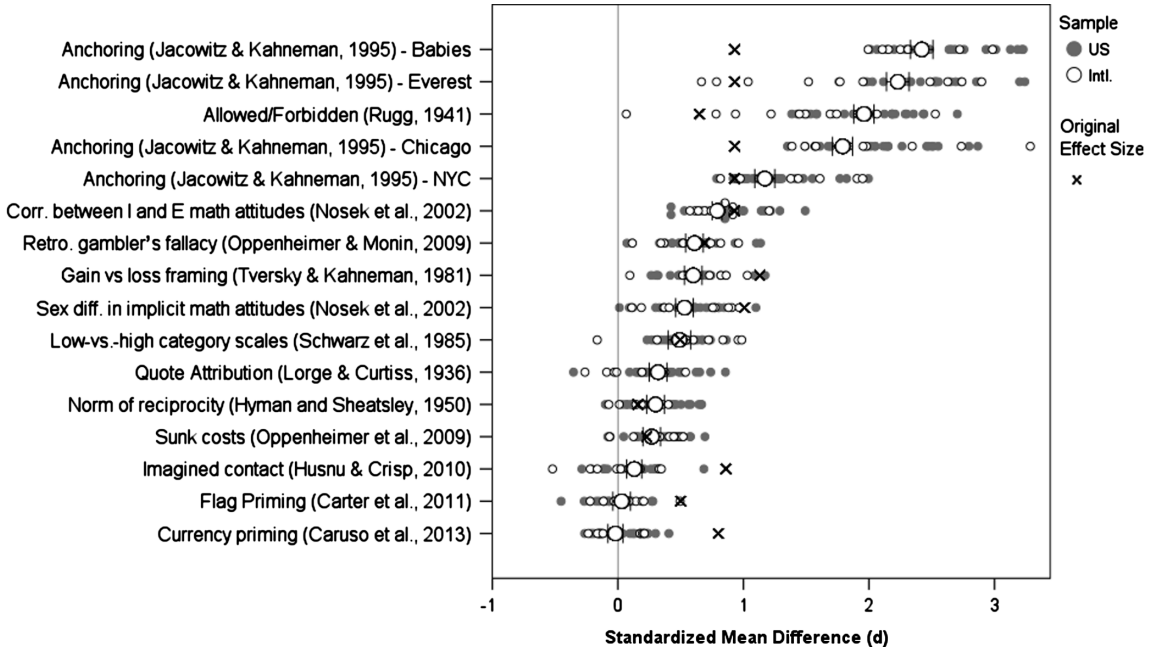

Figure 1. Replication results organized by effect. "X" indicates the effect size obtained in the original study. Large circles represent the aggregate effect size obtained across all participants. Error bars represent 99% noncentral confidence intervals around the effects. Small circles represent the effect sizes obtained within each site (black and white circles for US and international replications, respectively).

Procedure

The experiments were implemented on the Project Implicit infrastructure and all data were automatically recorded in a central database with a code identifying the sample source. After a paragraph of introduction, the studies were presented in a randomized order, except that the math IAT and associated explicit measures were always the final study. After the studies, participants completed an instructional manipulation check (IMC; Oppenheimer et al., 2009), a short demographic questionnaire, and then the moderator measures for flag priming. See Table S1[2] for IMC and summary demographic information by site. The IMC was not analyzed further for this report. Each replication team had a private link for their participants, and they coordinated their own data collection. Experimenters in laboratory studies were not aware of participant condition for each task, and did not interact with participants during data collection unless participants had questions. Investigators who led replications at specific sites completed a questionnaire about the experimental setting (responses summarized in Table S1), and details and videos of each setting along with the actual materials, links to run the study, supplemental tables, datasets, and original proposal are available at https://osf.io/ydpbf/.

Confirmatory Analysis Plan

Prior to data collection we specified a confirmatory analysis plan. All confirmatory analyses are reported either in text or in supplementary materials. A few of the tasks produced highly erratic distributions (particularly anchoring) requiring revisions to those analysis plans. A summary of differences between the original plans and actual analysis is reported in the supplementary materials.

Results

Summary Results

Figure 1 presents an aggregate summary of replications of the 13 effects, presenting each of the four anchoring effects separately. Table 2 presents the original effect size, median effect size, weighted and unweighted effect size and 99% confidence intervals, and proportion of samples that rejected the null hypothesis in the expected and unexpected direction. In the aggregate, 10 of the 13 studies replicated the original results with varying distance from the original

[2] Table names that begin with the prefix "S" (e.g., Table S1) refer to tables that can be found in the supplementary materials. Tables with no prefix are in this paper.

Table 2. Summary confirmatory results for original and replicated effects

	Original study			Unweighted		Weighted		Null hypothesis significance tests by sample (N = 36)			Null hypothesis significance tests of aggregate			
Effect	ES	95% CI lower, upper	Median replication ES	Replication ES	99% CI lower, upper	Replication ES	99% CI lower, upper	Proportion $p < .05$, opposite direction	Proportion $p < .05$, same direction	Proportion *ns*	Key statistics	*df*	*N*	*p*
Anchoring – babies born	0.93	.51, 1.33	2.43	2.60	2.41, 2.79	2.42	2.33, 2.51	0.00	1.00	0.00	$t = 90.49$	5,607	5,609	<.001
Anchoring – Mt. Everest	0.93	.51, 1.33	2.00	2.45	2.12, 2.77	2.23	2.14, 2.32	0.00	1.00	0.00	$t = 83.66$	5,625	5,627	<.001
Allowed/forbidden	0.65	.57, .73	1.88	1.87	1.58, 2.16	1.96	1.88, 2.04	0.00	0.97	0.03	$\chi^2 = 3{,}088.7$	1	6,292	<.001
Anchoring – Chicago	0.93	.51, 1.33	1.88	2.05	1.84, 2.25	1.79	1.71, 1.87	0.00	1.00	0.00	$t = 65.00$	5,282	5,284	<.001
Anchoring – distance to NYC	0.93	.51, 1.33	1.18	1.27	1.13, 1.40	1.17	1.09, 1.25	0.00	1.00	0.00	$t = 42.86$	5,360	5,362	<.001
Relations between I and E math attitudes	0.93	.77, 1.08	0.84	0.79	0.63, 0.96	0.79	0.75, 0.83	0.00	0.94	0.06	$r = .38$		5,623	<.001
Retrospective gambler fallacy	0.69	.16, 1.21	0.61	0.59	0.49, 0.70	0.61	0.54, 0.68	0.00	0.83	0.17	$t = 24.01$	5,940	5,942	<.001
Gain vs. loss framing	1.13	.89, 1.37	0.58	0.62	0.52, 0.71	0.60	0.53, 0.67	0.00	0.86	0.14	$\chi^2 = 516.4$	1	6,271	<.001
Sex differences in implicit math attitudes	1.01	.54, 1.48	0.59	0.56	0.45, 0.68	0.53	0.46, 0.60	0.00	0.71	0.29	$t = 19.28$	5,840	5,842	<.001
Low vs. high category scales	0.50	.15, .84	0.50	0.51	0.42, 0.61	0.49	0.40, 0.58	0.00	0.67	0.33	$\chi^2 = 342.4$	1	5,899	<.001
Quote attribution	na		0.30	0.31	0.19, 0.42	0.32	0.25, 0.39	0.00	0.47	0.53	$t = 12.79$	6,323	6,325	<.001
Norm of reciprocity	0.16	.06, .27	0.27	0.27	0.18, 0.36	0.30	0.23, 0.37	0.00	0.36	0.64	$\chi^2 = 135.3$	1	6,276	<.001
Sunk costs	0.23	−.04, .50	0.32	0.31	0.22, 0.39	0.27	0.20, 0.34	0.00	0.50	0.50	$t = 10.83$	6,328	6,330	<.001
Imagined contact	0.86	.14, 1.57	0.12	0.10	0.00, 0.19	0.13	0.07, 0.19	0.03	0.11	0.86	$t = 5.05$	6,334	6,336	<.001
Flag priming	0.50	.01, .99	0.02	0.01	−0.07, 0.08	0.03	−0.04, 0.10	0.04	0.00	0.96	$t = 0.88$	4,894	4,896	0.38
Currency priming	0.80	.05, 1.54	0.00	0.01	−0.06, 0.09	−0.02	−0.08, 0.04	0.00	0.03	0.97	$t = -0.79$	6,331	6,333	0.83

Notes. All effect sizes (ES) presented in Cohen's d units. Weighted statistics are computed on the whole aggregated dataset (N > 6,000); Unweighted statistics are computed on the disaggregated dataset (N = 36). 95% CI's for original effect sizes used cell sample sizes when available and assumed equal distribution across conditions when not available. The original anchoring article did not provide sufficient information to calculate effect sizes for individual scenarios, therefore an overall effect size is reported. The Anchoring original effect size is a mean point-biserial correlation computed across 15 different questions in a test-retest design, whereas the present replication adopted a between-subjects design with random assignments. One sample was removed from sex difference and relations between implicit and explicit math attitudes because of a systemic error in that laboratory's recording of reaction times. Flag priming includes only US samples. Confidence intervals around the unweighted mean are based on the central normal distribution. Confidence intervals around the weighted effect size are based on noncentral distributions.

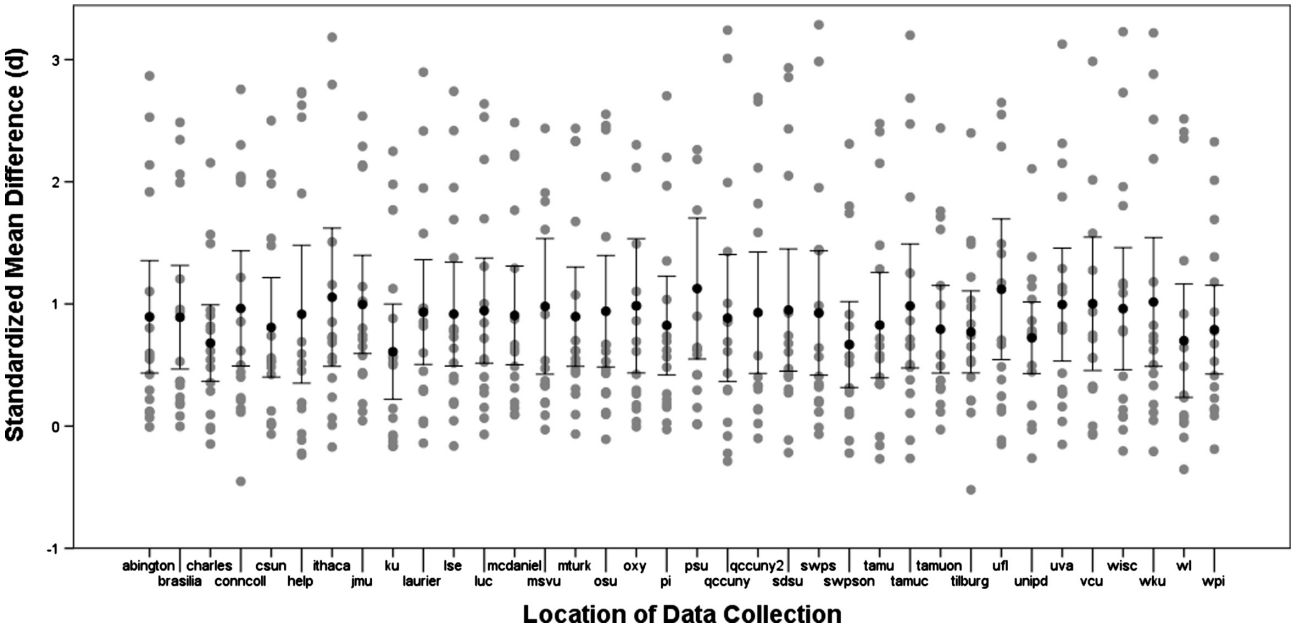

Figure 2. Replication results organized by site. Gray circles represent the effect size obtained for each effect within a site. Black circles represent the mean effect size obtained within a site. Error bars represent 95% confidence interval around the mean.

effect size. One study, imagined contact, showed a significant effect in the expected direction in just 4 of the 36 samples (and once in the wrong direction), but the confidence intervals for the aggregate effect size suggest that it is slightly different than zero. Two studies – flag priming and currency priming – did not replicate the original effects. Each of these had just one p-value < .05 and it was in the wrong direction for flag priming. The aggregate effect size was near zero whether using the median, weighted mean, or unweighted mean. All confidence intervals included zero. Figure 1 presents all 36 samples for flag priming, but only US data collections were counted for the confirmatory analysis (see Table 2). International samples also did not show a flag priming effect (weighted mean d = .03, 99% CI [−.04, .10]). To rule out the possibility that the priming effects were contaminated by the contents of other experimental materials, we reexamined only those participants who completed these tasks first. Again, there was no effect (Flag Priming: $t(431) = 0.33$, p = .75, 95% CI [−.171, .240], Cohen's d = .03; Currency Priming: $t(605) = -0.56$, p = .57, 95% CI [−.201, .112], Cohen's d = .05).[3]

When an effect size for the original study could be calculated, it is presented as an "X" in Figure 1. For three effects (contact, flag priming, and currency priming), the original effect is larger than for any sample in the present study, with the observed median or mean effect at or below the lower bound of the 95% confidence interval for the original effect.[4] Though the sex difference in implicit math attitudes effect was within the 95% confidence interval of the original result, the replication estimate combined with another large-scale replication (Nosek & Smyth, 2011) suggests that the original effect was an overestimate.

Variation Across Samples and Settings

Figure 1 demonstrates substantial variation for some of the observed effects. That variation could be a function of the true effect size, random error, sample differences, or setting differences. Comparing the intra-class correlation of samples across effects (ICC = .005; $F(35, 385) = 1.06$, p = .38, 95% CI [−.027, .065]) with the intra-class correlation of effects across samples (ICC = .75; $F(12,420) = 110.62$, p < .001, 95% CI [.60, .89]) suggests that very little in the variability of effect sizes can be attributed to the samples, and substantial variability is attributable to the effect under investigation. To illustrate, Figure 2 shows the same data as Figure 1 organized by sample rather than by effect. There is almost no variation in the average effect size across samples.

However, it is possible that particular samples would elicit larger magnitudes for some effects and smaller magnitudes for others. That might be missed by the aggregate analyses. Table 3 presents tests of whether the heterogeneity of effect sizes for each effect exceeds what is expected by measurement error. Cochran's Q and I^2 statistics

[3] None of the effects was moderated by which position in the study procedure it was administered.
[4] The original anchoring report did not distinguish between topics so the aggregate effect size is reported.

Table 3. Tests of effect size heterogeneity

	Heterogeneity statistics				Moderation tests					
Effect	*Q*	*DF*	*p*	I^2	US or international	*p*	η_p^2	Laboratory or online	*p*	η_p^2
Anchoring – babies born	59.71	35	0.01	0.402	0.16	0.69	0.00	16.14	<0.01	0.00
Anchoring – Mt. Everest	152.34	35	<.0001	0.754	94.33	<0.01	0.02	119.56	<0.01	0.02
Allowed/forbidden	180.40	35	<.0001	0.756	70.37	<0.01	0.01	0.55	0.46	0.00
Anchoring – Chicago	312.75	35	<.0001	0.913	0.62	0.43	0.00	32.95	<0.01	0.01
Anchoring – distance to NYC	88.16	35	<.0001	0.643	9.35	<0.01	0.00	15.74	<0.01	0.00
Relations between I and E math attitudes	54.84	34	<.0001	0.401	0.41*	0.52	<.001*	2.80*	0.09	<.001*
Retrospective gambler fallacy	50.83	35	0.04	0.229	0.40	0.53	0.00	0.34	0.56	0.00
Gain vs. loss framing	37.01	35	0.37	0.0001	0.09	0.76	0.00	1.11	0.29	0.00
Sex differences in implicit math attitudes	47.60	34	0.06	0.201	0.82	0.37	0.00	1.07	0.30	0.00
Low vs. high category scales	36.02	35	0.42	0.192	0.16	0.69	0.00	0.02	0.88	0.00
Quote attribution	67.69	35	<.001	0.521	8.81	<0.01	0.001	0.50	0.48	0.00
Norm of reciprocity	38.89	35	0.30	0.172	5.76	0.02	0.00	0.64	0.43	0.00
Sunk costs	35.55	35	0.44	0.092	0.58	0.45	0.00	0.25	0.62	0.00
Imagined contact	45.87	35	0.10	0.206	0.53	0.47	0.00	4.88	0.03	0.00
Flag priming	30.33	35	0.69	0	0.53	0.47	0.00	1.85	0.17	0.00
Currency priming	28.41	35	0.78	0	1.00	0.32	0.00	0.11	0.74	0.00

Notes. Tasks ordered from largest to smallest observed effect size (see Table 2). Heterogeneity tests conducted with R-package metafor. REML was used for estimation for all tests. One sample was removed from sex difference and relations between implicit and explicit math attitudes because of a systemic error in that laboratory's recording of reaction times.

*Moderator statistics are *F* value of the interaction of condition and the moderator from an ANOVA with condition, country, and location as independent variables with the exception of relations between impl. and expl. math attitudes for is reported the *F* value associated with the change in R squared after the product term between the independent variable and the moderator is added in a hierarchical linear regression model. Details of all analyses are available in the supplement.

revealed that heterogeneity of effect sizes was largely observed among the very large effects – anchoring, allowed-forbidden, and relations between implicit and explicit attitudes. Only one other effect – quote attribution – showed substantial heterogeneity. This appears to be partly attributable to this effect occurring more strongly in US samples and to a lesser degree in international samples.

To test for moderation by key characteristics of the setting, we conducted a Condition × Country (US or other) × Location (lab or online) ANOVA for each effect. Table 3 presents the essential Condition × Country and Condition × Location effects. Full model results are available in supplementary materials. A total of 10 of the 32 moderation tests were significant, and seven of those were among the largest effects – anchoring and allowed-forbidden. Even including those, none of the moderation effect sizes exceeded a η_p^2 of .022. The heterogeneity in anchoring effects may be attributable to differences in knowledge of the height of Mt Everest, distance to NYC, or population of Chicago between the samples. Overall, whether the sample was collected in the US or elsewhere, or whether data collection occurred online or in the laboratory, had little systematic effect on the observed results.

Additional possible moderators of the flag priming effect were suggested by the original authors. On the US participants only ($N \sim 4{,}670$), with five hierarchical regression models, we tested whether the items moderated the effect of the manipulation. They did not (*p*'s = .48, .80, .62, .07, .05, all $\Delta R^2 < .001$). Details are available in the online supplement.

Discussion

A large-scale replication with 36 samples successfully replicated eleven of 13 classic and contemporary effects in psychological science, some of which are well-known to be robust, and others that have been replicated infrequently or not at all. The original studies produced underestimates of some effects (e.g., anchoring-and-adjustment and allowed versus forbidden message framing), and overestimates of other effects (e.g., imagined contact producing willingness to interact with outgroups in the future). Two effects – flag priming influencing conservatism and currency priming influencing system justification – did not replicate.

A primary goal of this investigation was to examine the heterogeneity of effect sizes by the wide variety of samples and settings, and to provide an example of a paradigm for testing such variation. Some studies were conducted online, others in the laboratory. Some studies were conducted in the United States, others elsewhere. And, a wide variety of educational institutions took part. Surprisingly, these factors did not produce highly heterogeneous effect sizes.

Intraclass correlations suggested that most of the variation in effects was due to the effect under investigation and almost none to the particular sample used. Focused tests of moderating influences elicited sporadic and small effects of the setting, while tests of heterogeneity suggested that most of the variation in effects is attributable to measurement error. Further, heterogeneity was mostly restricted to the largest effects in the sample – counter to an intuition that small effects would be the most likely to be variable across sample and setting. Further, the lack of heterogeneity is particularly interesting considering that there is substantial interest and commentary about the contingency of effects on our two moderators, lab versus online (Gosling, Vazire, Srivastava, & John, 2004; Paolacci, Chandler, & Ipeirotis, 2010), and cultural variation across nations (Henrich et al., 2010).

All told, the main conclusion from this small sample of studies is that, to predict effect size, it is much more important to know what effect is being studied than to know the sample or setting in which it is being studied. The key virtue of the present investigation is that the study procedure was highly standardized across data collection settings. This minimized the likelihood that factors other than sample and setting contributed to systematic variation in effects. At the same time, this conclusion is surely constrained by the small, nonrandom sample of studies represented here. Additionally, the replication sites included in this project cannot capture all possible cultural variation, and most societies sampled were relatively Western, Educated, Industrialized, Rich, and Democratic (WEIRD; Henrich et al., 2010). Nonetheless, the present investigation suggests that we should not necessarily assume that there are differences between samples; indeed, even when moderation was observed in this sample, the effects were still quite robust in each setting.

The present investigation provides a summary analysis of a very large, rich dataset. This dataset will be useful for additional exploratory analysis about replicability in general, and these effects in particular. The data are available for download at the Open Science Framework (https://osf.io/ydpbf/).

Conclusion

This investigation offered novel insights into variation in the replicability of psychological effects, and specific information about the replicability of 13 effects. This methodology – crowdsourcing dozens of laboratories running an identical procedure – can be adapted for a variety of investigations. It allows for increased confidence in the existence of an effect and for the investigation of an effect's dependence on the particular circumstances of data collection (Open Science Collaboration, 2014). Further, a consortium of laboratories could provide mutual support for each other by conducting similar large-scale investigations on original research questions, not just replications. Thus, collective effort could accelerate the identification and verification of extant and novel psychological effects.

Note From the Editors

Commentaries and a rejoinder on this paper are available (Crisp, Miles, & Husnu, 2014; Ferguson, Carter, & Hassin, 2014; Kahneman, 2014; Klein et al., 2014; Monin & Oppenheimer, 2014; Schwarz & Strack, 2014; doi: 10.1027/1864-9335/a000202).

Acknowledgments

We thank Eugene Caruso, Melissa Ferguson, Daniel Oppenheimer, and Norbert Schwarz for their feedback on the design of the materials. This project was supported by grants to the second and fifty-first authors from Project Implicit. Ratliff and Nosek are consultants of Project Implicit, Inc., a nonprofit organization that includes in its mission "to develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender or other factors." Author contributions: Designed research: R. K., B. N., K. R.; Translated materials: S. B., K. B., M. Brandt, B. B., Z. C., N. F., E. G., F. H., H. I., R. K., R. P., A. V., M. Vianello, M. Vranka; Performed research: R. A., S. B., M. Bernstein, K. B., M. Brandt, C. B., Z. C., J. C., W. C., W. D., T. D., M. E., N. F., D. F., E. G., J. A. H., J. F. H., S. J. H., J. H., H. I., M. J., J. J., H. K., R. K., L. K., J. K., C. L., R. M., W. M., A. N., J. N., G. P., R. P., K. R., A. R., K. S., J. L. S., R. S., T. S., J. S., L. V., D. T., A. V., L. V., M. Vranka, A. L. W., J. W.; Analyzed data: M. Vianello, F. H., R. K.; Wrote paper: B. N., K. R., R. K., M. Vianello, J. C. We report all data exclusions, manipulations, measures, and how we determined our sample sizes either in text or the online supplement. All materials, data, videos of the procedure, and the original preregistered design are available at the project page https://osf.io/ydpbf/.

References

Aronson, E. (1992). The return of the repressed: Dissonance theory makes a comeback. *Psychological Inquiry, 3*, 303–311.

Bodenhausen, G. V. (1990). Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination. *Psychological Science, 1*, 319–322.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van 't Veer, A. (2013). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224.

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*, 1011–1018.

Caruso, E. M., Vohs, K. D., Baxter, B., & Waytz, A. (2013). Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General, 142*, 301–306.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*, 40–48.

Collins, H. M. (1974). The TEA set: Tacit knowledge and scientific networks. *Science Studies, 4*, 165–185.

Crisp, R. J., Miles, E., & Husnu, S. (2014). Support for the replicability of imagined contact effects. Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/1000202

Ferguson, M. J., Carter, T. J., & Hassin, R. R. (2014). Commentary on the attempt to replicate the effect of the American flag on increased republican attitudes. Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/1000202

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist, 59*, 93.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review, 93*, 216–229.

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

Henry, J. D., MacLeod, M. S., Phillips, L. H., & Crawford, J. R. (2004). A meta-analytic review of prospective memory and aging. *Psychology and Aging, 19*, 27.

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: 21st yearbook of the National Council of Social Studies* (pp. 11–34). New York, NY: National Council of Social Studies.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124.

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

Kahneman, D. (2014). A new etiquette for replication. Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/1000202

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*, 823–837.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Theory building through replication: Response to commentaries on the ''Many Labs'' replication project. Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/1000202

Knox, R. E., & Inkster, J. A. (1968). Postdecision dissonance at post time. *Journal of Personality and Social Psychology, 8*, 319.

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

Monin, B., & Oppenheimer, D. M. (2014). The limits of direct replications and the virtues of stimulus sampling. Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/1000202

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

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, Me = female, therefore math ≠ Me. *Journal of Personality and Social Psychology, 83*, 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*, 1125–1156.

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

Open Science Collaboration. (2014). The reproducibility project: A model of large-scale collaboration for empirical research on reproducibility. In V. Stodden, F. Leisch, & R. Peng (Eds.), *Implementing reproducible computational research (A volume in the R series)* (pp. 299–323). New York, NY: Taylor & Francis.

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., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411–419.

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

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100.

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.

Schwarz, N., & Strack, F. (2014). Does merely going through the same moves make for a ''direct'' replication? Concepts, contexts, and operationalizations. Commentaries and rejoinder on Klein et al. (2014). *Social Psychology*. Advance online publication. doi: 10.1027/1864-9335/1000202

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing Science, 4*, 199–214.

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*, 453–458.

Received March 4, 2013
Accepted November 26, 2013
Published online May 19, 2014

Richard A. Klein

Department of Psychology
University of Florida
Gainesville, FL 32611
USA
E-mail raklein@ufl.edu