Many Labs 2: Investigating Variation in Replicability Across Sample and Setting

Richard A. Klein	Université Grenoble Alpes	raklein22@gmail.com
Michelangelo Vianello	University of Padua	michelangelo.vianello@unipd.it
Fred Hasselman	Radboud University Nijmegen	f.hasselman@bsi.ru.nl
Byron G. Adams	Tilburg University & University of Johannesburg	b.g.adams@uvt.nl
Reginald B. Adams	The Pennsylvania State University	radams@psu.edu
Sinan Alper	Baskent University	sinanalper@baskent.edu.tr
Mark Aveyard	American University of Sharjah	maveyard@aus.edu
Jordan R. Axt	University of Virginia	jra3ee@virginia.edu
Štěpán Bahník	University of Economics, Prague	bahniks@seznam.cz
Rishtee Batra	Indian School of Business	Rishtee_Batra@isb.edu
Mihály Berkics	Eötvös Loránd University, Budapest	berkics.mihaly@ppk.elte.hu
Michael J. Bernstein	Penn State University Abington	mjb70@psu.edu
Daniel Berry	California State University, San Marcos	drberry@csusm.edu
Olga Bialobrzeska	SWPS University of Social Sciences and Humanities, Warsaw	obialobrzeska@swps.edu.pl
Evans Binan	University of Jos, Jos Nigeria	evansbinan@gmail.com
Konrad Bocian	SWPS University of Social Sciences and Humanities, Faculty in Sopot	kbocian1@swps.edu.pl
Mark J. Brandt	Tilburg University	M.j.brandt@tilburguniversity.edu
Robert Busching	University of Potsdam	busching@uni-potsdam.de
Anna Cabak Rédei	Lund University	anna.cabak_redei@semiotik.lu.se
Huajian Cai	Chinese Academy of Science	huajian.cai@gmail.com,
Fanny Cambier	Université catholique de Louvain	fanny.cambier@uclouvain.be
Katarzyna Cantarero	SWPS University of Social Sciences and Humanities, Faculty in Sopot	kcantarero@swps.edu.pl
Cheryl L. Carmichael	Brooklyn College & Graduate Center, CUNY	ccarmichael@brooklyn.cuny.edu
Francisco Ceric	Universidad del Desarrollo (Santiago, Chile)	fceric@udd.cl
David C. Cicero	University of Hawaii at Manoa	dcicero@hawaii.edu
Jesse Chandler	University of Michigan; PRIME Research	jjchandl@umich.edu
Armand Chatard	Poitiers University and CNRS	armand.chatard@univ-poitiers.fr
Eva E. Chen	The Hong Kong University of Science and Technology	evaechen@ust.hk
Jen-Ho Chang	Academia Sinica	jenhoc@gate.sinica.edu.tw
Winnee Cheong	HELP University, Malaysia	winnee.cheong@gmail.com
Sharon Coen	University of Salford	s.coen@salford.ac.uk
Jennifer A. Coleman	Virginia Commonwealth University	colemanj3@vcu.edu
Brian Collisson	Azusa Pacific University	bcollisson@marian.edu
Morgan A. Conway	University of Florida	morgan.conway@ufl.edu
Katherine S. Corker	Grand Valley State University	k.corker@gmail.com
Paul G. Curran	Grand Valley State University	curranpa@gvsu.edu
Fiery Cushman	Harvard University	cushman@wjh.harvard.edu
Zubairu K. Dagona	University of Jos, Jos Nigeria	dagonazk@gmail.com
Ilker Dalgar	Middle East Technical University	ilkerdalgar@gmail.com
Anna Dalla Rosa	University of Padua	anna.dallarosa@unipd.it
William E. Davis	Mount Holyoke College	dbillium@gmail.com
Maaike de Bruijn	Tilburg University	m.debruijn@tilburguniversity.edu
Leander De Schutter	WHU - Otto Beisheim School of Management	leander.deschutter@whu.edu
Thierry Devos	San Diego State University	tdevos@mail.sdsu.edu
Canay Doğulu	Baskent University	canaydogulu@gmail.com
Nerisa Dozo	The University of Queensland	nerisa.dozo@gmail.com
Kristin Nicole Dukes	Simmons College	kristin.dukes@simmons.edu
Yarrow Dunham	Yale University	yarrow.dunham@yale.edu
Kevin Durrheim	University of KwaZulu-Natal	durrheim@ukzn.ac.za
Charles R. Ebersole	University of Virginia	cebersole@virginia.edu
John E. Edlund	Rochester Institute of Technology	john.edlund@rit.edu
Alexander Scott English	Shanghai International Studies University	AlexEnglish@shisu.edu.cn
Anja Eller	National Autonomous University of Mexico	eller@unam.mx

Carolyn Finck	Universidad de los Andes, Colombia	cfinck@uniandes.edu.co
Natalia Frankowska	SWPS University of Social Sciences and Humanities, Warsaw	nfrankowska@swps.edu.pl
Miguel-Ángel Freyre	National Autonomous University of Mexico	migfreyre@gmail.com
Mike Friedman	Université catholique de Louvain	mike.d.friedman@gmail.com
Elisa Maria Galliani	University of Padua	elisamaria.galliani@unipd.it
Joshua C. Gandi	University of Jos, Jos Nigeria	jcgandi@rocketmail.com
Tanuka Ghoshal	Indian School of Business	Tanuka_Ghoshal@isb.edu
Steffen R. Giessner	Rotterdam School of Management, Erasmus University	sgiessner@rsm.nl
Tripat Gill	Wilfrid Laurier University	tgill@wlu.ca
Timo Gnambs	Leibniz Institute for Educational Trajectories	timo.gnambs@lifbi.de
Ángel Gómez	Universidad Nacional de Educación a Distancia	agomez@psi.uned.es
Roberto González	Pontificia Universidad Católica de Chile	rgonzale@uc.cl
Jesse Graham	Eccles School of Business, University of Utah	jesse.graham@eccles.utah.edu
Jon E. Grahe	Pacific Lutheran University	graheje@plu.edu
Ivan Grahek	Ghent University	ivan.grahek@ugent.be
Eva G. T. Green	University of Lausanne	eva.green@unil.ch
Kakul Hai	Manipal University Jaipur	kakulhai@gmail.com
Matthew Haigh	Northumbria University	matthew.haigh@northumbria.ac.uk
Elizabeth L. Haines	William Paterson University	hainese@wpunj.edu
Michael P. Hall	University of Michigan	mikeph@umich.edu
Marie E. Heffernan	University of Illinois at Urbana-Champaign	marieheffernan@gmail.com
Joshua A. Hicks	Texas A&M University	joshua.hicks@tamu.edu
Petr Houdek	Jan Evangelista Purkyne University	petr.houdek@gmail.com
Jeffrey R. Huntsinger	Loyola University Chicago	jhuntsinger@luc.edu
Ho Phi Huynh	Texas A&M University - San Antonio	hophih@gmail.com
Hans IJzerman	Université Grenoble Alpes	h.ijzerman@gmail.com
Yoel Inbar	University of Toronto Scarborough	yi38@cornell.edu
Åse H. Innes-Ker	Lund University	ase.innes-ker@psy.lu.se
William Jiménez-Leal	Universidad de los Andes, Colombia	w.jimenezleal@uniandes.edu.co
Melissa-Sue John	Worcester Polytechnic Institute	mjohn@wpi.edu
Jennifer A. Joy-Gaba	Virginia Commonwealth University	jennifer.joygaba@gmail.com
Anna Kende	Eötvös Loránd University, Budapest	kende.anna@ppk.elte.hu
Roza G. Kamiloğlu	University of Amsterdam	rozagizem@gmail.com
Heather Barry Kappes	London School of Economics and Political Science	h.kappes@lse.ac.uk
Serdar Karabati	Bilgi University, Istanbul	serdar.karabati@bilgi.edu.tr
Haruna Karick	SWPS Warsaw Poland/University of Jos, Jos Nigeria	harunakarick@gmail.com
Victor N. Keller	University of Brasilia	vnfskeller@gmail.com
Nicolas Kervyn	Université catholique de Louvain	nicolas.o.kervyn@uclouvain.be
Goran Knežević	Department of psychology, University of Belgrade	gknezevi@f.bg.ac.rs
Carrie Kovacs	Johannes Kepler University Linz	carrie.kovacs@jku.at
Lacy E. Krueger	Texas A&M University-Commerce	lacy.krueger@tamuc.edu
German Kurapov	Tilburg University	g.i.kurapov@tilburguniversity.edu
Jamie Kurtz	James Madison University	jamiekurtz@gmail.com
Daniël Lakens	Eindhoven University of Technology	d.lakens@tue.nl
Ljiljana B. Lazarević	Institute of psychology, University of Belgrade	ljiljana.lazarevic@f.bg.ac.rs
Carmel A. Levitan	Occidental College	levitan@oxy.edu
Neil A. Lewis, Jr.	Cornell University	nlewisjr@cornell.edu
Samuel Lins	University of Porto	samuellins@fpce.up.pt
Nikolette P. Lipsey	University of Florida	nlipsey5@gmail.com
Joy Losee	University of Florida	jl01745@ufl.edu
Esther Maassen	Tilburg University	e.maassen@tilburguniversity.edu
Angela T. Maitner	American University of Sharjah	amaitner@aus.edu
Winfrida Malingumu	Open University of Tanzania	wimnyamka@yahoo.co.uk
Robyn K. Mallett	Loyola University Chicago	rmallett@luc.edu
Satia A. Marotta	Tufts University	satia.marotta@tufts.edu
Janko Međedović	Institute of Criminological and Sociological Research, Belgrade	

	and Faculty of Media and Communications, Singidunum University	janko.medjedovic@fmk.edu.rs
Fernando Mena Pacheco	Universidad Latina de Costa Rica	fernando.mena@ulatina.cr
Taciano L. Milfont	Victoria University of Wellington	taciano.milfont@vuw.ac.nz
Wendy L. Morris	McDaniel College	wmorris@mcdaniel.edu
Sean Murphy	The University of Melbourne	seanchrismurphy@gmail.com
Andriy Myachykov	Northumbria University	andriy.myachykov@northumbria.ac.uk
Nick Neave	Northumbria University	nick.neave@northumbria.ac.uk
Koen Neijenhuijs	VU Amsterdam (previously Radboud University Nijmegen)	k.i.neijenhuijs@vu.nl
Anthony J. Nelson	The Pennsylvania State University	ajn157@psu.edu
Félix Neto	Universidade do Porto	fneto@fpce.up.pt
Austin Lee Nichols	University of Navarra	anichols@unav.es
Aaron Ocampo	Universidad Latina de Costa Rica	Ocampo.aaron@gmail.com
Susan L. O'Donnell	George Fox University	sodonnell@georgefox.edu
Elsie Ong	The Open University of Hong Kong	eong@ouhk.edu.hk
Malgorzata Osowiecka	SWPS University of Social Sciences and Humanities, Warsaw	malgorzataosowiecka@gmail.com
Gábor Orosz	Eötvös Loránd University, Budapest	orosz.gabor@ppk.elte.hu
Grant Packard	Wilfrid Laurier University	gpackard@wlu.ca
Rolando Pérez-Sánchez	University of Costa Rica	rolarez@gmail.com
Boban Petrović	Institute of Criminological and Sociological Research, Belgrade	bobanpetrovi@gmail.com
Ronaldo Pilati	University of Brasilia	rpilati@gmail.com
Brad Pinter	The Pennsylvania State University	tbp1@psu.edu
Lysandra Podesta	Radboud University Nijmegen	l.podesta@pwo.ru.nl
Gabrielle Pogge	University of Florida	gcm0402@ufl.edu
Monique M.H. Pollmann	Tilburg University	m.m.h.pollmann@tilburguniversity.edu
Abraham M. Rutchick	California State University Northridge	abraham.rutchick@csun.edu
Alexander Saeri	The University of Queensland	alexander@aksaeri.com
Patricio Saavedra	Pontificia Universidad Catolica de Chile	pj.saavedram@gmail.com
Erika Salomon	University of Illinois at Urbana-Champaign	salomon3@illinois.edu
Kathleen Schmidt	Southern Illinois University Carbondale	kathleen.schmidt@siu.edu
Felix D. Schönbrodt	Ludwig-Maximilians-Universität München	felix.schoenbrodt@psy.lmu.de
Maciej B. Sekerdej	Jagiellonian University	maciek@apple.phils.uj.edu.pl
David Sirlopú	Universidad del Desarrollo (Concepción, Chile)	dsirlopu@udd.cl
Jeanine L. M. Skorinko	Worcester Polytechnic Institute	skorinko@wpi.edu
Michael A. Smith	Northumbria University	michael4.smith@northumbria.ac.uk
Vanessa Smith-Castro	University of Costa Rica	vanessa.smith@ucr.ac.cr
Karin Smolders	Eindhoven University of Technology	K.C.H.J.Smolders@tue.nl
Agata Sobkow	SWPS University of Social Sciences and Humanities, Wroclaw Faculty of Psychology	asobkow@swps.edu.pl
Walter Sowden	Center for Military Psychiatry & Neuroscience, Walter Reed Army Institute of Research,	wjsowden@gmail.com
Manini Srivastava	University of Lucknow	maninigarima@gmail.com
Oskar K. Sundfelt	Lund University	oskar.sundfelt@gmail.com
Philipp Spachtholz	University of Regensburg	philipp.spachtholz@ur.de
Troy G. Steiner	The Pennsylvania State University	tgs5057@psu.edu
Jeroen Stouten	KULeuven	jeroen.stouten@kuleuven.be
Chris N. H. Street	University of Huddersfield	c.street@hud.ac.uk
Stephanie Szeto	The Open University of Hong Kong	s.s.szeto@edu.salford.ac.uk
Ewa Szumowska	Jagiellonian University in Krakow	ewa.szumowska@uj.edu.pl
Andrew Tang	The Open University of Hong Kong	acwtang@ouhk.edu.hk
Norbert Tanzer	University of Graz	norbert.tanzer@uni-graz.at
Morgan Tear	The University of Queensland	morgantear@gmail.com
Manuela Thomae	University of Winchester	manuela.thomae@winchester.ac.uk
Jakub Traczyk	SWPS University of Social Sciences and Humanities, Wroclaw Faculty of Psychology	jtraczyk@swps.edu.pl
David Torres	Universidad de Iberoamerica	datofez@gmail.com

Jordan Theriault	Boston College	jordan.theriault@bc.edu
Joshua M. Tybur	VU Amsterdam	j.m.tybur@vu.nl
Adrienn Ujhelyi	Eötvös Loránd University, Budapest	ujhelyi.adrienn@ppk.elte.hu
Robbie C.M. van Aert	Tilburg University, Netherlands	r.c.m.vanaert@uvt.nl
Marcel A.L.M. van Assen	Tilburg University, Netherlands	m.a.l.m.vanassen@uvt.nl
Paul A. M. van Lange	VU Amsterdam	p.a.m.van.lange@vu.nl
Marije van der Hulst	Erasmus MC Rotterdam (previously Radboud University Nijmegen)	m.vanderhulst@erasmusmc.nl
Anna Elisabeth van 't Veer	Leiden University, Netherlands	a.e.van.t.veer@fsw.leidenuniv.nl
Alejandro Vásquez Echeverría	Universidad de la República, Uruguay	avasquez@psico.edu.uy
Leigh Ann Vaughn	Ithaca College	lvaughn@ithaca.edu
Alexandra Vázquez	Universidad Nacional de Educación a Distancia	alx.vazquez@psi.uned.es
Luis Diego Vega	Universidad Latina de Costa Rica	luis.vegaa@ulatina.cr
Catherine Verniers	Paris Descartes University - Sorbonne Paris Cité	catherine.verniers@parisdescartes.fr
Mark Verschoor	Tilburg University	m.verschoor@rug.nl
Ingrid Voermans	Radboud University	ingrid_voermans@hotmail.com
Marek A. Vranka	Charles University	marek.vranka@ff.cuni.cz
Marieke de Vries	Radboud University	Marieke.deVries@ru.nl
Cheryl Welch	James Madison University	welch2ca@dukes.jmu.edu
Aaron L. Wichman	Western Kentucky University	aaron.wichman@wku.edu
Lisa A. Williams	University of New South Wales	lwilliams@unsw.edu.au
Michael Wood	University of Winchester	michael.wood@winchester.ac.uk
Julie A. Woodzicka	Washington and Lee University	woodzickaj@wlu.edu
Marta K. Wronska	SWPS University of Social Sciences and Humanities, Faculty in Sopot	wronska.marta@gmail.com
Liane Young	Boston College	liane.young@bc.edu
John M. Zelenski	Carleton University	john_zelenski@carleton.ca
Zeng Zhijia	Guangdong Literature & Art Vocational College	Hpzhijia@163.com
Brian A. Nosek	University of Virginia; Center for Open Science	nosek@virginia.edu

Authors' note: This research was supported by the Center for Open Science and from a grant through the Association for Psychological Science from the Laura and John Arnold Foundation. Correspondence concerning this paper should be addressed to Richard A. Klein, Université Grenoble Alpes, raklein22@gmail.com.

Abstract

We conducted preregistered replications of 28 classic and contemporary published findings with protocols that were peer reviewed in advance to examine variation in effect magnitudes across sample and setting. Each protocol was administered to approximately half of 125 samples and 15,305 total participants from 36 countries and territories. Using conventional statistical significance ($p < .05$), fifteen (54%) of the replications provided evidence in the same direction and statistically significant as the original finding. With a strict significance criterion ($p < .0001$), fourteen (50%) provide such evidence reflecting the extremely high powered design. Seven (25%) of the replications had effect sizes larger than the original finding and 21 (75%) had effect sizes smaller than the original finding. The median comparable Cohen's *d* effect sizes for original findings was 0.60 and for replications was 0.15. Sixteen replications (57%) had small effect sizes (< .20) and 9 (32%) were in the opposite direction from the original finding. Across settings, 11 (39%) showed significant heterogeneity using the Q statistic and most of those were among the findings eliciting the largest overall effect sizes; only one effect that was near zero in the aggregate showed significant heterogeneity. Only one effect showed a Tau > 0.20 indicating moderate heterogeneity. Nine others had a Tau near or slightly above 0.10 indicating slight heterogeneity. In moderation tests, very little heterogeneity was attributable to task order, administration in lab versus online, and exploratory WEIRD versus less WEIRD culture comparisons. Cumulatively, variability in observed effect sizes was more attributable to the effect being studied than the sample or setting in which it was studied.

Word count = 265

Keywords = social psychology; cognitive psychology; replication; culture; individual differences; sampling effects; situational effects; meta-analysis

Many Labs 2: Investigating Variation in Replicability Across Sample and Setting

Suppose a researcher, Josh, conducts an experiment finding that experiencing threat reduces academic performance compared to a control condition. Another researcher, Nina, conducts the same study at her institution and finds no effect. Person and situation explanations may come to mind immediately: (1) Nina used a sample that might differ in important ways from Josh's sample, and (2) the situational context in Nina's lab might differ in theoretically important but non-obvious ways from Josh's lab. Both could be true simultaneously. A less interesting, but real, possibility is that one of them made an error in design or procedure that the other did not. Finally, it is possible that the different effects are a function of sampling error: Nina's result could be a false negative, or Josh's result could be a false positive. The present research contributes evidence toward understanding the contribution of variation in sample and setting for observing psychological effects.

Variation in effects: Person, situation, or sampling error?

There is a history of research evidence for effects of variation by particular person characteristics, in particular situations, and for particular experimental effects (Lewin, 1936; Ross & Nisbett, 1991). For example, people tend to attribute behavior to characteristics of the person rather than characteristics of the situation (e.g., Gilbert & Malone, 1995; Jones & Harris, 1967), but some evidence suggests that this effect is stronger in western than eastern cultures (Miyamoto & Kitayama, 2002). A common model of investigating psychological processes is to identify an effect, and then investigate moderating influences that make the effect stronger or weaker. As such, when one confronts different outcomes from similar experiments, the readily available conclusion is that a moderating influence accounts for the difference. However, if effects vary less across sample and setting than assumed in the psychological literature, then the

assumptions of moderation may be overapplied and the role of sampling error underestimated.

If effects are highly variable across sample and setting, then variation in effect sizes will routinely exceed what would be expected from sampling error. In this circumstance, the lack of consistency between Josh and Nina's results is unlikely to influence beliefs about the original effect. Moreover, if there are many influential factors, then it is difficult to isolate moderators and identify the necessary conditions to obtain the effect. In this case, the lack of consistency between Josh and Nina's results might produce collective indifference -- there are just too many variables to know why there was a difference, so their different results produce no change in perceived understanding of the phenomenon.

Alternatively, variations in effect sizes may not exceed expectations due to sampling error. In this case, the observed differences in effects are not indicating moderating influences of sample or setting. This would indicate imprecision in effect estimation is the sole source of variation and require no causal explanation. For Josh and Nina, the possibility that the variation is sampling error rather than evidence for moderation is not necessarily easy to assess, especially if their studies had small samples (Morey & Lakens, 2016). With small samples, Josh's positive result and Nina's null result will likely have confidence intervals that overlap each other leaving little to conclude other than "more data are needed".

The difference between these interpretations is substantial, but there is little *direct* evidence regarding the extent to which persons and situations--or samples and settings--influence the size of psychological effects *in general* (but see Coppock, in press; Krupnikov & Levine, 2014; Mullinix, Leeper, Druckman, & Freese, 2015). The default assumption is that psychological effects are awash in interactions among many variables. The present report follows initial evidence from the "Many Labs" projects (Ebersole et al., 2016; Klein et al., 2014).

The first Many Labs project replicated 13 classic and contemporary psychological effects with 36 different samples/settings (N = 6,344). The results of that study showed that: (a) variation in sample and setting had little impact on observed effect magnitudes, (b) when there was variation in effect magnitude across samples, it occurred in studies with large effects, not in studies with small effects, (c) overall, effect size estimates were more related to the effect of study rather than the sample or setting in which it was studied, and (d) this held even for lab-based versus web-based data collections, and across nations.

A limitation of the first "Many Labs" is that there was a small number of effects and there was no reason to presume them to vary substantially across sample and setting. It is possible that those effects are more robust and homogenous than the typical behavioral phenomena, or that the populations were more homogenous than initially expected. The present research represents a major expansion of the "Many Labs" study design with (1) more effects, (2) inclusion of some effects that are presumed to vary across sample or setting, (3) more labs, and (4) diverse samples. The selected effects are not random nor are they representative, but they do cover a wide range of topics. This study provides preliminary evidence for the extent to which variation in effect magnitudes is attributable to sample and setting, versus sampling error.

Other Influences on Observed Effects

Across systematic replication efforts in the social-behavioral sciences, there is accumulating evidence that fewer published effects replicate than might be expected, and that replication effect sizes are typically smaller than original effect sizes (Camerer et al., 2016, 2018; Ebersole et al., 2016; Klein et al., 2014; Open Science Collaboration, 2015). For example, Camerer et al. (2018) successfully replicated 13 of 21 social science studies published in Science and Nature. Among failures to replicate, the average effect size was approximately 0, but even

among successful replications, the average replication effect size was about 75% of what was observed in the original experiments. Failures to replicate could be due to errors in the replication, or because of unanticipated moderation by changes to sample and setting as is investigated here. They can also occur because of pervasive low-powered research plus publication bias that is more likely to select positive than negative results for publication (Button et al., 2013; Cohen, 1962; Greenwald, 1975; Rosenthal, 1979), and because of questionable research practices, or *p*-hacking, that can inflate the likelihood of obtaining false positives (John et al., 2012; Simmons et al., 2011). These are not investigated directly in this research, but they could contribute to observing failures to replicate and to weaker effect sizes than observed in the original research.

Origins of Study Design

To obtain a candidate list of effects, we held a round of open nomination and invited submissions for any effect that fit the defined criteria (see the Coordinating Proposal available on the OSF: https://osf.io/uazdm/). Those nominations were supplemented by ideas from the project team, and from direct queries for suggestions to independent experts in psychological science.

The nominated studies were evaluated individually on the following criteria: (1) feasibility of implementation through a web browser, (2) brevity of study procedures (shorter procedures desired), (3) citation impact of the effect (higher impact desired), (4) identifiability of a meaningful two-condition experimental design or simple correlation as the target of replication (with an emphasis on experiments), (5) general interest value of the effect, and (6) applicability to samples of adults. The nominated studies were evaluated collectively to assure diversity on the following criteria: (1) effects known to be observable in multiple samples and settings and others

for which reliability of the effect is unknown[1], (2) effects known to be sensitive to sample or setting and others for which variation is unknown or assumed to be minimal, (3) classic and contemporary effects, (4) breadth of topical areas in social and cognitive psychology, (5) the research groups who conducted the study, and (6) publication outlet.

More than 100 effects were nominated as potentially fitting these criteria. A subset of the project team reviewed these effects to maximize the number of included effects and diversity of the total slate on these criteria. No specific researcher was selected for replication because of beliefs or concerns about their research or the effects they have reported, but some areas and authors were included more than once because of producing short, simple, interesting effects that met the selection criteria.

Once selected for inclusion, a member of the research team contacted the corresponding author (if alive) to obtain original study materials and get advice about adapting the procedure for this use. In particular, original authors were asked if there were moderators or other limitations to obtaining the result that would be useful for the team to understand in advance and, perhaps, anticipate in data collection.

In some cases, correspondence with original authors identified limitations of the selected effect that reduced its applicability for the present design. In those cases, we worked with the original authors to identify alternative studies or decided to remove the effect entirely from the selected set, and replaced it with one of the available alternatives.

We split the studies into two slates that would require about 30 minutes each. We included 32 effects in total before peer review and pilot testing. In only one instance did original authors express strong concerns about inclusion in the study. Because we make no claim about

[1] Because the project goal was to examine variability in effect magnitudes across samples and settings, we were not interested in including studies that were known or suspected to be unreplicable.

the sample of studies being randomly selected or representative, we removed the effect from the project. With the remaining 31 effects, we pilot tested both slates with participation across the authors and members of their labs to ensure that each slate could be completed within 30 minutes. We observed that we underestimated the time required for a few effects. As a consequence, we had to remove three effects (Ashton-James, Maddux, Galinsky, & Chartrand, 2009; Srull & Wyer, 1979; Todd, Hanko, Galinsky, & Mussweiler, 2011), shorten or remove a few individual difference measures, and slightly reorganize the slates to achieve the final 28 included effects. We divided the studies across slate to be balanced on the criteria above and to avoid substantial overlap in topics.

Following the Registered Report model (Nosek & Lakens, 2014), prior to data collection the materials and protocols were formally peer reviewed in a process conducted by the journal editor.

Disclosures

Preregistration. The accepted design was preregistered at https://osf.io/ejcfw/.

Data, materials, and online resources. Comprehensive materials, data, and supplementary information about the project are available at https://osf.io/8cd4r/. Any deviations from the preregistered design in study description or implementation are recorded in supplementary materials (https://osf.io/7mqba/). Any changes to analysis plans are noted with justification and comparisons between original and revised analytic approaches, also available in supplementary materials (https://osf.io/4rbh9/), see Table 1 for a summary. A guide to the data analysis code is available at: https://manylabsopenscience.github.io/.

Measures. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Subjects. The research was conducted in accordance with the Declaration of Helsinki and following local requirements for Institutional Review Board approvals for each of the data collection sites.

Conflicts of Interest. Brian Nosek is Executive Director of the non-profit Center for Open Science which has a mission to increase openness, integrity, and reproducibility of research.

Author Contributions. Coordinated project: Fred Hasselman, Richard Klein, Brian Nosek, Michelangelo Vianello.

Designed the study: Štěpán Bahník, Jesse Chandler, Katherine Corker, Fred Hasselman, Hans IJzerman, Richard Klein, Brian Nosek, Kathleen Schmidt, Marcel van Assen, Leigh Ann Vaughn, Michelangelo Vianello, Aaron Wichman.

Developed materials: Jordan Axt, Štěpán Bahník, John Conway, Paul Curran, Richard Klein, Kathleen Schmidt

Wrote proposal: Jordan Axt, Štěpán Bahník, Mihály Berkics, Jesse Chandler, Eva E. Chen, Sharon Coen, John Conway, Katherine Corker, William E. Davis, Timo Gnambs, Fred Hasselman, Hans IJzerman, Richard Klein, Carmel Levitan, Wendy Morris, Brian Nosek, Kathleen Schmidt, Vanessa Smith-Castro, Jeroen Stouten, Marcel van Assen, Leigh Ann Vaughn, Michelangelo Vianello, Aaron Wichman

Collected data: Byron G. Adams, Reginald B. Adams, Sinan Alper, Mark Aveyard, Štepán Bahník, Rishtee Batra, Mihály Berkics, Michael J. Bernstein, Daniel Berry, Olga Bialobrzeska, Evans Binan, Konrad Bocian, Mark J. Brandt, Robert Busching, Anna Cabak Rédei, Huajian Cai, Fanny Cambier, Katarzyna Cantarero, Cheryl L. Carmichael, Francisco Ceric, David C. Cicero, Jesse Chandler, Armand Chatard, Eva E. Chen, Jen-Ho Chang, Winnee

Cheong, Sharon Coen, Jennifer A. Coleman, Brian Collisson, Katherine S. Corker, Paul G. Curran, Fiery Cushman, Zubairu K. Dagona, Ilker Dalgar, Anna Dalla Rosa, William E. Davis, Maaike de Bruijn, Leander De Schutter, Thierry Devos, Canay Dogulu, Nerisa Dozo, Kristin Nicole Dukes, Yarrow Dunham, Kevin Durrheim, Charles R. Ebersole, John E. Edlund, Alexander Scott English, Anja Eller, Carolyn Finck, Natalia Frankowska, Miguel-Ángel Freyre, Mike Friedman, Jennifer A. Joy-Gaba, Elisa Maria Galliani, Joshua C. Gandi, Tanuka Ghoshal, Steffen R. Giessner, Tripat Gill, Timo Gnambs, Ángel Gómez , Roberto González, Jesse Graham, Jon E. Grahe, Ivan Grahek, Kakul Hai, Eva G. T. Green, Matthew Haigh, Elizabeth L. Haines , Michael P. Hall, Fred Hasselman, Marie E. Heffernan, Joshua A. Hicks, Petr Houdek, Jeffrey R. Huntsinger, Ho Phi Huynh, Melissa-Sue John, Hans IJzerman, Yoel Inbar, Anna Kende, Åse H. Innes-Ker, William Jiménez-Leal, Roza G. Kamiloglu, Heather Barry Kappes, Serdar Karabati, Haruna Karick, Victor N. Keller, Nicolas Kervyn, Richard A. Klein, Goran Kneževic, Carrie Kovacs, Lacy E. Krueger, German Kurapov, Jamie Kurtz, Daniël Lakens, Ljiljana B. Lazarevic, Carmel A. Levitan, Samuel Lins, Melissa-Sue John, Esther Maassen, Angela T. Maitner, Winfrida Malingumu, Robyn K. Mallett, Satia A. Marotta, Janko Mededovic, Fernando Mena Pacheco, Taciano L. Milfont, Wendy L. Morris, Sean Murphy, Andriy Myachykov, Nick Neave, Koen Neijenhuijs, Anthony J. Nelson, Félix Neto , Austin Lee Nichols, Aaron Ocampo, Susan L. O'Donnell, Elsie Ong, Malgorzata Osowiecka, Gábor Orosz, Grant Packard, Rolando Pérez-Sánchez, Boban Petrovic, Ronaldo Pilati, Brad Pinter, Lysandra Podesta, Monique M.H. Pollmann, Abraham M. Rutchick, Alexander Saeri, Patricio Saavedra, Erika Salomon, Felix D. Schönbrodt, Maciej B. Sekerdej, David Sirlopú, Jeanine L. M. Skorinko, Michael A. Smith, Vanessa Smith-Castro, Karin Smolders, Agata Sobkow, Walter Sowden, Manini Srivastava, Oskar K. Sundfelt, Philipp Spachtholz, Troy G. Steiner, Jeroen

Stouten, Chris N. H. Street, Stephanie Szeto, Ewa Szumowska, Andrew Tang, Norbert Tanzer, Morgan Tear, Manuela Thomae, Jakub Traczyk, David Torres, Jordan Theriault, Joshua M. Tybur, Adrienn Ujhelyi, Robbie C.M. van Aert, Marcel A.L.M. van Assen, Paul A. M. van Lange, Marije van der Hulst, Anna Elisabeth van 't Veer, Alejandro Vásquez Echeverría, Leigh Ann Vaughn, Alexandra Vázquez, Luis Diego Vega, Catherine Verniers, Mark Verschoor, Ingrid Voermans, Marek A. Vranka, Marieke de Vries, Cheryl Welch, Aaron L. Wichman, Lisa A. Williams, Michael Wood, Julie A. Woodzicka, Marta K. Wronska, Liane Young, John M. Zelenski, Zeng Zhijia

Analyzed data: Fred Hasselman, Michelangelo Vianello, Richard Klein, with support from Katie Corker, Brian Nosek, Robbie C.M. van Aert, Marcel A.L.M. van Assen

Designed figures: Fred Hasselman, Brian Nosek

Wrote report: Richard Klein, Brian Nosek, Michelangelo Vianello

Commented, edited, and approved report: All authors

Acknowledgements. We thank Cameron Anderson, Adam Baimel, Galen Bodenhausen, Emma Buchtel, Zeynep Cemalcilar, Clayton Critcher, Fiery Cushman, Itamar Gati, Steffen Giessner, Jesse Graham, Kurt Gray, Christopher Hsee, Yanli Huang, Yoel Inbar, Daniel Kahneman, Aaron Kay, Shinobu Kitayama, Joshua Knobe, Michael Kubovy, Yuri Miyamoto, Ara Norenzayan, Jane Risen, Lee Ross, Yuval Rottenstreich, Krishna Savani, Norbert Schwarz, Eldar Shafir, Chi-Shing Tse, Paul van Lange, Liane Young, Lisa Zaval, and Chenbo Zhong for helping develop and review materials, and for providing additional details from original studies when needed.

Prior versions: None.

Methods

Participants

An open invitation to participate as a data collection site in Many Labs 2 was issued in early 2014. To be eligible for inclusion, participating labs agreed to administer their assigned study procedure to at least 80 participants and to collect as many as was feasible. Lab decisions to stop data collection were based on their access to participants and time constraints. None had opportunity to observe the outcomes prior to conclusion of data collection. All contributors who met the design and data collection requirements received authorship on the final report. Upon completion of data collection there were 125 total samples (64 for Slate 1 and 61 for Slate 2, which includes 15 sites that collected data for both slates) with a cumulative sample size of 15,305 (mean = 122.44, median = 99, SD = 92.71, Range = 16 to 841).

Samples included 79 in-person (typically lab-based) and 46 web-based data collections. 39 samples were from the United States, and the 86 others were from Australia (2), Austria (2), Belgium (2), Brazil (1), Canada (4), Chile (3), China (5), Colombia (1), Costa Rica (2), Czech Republic (3), France (2), Germany (4), Hong Kong, China (3), Hungary (1), India (5), Italy (1), Japan (1), Malaysia (1), Mexico (1), The Netherlands (9), New Zealand (2), Nigeria (1), Poland (6), Portugal (1), Serbia (3), South Africa (3), Spain (2), Sweden (1), Switzerland (1), Taiwan (1), Tanzania (2), Turkey (3), The United Arab Emirates (2), The United Kingdom (4), and Uruguay (1). Details about each site of data collection are available here: https://osf.io/uv4qx/.

Of those that responded to demographics questions, in Slate 1 34.5% were men, 64.4% were women, 0.3% selected "Other", and 0.8% selected "Prefer not to answer". The average age for Slate 1 was 22.37 (SD = 7.09)[2]. For Slate 2, 35.9% were men, 62.9% were women, 0.4% selected "Other", and 0.8% selected "Prefer not to answer". The average age for Slate 2 was

[2] Excluding age responses > 100

23.34 (SD = 8.28)[3]. Variation in demographic characteristics across samples is documented at https://osf.io/g3bza/.

Procedure

The study was administered over the Internet for standardization across locations. At some locations, participants completed the survey in a lab or room on computers or tablets, whereas in other locations the participants completed the survey entirely online at their own convenience. Surveys were created in Qualtrics software (qualtrics.com) and unique links to run the studies were sent to each data collection team to track the origin of data. Each site was assigned an identifier. These identifiers can be found under the "source" variable in the public dataset.

Data were deposited to a central database and analyzed together. Each team created a video simulation of study administration to illustrate the features of the data collection setting. For languages other than English, labs completed a translation and back translation of the study materials to check against original meaning (cf. Brislin, 1970). Labs decided themselves the appropriate language for their sample and adapted materials for content appropriateness for the national sample (e.g., editing monetary units).

Assignment of labs to slates maximized national diversity for both slates. If there was one lab for a country, it was randomly assigned to a slate using random.org. If there was more than one lab for a country, then labs were randomly assigned to slate using random.org with the exception that they were evenly distributed across slates as closely as possible (e.g., 2 in each slate if there were 4 countries). Nearing data collection, we recruited some additional Asian sites specifically for Slate 1 to increase sample diversity. The slates were administered by a single experiment script that began with informed consent, then presented the effects in that slate in a

[3] Excluding age responses > 100

fully randomized order at the level of participants, followed by the individual difference measures in randomized order, and then closing with demographics measures and debriefing.

Demographics

The demographics below were included to characterize each sample and for possible moderator investigations. Participants were free to decline to answer any question.

Age. Participants noted their age in years in an open-response box.

Sex. Participants selected "male", "female", "other", or "prefer not to answer" to indicate their biological sex.

Race/ethnicity. Participants indicated race/ethnicity by selecting from a drop-down menu populated with options determined by the replication lead for each site. Participants could also select "other" and write an open-response. Note that response items were not standardized as some countries have very different conceptualizations of race/ethnicity.

Cultural origins. Three items assessing cultural origins used a drop-down menu populated by a list of countries or territories, and an "other" option with an open-response box. The three items were: (1) In which country/region were you born?, (2) In which country/region was your primary caregiver (e.g., parent, grandparent) born?, and (3) If you had a second primary caregiver, in which country/region was he or she born?

Hometown. A single item "What is the name of your home town/city?" with an open response blank was included as a potential moderator for the Huang et al. (2014) effect.

Wealth in hometown. A single item "Where do wealthier people live in your home town/city?" with North, South, and Neither as response options was included as a potential moderator of the Huang et al. (2014) effect. This item appeared in Slate 1 only.

Political ideology. Participants rated their political ideology on a scale with response

options of: strongly left-wing, moderately left-wing, slightly left-wing, moderate, slightly right-wing, moderately right-wing, strongly right-wing. Instructions were adapted for each country of administration to ensure relevance of the ideology dimension to the local context. For example, the U.S. instructions read: "Please rate your political ideology on the following scale. In the United States, 'liberal' is usually used to refer to left-wing and 'conservative' is usually used to refer to right-wing."

Education. Participants reported their educational attainment on a single item "What is the highest educational level that you have attained?" using a 6-point response scale: 1 = no formal education, 2 = completed primary/elementary school, 3 = completed secondary school/high school, 4 = some university/college, 5 = completed university/college degree, 6 = completed advanced degree.

Socio-economic status. Socio-economic status was measured with the ladder technique (Adler et al., 1994). Participants indicated their standing in their community relative to other people in the community with which they most identify on a ladder with ten steps where 1 indicates people at the bottom having the lowest standing in the community and 10 referring to people at the top having the highest standing. Previous research demonstrated good convergent validities of this item with objective criteria of individual social status and also construct validity with regard to several psychological and physiological health indicators (e.g., Adler, Epel, Castellazzo, & Ickovics, 2000; Cohen, Alper, Doyle, Adler, Treanor, & Taylor, 2008). This ladder was also used in Effect 12 in Slate 1 (Anderson, Kraus, Galinsky, & Keltner, 2012, Study 3). Participants in that slate answered the ladder item as part of the Effect 12 materials and did not receive the item a second time.

Data quality. Recent research in the area of careless or insufficient effort responding has

moved toward refining implementation of established scales embedded in data collection to check for aberrant response patterns (Huang et al., 2014; Meade & Craig, 2012). We included two items at the end of the study, just prior to demographic items. The first item asked participants "In your honest opinion, should we use your data in our analyses in this study?" with *yes* and *no* as response options (Meade & Craig, 2012). The second item was an Instructional Manipulation Check (IMC; Oppenheimer, Meyvis, & Davidenko, 2009), in which an ostensibly simple demographic question ("Where are you completing this study?") is preceded by a long block of text that contains, in part, alternative instructions for the participant to complete to demonstrate they are paying attention ("Instead, simply check all four boxes and then press "continue" to proceed to the next screen").

Individual Difference Measures

The following individual difference measures were included to allow future tests of effect size moderation.

Cognitive reflection (Finucane & Gullion, 2010). The cognitive reflection task (CRT; Frederick, 2005) assesses individuals' ability to suppress an intuitive (wrong) response in favor of a deliberative (correct) answer. The items on the original CRT are widely known, and the measure is vulnerable to practice effects (Chandler, Mueller, & Paolacci, 2014). As such, we used an updated version that is logically equivalent and correlates highly with the items on the original CRT (Finucane & Gullion, 2010). The three items are: (1) "If it takes 2 nurses 2 minutes to measure the blood pressure of 2 patients, how long would it take 200 nurses to measure the blood pressure of 200 patients?"; (2) "Soup and salad cost $5.50 in total. The soup costs a dollar more than the salad. How much does the salad cost?"; and, (3) "Sally is making tea. Every hour, the concentration of the tea doubles. If it takes 6 hours for the tea to be ready, how long would it

take for the tea to reach half of the final concentration?" Also, we constrained the total time available to answer the three questions to 75 seconds. This likely lowered overall performance on average as it was somewhat less time than some participants took in pretesting.

Subjective well-being (Veenhoven, 2009). Subjective well-being was measured with a single item "All things considered, how satisfied are you with your life as a whole these days?" on a response scale from 1 "dissatisfied" to 10 "satisfied". Similar items are included into numerous large-scale social surveys (cf. Veenhoven, 2009) and have shown satisfactory reliabilities (e.g., Lucas & Donnellan, 2012) and validities (Cheung & Lucas, 2014; Oswald & Wu, 2010; Sandvik, Diener, & Seidlitz, 1993).

Global self-esteem (Robins, Hendin, & Trzesniewski, 2001). Global self-esteem was measured using a Single-Item Self-Esteem Scale (SISE) designed as an alternative to using the Rosenberg Self-Esteem Scale (1965). The SISE consists of a single item: "I have high self-esteem". Participants respond on a 5-point Likert scale, ranging from 1 = *not very true of me* to 5 = *very true of me*. Robins, Hendings, and Trzesniewski (2001) reported strong convergent validity with the Rosenberg Self-Esteem Scale (with *r*s ranging from 0.70 to 0.80) among adults. Also, the scale had similar predictive validity as the Rosenberg Self-Esteem Scale.

TIPI for Big-Five personality (Gosling, Rentfrow, & Swann, 2003). The five basic traits of human personality (Goldberg, 1981) -- conscientiousness, agreeableness, neuroticism / emotional stability, openness / intellect, and extraversion -- were measured with the Ten Item Personality Inventory (Gosling et al., 2003). Each trait was assessed with two items on seven point response scales from 1 = *disagree strongly* to 7 = *agree strongly*. The five scales show satisfactory retest reliabilities (cf. Gnambs, 2014) and substantial convergent validities with longer Big Five instruments (e.g., Ehrhart et al., 2009; Gosling et al., 2003; Rojas & Widiger,

2014).

Mood (Cohen, Sherman, Bastardi, Hsu, McGoey, & Ross, 2007). There exist many assessments of mood. We selected the single-item from Cohen and colleagues (2007). Respondents answer "How would you describe your mood right now?" on a 5-point response scale: 1 = extremely bad, 2 = bad, 3 = neutral, 4 = good, 5 = extremely good.

Disgust Sensitivity Scale--Contamination Subscale (DS-R; Olatunji et al., 2007). The DS-R is a 25-item revision of the original Disgust Sensitivity Scale (Haidt, McCauley, & Rozin, 1994). Subscales of the DS-R were determined by factor analysis. The contamination subscale includes the 5 items related to concerns about bodily contamination. For length considerations, only the contamination subscale was included for Effect 8 in Slate 1. No part of this scale appeared in Slate 2.

The 28 Effects

Before describing the main results examining heterogeneity across samples and settings, we describe each of the 28 selected effects. We provide a summary of the main idea of the original research and the sample size, inferential test, and effect size that is the key result for replication. Then, we summarize the confirmatory aggregate result of the replication. The aggregate result is tested by pooling the data of all available samples, ignoring sample origin. An aggregate result was labelled consistent with the original finding if they were in the same direction and statistically significant as the original study conducted in a western, educated, industrialized, rich, democratic society (Henrich et al., 2010). In four cases, the original study focused on cultural differences in the key effect. Our main replication result is the aggregate effect size regardless of cultural context. Whether effects vary by setting (or cultural context more generally) is examined in the heterogeneity analyses in the results section. If there was

opportunity to test the original cultural difference with similar samples, they are reported as additional results in reports of the individual effects. For some of the effects, moderating influences were anticipated in advance by the original authors that could affect comparison of the original and replication effect sizes. If any were planned, we report the *a priori* identified additional, moderator, or subset analyses.

For readers interested in the global results of this replication project this long section detailing each individual replication can be skipped. Systematic tests of variation by sample using meta-analysis follow the section of describing results of individual findings. Heterogeneity was assessed using the *Q*, *Tau*, and I^2 measures (Borenstein et al., 2009).

SLATE 1

1. Direction and Socioeconomic status: LIVING IN THE NORTH IS NOT NECESSARILY FAVORABLE: DIFFERENT METAPHORIC ASSOCIATIONS BETWEEN CARDINAL DIRECTION AND VALENCE IN HONG KONG AND IN THE UNITED STATES (Huang, Tse & Cho, 2014, Study 1a)

People in the United States and Hong Kong have different demographic knowledge that may shape their metaphoric association between valence and cardinal direction (North/South). 180 participants from the United States and Hong Kong participated. Participants were presented with a blank map of a fictional city and were randomly assigned to indicate on the map where either a high-SES or low-SES person might live. There was an interaction between SES (high vs. low) and population (US vs. HK), $F(1,176) = 20.39$, $MS_E = 5.63$, $p < .001$, $\eta_p^2 = 0.10$, $d = .68$, 95% CI [.38, .98] . US participants expected the high-SES person to live further north (M = +0.98, SD = 1.85) than the low-SES person (M = -0.69, SD = 2.19), $t(78) = 3.69$, $p < .001$, d = .83, 95% CI [.37, 1.30]. Conversely, HK participants expected the low-SES person to live

further north (M = +0.63, SD = 2.75) than the high-SES person (M = -0.92, SD = 2.47), t(98) = -2.95, p = .004, d = -.59, 95% CI [-.99, -.19]. The authors explained that wealth in Hong Kong is concentrated in the south of the city, and wealth in cities in the United States is more commonly concentrated in the north of the city. As a consequence, cultures differ in their assumptions of wealth concentration in fictional cities.

Replication. The coordinates of participants' click on the fictional map were recorded (X, Y) from the top-left of the image, and then recentered in the analysis such that clicks in the north half of the map were positive and clicks in the southern half of the map were negative. Across all samples (N = 6,591), participants in the high-SES condition (M = 11.70, SD = 84.31) selected a further north location than participants in the low-SES condition (M = -22.70, SD = 88.78; t(6,554.05) = 16.12, p = 2.15e-57, d = 0.40, 95% CI [0.35, 0.45].

The original authors suggested we may only replicate the pattern for "Western" participants for whom up and North are aligned with the predicted "good" and high-SES evaluation. As suggested by the original authors, the focal test for replicating the effect for "Western" participants was completed by selecting only participants across all samples who indicated wealth tended to be in the north in their hometown. These participants expected the high-SES person to live further north (M = 43.22, SD = 84.43) than the low-SES person (M = -40.63, SD = 84.99; t(1,692) = 20.36, p = 1.24e-82, d = 0.99; 95% CI [0.89, 1.09]). This result is consistent with the hypothesis that people reporting that wealthier people tend to live in the North in their hometown also guess that wealthier people will tend to live in the North in a fictional city, and is a substantially larger effect compared to examining the sample as a whole.

Follow-up analyses. The original study compared Hong Kong and U.S. participants. In the replication, Hong Kong participants expected the high-SES person to live further south (M =

-37.44, *SD* = 84.29) than the low-SES person (*M* = 12.43, *SD* = 95.03; *t*(140) = -3.30, *p* = 0.001, *d* = -0.55; 95% CI [-0.89, -0.22]). U.S. participants expected the high-SES person to live further north (*M* = 41.55, *SD* = 80.73) than the low-SES person (*M* = -42.63, *SD* = 82.41; *t*(2,199) = 24.20, *p* = 6.53e-115, *d* = 1.03; 95% CI [0.94, 1.12]). This result is consistent with the finding from the original study demonstrating cultural differences in perceived location of wealth in a fictional city correlating with location of wealth in one's hometown.

For most participants, the study was completed on a vertically oriented monitor display as opposed to completing a paper survey on a desk as in the original study. The original authors suggested *a priori* this may be important because associations between "up" and "good" or "down" and "bad" may interfere with any North/South associations. At ten data collection sites (*N* = 582), we assigned some participants to complete the slate on Microsoft Surface tablets resting on the table for horizontal administration. This addressed the original authors' hypothesis that the vertical orientation of the monitor would interfere with observing the relationship between cardinal direction on the map and perceived location of wealth. With just the participants using the horizontal tablets, those that said wealth tended to be in the north in their hometown (*n* = 156) expected the high-SES person to live further north (*M* = 38.66, *SD* = 80.43) than the low-SES person (*M* = -43.92, *SD* = 80.32; *t*(154) = 6.38, *p* = 1.95e-09, *d* = 1.03; 95% CI [.69, 1.36]). By comparison, within this horizontal tablet group, those that said wealth tended to be in the south in their hometown (*n* = 87) expected the high-SES person to live further south (*M* = -33.58, *SD* = 72.89) than the low-SES person (*M* = -4.11, *SD* = 88.33; *t*(85) =-1.63, *p* = .11, *d* = -.36; 95% CI [-.79, .08]). The effect sizes with just these subsamples are very similar to the effect sizes with the whole sample, suggesting that display orientation did not moderate this effect.

2. Structure and goal pursuit: A FUNCTIONAL BASIS FOR STRUCTURE-SEEKING: EXPOSURE TO STRUCTURE PROMOTES WILLINGNESS TO ENGAGE IN MOTIVATED ACTION (Kay, Laurin, Fitzsimons, & Landau, 2014, Study 2)

In Kay, Laurin, Fitzsimons, and Landau (2014), 67 participants generated what they felt was their most important goal. Participants then read one of two scenarios where a natural event (leaves growing on trees) was described as being a structured or random event. For example, in the structured condition, a sentence read "The way trees produce leaves is one of the many examples of the orderly patterns created by nature…", but in the random condition it read "The way trees produce leaves is one of the many examples of the natural randomness that surrounds us…". Next, participants answered three questions about their most important goal on a scale from "1 = *not very*" to "7 = *extremely*". The first measured subjective value of the goal and the other two measured willingness to engage in goal pursuit. Those exposed to a structured event (M = 5.26, SD = 0.88) were more willing to pursue their goal compared to those exposed to a random event (M = 4.72, SD = 1.32; $t(65) = 2.00$, $p = 0.05$, $d =$ 0.49, 95% CI [0.001, 0.980]).

In the overall replication sample (N = 6,506), those exposed to a structured event (M = 5.48, SD = 1.45) were not significantly more willing to pursue their goal compared to those exposed to a random event (M = 5.51, SD = 1.39; $t(6{,}498.63) = -0.94$, $p = 0.35$, $d = -.02$, 95% CI [-0.07, 0.03]. This result does not support the hypothesis that willingness to pursue goals is higher after exposure to structured versus random events.

3. Disfluency engages analytical processing: OVERCOMING INTUITION: METACOGNITIVE DIFFICULTY ACTIVATES ANALYTIC REASONING (Alter, Oppenheimer, Epley, & Eyre, 2007, Study 4)

Alter and colleagues (2007) investigated whether a deliberate, analytic processing style can be activated by incidental disfluency cues that suggest task difficulty. Forty-one participants attempted to solve syllogisms presented in either a hard- or easy-to-read font. The hard-to-read font served as an incidental induction of disfluency. Participants in the hard-to-read condition answered more moderately difficult syllogisms correctly (64%) than participants in the easy-to-read condition (42%; $t(39) = 2.01$, $p = 0.051$, $d = 0.64$, 95% CI [-0.004, 1.27]).

The original study focused on the two moderately difficult items from the six administered. Our confirmatory analysis strategy was sensitive to potential differences across samples in ability on syllogisms. We first determined which syllogisms were moderately difficult to participants by excluding any of the six items, within each sample, that were answered correctly by fewer than 25% of participants or more than 75% of participants across conditions. The remaining syllogisms were the basis of computing mean syllogism performance for each participant.

Following Alter et al. (2007), the easy-to-read font was *black Myriad Web 12-point* and the hard-to-read font was *10% grey italicized Myriad Web 10-point*. For a direct comparison with the original effect size, the original authors suggested that only English in-lab samples be used for two reasons: (1) we could not adequately control for online participants "zooming in" on the page or otherwise making the font more readable, and (2) we anticipated having to substitute the font in some translated versions because the original font (Myriad Web) may not support all languages[4]. In this subsample ($N = 2,580$), participants in the hard-to-read condition answered a similar number of syllogisms correct ($M = 1.10$, $SD = 0.88$) as participants in the easy-to-read condition ($M = 1.13$, $SD = 0.91$; $t(2,578) = -0.79$, $p = 0.43$, $d = -0.03$, 95% CI [-0.11, 0.05]). As a secondary analysis that mirrored the original, we used the same two

[4] Myriad Web did support all included languages and was used consistently across locations.

syllogisms from Alter et al (2007). Participants in the hard-to-read condition answered a similar number of syllogisms correctly (M = 0.80, SD = 0.79) as participants in the easy-to-read condition (M = 0.84, SD = 0.81; t(2,578) = -1.19, p = 0.23, d = -0.05, 95% CI [-0.12, 0.03]).[5] These results do not support the hypothesis that syllogism performance would be higher when the font is harder to read versus easier to read; the difference was slightly in the opposite direction and not distinguishable from zero (d = -0.03, 95% CI [-0.11, 0.05] versus original d=0.64).

Follow-up analyses. In the aggregate replication sample (N = 6,935), participants in the hard-to-read condition answered a similar number of syllogisms correctly (M = 1.03, SD = 0.86) as participants in the easy-to-read condition (M = 1.06, SD = 0.87; t(6,933) = -1.37, p = 0.17, d = -0.03, 95% CI [-0.08, 0.01]). Finally, in the whole sample, using the same two syllogisms from Alter et al. (2007), participants in the hard-to-read condition answered a similar number of syllogisms correctly (M = 0.75, SD = 0.76) as participants in the easy-to-read condition (M = 0.79, SD = 0.77; t(6,933) = -2.07, p = 0.039, d = -0.05, 95% CI [-0.097, -0.003]). These follow-up analyses do not qualify the conclusion from the focal tests.

4. Moral Foundations: LIBERALS AND CONSERVATIVES RELY ON DIFFERENT SETS OF MORAL FOUNDATIONS (Graham, Haidt, & Nosek, 2009, Study 1)

People on the political left (liberal) and political right (conservative) have distinct policy

[5] The original authors also hypothesized that this effect is sensitive to task order. If people are already thinking carefully (or if they're fatigued), the disfluency manipulation might not change how deeply they engage with the task. As such, the effect may be most detectable when it is done first. Considering only participants who did this task first (N = 988), participants in the hard-to-read condition answered a similar number of syllogisms correct (M = 0.49, SD = 0.77) as participants in the easy-to-read condition (M = 0.49, SD = 0.81; t(986) = -0.08, p = 0.94, d = -0.01, 95% CI [-0.13, 0.12]). Finally, using the same two syllogisms from Alter et al (2007), participants in the hard-to-read condition answered a similar number of syllogisms correctly (M = 0.37, SD = 0.65) as participants in the easy-to-read condition (M = 0.35, SD = 0.66; t(986) = 0.39, p = 0.70, d = 0.02, 95% CI [-0.10, 0.15])

preferences and may also have different moral intuitions and principles. 1,548 participants across the ideological spectrum rated whether different concepts such as *purity* or *fairness* were relevant for deciding whether something was right or wrong. Items that emphasized concerns of harm or fairness (individualizing foundations) were deemed more relevant for moral judgment by the political left than right (r = -0.21, d = -0.43, 95% CI [-0.55, -0.32]), whereas items that emphasized concerns for the ingroup, authority, or purity (binding foundations) were deemed more relevant for moral judgment by the political right than left (r = 0.25, d = 0.52, 95% CI [0.40, 0.63])[6]. Participants rated the relevance to moral judgment of 15 items (3 for each foundation) in a randomized order on a 6-point scale from "not at all relevant" to "extremely relevant".

The primary target of replication was the relationship of political ideology with the "binding" foundations. In the aggregate sample (N = 6,966), items that emphasized concerns for the ingroup, authority, or purity were deemed more relevant for moral judgment by the political right than political left (r = 0.14, p = 6.05e-34, d = 0.29, 95% CI [0.25, 0.34], q = 0.15, 95% CI [0.12, 0.17]). This result is consistent with the hypothesis that "binding" foundations are perceived as more morally relevant by members of the political right than the political left. The overall effect size was smaller than the original result (d = 0.29, 95% CI [0.25, 0.34] versus original d=0.52).

Follow-up analyses. The relationship of political ideology with the "individualizing" foundations was a secondary replication. In the aggregate sample (N = 6,970), items that emphasized concerns of harm or fairness were deemed more relevant for moral judgment by the

[6] Zero-order Pearson correlations are not provided in the original article. They have been computed on the raw public data and are based on N = 1,209 participants with pairwise complete values: https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/12658&studyListingIndex=0_775f45d232bb5e430d0024139e25

political left than political right (r = -0.13, p = 2.54e-29, d = -0.27, 95% CI [-0.32, -0.22], q = -0.13, 95% CI [-0.16, -0.11]). This result is consistent with the hypothesis that "individualizing" foundations are perceived as more morally relevant by members of the political left than the political right. The overall effect size was smaller than the original result (d = -0.27, 95% CI [-0.32, -0.22] versus original d = -0.43).

5. Affect and Risk: MONEY, KISSES, AND ELECTRIC SHOCKS: ON THE AFFECTIVE PSYCHOLOGY OF RISK (Rottenstreich & Hsee, 2001, Study 1)

Forty participants chose whether they would prefer an affectively attractive option (a kiss from a favorite movie star) or a financially attractive option ($50). In one condition, participants made the choice imagining a low probability (1%) of getting the outcome. In the other condition, participants imagined that the outcome was certain, they just needed to choose which one. When the outcome was unlikely 70% preferred the affectively attractive option, when the outcome was certain 35% preferred the affectively attractive option (χ^2(1, N=40) = 4.91), p = 0.0267, Kramers φ = 0.35, d = 0.74, 95% CI [<0.001, 1.74]). This result supported the hypothesis that positive affect has greater influence on judgments about uncertain outcomes than judgments about definite outcomes.

In the aggregate replication sample (N = 7,218), when the outcome was unlikely, 47% preferred the affectively attractive choice, and when the outcome was certain, 51% preferred the affectively attractive choice (p = 0.002, OR = 0.87, d = -0.08, 95% CI [-0.13, -0.03]). This result is opposite of the hypothesis that affectively attractive choices are more preferred when they are uncertain versus definite. The overall effect size was much smaller and in the opposite direction of the original study (d = -0.08, 95% CI [-0.13, -0.03] versus original d = 0.74).

6. Priming consumerism: CUING CONSUMERISM SITUATIONAL MATERIALISM

UNDERMINES PERSONAL AND SOCIAL WELL-BEING (Bauer, Wilkie, Kim, & Bodenhausen, 2012, Study 4)

Bauer and colleagues (2012) examined whether being in a consumer mindset would lead to less trust towards others. In Study 4, 77 participants read about a hypothetical water conservation dilemma in which they were involved. Participants were randomly assigned to either a condition that referred to the participant and others in the scenario as "consumers" or as "individuals." Participants in the consumer condition reported less trust toward others (1= *not at all*, 7 = *very much*) to conserve water (M = 4.08, SD = 1.56) compared to the control condition (M = 5.33, SD = 1.30), $t(76) = 3.86$, $p = 0.001$, $d = 0.87$, 95% CI [0.41, 1.34]).

In the aggregate replication sample (N = 6,608), participants in the consumer condition reported slightly less trust toward others to conserve water (M = 3.92, SD = 1.44) compared to the control condition (M = 4.10, SD = 1.45), $t(6{,}606) = 4.93$, p = 8.62e-7, $d = 0.12$, 95% CI [0.07, 0.17]). This result is consistent with the hypothesis that trust is lower when thinking of others as consumers versus thinking of others as individuals. The overall effect size was much smaller than the original result (d = 0.12, 95% CI [0.07, 0.17] versus original d = 0.87).

Follow-up analyses. The original experiment included four additional dependent variables. Comparing with the original study, the replication showed weaker effects in the same direction for (1) responsibility for the crisis (original d = 0.47; replication d = 0.10, 95% CI [0.05, 0.15]), (2) obligation to cut water usage (original d = 0.29; replication d = 0.08, 95% CI [0.03, 0.13]), (3) how much they viewed others as partners (original d = 0.53; replication d = 0.12, 95% CI [0.07, 0.16]), and (4) how much others should use less water (original d = 0.25; replication d = 0.01, 95% CI [-0.04, 0.06]).

7. Correspondence bias: CULTURAL VARIATION IN CORRESPONDENCE BIAS: THE

CRITICAL ROLE OF ATTITUDE DIAGNOSTICITY OF SOCIALLY CONSTRAINED BEHAVIOR (Miyamoto & Kitayama, 2002, Study 1)

Miyamoto and Kitayama (2002) examined whether Americans would be more likely than Japanese to show a bias toward ascribing to an actor an attitude corresponding to the actor's behavior, referred to as correspondence bias (Jones & Harris, 1967). In their Study 1, 49 Japanese and 58 American undergraduates learned they would read a university student's essay about the death penalty and infer the student's true attitude toward the issue. The essay was either in favor or against the death penalty, and it was designed to be diagnostic or not very diagnostic of a strong attitude. After reading the essay, participants learned that the student was assigned to argue the pro- or anti-position. Then, participants estimated the essay writer's actual attitude toward capital punishment and the extent to which they thought the student's behavior was constrained by the assignment.

Controlling for perceived constraint, analyses compared perceived attitudes of pro- versus anti-capital punishment essay writers. American participants perceived a large difference in actual attitudes when the essay writer had been assigned to write a pro-capital punishment essay ($M = 10.82$, $SD = 3.47$) versus anti-capital punishment essay ($M = 3.30$, $SD = 2.62$; $t(56) = 6.66$, $p < 0.001$, $d = 1.75$, 95% CI [1.14, 2.35]). Japanese participants perceived less of a difference in actual attitudes when the essay writer had been assigned to write a pro-capital punishment essay ($M = 9.27$, $SD = 2.88$) versus an anti-capital punishment essay ($M = 7.02$, $SD = 3.06$); $t(47) = 1.84$, $p = 0.069$, $d = 0.53$.

In the aggregate replication sample ($N = 7,197$), controlling for perceived constraint, participants perceived a difference in actual attitudes when the essay writer had been assigned to write a pro-capital punishment essay ($M = 10.98$, $SD = 3.69$) versus anti-capital punishment

essay (M = 4.45, SD = 3.51; F(2, 7194) = 3042.00, p < 2.2e-16, d = 1.82, 95% CI [1.76, 1.87]). This finding is consistent with the correspondence bias hypothesis--participants inferred the essay writer's attitude based, in part, on the observed behavior. Whether the magnitude of this effect varies cross-culturally is examined in the aggregate analysis section.

Follow-up analyses. For the primary replication, participants estimated the writer's true attitude toward capital punishment to be similar to the position that they were assigned to defend. Participants also expected writers would freely express attitudes consistent with the position to which they were assigned (pro-capital punishment M = 10.17, SD = 3.84; anti-capital punishment M = 4.96, SD = 3.61; t(7,187) = 59.44, p = 2.2e-16, d = 1.40, 95% CI [1.35, 1.45].

Two possible moderators were included in the design: perceived attitude of the average student in the writer's country (tailored to be the same country as the participant) and perceived persuasiveness of the essay. In the aggregate replication sample (N = 7,211), controlling for perceived constraint, we did not observe an interaction between condition and perceived attitude of the average student in the writer's country on estimations of the writer's true attitude toward capital punishment (t(7,178) = 0.55, p = 0.58, d = 0.013, 95% CI [-0.03, 0.06]). Also, in the aggregate replication sample (N = 7,211), controlling for perceived constraint, we did observe an interaction between condition and perceived persuasiveness of the essay on estimations of the writer's true attitude toward capital punishment (t(7,170) = 16.25, p = 2.3e-58, d = 0.38, 95% CI [0.34, 0.43]). The effect of condition on estimations of the writer's true attitude toward capital punishment was stronger for higher levels of perceived persuasiveness of the essay.

8. Disgust predicts homophobia: DISGUST SENSITIVITY PREDICTS INTUITIVE DISAPPROVAL OF GAYS (Inbar, Pizarro, Knobe, & Bloom, 2009, Study 1)

Behaviors that are deemed morally wrong may be judged as more intentional (Knobe,

2006). Thus, people who judge the portrayal of gay sexual activity in the media as an intentional act may find homosexuality morally reprehensible. In Inbar et al. (2009), 44 participants read a vignette about a director's action and judged him as more intentional when he encouraged gay kissing (M = 4.36, SD = 1.51) than when he encouraged kissing (M = 2.91, SD = 2.01; β = 0.41, $t(39) = 3.39$, $p = 0.002$, $r = 0.48$). Disgust sensitivity was related to judgments of greater intentionality in the gay kissing condition, $\beta = 0.79$, $t(19) = 4.49$, $p = 0.0003$, $r = 0.72$ and not the kissing condition, $\beta = -0.20$, $t(19) = -0.88$, $p = 0.38$, $r = 0.20$. The correlation in gay kissing condition was stronger than the correlation in the kissing condition, $z = 2.11$, $p = 0.03$, $q = .70$, 95% CI [.05, 1.35]. The authors concluded that individuals prone to disgust are more likely to interpret the gay kissing inclusion as intentional indicating that they intuitively disapprove of homosexuality.

The relationship between disgust sensitivity and intentionality ratings was the target of direct replication. In the aggregate replication sample (N = 7,117), participants did not judge the director's action as more intentional when he encouraged gay kissing (M = 3.48, SD = 1.87) than when he encouraged kissing (M = 3.51, SD = 1.84; $t(7{,}115) = -0.74$, $p = 0.457$, d = -0.02, 95% CI [-0.06, 0.03]). Disgust sensitivity was related to judgments of greater intentionality in both the gay kissing condition, $r = 0.12$, p = 1.2e-13, and the kissing condition, $r = 0.07$, p =2.48e-5. The correlation in the gay kissing condition was similar to the correlation in the kissing condition, z = 2.62, $p = 0.02$, $q = 0.05$, 95% CI [0.01, 0.10]. These data are inconsistent with the original finding that disgust sensitivity and intentionality are more strongly related when considering gay kissing than kissing, and the effect size was much smaller ($q = 0.05$, 95% CI [0.01, 0.10] versus original $q = 0.70$). Disgust sensitivity was very weakly related to intentionality and there was no mean difference in intentionality between gay kissing and kissing conditions.

Follow-up analyses. The original study included two other outcome measures. These were examined as secondary replications following the same analysis strategy. For the first, disgust sensitivity was only slightly more related to yes or no answers to "Is there anything wrong with homosexual men French kissing in public?" (r = -0.20, p < 2.2e-16) than "Is there anything wrong with couples French kissing in public?" (r = -0.16, p < 2.2e-16; z = -1.66, p = 0.096, q = -0.04, 95% CI [-0.09, 0.01]). For the second, disgust sensitivity was only slightly more related to answers to "Was it wrong of the director to make a video that he knew would encourage homosexual men to French kiss in public?" (r = 0.27, p < 2.2e-16) than to "Was it wrong of the director to make a video that he knew would encourage couples to French kiss in public?"(r = 0.22, p < 2.2e-16; z = 2.28, p = 0.02, q = 0.05, 95% CI [0.01, 0.10]).

9. Incidental anchors: INCIDENTAL ENVIRONMENTAL ANCHORS (Critcher & Gilovich, 2008, Study 2)

In Critcher and Gilovich (2008), 207 participants predicted the relative popularity between geographic regions of a new cell phone that was entering the marketplace. In one condition, the smartphone was called the P97; in the other condition, the smartphone was called the P17. Participants in the P97 condition estimated a greater proportion of sales in the U.S. (M = 58.1%, SD = 19.6%) than did participants in the P17 condition (M = 51.9%, SD = 21.7%; t(197.5) = 2.12, p = 0.03, d = 0.30, 95% CI [0.02, 0.58]). This supported the hypothesis that judgment can be influenced by incidental anchors in the environment. The mere presence of a high or low number in the name of the cell phone influenced estimates of sales of the phone.

In the aggregate replication sample (N = 6,826), participants in the P97 condition estimated approximately the same proportion of sales in their region (M = 49.87%, SD = 21.86%) as did participants in the P17 condition (M = 48.98%, SD = 22.14%; t(6824) = 1.68, p =

0.09, d = 0.04, 95% CI [-0.01, 0.09]). This result does not support the hypothesis that sales estimates would be influenced by incidental anchors. The effect size was in the same direction, but much smaller (d = 0.04, 95% CI [-0.01, 0.09] versus original d = 0.30) and indistinguishable from zero.

Follow-up analyses. The original authors avoided administering these studies on computer, rather than with paper and pencil, to avoid the possibility that the numeric keys on the keyboard might serve as primes. We administered this task with paper and pencil at 11 sites. Using just the paper-pencil sites (N = 1,112), participants in the P97 condition estimated a slightly smaller proportion of sales in their region (M = 53.02%, SD = 20.15%) than did participants in the P17 condition (M = 53.28%, SD = 20.17%; $t(1110) = -0.22$, $p = 0.83$, d = -0.01, 95% CI [-0.13, 0.10]). This difference is in the opposite direction of the original finding, but not reliably different from zero.

10. Social Value Orientations: DEVELOPMENT OF PROSOCIAL, INDIVIDUALISTIC, AND COMPETITIVE ORIENTATIONS: THEORY AND PRELIMINARY EVIDENCE (Van Lange, Otten, De Bruin, & Joireman, 1997, Study 3)

Van Lange and colleagues (1997) proposed that social value orientations (SVOs) are rooted in social interaction experiences, among them the number of one's siblings. In one of the four studies (Study 3), they examine the association between SVO and family size, thereby providing a test of two competing hypotheses. One hypothesis states larger families, resources have to be shared more frequently, facilitating cooperation and the development of a prosocial orientation (*sibling-prosocial hypothesis*). Another hypothesis, rooted in group size effects, states that greater family size may undermine trust and expected cooperation from others, and may therefore inhibit the development of prosocial orientation (*sibling-proself hypothesis*). In their

Study 3, 631 participants reported how many siblings they had and completed a SVO measure called the triple dominance measure to identify them as prosocials, individualists, or competitors. Prosocials had more siblings (M = 2.03, SD = 1.56) than individualists (M = 1.63, SD = 1.00) and competitors (M = 1.71, SD = 1.35; $F(2, 535) = 4.82$, $p = 0.01$, ds = 0.287, 95% CI [0.095, 0.478] and 0.210, 95% CI [-0.045, 0.465] respectively). Planned comparisons revealed a significant contrast between prosocials versus individualists and competitors (F(1,535) = 9.14, p = 0.003, d = 0.19, 95% CI [<0.01, 0.47]). The original demonstration used a triple dominance measure of social value orientation with three categorical values. In discussion with the original first author, an alternative measure, the SVO slider (Murphy et al., 2011) was identified as a useful replacement to yield a continuous distribution of scores.

The current replication focuses only on the observed direct positive correlation between greater prosocial orientation and number of siblings. In the aggregate replication sample (N = 6,234), number of siblings was not related to prosocial orientation (r = -0.02, p = 0.18, 95% CI [-0.04, 0.01]). This result does not support the hypothesis that more siblings is positively related with prosocial orientation. Direct comparison of effect size is not possible because of changes in measures, but the replication effect size was near zero.

11. Trolley Dilemma 1: A DISSOCIATION BETWEEN MORAL JUDGMENTS AND JUSTIFICATIONS (Hauser, Cushman, Young, Jin, & Mikhail, 2007, Scenarios 1+2)

The principle of the double effect suggests that acts that harm others are judged as more morally permissible if the act is a foreseen side effect rather than the means to the greater good. Hauser and colleagues (2007) compared participant reactions to two scenarios to test this principle. As a *foreseen side effect* scenario, a person on an out-of-control train changes the train's trajectory so that the train kills one person instead of five. As a *greater good* scenario, a

person pushes a fat man in front of a train, killing him, to save five people. While 89% of participants judged the action in the foreseen side effect scenario as permissible (95% CI [0.87, 0.91]), only 11% of participants in the greater good scenario judged it as permissible (95% CI [0.09, 0.13]). The difference between the proportions was significant. (χ^2 [1, N = 2646] = 1615.96, $p < 0.001$), $w = 0.78$, $d = 2.50$, 95% CI [2.22, 2.86], providing evidence for the principle of the double effect.

In the aggregate replication sample (N = 6,842 after removing participants that responded in less than 4 seconds), 71% of participants judged the action in the foreseen side effect scenario as permissible, but only 17% of participants in the greater good scenario judged it as permissible. The difference between the proportions was significant (p = 2.2e-16), OR = 11.54, d = 1.35, 95% CI [1.28, 1.41]. The replication results were consistent with the hypothesis of the double effect, and the effect was about half the magnitude of the original (d = 1.35, 95% CI [1.28, 1.41] versus original d = 2.50).

Follow-up analyses. Variations of the trolley problem are well-known. The original authors suggested the effect may be weaker for participants who have previously been exposed to this sort of task. We included an additional item assessing participants' prior knowledge of the task. Among the 3,069 participants reporting that they were not familiar with the task, the effect size was d = 1.47, 95% CI [1.38, 1.57]; and among the 4,107 familiar with the task, the effect size was d = 1.20, 95% CI [1.12, 1.28]. This suggests moderation by task familiarity, but the effect was very strong regardless of familiarity.

12. Sociometric status and well-being: THE LOCAL-LADDER EFFECT AND SUBJECTIVE WELL-BEING (Anderson, Kraus, Galinsky, & Keltner, 2012, Study 3).

Anderson and colleagues (2012) examined the relationship between sociometric status

(SMS), socioeconomic status (SES), and subjective well-being. According to the authors, SMS refers to interpersonal wealth, whereas SES measures fiscal wealth. Study 3 examined whether SMS has stronger ties than SES to well-being. In a 2 x 2 between subjects design, 228 Mechanical Turk participants were presented with descriptions of people who were either relatively high or low on either socioeconomic or sociometric status, and then made upward or downward social comparisons (e.g., participants in the high sociometric status condition imagined and compared themselves with a low sociometric status person). Then, participants wrote about what it would be like to interact with such people, and then reported subjective well-being. Results showed a significant 2 x 2 interaction ($F(1, 224) = 4.73$, $p = 0.03$) such that participants in the high sociometric status condition had higher subjective well-being than those in the low sociometric status condition, $t(115) = 3.05$, $p = 0.003$, $d = 0.57$, 95% CI [0.20, 0.93]. There were no differences between the two socioeconomic conditions, $t(109) = 0.06$, $p = 0.96$, $d = 0.01$.

For replication, we used only the high- and low-sociometric status conditions and excluded the high- and low-socioeconomic status conditions that showed no differences in the original study. In the aggregate replication sample ($N = 6{,}905$), participants in the high sociometric status condition ($M = -0.01$, $SD = 0.67$) had slightly *lower* subjective well-being than those in the low sociometric status condition ($M = 0.01$, $SD = 0.66$), $t(6{,}903) = -1.76$, $p = 0.08$, $d = -0.04$, 95% CI [-0.09, 0.004]. This result did not support the hypothesis that subjective well-being would be higher for participants exposed to higher versus lower sociometric status descriptions. The effect size was much smaller and slightly in the opposite direction ($d = -0.04$, 95% CI [-0.09, 0.004] versus original $d = 0.57$).

13. False Consensus - Supermarket: THE "FALSE CONSENSUS EFFECT": AN

EGOCENTRIC BIAS IN SOCIAL PERCEPTION AND ATTRIBUTION PROCESSES

(Ross, Greene, & House, 1977, Study 1, Supermarket Scenario)

People perceive a "false consensus" about the commonness of their responses among others (Ross, Greene, & House, 1977). Thus, estimates of the prevalence of a particular belief, opinion, or behavior are biased in the direction of the perceiver's beliefs, opinions, and behaviors. Ross and colleagues (1977, Study 1) presented 320 college undergraduates with one of four hypothetical events that culminated in a clear dichotomous choice of action. Participants first estimated what percentage of peers would choose each option, and then indicated their own choice. For each of the four scenarios, participants that chose the first option believed that a higher percentage of others would also choose that option (M = 75.4%) than participants that chose the second option (M = 54.9%), $F(1, 312) = 49.1$, $p < 0.001$, $d = 0.79$, 95% CI [0.56, 1.02] for the main effect of experimental condition; meta-analysis (random effects model) of scenario effect sizes: $d = 0.66$). A later meta-analysis suggests that this effect is robust and moderate in size across a variety of paradigms ($r = 0.31$, Mullen et al., 1985).

This study was replicated in Slate 1 and Slate 2 using different scenarios. In Slate 1, participants were presented with the "supermarket" vignette from the original study ($F(1, 78) = 17.7$, $d = 0.99$, 95% CI [0.00, 2.29]) . All participants who provided percent estimates between 0-100 and responded to all three items were included in the analysis. In the aggregate replication sample (N = 7,205), participants that chose the first option believed that a higher percentage of others would also choose that option (M = 69.19%) than participants that chose the second option (M = 43.35%), $t(6,420.77) = 49.93$, $p < 2.2e-16$, $d = 1.18$, 95% CI [1.13, 1.23]. This result is consistent with the hypothesis that participants' choices would be positively correlated with their perception of the percentage of others that would make the same choice.

SLATE 2

14. False Consensus - Traffic Ticket: THE "FALSE CONSENSUS EFFECT": AN EGOCENTRIC BIAS IN SOCIAL PERCEPTION AND ATTRIBUTION PROCESSES (Ross, Greene, & House, 1977, Study 1, Traffic Ticket Scenario)

The original study was presented in Effect 13 in Slate 1 above. In Slate 2, participants were presented with the "traffic ticket" vignette from the original study ($F(1, 78) = 12.8$, $d = 0.80$, 95% CI [0.00, 1.87]). All participants who provided percent estimates between 0-100 and responded to all three items were included in the analysis. In the aggregate replication sample ($N = 7{,}827$), participants that chose the first option believed that a higher percentage of others would also choose that option ($M = 72.48\%$) than participants that chose the second option ($M = 48.76\%$), $t(6{,}728.25) = 41.74$, $p < 2.2\text{e-}16$, $d = 0.95$, 95% CI [0.90, 1.00]. This result is consistent with the hypothesis that participants' choices would be positively correlated with their perception of the percentage of others that would make the same choice.

15. Vertical position and power: HIGH IN THE HIERARCHY: HOW VERTICAL LOCATION AND JUDGMENTS OF LEADERS' POWER ARE INTERRELATED (Giessner & Schubert, 2007, Study 1a)

Sixty-four participants formed an impression of a manager based on few pieces of information including a organization chart with a vertical line connecting the manager on top with his team below. Participants were randomly assigned to one of two conditions in which the line was either short (2 cm) or long (7 cm). Then, participants evaluated the manager on a variety of qualities including the manager's power. Participants in the long line condition ($M = 5.01$, $SD = 0.60$) perceived the manager to have greater power than did participants in the short line condition ($M = 4.62$, $SD = 0.81$; $t(62) = 2.20$, $p = 0.03$, $d = 0.55$, 95% CI [0.05, 1.05]. This

result was interpreted as showing that people associated vertical position with power, with a higher position indicating greater power.

Participants' responses to the five items assessing the manager's power were averaged. In the aggregate replication sample (N = 7,890), participants in the long line condition (M = 4.97, SD = 1.09) perceived the manager to have similar power as did participants in the short line condition (M = 4.93, SD = 1.07; t(7,888) = 1.40, p = 0.16, d = 0.03, 95% CI [-0.01, 0.08]. This result does not support the hypothesis that perceived power would be higher with greater versus less physical distance. The replication effect size was in the same direction, but much smaller than the original (d = 0.03, 95% CI [-0.01, 0.08] versus original d = 0.55).

16. Framing decisions: THE FRAMING OF DECISIONS AND THE PSYCHOLOGY OF CHOICE (Tversky & Kahneman, 1981, Study 10)

In Tversky and Kahneman (1981), 181 participants considered a scenario in which they were buying two items, one relatively cheap ($15) and one relatively costly ($125). Ninety-three participants were assigned to a condition in which the cheap item could be purchased for $5 less by going to a different branch of the store 20 minutes away. Eighty-eight participants saw another condition in which the costly item could be purchased for $5 less at the other branch. Therefore, the total cost for the two items, and the cost savings for traveling to the other branch, was the same across conditions. Participants were more likely to say that they would go to the other branch when the cheap item was on sale (68%) than when the costly item was on sale (29%, Z = 5.14, p = 7.4e-7, OR = 4.96, 95% CI [2.55, 9.90]). This suggests that the decision of whether to travel was influenced by the base cost of the discounted item rather than the total cost.

For the replication, in consultation with one of the original authors, dollar amounts were adjusted to be more appropriate for 2014 (i.e., when the replication study was conducted). The

stimuli were also replaced with consumer items that were relevant in 2014 and plausibly sold by a single salesperson (a ceramic vase and a wall hanging). In the aggregate replication sample (N = 7,228), participants were more likely to say that they would go to the other branch when the cheap item was on sale (49%) than when the costly item was on sale (32%, p = 1.01e-50, d = 0.40 95% CI [.35, .45]; OR = 2.06, 95% CI [1.87, 2.27]). These results are consistent with the hypothesis that the base cost of the discounted item influenced willingness to travel, though the effect was less than half the size of the original (OR = 2.06, 95% CI [1.87, 2.27] versus original OR = 4.96).

17. Trolley dilemma 2: A DISSOCIATION BETWEEN MORAL JUDGMENTS AND JUSTIFICATIONS (Hauser et al., 2007, Study 1, Scenarios 3+4)

This study was presented in Effect 11 in Slate 1 using different scenarios. In Slate 2, participants were presented with the "Ned" and "Oscar" scenarios as the *greater good* and *foreseen side effect* scenarios. In the original study, when these two effects were compared, 72% of subjects judged the action in the foreseen side effect scenario as permissible (95% CI [0.69, 0.74]), and 56% of subjects judged the action in the means to a greater good scenario as permissible (95% CI [0.53, 0.59]). The difference between the proportions was significant, χ^2[1, N = 2612] = 72.35, $p < 0.001$, w = 0.17, d = 0.34, 95% CI [0.26, 0.42].

In the aggregate replication sample (N = 7,923) after removing participants who responded in less than 4 seconds, 64% of participants judged the action in the foreseen side effect scenario as permissible and 53% of participants in the greater good scenario judged it as permissible (95%). The difference between the proportions was significant (p = 4.66e-23, d = 0.25, 95% CI [0.20, 0.30]; OR = 1.58, 95% CI [1.44, 1.72]) . These results are consistent with the principle of double effect with a somewhat smaller effect size in the replication compared to

the original study (0.25, 95% CI [0.20, 0.30] versus original d = 0.34).

Follow-up analyses. Again, we included an additional item assessing participants' prior knowledge of the task. Among the 3,558 participants reporting that they were not familiar with the task, the effect size was d = 0.27, 95% CI [0.20, 0.34]; and among the 4,297 familiar with the task, the effect size was d = 0.24, 95% CI [0.17, 0.30]. In this case, familiarity did not moderate the observed effect size.

18. Tempting fate: WHY PEOPLE ARE RELUCTANT TO TEMPT FATE (Risen & Gilovich, 2008, Study 2)

Risen and Gilovich (2008) explored the belief that tempting fate increases bad outcomes. The authors tested whether people judge the likelihood of a negative outcome to be higher when they imagined themselves or a classmate tempting fate, compared to when they do not tempt fate. One hundred and twenty participants read a scenario in which either they or a classmate ("Jon") tempt fate (e.g., by not reading before class), or do not tempt fate (e.g., by coming to class prepared). Participants then estimated how likely it is that the protagonist (themselves or Jon) would be called on by the professor. The predicted main effect of tempting fate emerged, as participants judged the likelihood of being called on to be higher when the protagonist had tempted fate (M = 3.43, SD = 2.34) than when the protagonist had not tempted fate (M = 2.53, SD = 2.24; t(116) = 2.15, p = 0.034, d = 0.39, 95% CI [0.03, 0.75]). The original study design included self and other scenarios. No self-other differences were found. With the original authors' approval, we limited the study to the two self conditions.

In the aggregate replication sample (N = 8,000), participants judged the likelihood of being called on to be higher when they had tempted fate (M = 4.58, SD = 2.44) than when they had not tempted fate (M = 4.14, SD = 2.45; t(7,998) = 8.08, p = 7.70e-16, d = 0.18, 95% CI

[0.14, 0.22]). This is consistent with the hypothesis that tempting fate would increase the likelihood of being called on, though the effect size was less than half the size of the original study (d = 0.18, 95% CI [0.14, 0.22] versus original d = 0.39).

For the key confirmatory test, the original authors suggested that the sample should include only undergraduate student samples given the nature of the question. In that subsample (N = 4,599), participants judged the likelihood of being called on to be higher when they had tempted fate (M = 4.61, SD =2.42) than when they had not tempted fate (M = 4.07, SD = 2.36; t(4,597) =7.57, p = 4.4e-14, d = 0.22, 95% CI [0.17, 0.28]). The observed effect size (0.22) was very similar to what was observed with the whole sample (0.18).

Follow-up analyses. During peer review of the design and analysis plan, gender was suggested as a possible moderator of the effect. Using the undergraduate subsample, we conducted a 2 x 2 ANOVA with condition and gender as factors. In addition to the main effect of condition, there was a main effect of gender (F(1, 4524) = 31.80, p = 1.81e-8, d = 0.17, 95% CI [0.09, 0.25]) indicating that females judged the likelihood of being called on to be higher than males. There was also a very weak interaction of condition and gender (F(1, 4524) = 5.10, p = 0.024, d = 0.07, 95% CI [0.04, 0.13]).

19. Actions are choices?: WHAT COUNTS AS A CHOICE? U.S. AMERICANS ARE MORE LIKELY THAN INDIANS TO CONSTRUE ACTIONS AS CHOICES (Savani, Markus, Naidu, Kumar, & Berlia, 2010, Study 5)

Savani and colleagues (2010) examined cultural asymmetry in people's construal of behavior as choices. In Study 5, 218 participants (90 Americans, 128 Indians) were randomly assigned to either recall personal actions or interpersonal actions, and then to indicate whether the actions constituted choices. In a logistic HLM model with construal of choice as the

dependent measure, culture and condition as participant-level predictors, and importance as a trial-level covariate, the authors found no main effect of condition across cultures: β = –0.13, OR = 0.88, d = 0.10, t(101) = 0.71, p = 0.48. Among Americans, there was no difference between construing personal (M = 0.83, SD = 0.15) and interpersonal actions (M = 0.82, SD = 0.14) as choices, t(88) = 0.39, p = 0.65, d = 0.04. However, Indians were less likely to construe personal actions (M = 0.61, SD = 0.26) than interpersonal actions (M = 0.71, SD = 0.26) as choices, t(126) = -3.69, p = 0.0002, d = -0.65, 95% CI [-1.00, -.30].

For the replication, we conducted the same hierarchical logistic regression analysis with choice (binary) as the dependent variable, the importance of decision (ordered categorical) as a trial-level covariate nested within participants, and condition (categorical) as a participant-level factor indicating whether a participant was in the personal or interpersonal condition. The effect of interest was the odds of construing an action as a choice, depending on the condition a participant was in, controlling for the reported importance of the action.

After excluding participants collected outside of university labs, based on recommendation of the original authors, and those who did not respond to all choice and importance of choice questions (N = 3,506), we found a significant main effect of condition (β = -0.43, SE = 0.03, z = -12.54, p < 2e-16, d = -0.24, 95% CI [-0.27, -0.21]). Additional exploratory analyses revealed a significant interaction between condition and importance of decision (β = -0.08, SE = 0.02, z = -4.23, p = 2.37e-5). Participants were less likely to construe personal (M = 0.74, SD = 0.44) than interpersonal (M = 0.82, SD = 0.39) actions as choices, and this effect was stronger at higher ratings of the importance of the choice. This small effect (d = -0.24, 95% CI [-0.27, -0.21]) differs from the original null (0.04) among Americans and is in the same direction but smaller than the original effect among Indians (-0.65), but the present sample was highly

diverse.

For the key confirmatory test of the original result among Indians, we selected participants from university labs in India who responded to all choice and importance of choice questions (N = 122). In this subsample, we found no main effect of condition (β = -0.06, SE = 0.17, z = -0.34, p = 0.73, d = -0.03, 95% CI [-0.18, 0.11]) and a significant interaction effect between condition and importance of decision (β = 0.35, SE = 0.09; z = 3.79, p = 1.0e-4, d = 0.19, 95% CI [0.05, 0.34]. Indian participants were equally likely to construe personal actions (M = 0.63, SD = 0.48) and interpersonal actions (M = 0.63, SD = 0.48) as choices. The main effect, controlling for importance, does not support the original hypothesis that Indians are less likely to construe personal actions than interpersonal actions as choices, despite there being such a main effect in the full sample. Like in the full sample, this effect was moderated by ratings of the importance of the choice, such that interpersonal actions were more likely to be construed as choices at lower levels of importance, whereas personal actions were more likely to be construed as choices at higher levels of importance. This moderation result was not reported in the original paper.

Follow-up analyses. The original authors suggested that only university samples should be included for the main analyses above. Among the whole sample, after excluding only participants who did not respond to all choice and importance of choice questions (N = 5,882), we found a significant effect of condition (β = -0.33, SE = 0.03, z = -11.54, p < 2.0e-16, d = -0.18, 95% CI [-0.21, -0.16]) and a significant interaction effect between condition and importance of choice (β = -0.06, SE = 0.014, z = -4.46, p = 8.04e-6, d = -0.03, 95% CI [-0.06, -0.01]). In the whole sample, participants were less likely to construe personal (M = 0.74, SD = 0.44) than interpersonal (M = 0.79, SD = 0.40) actions as choices, and this effect was stronger at

higher ratings of the importance of the choice.

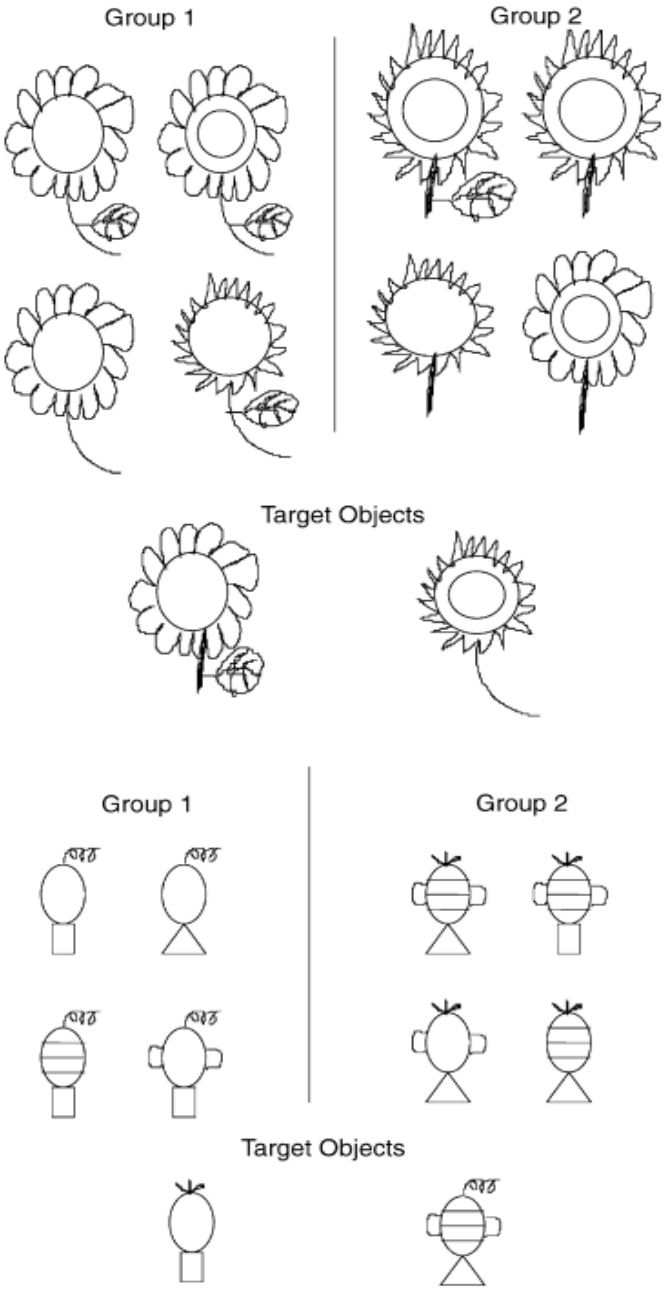
20. Formal versus intuitive reasoning: CULTURAL PREFERENCES FOR FORMAL VERSUS INTUITIVE REASONING (Norenzayan, Smith, Kim, & Nisbett, 2002, Study 2)

The way people living in the West think may be more rule-based compared to the way people living in East Asia think. Fifty-two European American (27 men, 25 women), 52 Asian American (28 men, 24 women) and 53 East Asian participants (27 men, 26 women) were randomly assigned to either a classification (decide "which group the target object belongs to"; two-thirds of sample) or similarity judgment (decide "which group the target object is most similar to"; one-third of sample) condition.

All participants categorized targets into two alternative groups, each consisting of four exemplars. Both targets and group exemplars were defined according to four binary features (e.g., long-stemmed or short-stemmed flowers). In Group 1, all exemplars had one feature in common with each other and with the target. In Group 2, there was no feature in common among all exemplars and the target, but one exemplar had three features in common with the target and three exemplars had two features in common with the target (see Figure 1). As a consequence, Group 2 looked more similar to the target, but there was no feature that could be used as a rule to categorize the target as a member of the group. But, for Group 1, a single feature common to all could be used as a rule for classification. Each set of targets and groups had a mirror-image target so that one group could be used for rule-based classification for one target, and the other group could be used for rule-based classification for the other target. When asked "which Group the target object *belongs* to" participants across all three cultures preferred to classify based on rule (M = 69% European Americans; M = 70% East Asians) rather than on family resemblance ($F(1, 100) = 44.40$, $p < 0.001$, $r = 0.55$). When asked "which group the

target object is more *similar* to", European Americans gave many more responses based on the unidimensional rule (M_{rule} = 69%) than on family resemblance (M_{family} = 31%), $t(17) = 3.68$, p = 0.002, d = 1.65, 95% CI [0.59, 2.67]. In contrast, East Asians gave fewer rule-based responses

Figure 1. Examples of Targets and Groups for Replication of Norenzayan et al. (2002)



than family resemblance responses (M_{rule} = 41% vs. M_{family} = 59%), $t(17) = -2.09$, $p = 0.05$, d = -

.93, 95% CI [-1.85, 0.01]. The responses of Asian Americans were intermediate, with participants indicating no preference for one rule over the other (M_{rule} = 46% vs. M_{family} = 54%), $t < 1$.

For replication, we preregistered to compare the percentage of rule-based responses between the "belong to" and "similar to" conditions for which European Americans showed no difference (d = 0.00, 95% CI [-0.15, 0.15]) and East Asians showed a greater likelihood of selecting rule-based responses in the "belong to" than the "similar to" condition (d = 0.67, 95% CI [0.52, 0.81). Note that this preregistered plan was a comparison across the experimental conditions, whereas Norenzayan et al. focused their analysis and theoretical interest on between cultural groups comparisons within experimental conditions.

For the replication analysis, we computed for each participant the percentage of rule-based responses and tested whether the means of the two experimental groups ("belong to" versus "similar to") on this DV were equal with a *t*-test for independent samples. In the aggregate replication sample (N = 7,396), participants who were asked "which Group the target object *belongs* to" were more likely to classify based on a rule (M = 64%, SD = 25%) than family resemblance (36%), and participants asked "which group the target object is more *similar* to" were more likely to classify based on family resemblance (56%) than a rule (M = 44%, SD = 21%). The likelihood of using a rule was higher in the "belong to" condition compared to the "similar to" condition (t(7,227.59) = 37.05, p = 3.04e-275, d = 0.86, 95% CI [0.81, 0.91]). This is in the same direction as the original aggregate result with a somewhat larger effect size--People were more likely to categorize based on a rule when considering what the target "belongs to" and more likely to categorize based on family resemblance when considering what the target is

"similar to".[7]

Follow-up analyses. For this effect in particular, we identified *a priori* that it and Tversky and Gati (1978) both involved similarity judgments and thus order of these may be particularly relevant. We compared whether this effect was moderated by Tversky and Gati (1978) appearing before or after, and observed very weak moderation by order ($t(7,392) = 2.34$, $p = 0.02$, $d = 0.05$, 95% CI [0.01, 0.10]).

21. Less is better: WHEN LOW-VALUE OPTIONS ARE VALUED MORE HIGHLY THAN HIGH-VALUE OPTIONS (Hsee, 1998, Study 1)

Hsee (1998) demonstrated the less-is-better effect wherein a less expensive gift can be perceived as more generous than a more expensive gift when the less expensive gift is relatively higher priced compared to other items in its category, and the more expensive item is a low-priced item compared to other items in its category. Eighty-three participants imagined that they were about to study abroad and had received a goodbye gift from a friend. In one condition, participants imagined receiving a $45 scarf bought in a store where the prices of scarves ranged from $5 to $50. In the other condition, participants imagined receiving a $55 coat bought in a store where the prices of coats ranged from $50 to $500. Participants in the scarf condition considered their gift giver significantly more generous ($M = 5.63$) than those in the coat condition ($M = 5.00$; $t(82) = 3.13$, $p = 0.002$, $d = 0.69$, 95% CI [0.24, 1.13]), despite the gift being objectively less expensive.

[7] Norenzayan et al.'s original study had two key predictions: (1) all cultural groups would show more rule-based responding in the "belong to" than in the "similar to" conditions, and (2) the European American sample would show more rule-based responding across both conditions. They observed evidence for the first prediction, and evidence for the second prediction only in the "similar to" condition. Across the replication samples, we also observed greater likelihood of selecting rule-based responses in the "belong to" condition compared to the "similar to" condition, but the WEIRD samples showed more use of rule-based responses than less WEIRD samples in the "belong to" condition (WEIRD M = 65.2%, less WEIRD M = 59.8%), and less use of rule-based responses than less WEIRD samples in the "similar to" condition (WEIRD M = 42.8%, less WEIRD M = 48.4%).

In the replication, the dollar values were approximately adjusted for inflation. We converted this amount to local currencies at sites where the U.S. dollars would be less familiar to participants. In the aggregate replication sample (N = 7,646), participants in the scarf condition considered their gift giver significantly more generous (M = 5.50, SD = 0.89) than those in the coat condition (M = 4.61, SD = 1.34; t(6,569.67) = 34.20, p = 4.5e-236, d = 0.78, 95% CI [0.74, 0.83]). This result is consistent with the less-is-better effect, and yielded a slightly larger effect size than the original demonstration (d = 0.78, 95% CI [0.74, 0.83] versus original d = 0.69).

22. Moral typecasting: DIVERGENT PERCEPTIONS OF MORAL AGENTS AND MORAL PATIENTS (Gray & Wegner, 2009, Study 1a)

Gray and Wegner (2009) examined the attribution of intentionality and responsibility as a function of perceived moral agency--the ability to direct and control one's moral decisions. In Study 1a, 69 participants read about an event involving a person high on moral agency (an adult man) and a person low on moral agency (a baby). In one condition, the man knocked over a tray of glasses resulting in harm to the baby. In the other condition, the baby knocked over the tray of glasses resulting in harm to the man. Participants then rated the degree to which the person who committed the act was responsible, how intentional the act was, and how much pain was felt by the victim. The adult man (M = 5.29, SD = 1.86) was evaluated as more responsible for committing the act than the baby (M = 3.86, SD = 1.64, t(68) = 3.32, p = 0.001, d = 0.80, 95% CI [0.31, 1.29]). Likewise, the adult man (M = 4.05, SD = 2.05) was rated as acting more intentionally than the baby (M = 3.07, SD = 1.55, t(68) = 2.20, p = 0.03, d = 0.53). Finally, when on the receiving end of the act, the adult man (M = 4.63, SD = 1.15) was viewed as feeling less pain compared to a baby (M = 5.76, SD = 1.55, t(68) = 3.49, p = 0.001, d = 0.85).

The effect of condition on perceived responsibility was identified as the primary

relationship for replication. In the aggregate replication sample (N = 8,002), the adult man (M = 5.41, SD = 1.63) was evaluated as more responsible for committing the act than the baby (M = 3.77, SD = 1.79, t(7,913.89) = 42.62, p < 3.32e-285, d = 0.95, 95% CI [0.91, 1.00]). This result is consistent with the hypothesis that an adult would be perceived as more responsible for harming a baby than a baby would be for harming an adult. The effect size in the replication was slightly larger than the original result (d = 0.95, 95% CI [0.91, 1.00] versus original d = 0.80).

Follow-up analyses. There were two additional dependent variables for secondary analysis: intentionality and felt pain by the victim. The adult man (M = 3.62, SD = 1.89) was rated as acting more intentionally than the baby (M = 2.73, SD = 1.64, t(7,864.62) = 22.51, p = 8.3e-109, d = 0.50, 95% CI [0.46, 0.55]). And, when on the receiving end of the act, the adult man (M = 4.66 SD = 1.25) was viewed as feeling less pain compared to a baby (M = 5.44, SD = 1.25, t(7,989) = 27.54, p = 1.5e-159, d = 0.62, 95% CI [0.57, 0.66]).

23. Morality and physical cleansing: WASHING AWAY YOUR SINS: THREATENED MORALITY AND PHYSICAL CLEANSING (Zhong & Liljenquist, 2006, Study 2)

Zhong and Liljenquist (2006) investigated whether moral violations can induce a desire for cleansing. In Study 2, under the guise of a study assessing personality from handwriting, 27 participants hand-copied a first-person account of an ethical act (helping a co-worker) or unethical act (sabotaging a co-worker). Then, participants rated the desirability of five cleaning products and five non-cleaning products. Participants who copied the unethical account (M = 4.95, SD = 0.84) reported that the cleansing products were more desirable than participants who copied the ethical account (M = 3.75, SD = 1.32; F(1, 25) = 6.99, p = 0.01, d = 1.02, 95% CI [<0.01, 2.44]). There was no difference between the unethical (M = 3.85, SD = 1.21) and ethical (M = 3.91, SD = 1.03) conditions in ratings of non-cleansing products (F(1, 25) = 0.02, p = 0.89,

$d = 0.05$).

The effect of interest for replication is whether condition affects ratings of the cleaning products. In the aggregate replication sample ($N = 7{,}001$), after removing participants that copied less than half of the target article, participants who copied the unethical account ($M = 3.95$, $SD = 1.43$) reported that the cleansing products were similarly desirable as participants who copied the ethical account ($M = 3.95$, $SD = 1.45$; $t(6{,}999) = -0.11$, $p = 0.91$, $d = 0.00$, 95% CI [-0.05, 0.04]). This result is not consistent with the hypothesis that copying an unethical action would increase desirability of cleaning products compared to copying an ethical action.

Follow-up analyses. The original study observed no difference by condition in ratings of non-cleansing products. In a 2 (condition) x 2 (cleansing, non-cleansing) linear mixed effects model with subjects as a random effect of the replication data, there was no interaction of copying an ethical or unethical passage on desirability of cleansing or non-cleansing products ($t(6{,}999) = -0.57$, $p = 0.57$, $d = -0.01$, 95% CI [-0.06, 0.03]). Moreover, there was no difference between the ethical ($M = 3.12$, $SD = 1.08$) and unethical ($M = 3.11$, $SD = 1.05$) conditions in ratings of non-cleansing products ($t(6{,}999) = 0.63$, $p = 0.53$, $d = 0.02$, 95% CI [-0.03, 0.06]).

24. Assimilation and contrast: ASSIMILATION AND CONTRAST EFFECTS IN PART-WHOLE QUESTION SEQUENCES: A CONVERSATIONAL LOGIC ANALYSIS (Schwarz, Strack, & Mai, 1991, Study 1)

One hundred participants answered a question about life satisfaction in a specific domain "How satisfied are you with your relationship?" and a question about life satisfaction in general "How satisfied are you with your life-as-a-whole?" Participants were randomly assigned to the order of answering the specific and general questions. When the specific question was asked first, the correlation between the responses to the two questions was strong ($r = 0.67$, $p < 0.05$).

When the specific question was asked second, the correlation between them was weaker (r = 0.32, $p < 0.05$). The difference between these correlations was significant, $z = 2.32$, $p < 0.01$, q = 0.48, 95% CI [0.07, 0.88].

The authors suggest that the specific-first condition makes the relationship more accessible such that participants are more likely to incorporate information about their relationship when evaluating a more general question about their life satisfaction. Because responses to the two items are linked by the accessibility of relationship information, they should be correlated. In contrast, in the specific-second condition, relationship satisfaction is not necessarily accessible and participants may draw on any number of different areas to generate their overall life satisfaction response. Thus, the correlation between the items is weaker than in the specific-first condition.

In the aggregate replication sample (N = 7,460), when the specific question was asked first, the correlation between the responses to the two questions was moderate (r = 0.38). When the specific question was asked second, the correlation between them was slightly stronger (r = 0.44). The difference between these correlations was significant, $z = -3.03$, $p = 0.002$, $q = -0.07$, 95% CI [-0.12, -0.02]. The replication effect size was much smaller and in the opposite direction of the original result ($q = -0.07$, 95% CI [-0.12, -0.02] versus 0.48).

Follow-up analysis. In the original procedure, no other measures preceded the questions. The effect is about the influence of question context, so it is reasonable to presume that task order will have an impact on the estimated effect. As such, the most direct comparison with the original is for the conditions in which this task is administered first. In that subsample (N = 470), when the specific question was asked first, the correlation between the responses to the two questions was strong (r = 0.41). When the specific question was asked second, the

correlation between them was the same (r = 0.41). The difference between these correlations was not significant, z = 0.01, p = 0.99, q = 0.00, 95% CI [-0.18, 0.18].

25. Choosing versus rejecting: WHY SOME OPTIONS ARE BOTH BETTER AND WORSE THAN OTHERS (Shafir, 1993, Study 1)

One hundred and seventy participants imagined that they were on the jury of a custody case and had to choose between two parents. One of the parents had both more strongly positive and more strongly negative characteristics (extreme) than the other parent (average). Participants were randomly assigned to either decide to *award* custody to one parent or to *deny* custody to one parent. Participants were more likely to both award (64%) and deny (55%) custody to the extreme parent than the average parent, the sum of probabilities being significantly greater than 100% (z = 2.48, p = 0.013, d = 0.35, 95% CI [-0.04, 0.68]). This finding was consistent with the hypothesis that negative features are weighted more strongly when people are rejecting options, and positive features are weighted more strongly when people are selecting options (Shafir, 1993).

In the aggregate replication sample (N = 7,901), participants were less likely to both award (45.5%) and deny (47.6%) custody to the extreme parent than the average parent, and the sum of probabilities (93%) was significantly smaller than the 100% we would expect if choosing and rejecting were complementary (z = -6.10 p = 1.1e-9, d = -0.13, 95% CI [-0.18, -0.09]). This result was slightly in the opposite direction of the original finding and it is incompatible with the hypothesis that negative features are weighted more strongly when rejecting options and positive features are weighted more strongly when selecting options.

26. Priming heat affects climate beliefs: HOW WARM DAYS INCREASE BELIEF IN GLOBAL WARMING (Zaval, Keenan, Johnson, & Weber, 2014, Study 3A)

Zaval et al. (2014) investigated how beliefs in climate change could be influenced by immediately available information about temperature. In Study 3A, 300 Mechanical Turk workers reported their beliefs about global warming after completing one of three scrambled sentence tasks in which there was a theme of words priming the concepts heat, cold, or a no theme control condition. There was a significant effect of condition on both global warming belief, $F(2, 288) = 3.88$, $p = 0.02$, and concern, $F(2, 288) = 4.74$, $p = 0.01$, controlling for demographic and actual temperature data. Post hoc pairwise comparisons revealed that participants in the heat-priming condition expressed stronger belief ($M = 2.7$, $SD = 1.1$) in global warming than participants in the cold-priming ($M = 2.4$, $SD = 1.1$; $t(191) = 1.9$, $p = 0.06$, $d = 0.27$, 95% CI [0.05, 0.49]) or control conditions ($M = 2.3$, $SD = 1.1$; $t(193) = 2.23$, $p = 0.03$, $d = 0.37$, 95% CI [0.14, 0.59]). Likewise, participants in the heat-priming condition expressed greater concern ($M = 2.4$, $SD = 1.0$) about global warming than participants in the cold-priming ($M = 2.1$, $SD = 1.0$; $t(191) = 2.15$, $p = 0.03$, $d = 0.31$, 95% CI [0.03, 0.59]) or control conditions ($M = 2.1$, $SD = 1.0$; $t(193) = 2.23$, $p = 0.02$, $d = 0.31$, 95% CI [0.02, 0.59]).

For direct replication, mean differences in concern about global warming between heat and cold-priming conditions were evaluated. In the aggregate replication sample after excluding participants that made errors in the sentence unscrambling (remaining $N = 4{,}204$), participants in the heat-priming condition ($M = 2.47$, $SD = 0.90$) expressed similar concern about global warming compared to participants in the cold-priming condition ($M = 2.50$, $SD = 0.89$; $t(4{,}202) = -1.09$, $p = 0.27$, $d = -0.03$, 95% CI [-0.09, 0.03]). This result is not consistent with the hypothesis that temperature priming alters concern about global warming. The effect size was much weaker and slightly in the opposite direction compared to the original finding ($d = -0.03$, 95% CI [-0.09, 0.03] versus original $d = 0.31$).

Translations of the scrambled sentence manipulation may have disrupted the effectiveness of the manipulation. As such, the most direct comparison with the original effect size is with the English-only test administrations. With this subsample (N = 2,939), participants in the heat-priming condition (M = 2.40, SD = 0.90) also expressed similar concern about global warming compared to participants in the cold-priming (M = 2.44, SD = 0.89; t(2,937) = -0.18, p = 0.24, d = -0.04, 95% CI [-0.12, 0.03]).

Follow-up analyses. Belief in global warming was included as a secondary dependent variable. In the aggregate replication sample (N = 4,212), participants in the heat-priming condition (M = 3.25, SD = 0.84) expressed similar belief about global warming than participants in the cold-priming condition (M = 3.25, SD = 0.82; t(4,210) = 0.50, p = 0.62, d = 0.00, 95% CI [-0.06, 0.06]). With the subsample of English test administrations, participants in the heat-priming condition expressed similar belief (M = 3.25, SD = 0.86) in global warming as participants in the cold-priming (M = 3.23, SD = 0.85; t(2,940) = 1.40, p = 0.16, d = 0.02, 95% CI [-0.05, 0.09]). Neither of these follow-up analyses were consistent with the original study indicating an effect of temperature priming on beliefs in global warming.

27. Intentionality for side effects: INTENTIONAL ACTION AND SIDE EFFECTS IN ORDINARY LANGUAGE (Knobe, 2003, Study 1)

Knobe (2003) investigated whether helpful or harmful side effects were differently perceived to be intentional. Consider, for example, an agent who knows that their behavior will have a particular side effect, but does not care whether the side effect does or does not occur. If the agent chooses to go ahead with the behavior and the side effect occurs, do people believe that the agent brought about the side effect *intentionally*? Knobe (2003) had participants read vignettes about such situations and found that participants were more likely to believe the agent

brought about the side effect intentionally when the side effect was harmful compared to when it was helpful. 82% of participants in the harm condition said that the agent brought about the side-effect intentionally, whereas 23% in the help condition said that the agent brought about the side-effect intentionally ($X^2(1, N = 78) = 27.2$, $p < 0.001$, $d = 1.45$, 95% CI [0.79, 2.77]). Agents who brought about harmful side effects were also rated as being more blameworthy than agents who brought about helpful side effects were rated as being praiseworthy $t(120) = 8.4$, $p < 0.001$, $d = 1.55$, 95% CI [1.14, 1.95]. The total amount of blame or praise attributed to the agent was associated with believing the agent brought about the side effect intentionally $r(120) = 0.53$, $p < 0.001$, $d = 0.63$, 95% CI [0.26, 0.99].

Ratings of intentionality in the harm and help conditions were compared for the direct replication using a 7-point scale rather than a dichotomous judgment. In the aggregate replication sample ($N = 7{,}982$), participants in the harm condition ($M = 5.34$, $SD = 1.94$) said that the agent brought about the side-effect intentionally to a greater extent than did participants in the help condition ($M = 2.17$, $SD = 1.69$; $t(7{,}843.86) = 78.11$, $p < 1.68\text{e-}305$, $d = 1.75$, 95% CI [1.70, 1.80]). This is consistent with the original result with a somewhat stronger effect in the replication ($d = 1.75$, 95% CI [1.70, 1.80] versus original $d = 1.45$).

Follow-up analyses. Blame and praise ratings were assessed as a secondary replication. Agents who brought about harmful side effects were rated as being more blameworthy ($M = 6.03$, $SD = 1.26$) than agents who brought about helpful side effects were rated as being praiseworthy ($M = 2.54$, $SD = 1.60$; $t(7{,}553.82) = 108.15$, $p < 1.68\text{e-}305$, $d = 2.42$, 95% CI [2.36, 2.48]). This is also consistent with the original result with a notably larger effect size (2.42 versus 1.55).

28. Directionality and similarity: STUDIES OF SIMILARITY (Tversky & Gati, 1978,

Study 2)

Tversky and Gati (1978) investigated the relationship between directionality and similarity. 144 participants made 21 similarity ratings of country pairs in which one country (e.g., U.S.A.) was pre-tested as more prominent than the other (e.g., Mexico). The pair was presented with either the more prominent country first (U.S.A.-Mexico) or the less prominent country first (Mexico-U.S.A.). Two versions of the survey with 21 pairs were created that presented the more prominent country first "about an equal number of times", with the same pair of countries being manipulated between-subjects. Results indicated that participant similarity ratings were higher when less prominent countries were displayed first compared to when more prominent countries were displayed first, $t(153) = 2.99$, $p = 0.003$, $d = 0.48$, 95% CI [0.16, 0.80], and that higher similarity ratings were given to the version of each pair that listed the more prominent country second, $t(20) = 2.92$, $p = 0.001$, $d = 0.64$, 95% CI [0.16, 1.10].

A follow-up study ($N = 46$) with the same design examined ratings of differences rather than similarities. Following the prior result, participant difference ratings were higher when the more prominent countries were displayed first compared to the less prominent countries displayed first, $t(45) = 2.24$, $p < 0.05$, $d = 0.66$, 95% CI [0.06, 1.25] and higher difference ratings were given to the version of each pair that listed the more prominent country first, $t(20) = 2.72$, $p < 0.01$, $d = 0.59$, 95% CI [0.12, 1.05].

For replication, participants were randomly assigned to one of the two counterbalancing conditions as described above, and were randomly assigned to rate either similarities or differences between the two countries. Following the original study, we considered the similarity and difference judgements as two independent samples. Therefore, each site has about half as much data for its critical test as other effects. The similarity ratings were the primary test for

direct replication, difference ratings were a secondary analysis.

On the aggregate similarities replication sample (N = 3,549), we created an asymmetry score for each subject, calculated as the average similarity for comparisons where the prominent country appeared second minus the average for the comparisons where the prominent country appeared first. Across participants, the asymmetry score was not different from zero (t(3,548) = 0.60, p = 0.55, d = 0.01, 95% CI [-0.02, 0.04], meaning that the order of presentation of prominent countries did not influence their evaluations of similarity. Distinct from the critical test, we observed that the average similarity ratings of one counterbalancing condition were M_{1a}= 8.78 (SD_{1a} = 2.44) and M_{2b}= 8.84 (SD_{2b} = 2.43) when the more prominent country was presented first and second, respectively, whereas the average similarity rating of the other counterbalancing condition was higher M_{1b} = 10.14 (SD_{1b} = 2.42) and M_{2a} = 10.09 (SD_{2a} = 2.44). In summary, there was no evidence of the key effect of country order (prominent first vs prominent second) and similarity ratings were different between counterbalancing conditions, a procedural effect.

Then, we reproduced the original by-item analysis. Participants similarity ratings were nearly identical when the less prominent country was displayed first (M = 9.42, SD = 2.61) compared to the more prominent country displayed first, M = 9.43, SD = 2.57; t(20) = -0.29, p = 0.78, d = -0.04, 95% CI [-0.35, 0.26]. Overall, the replication results were near zero, or slightly in the opposite direction of the original findings.

Follow-up analyses. We reproduced the original analyses on the differences data (N = 3,582). The asymmetry score for each subject was not different from zero (t(3,581) = 1.70, p = 0.09, d = .03, 95% CI [-0.004, 0.061], meaning that the order of presentation of prominent countries did not influence their evaluations of similarity.

The by-item analysis showed that participant difference ratings were very similar when the more prominent country was displayed first (M = 11.19, SD = 2.54) compared to the less prominent country displayed first, M = 11.25, SD = 2.54; $t(20) = 1.1$, $p = 0.29$, $d = 0.17$, 95% CI [-0.14, 0.47].

Order effects in general are reported in the next section. For this effect in particular, we identified *a priori* that it and Norenzayan et al. (2002) both involved similarity judgments and thus order of these may be particularly relevant. We compared whether the asymmetry score was moderated by Norenzayan et al. appearing before or after, and observed no moderation for the primary similarities test ($t(3,547) = -0.48$, $p = 0.63$, $d = -0.02$, 95% CI [-0.08, 0.05]) and for the secondary differences test ($t(3,580) = -0.23$, $p = 0.82$, $d = -0.01$, 95% CI [-0.07, 0.06]) .

Results

Table 2 presents the original study effect size, median effect size of replication studies, weighted means of replication effect sizes with 95% confidence intervals after pooling data of all samples, and proportion of samples that rejected the null hypothesis in the expected direction, rejected the null hypothesis in the unexpected direction, or did not reject the null hypothesis. Effects are ordered from the largest global replication effect size consistent with the original study first to the smallest or opposite direction effects. Importantly, we separated those studies that had shown cultural differences in original research into two rows to avoid aggregating results when effects might be anticipated in one sample and not another. However, the differences observed between samples in the original research may not be expected to replicate between our aggregate comparisons across many cultural contexts. As such, we avoid drawing conclusions about replication or not of original cultural differences beyond what is discussed in the individual finding reports and in our aggregate observation of variability across samples.

Overall, 14 of the 28 effects (50%) showed significant evidence in the same direction as the original finding, 1 was weakly consistent (4%),[8] and 13 (46%) showed a null effect or evidence in the opposite direction of the original finding.[9] Larger aggregate effects tended to have a higher proportion of significant positive results than smaller aggregate effects, as would be expected based on power of the individual samples to detect the observed aggregate effect size. Eight effects had almost 90% to 100% significant effects in the individual samples, and six findings that were positive results in the aggregate had between 11% and 46% significant effects in the individual samples. As would be expected, effects that were null in the aggregate also tended to have more than 90% of the individual samples showing null effects with occasional significant results both in the same and opposite direction as the original findings. Most observed pooled effect sizes (21 of 28; 75%) were smaller than the original WEIRD findings, but some (7 of 28; 25%) were larger.

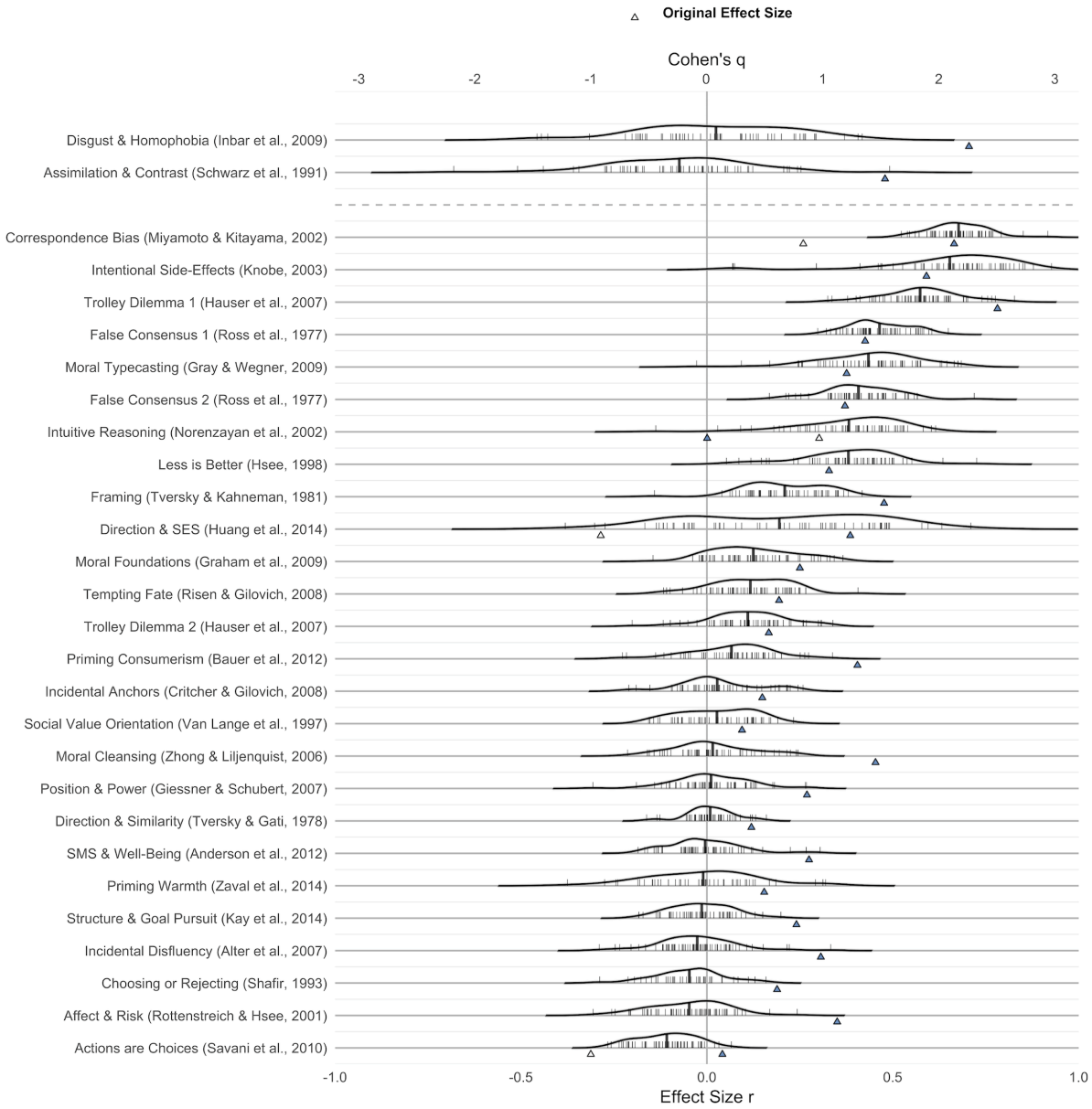
Figure 2 provides summary illustration of the 28 studies including: (1) aggregate effect size estimate in thick vertical lines, (2) the effect size estimates for each individual sample using thin vertical lines, and (3) the original study's effect size estimates as triangles with samples from WEIRD cultures identified with grey triangles and samples from less WEIRD cultures identified with white triangles (4 original studies had samples from two cultures). An alternate

[8] Inbar et al. (2009) was categorized as "weakly consistent". The key correlation comparison had a *p*-value of .02 in the same direction as the original study, and a mean difference in perceived intentionality between the experimental conditions was not replicated ($p = .457$). In the original study, the mean difference was accounted for by the difference in correlations.

[9] For the four original studies that used two samples to make cultural comparisons, we defined the positive direction using the effect size observed in the original sample that was more western, educated, industrialized, rich, and democratic.included WEIRD and less WEIRD samples, "consistent" is defined as the WEIRD samples in the replications showing a significant effect in the same direction as the WEIRD sample in the original study.

Figure S2 showing separate distributions for WEIRD and less WEIRD samples is available in supplementary materials.

Figure 2. Effect size distributions for all samples for 28 effects



Note: Effect size for each sample plotted as a short vertical line; aggregate estimate as longer, thick vertical line. Samples with less than 15 participants due to exclusions are not plotted, and

some samples were excluded because of errors in administration. A detailed accounting of all exclusions is available at https://manylabsopenscience.github.io/ML2_data_cleaning. Positive effect sizes indicate effects consistent with the direction of the original finding in the original western sample. Original effect sizes appear as grey-filled triangles. If the original study had a cultural comparison, the non-western sample appears as a second clear triangle. For Inbar et al. (2009) and Schwartz et al. (1991), values represent a Cohen's *q* estimate of the difference between 2 correlations.

Variation Across Samples and Settings

Our central interest was the variation in effect estimates across all samples and settings. In a linear mixed model with samples and studies as random effects, we compared the intra-class correlation of samples across effects (ICC = 0.782) which was quite large, with the intra-class correlation of effects across samples (ICC = 0.004) which was near zero. In other words, to predict effect sizes across the 28 findings and dozens of samples studied here, it is very useful to know the effect being studied and barely useful to know the sample in which it is being studied.

Next, we examined whether *specific* effects are sensitive to variation in sample or setting. For each of the 28 replication studies, we examined variability in effect sizes using a random effects meta-analysis (with restricted maximum likelihood as estimator for the between-study variance) and established heterogeneity estimates - Tau, Q and I^2 - to determine if the amount of variability across samples exceeds that expected by measurement error. Because the study procedures are nearly identical (except for language translations), variation exceeding measurement error is likely to be due to effects of sample or setting, and interactions between samples and the materials. Eleven of the 28 effects (39%) showed significant heterogeneity with the Q-test ($p < .001$). Notably, of those showing such variability, 8 were among the 10 largest effect sizes. Only one of the non-significant replication effects showed significant heterogeneity

using Q (Van Lange et al., 1997). The I^2 statistic indicated substantial heterogeneity for some of the tests, with 10 (36%) showing at least medium heterogeneity ($I^2 \geqslant 50\%$), and two showing heterogeneity larger than 75% (Huang et al., 2014 and Knobe, 2003; see Table 3). Note, however, that estimation of heterogeneity is rather imprecise, as evidenced by many large confidence intervals of I^2, particularly for the cases with low estimates of heterogeneity. Ten of the 14 smaller I^2 effects had a lower bound of 0. Also, the I^2 statistic increases if sample size increases, so the large samples may be an explanation for the large I^2 statistics that were observed (Rücker, Schwarzer, Carpenter, & Schumacher, 2008). As in the first Many Labs project (Klein et al., 2014), heterogeneity was more likely to occur for large effects than small effects. The Spearman rank-order correlation between aggregate effect size of the pooled analysis and I^2 values is $r = 0.56$.

Finally, with Tau, only one effect (Huang et al., 2014) showed a substantial standard deviation among effect sizes (0.24), and 8 others showed modest heterogeneity near 0.10. Most of the effects, 19 of 28 (68%) showed near zero heterogeneity as estimated by Tau. Overall, this indicates that many effects showed minimal heterogeneity and, when it was observed, it was quite small.

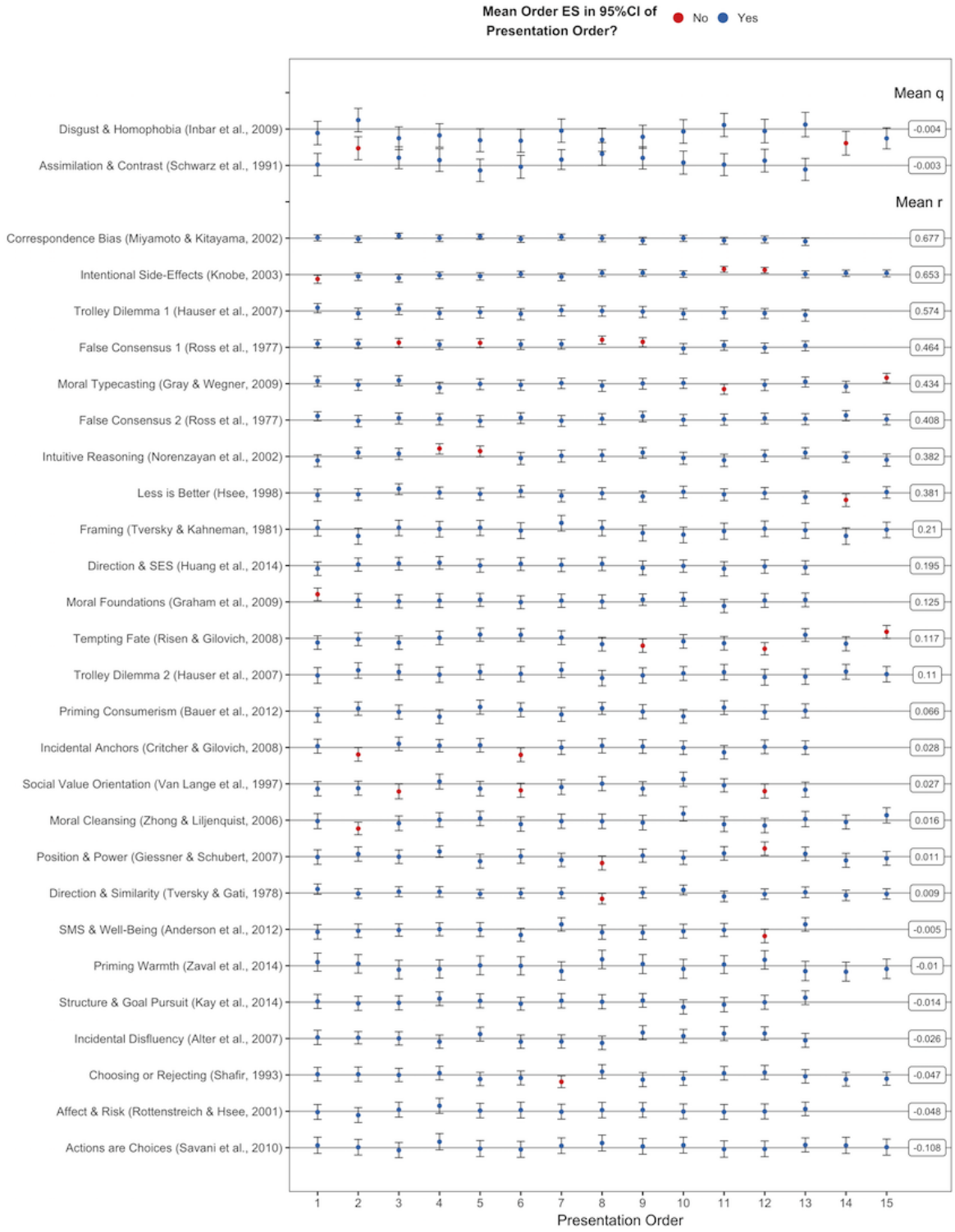
Table 3 summarizes moderation tests between lab and online samples. Just one result showed a significant difference between lab and online samples (Zhong & Liljenquist, 2006). For this finding, the overall result was not different from zero and approximately 95% of their individual samples showed null effects. This suggests some caution in concluding evidence for moderation by lab versus online data collection.

Exploratory analysis. For exploratory cultural comparisons, we computed a WEIRDness (Henrich, Heine, & Norenzayan, 2010) score for each sample based on their country

of origin using public country rankings. Western and Developed countries were given a score of 1 whereas Eastern and Emerging countries were given a score of 0. The list of developed countries, the scores at the Education Index and the scores at the Industrial Development Report were obtained from the United Nations official website. Democratization scores were obtained from the Global Democracy Ranking. We then computed a global WEIRDness score taking the mean across scores at the sample level. Details on the computation and specific links to the country rankings are available here: https://osf.io/b7qrt/. Samples were then categorized as WEIRD (Western, Educated, Industrialized, Rich, Democratic societies; Slate 1 n = 42, Slate 2 n = 44) and less WEIRD (Slate 1 n = 22, Slate 2 n = 17) based on whether their country WEIRDness score was higher or lower than the observed WEIRDness mean score across samples (see Figure 4).

Table 3 also presents heterogeneity statistics for comparing WEIRD and less WEIRD cultures. For 13 of the 14 effects that were reliable and in the same direction as the original study, the finding was observed for both WEIRD and less WEIRD samples with similar effect magnitudes. The only exception was Huang et al. (2014) for which less WEIRD samples showed no overall effect and wide variability across samples. This is relatively consistent with the original study in which Hong Kong participants showed an effect in the opposite direction as U.S. participants presumably because observed differences in whether wealth people tended to live in the North or South between the samples. It is likely that there is wide variability in whether wealthy people tend to live in the North or South of the many different settings within WEIRD and less WEIRD samples of the present study producing this high observed variability. Among the 14 effects that were near null in the aggregate, there was little evidence for the

Figure 3. Effect estimates for task order positions compared to the mean effect size for each of the 28 effects.



original finding in either WEIRD or less WEIRD samples. However, for Savani et al., both WEIRD and less WEIRD samples were in the direction of the original less WEIRD sample.

Ultimately, just three effects (Huang et al., 2014, Knobe, 2003, and Norenzayan et al., 2002) showed significant evidence for moderation by WEIRDness after correcting for multiple comparisons. However, for Norenzayan et al., the cultural difference was the inverse of the original result though their original study did not have a theoretical commitment regarding the cultural differences in the condition effect that we tested. Norenzayan et al. focused on rule-based responses across conditions, and predictions that their European American sample would show greater rule-based responses than the East Asian sample within each condition (see Footnote 7).

Influence of task order

The order of presentation could moderate effect sizes. Across the 30 minute session, effects may weaken if participants tire or if earlier procedures interfere with later procedures. We did not observe this in prior Many Labs investigations with the same design (Ebersole et al., 2016; Klein et al., 2014), but it remains a potential moderator. Order of administration was randomized, enabling a direct test of this possibility. Figure 3 shows each effect size in rank order across all locations from 1 (presented first) to 13 or 15 (presented last in its slate). Table 4 shows the aggregate effect size, effect size when the study was administered first, and effect size when the study was administered last. Across the 28 findings, we observed little systematic evidence that effects are stronger (or weaker) when administered first compared to last. Also, there was no evidence of linear, quadratic or cubic trends by task order (see supplements for analytic details: https://osf.io/z8dqs/). Considering all task positions for all 28 findings, the mean effect size fell outside of the 95% confidence interval for 29 of the 394 finding-position

estimates (7.4%) suggesting that there may be some order effects. However, the distribution of those unusual results appears to be relatively random across findings and positions (Figure 3).

Authors of four of the original articles (Alter et al., 2007; Giessner & Schubert, 2007; Miyamoto & Kitayama, 2002; Schwarz et al., 1991) noted *a priori* that their findings may be sensitive to order of administration. However, none of these showed evidence for systematic variation in effect magnitudes by task order. It is still possible that there are specific order effect influences, such as when a particular procedure immediately precedes another particular procedure; but these analyses confirm that the findings, in the aggregate, are robust to task order and, particularly, that task order cannot account for observation of null effects for any of the non-replicated results.

Discussion

We conducted preregistered replications of 28 published results with protocols that were peer reviewed in advance with data from about 125 samples, including thousands of participants from locations around the world. Using conventional statistical significance ($p < 0.05$), fifteen (54%) of the replications provided evidence in the same direction and statistically significant consistent with the original finding. With a strict significance criterion ($p < 0.0001$), fourteen (50%) provide such evidence reflecting the extremely high powered design (for Inbar et al., 2009 the replication *p*-value was 0.02). Seven (25%) of the replications had effect sizes (Cohen's *d* or *q*) larger than the original finding and 21 (75%) had effect sizes smaller than the original finding. In WEIRD samples, the median Cohen's *d* effect size for original findings was 0.60 and for replications was 0.15 indicating a substantial decline (Open Science Collaboration, 2015).[10] Sixteen replications (57%) had small effect sizes (< .20) and 9 (32%) were in the opposite

[10] These medians exclude the two studies that used Cohen's *q* for effect size estimates. Including those, despite the different scaling of *d* and *q*, yields similar medians of 0.60 and 0.09 respectively.

direction from the original finding. Three of these had an aggregate replication effect size that was *significantly* in the opposite direction (Rottenstreich & Hsee, 2001; Schwarz, et al., 1991; Shafir, 1993) at $p < 0.05$ but only one at $p < 0.0001$ (Shafir, 1993).

There is no simple decision rule for declaring success or failure in replication or for detecting positive results (Benjamin et al., 2018; Camerer et al., 2018; Open Science Collaboration, 2015). In Table 5, we show a variety of possible decision criteria to decide whether the observed global effect size successfully replicated the original finding. Two used the replication sample size either with a loose criterion of $p < .05$ (54% success rate) or a strict criterion of $p < .0001$ (50% success rate). The others consider what the *p*-value would have been for the observed effect size if it had been obtained with the original study sample size (41% success rate), 2.5x the original study sample size (Simonsohn, 2015; 44% success rate), or (c) 50 participants per group -- a reasonably large sample compared to historical trends (Fraley & Vazire, 2014; 33% success rate). Nine of the effects (32%) were successful replications across all criteria and 13 (46%) were unsuccessful replications across all criteria.[11] Six findings (21%) varied in replication success depending on the criteria usually because the replication effect size was substantially smaller than the original effect size. The final column in Table 5 provides the sample size needed to detect the original finding with observed global effect size of the replications when alpha = 0.05 and power is 0.80. Findings that were highly replicable across all criteria were relatively large effect sizes and relatively efficient to investigate with modest samples (*N*'s 12 to 54 and one 200). Replicable findings that had somewhat weaker effect sizes in general or compared to the original study need more substantial sample sizes to study efficiently (*N*'s 200 to 2,184). Findings that were in the same direction as the original study but

[11] Replication success could not be computed for three criteria for one finding because of the test used (Savani et al., 2010) and for the 50 participants per group criterion for four others because of the test used. For simplicity, we considered only computed tests for this summary.

too weak to reject the null-hypothesis of no effect with our large samples would need massive samples to reject the null-hypothesis (*N*'s 6,283 to 313,958). Finally, null-hypothesis of the 10 findings that had effect sizes of 0 or in the opposite direction of the original cannot be rejected not matter what sample size is used.

The high proportion of failures to replicate with extremely large samples and weaker effect sizes compared to original studies is consistent with the accumulating evidence in systematic replication studies (Camerer et al., 2016, 2018; Ebersole et al., 2016; Klein et al., 2014a; Open Science Collaboration, 2015). We cannot identify whether these are due to errors in replication design, *p*-hacking in original studies, or publication bias with selecting for positive results despite pervasive low-powered research. However, it is notable that surveys and prediction markets with researchers predicting and betting on whether these studies would replicate were effective at predicting replication success. For example, the correlation between market price and replication success for Many Labs 2 studies was 0.755. These results are reported in a separate paper (Dreber et al., 2018), and replicate other studies using prediction markets and surveys to predict replication success (Camerer et al., 2016, 2018; Dreber et al., 2016). In any case, these findings provide further justification for improving transparency of research (Miguel et al., 2014; Nosek et al., 2015), and preregistering studies to make all findings discoverable even if they are not published and preregistering analysis plans to make clear the distinction between confirmatory tests and exploratory discoveries for improving statistical inference (Nosek et al., 2018; Wagenmakers et al., 2012).

The main purpose of the investigation was to assess variability in effect sizes by sample and setting. It is reasonable to expect that many psychological phenomena are moderated by variation in sample, setting, or procedural details, and that this may impact reproducibility

(Henrich et al., 2010; Klein et al., 2014a, 2014b; Markus & Kitayama, 1991; Schwarz & Strack, 2014; van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). However, while calculations of intra-class correlations showed very strong relations of effect sizes across the findings (ICC = 0.782), they showed near zero relations of effect sizes across samples (ICC = 0.004). Sensibly, knowing the effect being studied provides a lot more information on effect size than knowing the sample being studied. Just 11 of the 28 effects (39%) showed significant heterogeneity with the Q-test, and most of these were among the effects with the largest overall effect size. Only one of the near zero replication effect (Van Lange et al., 1997) showed significant heterogeneity with the Q-test. In other words, if no effect was observed overall, there was also very little evidence for heterogeneity among samples.

Using the I^2 statistic, for approximately one third of all effects being studied (36%) at least medium heterogeneity across samples was found, but almost all evaluations of heterogeneity had high uncertainty (i.e., wide confidence intervals). Taken at face value, the I^2 statistics in Table 3 indicate that heterogeneity in samples is high for some of the findings, even when there is little evidence for an effect. For example, for Zaval et al. (2014), the main effect was not distinguishable from zero and 89% of the individual samples showed non-significant effects, close to expectation of samples drawn from a null distribution, and yet, the I^2 is 37%. However, an average effect size of 0 together with a majority of studies with null results can co-exist with strong heterogeneity, as measured with I^2 (https://osf.io/frbuv/). I^2 compares variability in the dependent variable across studies with variability within studies. With increasing power of primary studies, I^2 will tend toward 100% if there is any evidence for heterogeneity in the sample no matter how small the effect. As such, these estimates likely reflect the extremely large sample sizes rather than the amount of heterogeneity in absolute terms.

By comparison, the estimates for Tau in Table 3 indicate a small standard deviation in effect sizes for all studies except one (Tau = 0.24; Huang et al., 2014). In fact, 19 of the 28 (68%) had an estimated Tau near 0 indicating minimal heterogeneity and 8 (29%) had an estimated Tau near .10 indicating a small amount of heterogeneity. This illustrates the key finding for this study. For some effects heterogeneity across samples is near zero. It is not so surprising that this is the case for effects that failed to replicate in general, but it was also occasionally observed for successful replications. More importantly, even among successful replications, when heterogeneity was observed, it was relatively weak. As a consequence, at least for the variation investigated here, heterogeneity across samples does not provide much explanatory power for failures to replicate.

Estimates of average effect size and effect size heterogeneity may have been affected by imperfect reliabilities of instruments measuring the outcome variables. For instance, Hunter and Schmidt (1990) show how imperfect reliabilities attenuate effect size estimates and suggest correcting for these imperfections when estimating effect size. As both original and replication studies did not correct for these imperfect reliabilities, systematic differences in effect size estimates between original and replication studies cannot be explained by imperfect reliabilities, unless the measurement instruments were systematically much less reliable in the replication than in the original studies; we have no evidence that this is the case. Differences across labs in reliabilities of measurement instruments may also result in overestimation of effect size heterogeneity in case of a true non-zero effect size. Insofar as these differences exist, our results likely overestimate heterogeneity as our analyses do not take imperfect reliabilities of variables into account.

For 12 of the 28 findings, moderators or sample subsets that may be necessary to observe the effect were identified *a priori* by original authors or other experts during the Registered Report review process. These effect-specific analyses were reported with the individual effects. For 7 of those 12, the pooled result was null or in the opposite direction of the original; for the other 5, the pooled results showed evidence for the original finding. Among the 12, just one (8% of the total; Hauser et al., 2007, Trolley Dilemma 1) showed evidence consistent with the hypothesized moderator/subset, and two (17%) showed weak or partial evidence (Miyamoto & Kitayama, 2002; Risen & Gilovich, 2008). The other nine (75%) showed little evidence that narrowing the datasets to the samples and settings deemed most relevant to testing the hypothesis had impact on the likelihood of observing the effects or their effect magnitude. This does not mean that moderating effects do not occur, but it may mean that psychological theory is not yet advanced enough to predict them reliably.

Another possible moderating influence, unique to the present design, was task order. Participants completed their slate of 13 to 15 effects in a randomized order. It is possible that tasks completed later in the sequence would be influenced by tasks completed earlier in the sequence, either because of the specific content of the task, or because of interference, fatigue, or other order-related influences (Ferguson, Carter, & Hassin, 2014; Kahneman, 2016; Schnall, 2014). Contrary to this prediction, we observed little evidence for systematic order effects for the 28 findings investigated. This replicates the lack of evidence for task order effects observed in Many Labs 1 (Klein et al., 2014) and Many Labs 3 (Ebersole et al., 2016). Across 51 total replication tests (28 reported here; 13 in Klein et al., 2014, and 10 in Ebersole et al., 2016) we have observed little evidence for reliable effects of task order. The idea that completing a study

first, in the middle, or at the end of a sequence has an impact on the magnitude of the observed effect is appealing and, so far, unsupported.

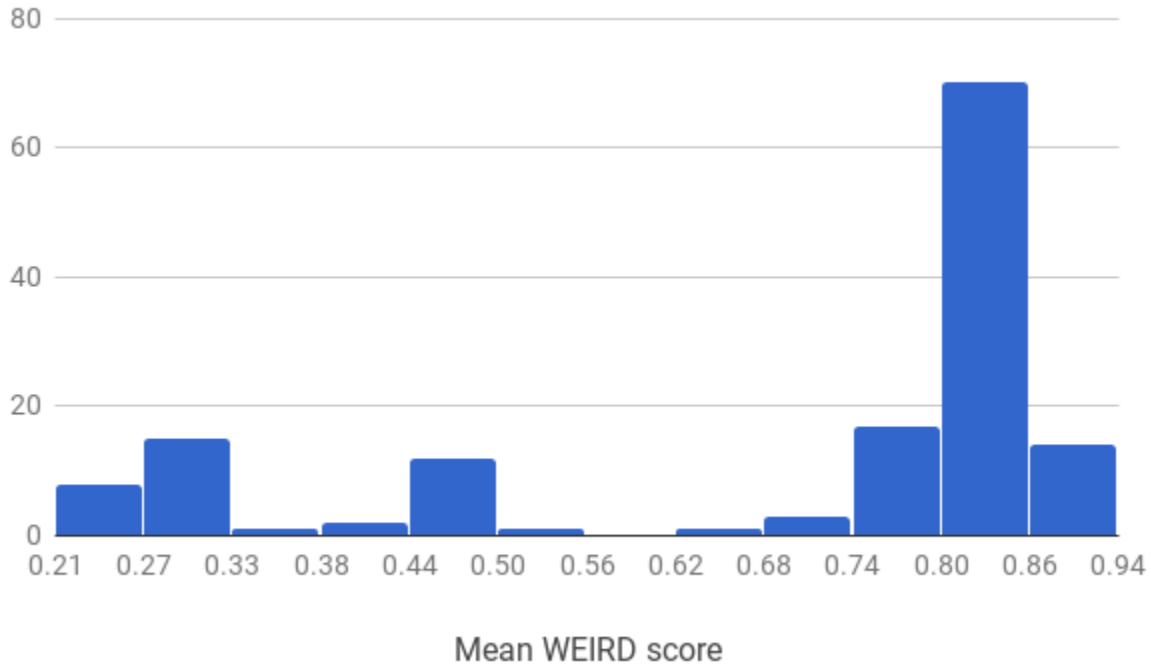
The same is true for effects of administration in lab versus on-line. Since the Internet became a source for behavioral research, there has been interest in the degree to which lab and on-line results are consistent with one another (Birnbaum, 2004; Dandurand, Shultz, & Onishi, 2008; Hilbig, 2016). As with task order, across Many Labs projects we have observed little evidence for an effect of mode of administration. There may be conditions under which lab versus online administration is consequential, but we did not observe meaningful evidence for its impact.

Finally, we included an exploratory analysis of the moderating influence of WEIRD versus less WEIRD cultural differences. We sampled from 125 highly heterogeneous sources (64 Slate 1, 61 Slate 2) to maximize the possibility of observing variation in effects based on sample characteristics--39 U.S. samples and 86 samples from 35 other countries and territories. Ultimately, just three effects (Huang et al., 2014; Knobe, 2003; Norenzayan et al., 2002) showed compelling evidence for differences between our WEIRD and less WEIRD samples.

However, our approach characterized cultural differences at the most general level possible--a dichotomy of WEIRDness--and ignored the rich diversity within that categorization. The distribution of WEIRD scores is such that the WEIRD samples are highly similar in WEIRDness, and the less WEIRD samples vary substantially in WEIRDness. Figure 4 illustrates the highly skewed distribution. Scores > 0.70 were categorized as WEIRD, the rest were less WEIRD. Our summary analyses also do not address the possibility of highly specific regional variations such as differences between U.S. and British samples, nor did they examine why differences were observed. Nor do these analyses investigate many interesting sampling

moderators available in this dataset such as individual differences, gender, and ethnicity. Some moderating influences could be evaluated using the present dataset; others will require new data collections to test. Also, a true examination of WEIRDness would need to more deliberately vary sampling across all dimensions -- Western, Educated, Industrialized, Rich, and Democratic. Further analyses of the present dataset may inspire hypotheses to test in future studies.

Figure 4. Histogram of Mean WEIRD Score by sample



Implications

It is practically a truism that human behavior is contingent on the cultural and personal characteristics of the participants under study and the setting in which they are studied. The depth with which this idea is embedded in present psychological theorizing is illustrated by the appeals to “hidden moderators” as counterclaims to failures to replicate without empirically

testing whether such moderators are operative (Baumeister & Vohs, 2016; Crisp, Miles, & Husnu, 2014; Gilbert et al., 2016; Ramscar, Shaoul, & Baayen, 2015; Schwarz & Clore, 2016; Stroebe & Strack, 2014; van Bavel et al., 2016). The present study suggests that dismissing failures to replicate as a consequence of such moderators without conducting independent tests of the hypothesized moderators is unwise. Collectively, we observed some evidence for effect-specific heterogeneity, particularly for larger effects, occasional evidence for cultural variation, and little evidence for procedural factors such as task order and lab versus online administration.

There have been a variety of failures to replicate effects that were quite large in the original investigation (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012; Hawkins, Fitzgerald, & Nosek, 2015; Johnson, Cheung, & Donnellan, 2014; Hagger et al., 2016). If effects are highly contingent on the sample and setting then they could be large and easily detected with some samples and negligible with other samples. We did not observe this. Rather, evidence for moderation or heterogeneity was mostly observed in the large, consistently detectable effects.

Further, we observed some heterogeneity between samples, but *a priori* predictions (e.g., original authors predictions of moderating influences) and prior findings (e.g., previously observed cultural differences) were minimally successful in accounting for it. For these findings at least, it appears that the cumulative evidence base has not yet matured to predict moderating influences reliably. Simultaneously, there is accumulating evidence that researchers can predict the likelihood that the effect of interest will replicate or not (Camerer et al., 2016, 2018; Dreber et al., 2016, 2018).

For many multi-study investigations, a common template is to identify an effect in a first study, and then report evidence for a variety of moderating influences in follow-up studies. A pessimistic interpretation would suggest that this template may be a consequence of practices

that inflate the likelihood of false positives. Consider the context in which positive results are perceived as more publishable than negative results (Greenwald, 1975), and common analytic practices may inadvertently increase the likelihood of obtaining false positives (Simmons et al., 2011). In a program of research, researchers might eventually obtain a significant result for a simple effect and call that Study 1. In follow-up studies, the authors might fail to observe the original effect and then initiate a search for moderators. Such post hoc searches necessarily increase the likelihood of false positives, but finding one may simultaneously reinforce belief in the original effect despite failing to replicate it. That is, identifying a moderator may feel like one is unpacking the phenomenon and explaining why the main effect "failed".

An ironic consequence is that the identification of a moderator may increase confidence and decrease credibility of the effect simultaneously. Investigating moderating influences is much harder than presently appreciated in practice. For one, a 2 x 2 x 2 ANOVA has a nominal false positive rate of ~30% for one or more of its seven tests (1 - 0.95^7). Correcting for multiple tests in multivariate analyses is rare (Cramer et al., 2016). Also, typical study designs are woefully underpowered for studying moderation (Frazier, Tix, & Barron, 2004; McClelland, 1997), perhaps because researchers intuitively overestimate the power of various research designs (Bakker, Hartgerink, Wicherts, & van der Maas, 2016). The combination of low power and lack of correction for multiple tests means that every study offers ample opportunity for seeming to detect moderating influences that are not there.

Ultimately, the main implication of the present findings is a plain one -- it is not sufficient to presume moderating influences to account for variation in observed results of a phenomenon. Invocation of cultural, sample, or procedural variation as an account for differences in observed effects could be a reasonable hypothesis, but is not a credible hypothesis

until it survives confrontation with a confirmatory test (Nosek, Ebersole, DeHaven, & Mellor, 2018).

Limitations

The present study has the strength of very large samples collected from a wide variety of sources and cultures. Nevertheless, the generalizability of these results to other psychological findings is unknown. Here, 50% of the examined findings reproduced the original results, roughly consistent with other large-scale investigations of reproducibility (Camerer et al., 2016, 2018; Ebersole et al., 2016; Klein et al., 2014; Open Science Collaboration, 2015). However, the findings selected for replication were not a random sample of any definable population, nor was it a large sample of findings. It may be surprising that just 50% of findings reproduced under these circumstances (original materials, peer review in advance, extremely high power, multiple samples), but that does not mean that 50% of all findings in psychology will reproduce, or fail to reproduce, under similar circumstances.

This study has the advantage over the prior work by having many tests and large samples for relatively precise estimation. Nevertheless, the failures to replicate do not necessarily mean that the tested hypotheses are incorrect. The lack of effect may be limited to the particular procedural conditions examined here. Future theory and evidence will need to account for why the effects are not observed in these circumstances if they are replicable in others. Conversely, the successful replications add substantial precision for effect estimation and extend the generalizability of those phenomena across a variety of samples and settings.

Data availability

The amassed dataset is very rich for exploring the individual effects, potential interactions between specific effects, and alternate ways to estimate heterogeneity and analyze

the aggregate data. Our analysis plan focused on the big picture and not, for example, exploring potential moderating influences on each of the individual effects. These are worthy analyses, but beyond the scope of a single paper. Follow-up investigations on these data could provide substantial additional insight. For commentaries solicited by *Advances in Methods and Practices in Psychological Science* we leveraged the extremely high-powered design of this study to demonstrate the productive interplay of exploratory and confirmatory analysis strategies. Commentators received a third of the dataset for analysis. Upon completion of the exploratory analysis, the analytic scripts were registered and applied to the holdout data for a mostly confirmatory test (Nosek et al., 2018). Analysts' decisions could be influenced by advance observation of the summary results in this paper, but use of the holdout sample reduces other potential biasing influences. Finally, the full dataset (plus the portions used for the exploratory/confirmatory commentaries) and all study materials are available at https://osf.io/8cd4r/ so that other teams can use it for their own investigations.

Conclusion

Our results suggest that variation across samples, settings, and procedures has modest explanatory power for understanding variation in effects for 28 findings. These results do not indicate that moderating influences never occur. Rather, they suggest that hypothesizing a moderator to account for observed differences between contexts is not equivalent to testing it with new data. The Many Labs paradigm allows testing across a broad range of contexts to probe the variability of psychological effects across samples. Such an approach is particularly valuable to understanding the extent to which given psychological findings represent general features of the human mind.

References

Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L., & Syme, S. L. (1994). Socioeconomic status and health: The challenge of the gradient. *American Psychologist, 49*, 15-24. doi:10.1037/0003-066X.49.1.15

Adler, N. E., Epel, E. S., Castellazzo, G., & Ickovics, J. R. (2000). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy white women. *Health Psychology, 19*, 586-592. doi: 10.1037/0278-6133.19.6.586

Anderson, C., Kraus, M. W., Galinsky, A. D., & Keltner, D. (2012). The local-ladder effect social status and subjective well-being. *Psychological Science*, *23*, 764-771. doi: 10.1177/0956797611434537

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136,* 569. doi: 10.1037/0096-3445.136.4.569

Ashton-James, C. E., Maddux, W. W., Galinsky, A. D., & Chartrand, T. L. (2009). Who I am depends on how I feel the role of affect in the expression of culture. *Psychological Science, 20*(3), 340-346.

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological Science, 27*(8), 1069-1077.

Bauer, M. A., Wilkie, J. E., Kim, J. K., & Bodenhausen, G. V. (2012). Cuing consumerism situational materialism undermines personal and social well-being. *Psychological Science, 23*, 517-523. doi: 10.1177/0956797611429579

Baumeister, R. F., & Vohs, K. D. (2016). Misguided effort with elusive implications. *Perspectives on Psychological Science, 11*(4), 574-575.

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology, 55,* 803-832. doi: 10.1146/annurev.psych.55.090902.141601

Brislin, R.W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*, 185-216. doi: 10.1177/135910457000100301

Camerer, C. F., Dreber, A., Forsell, E., Ho, T-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*, 1433-1436. Doi: 10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E-J., Wu, H. (2018). Evaluating Replicability of Social Science

Experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*. Doi: 10.1038/s41562-018-0399-z

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112-130. doi: 10.3758/s13428-013-0365-7

Cheung, F., & Lucas, R. E. (2014). Assessing the validity of single-item life satisfaction measures: results from three large samples. *Quality of Life Research*, *10*, 2809-2818. doi: 10.1007/s11136-014-0726-4

Cohen, G. L., Sherman, D. K., Bastardi, A., Hsu, L., McGoey, M., & Ross, L. (2007). Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation. *Journal of Personality and Social Psychology, 93*, 415-430. doi: 10.1037/0022-3514.93.3.415

Cohen, S., Alper, C. M., Doyle, W. J., Adler, N., Treanor, J. J., & Turner, R. B. (2008). Objective and subjective socioeconomic status and susceptibility to the common cold. *Health Psychology, 27*, 268-274. doi:10.1037/0278-6133.27.2.268

Coppock, A. (in press). Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Science Research Methods*.

Cramer, A. O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., ... & Wagenmakers, E. J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review, 23*(2), 640-647.

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, 45*, 299-311. doi: 10.1027/1864-9335/a000202.

Critcher, C. R., & Gilovich, T. (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making, 21*, 241-251. doi: 10.1002/bdm.586

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods, 40*(2), 428-434.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*, 71–75. doi: 10.1207/s15327752jpa4901_13

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind?. *PloS one, 7*(1), e29081.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2016). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, 112*, 15343-15347. Doi: 10.1073/pnas.1516179112

Dreber, A., Pfeiffer, T., Forsell, E., Viganola, D., Johannesson, M., Chen, Y., Wilson, B., Nosek, B. A., & Almenberg, J. (2018). Predicting replication outcomes in the Many Labs 2 study. Unpublished manuscript.

Ehrhart, M. G., Ehrhart, K. H., Roesch, S. C., Chung-Herrera, B. G., Nadler, K., & Bradshaw, K. (2009). Testing the latent factor structure and construct validity of the Ten-Item Personality Inventory. *Personality and Individual Differences*, *47*, 900-905. doi: 10.1016/j.paid.2009.07.012

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, 45*, 299-311. doi: 10.1027/1864-9335/a000202.

Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. *Psychology and Aging, 25*, 271. doi: 10.1037/a0019106

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology, 51*(1), 115.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*, 25-42. doi:10.1257/089533005775196732

Giessner, S. R., & Schubert, T. W. (2007). High in the hierarchy: How vertical location and judgments of leaders' power are interrelated. *Organizational Behavior and Human Decision Processes*, *104*, 30-44. doi: 10.1016/j.obhdp.2006.10.001

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*, 21-38. doi: 10.1037/0033-2909.117.1.21

Gnambs, T. (2014). A meta-analysis of dependability coefficients (test-retest reliabilities) for measures of the Big Five. *Journal of Research in Personality, 52*, 20-28. doi: 10.1016/j.jrp.2014.06.003

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029-1046. doi: 10.1037/a0015141

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141-165). Beverly Hills, CA: Sage.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504-528. doi: 10.1016/S0092-6566(03)00046-1

Gray, K., & Wegner, D. M. (2009). Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*, 505-520. doi: 10.1037/a0013748

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... Calvillo, D. P. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546-573.

Haidt, J., McCauley, C., & Rozin, P. (1994). Individual differences in sensitivity to disgust: A scale sampling seven domains of disgust elicitors. *Personality and Individual Differences, 16*, 701-713. doi: 10.1016/0191-8869(94)90212-7

Harter, S. (1985). *Manual for the Self-Perception Profile for Children (revision of the Perceived Competence Scale for Children)*. Denver, CO: University of Denver.

Hauser, M. D., Cushman, F. A., Young, L., Jin, R., & Mikhail, J. M. (2007). A dissociation between moral judgments and justifications. *Mind & Language, 22*, 1-21. doi: 10.1111/j.1468-0017.2006.00297.x

Hawkins, C. B., Fitzgerald, C., & Nosek, B. A. (2015). In search of an association between conception risk and prejudice. *Psychological Science, 26*, 249-252.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences 33*(2-3), 61-83.

Hilbig, B. E. (2016). Reaction time effects in lab-versus Web-based research: Experimental evidence. *Behavior Research Methods, 48*, 1718-1724.

Hsee, C. K. (1998). Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making, 11*, 107-121. doi:10.1002/(SICI)1099-0771(199806)11:2<107::AID-BDM292>3.0.CO;2-Y

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*, 299-311. doi: 10.1007/s10869-014-9357-6

Huang, Y., Tse, C. S., & Cho, K. W. (2014). Living in the north is not necessarily favorable: Different metaphoric associations between cardinal direction and valence in Hong Kong and in the United States. *European Journal of Social Psychology, 44*, 360-369. doi: 10.1002/ejsp.2013

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.

Inbar, Y., Pizarro, D., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion, 9*, 435-439. doi: 10.1037/a0015960

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments? A direct replication of Schnall, Benton, and Harvey (2008). *Social Psychology, 45,* 209–215. doi: 10.1027/1864-9335/a000186

Johnson-Laird, P. N., & Bara, B. G.(1984). Syllogistic inference. *Cognition, 16*, 1-61. doi: 10.1016/0010-0277(84)90035-0

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3*, 1-24. doi: 10.1016/0022-1031(67)90034-0

Kay, A. C., Laurin, K., Fitzsimons, G. M., & Landau, M. J. (2014). A functional basis for structure-seeking: Exposure to structure promotes willingness to engage in motivated action. *Journal of Experimental Psychology: General*, *143*, 486-491. doi: 10.1037/a0034462

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, S., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142-152. doi: 10.1027/1864-9335/a000178

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*, 190-193. doi: 10.1111/1467-8284.00419

Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies, 130,* 203-231. doi: 10.1007/s11098-004-4510-0

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, *22*, 787-794. doi: 10.1177/0956797611407929

Krupnikov, Y., & Levine, A. S. (2014). Cross-sample comparisons and external validity. *Journal of Experimental Political Science, 1*(1), 59-80.

Lewin, K. (1936) *Principles of topological psychology*. New York, NY: McGraw-Hill.

Lucas, R. E., & Donnellan, M. B. (2012). Estimating the reliability of single-item life satisfaction measures: Results from four national panel studies. *Social Indicators Research, 105*, 323-331. doi: 10.1007/s11205-011-9783-z

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*, 224-253. doi:10.1037/0033-295X.98.2.224

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2*, 3-19. doi: 10.1037//1082-989X.2.1.3

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455. doi: 10.1037/a0028085

Meier, B. P., Moller, A. C., Chen, J. J., & Riemer-Peltz, M. (2011). Spatial metaphor and real estate north-south location biases housing preference. *Social Psychological and Personality Science, 2,* 547-553. doi: 10.1177/1948550611401042

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., & Van der Laan, M. (2014). Promoting transparency in social science research. *Science, 343*, 30-31. Doi: 10.1126/science.1245317

Miyamoto, Y., & Kitayama, S. (2002). Cultural variation in correspondence bias: The critical role of attitude diagnosticity of socially constrained behavior. *Journal of Personality and Social Psychology, 83*, 1239-1248. doi: 10.1037/0022-3514.83.5.1239

Morey, R. D., & Lakens, D. (2016). Why most of psychology is statistically unfalsifiable. Unpublished manuscript. doi: 10.5281/zenodo.838685

Morsanyi, K., & Handley, S. J. (2012). Logic feels so good—I like it! Evidence for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 596-616. doi: 10.1037/a0026099

Mullen, B., Atkins, J. L., Champion, D. S., Edwards, C., Hardy, D., Story, J. E., & Vanderklok, M. (1985). The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology*, *21*, 262-283. doi: 10.1016/0022-1031(85)90020-4

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science, 2*(2), 109-138.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*, 771-781. doi:10.2139/ssrn.1804189

Mussweiler, T. (2001). 'Seek and ye shall find': Antecedents of assimilation and contrast in social comparison. *European Journal of Social Psychology, 31*, 499-509. doi: 10.1002/ejsp.75

Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, *26*, 653-684. doi: 10.1207/s15516709cog2605_4

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T. A., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Levy Paluck, E., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., & Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*, 1422-1425. Doi: 10.1126/science.aab2374

Nosek, B. A., Ebersole, C. R., DeHaven, A., Mellor, D. (2018). The Preregistration Revolution. *Proceedings for the National Academy of Sciences*. doi: 10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, *45*, 137–141. doi: 10.1027/1864-9335/a000192. doi: 10.1027/1864-9335/a000192

Olatunji, B. O., Williams, N. L., Tolin, D. F., Abramowitz, J. S., Sawchuk, C. N. Lohr, J. M., & Elwood, L. S. (2007). The Disgust Scale: Item analysis, factor structure, and suggestions for refinement. *Psychological Assessment, 19*, 281-297. doi: 10.1037/1040-3590.19.3.281

Open Science Collaboration. (2015). Estimating the Reproducibility of Psychological Science. *Science*, 349(6251), aac4716. DOI: 10.1126/science.aac4716.

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. doi: 10.1016/j.jesp.2009.03.009

Oswald, A. J., & Wu, S. (2010). Objective confirmation of subjective measures of human well-being: Evidence from the U.S.A. *Science, 327*, 576-579. doi:10.1126/science.1180606

Ramscar, M., Shaoul, C., Baayen, R. H., & Tbingen, E. K. U. (2015). Why many priming results don't (and won't) replicate: A quantitative analysis. *Unpublished manuscript*.

Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology, 95*, 293-307. doi: 10.1037/0022-3514.95.2.293

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 27*, 151-161. doi: 10.1177/0146167201272002

Rojas, S. L., & Widiger, T. A. (2014). Convergent and discriminant validity of the Five Factor Form. *Assessment, 21*, 143-157. doi: 10.1177/1073191113517260

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*, 279-301. doi: 10.1016/0022-1031(77)90049-X

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York: McGraw-Hill.

Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science, 12*, 185-190. doi: 10.1111/1467-9280.00334

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, 8:79. Doi: 10.1186/1471-2288-8-79

Sandvik, E., Diener, E., & Seidlitz, L. (1993). Subjective well-being: The convergence and stability of self-report and non-self-report measures. *Journal of Personality, 61*, 317-342. doi: 10.1111/j.1467-6494.1993.tb00283.x

Savani, K., Markus, H. R., Naidu, N. V. R., Kumar, S., & Berlia, N. (2010). What counts as a choice? U.S. Americans are more likely than Indians to construe actions as choices. *Psychological Science, 21*, 391-398. doi: 10.1177/0956797609359908

Schnall, S. 2014. Social media and the crowd-sourcing of social psychology. Cambridge Embodied Cognition and Emotion Laboratory Blog, Mar. 4. www.psychol.cam.ac.uk/cece/blog.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?. *Journal of Research in Personality, 47*, 609-612. doi: 10.1016/j.jrp.2013.05.009

Schwarz, N., & Clore, G. L. (2016). Evaluating psychological research requires more than attention to the n: A comment on Simonsohn's (2015)"small telescopes". *Psychological Science, 27*, 1407-1409. doi: 10.1177/0956797616653102

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, 45*, 299-311. doi: 10.1027/1864-9335/a000202.

Schwarz, N., Strack, F., & Mai, H. P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly*, *55*, 3-23. doi: 10.1086/269239

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition, 21*, 546-556. doi: 10.3758/BF03197186

Simonsohn, U. (2017). "Many Labs" overestimated the importance of hidden moderators. Retrieved from http://datacolada.org/63 on November 12, 2017.

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology, 37*, 1660-1672. doi: 10.1037/0022-3514.37.10.1660

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*, 59-71. doi: 10.1177/1745691613514450

Todd, A. R., Hanko, K., Galinsky, A. D., & Mussweiler, T. (2011). When focusing on differences leads to similar perspectives. *Psychological Science, 22*, 134-141. doi: 10.1177/0956797610392929

Tversky, A., & Gati, I. (1978). Studies of similarity. *Cognition and Categorization, 1*, 79-98.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453-458. doi: 10.1126/science.7455683

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, 113*, 6454-6459. doi: 10.1073/pnas.1521897113

Van Lange, P. A. M., Otten, W., De Bruin, E. M. N. & Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology, 73,* 733-746, doi: 10.1037/0022-3514.73.4.733

Veenhoven, R. (2009). The international scale interval study. In V. Møller & D. Huschka (Eds.), *Quality of life in the new millennium: Advances in quality-of-life studies, theory and research* (pp. 45-58). Dordrecht, Netherlands: Springer.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063–1070. doi: 10.1037/0022-3514.54.6.1063

Wilcox, R. R. (2009). Comparing Pearson correlations: Dealing with heteroscedasticity and nonnormality. *Communications in Statistics-Simulation and Computation, 38*(10), 2220-2234.

Zaval, L., Keenan, E. A., Johnson, E. J., & Weber, E. U. (2014). How warm days increase belief in global warming. *Nature Climate Change, 4*, 143-147. doi: 10.1038/nclimate2093

Zhong, C. B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science, 313*, 1451–1452. doi: 10.1126/science.1130726

Zielinski, T. A., Goodwin, G., & Halford, G. S. (2006). Relational complexity and logic: Categorical syllogisms revisited. *Unpublished manuscript*

Appendix. Individual Difference Measures and Original Articles of Included Effects. Citation counts from Google Scholar on July 24, 2017.

Effect #	Measures, Effects, and Citation	# citations	Study #
	Demographics and individual difference measures		
	Age, Sex, Race/ethnicity, Cultural origins (3 items), political ideology, education, Hometown, location of wealthier people in hometown (for Huang et al., 2014)	N/A	
	Well-being: Cantril, H. (1965). The patterns of human concerns. New Brunswick, NJ: Rutgers University Press.	3679	
	Cognitive reflection: Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. Psychology and Aging, 25, 271.	94	
	Self-Esteem: Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. Personality and Social Psychology Bulletin, 27, 151-161.	1687	
	Personality: Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in Personality, 37, 504-528.	3857	
	Instruction Manipulation Check: 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.	1009	
	Data quality: Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. Psychological Methods, 17, 437–455.	562	
	Subjective well-being: Veenhoven, R. (2009). The international scale interval study. In V. Møller & D. Huschka (Eds.), Quality of life in the new millennium: Advances in quality-of-life studies, theory and research (pp. 45-58). Dordrecht, Netherlands: Springer.	28	
	Mood: Cohen, G. L., Sherman, D. K., Bastardi, A., Hsu, L., McGoey, M., & Ross, L. (2007). Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation. Journal of Personality and Social Psychology, 93, 415-430.	147	
	Disgust Sensitivity, Contamination subscale (Slate 1 only): Olatunji, B. O., Williams, N. L., Tolin, D. F., Abramowitz, J. S., Sawchuk, C. N. Lohr, J. M., & Elwood, L. S. (2007). The Disgust Scale: Item analysis, factor structure, and suggestions for refinement. Psychological Assessment, 19, 281-297.	405	
Effect #	***Slate 1***	# citations	Study #

1	Huang, Y., Tse, C. S., & Cho, K. W. (2014). Living in the north is not necessarily favorable: Different metaphoric associations between cardinal direction and valence in Hong Kong and in the United States. European Journal of Social Psychology, 44, 360-369.	2	1a
2	Kay, A. C., Laurin, K., Fitzsimons, G. M., & Landau, M. J. (2014). A functional basis for structure-seeking: Exposure to structure promotes willingness to engage in motivated action. Journal of Experimental Psychology: General, 143, 486-491.	30	2
3	Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. Journal of Experimental Psychology: General, 136, 569.	598	4
4	Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology, 96, 1029–1046.	1494	1
5	Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. Psychological Science, 12, 185-190.	653	1
6	Bauer, M. A., Wilkie, J. E., Kim, J. K., & Bodenhausen, G. V. (2012). Cuing consumerism situational materialism undermines personal and social well-being. Psychological Science, 23, 517-523.	118	4
7	Miyamoto, Y, & Kitayama, S. (2002). Cultural variation in correspondence bias: The critical role of attitude diagnosticity of socially constrained behavior. Journal of Personality and Social Psychology, 83,1239-1248.	133	1
8	Inbar, Y., Pizarro, D., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. Emotion, 9, 435-439.	369	1
9	Critcher, C. R., & Gilovich, T. (2008). Incidental environmental anchors. Journal of Behavioral Decision Making, 21, 241-251.	123	2
10	Van Lange, P. A. M., Otten, W., De Bruin, E. M. N., & Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. Journal of Personality and Social Psychology, 4, 733 - 746.	992	3
11	Hauser, M. D., Cushman, F. A., Young, L., Jin, R., & Mikhail, J. M. (2007). A dissociation between moral judgments and justifications. Mind & Language, 22, 1-21.	575	1.1
12	Anderson, C., Kraus, M. W., Galinsky, A. D., & Keltner, D. (2012). The local-ladder effect social status and subjective well-being. Psychological science, 23, 764-771.	172	3
13	Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: An egocentric bias in social perception and attribution processes. Journal of Experimental Social Psychology, 13, 279-301.	2666	1.1
Effect #	***Slate 2***	# citations	Study #
14	Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: An	2666	1.2

	egocentric bias in social perception and attribution processes. Journal of Experimental Social Psychology, 13, 279-301.		
15	Giessner, S. R., & Schubert, T. W. (2007). High in the hierarchy: How vertical location and judgments of leaders' power are interrelated. Organizational Behavior and Human Decision Processes, 104, 30-44.	217	1a
16	Tversky, A., Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211, 453-458.	15808	10
17	Hauser, M. D., Cushman, F. A., Young, L., Jin, R., & Mikhail, J. M. (2007). A dissociation between moral judgments and justifications. Mind & Language, 22, 1-21.	575	1.2
18	Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. Journal of Personality and Social Psychology, 95, 293.	93	2
19	Savani, K., Markus, H. R., Naidu, N. V. R., Kumar, S., & Berlia, N. (2010). What counts as a choice? US Americans are more likely than Indians to construe actions as choices. Psychological Science, 21, 391-398.	88	5
20	Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. Cognitive Science, 26, 653-684.	431	2
21	Hsee, C. K. (1998). Less is better: When low-value options are valued more highly than high-value options. Journal of Behavioral Decision Making, 11, 107-121.	322	1
22	Gray, K., & Wegner, D. M. (2009). Moral typecasting: divergent perceptions of moral agents and moral patients. Journal of Personality and Social Psychology, 96, 505.	183	1a
23	Zhong, C. B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. Science, 313, 1451–1452.	823	2
24	Schwarz, N., Strack, F., & Mai, H. P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. Public Opinion Quarterly, 55, 3-23.	435	1
25	Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. Memory & Cognition, 21, 546-556.	547	1
26	Zaval, L., Keenan, E. A., Johnson, E. J., & Weber, E. U. (2014). How warm days increase belief in global warming. Nature Climate Change, 4, 143-147.	95	3a
27	Knobe, J. (2003). Intentional action and side effects in ordinary language. Analysis, 63, 190-193.	743	1
28	Tversky, A., & Gati, I. (1978). Studies of similarity. Cognition and categorization, 1, 79-98.	652	2

Table 1. Summary of changes to preregistered analysis plan for each of the 28 studies

Study #	Effect	Known differences from original study	Change to analysis plan
1	Direction & SES (Huang et al., 2014)	Original paper-pencil, tested orientation difference with tablets at some sites	None
2	Structure and goal-pursuit (Kay et al., 2014)	None known	None
3	Incidental disfluency (Alter et al., 2007)	Original paper-pencil	None
4	Moral Foundations (Graham et al., 2009)	Political ideology item changed from U.S.-centric "liberal-conservative" to regionally appropriate terms for left-right; Simplified analysis strategy	None
5	Affect and Risk (Rottenstreich & Hsee, 2001)	Original may have been paper-pencil	None
6	Priming consumerism (Bauer et al., 2012)	None known	None
7	Correspondence Bias (Miyamoto & Kitayama, 2002)	Original paper-pencil; Names and location altered to be familiar to each sample; Essay prompt matched to legal status of capital punishment in nation; Included 10 second lag to increase likelihood of reading essay; Removed "low diagnositicity" study conditions	None
8	Disgust & Homophobia (Inbar et al., 2009)	Used 5-item contamination subscale of modern 25-item disgust sensitivity scale instead of original 8-item measure.	None
9	Incidental Anchors (Critcher & Gilovich, 2008)	Original paper-pencil, tested difference with 10 sites conducting this task on paper-pencil; Matched markets with location of data collection; Updated pictures to modern smartphones	None
10	Social Value Orientation (Van Lange et al., 1997)	Original paper-pencil; Measured social value orientation with modern SVO slider instead of original categorical measure	None
11	Trolley Dilemma 1 (Hauser et al., 2007)	Used just a subset of the scenarios	Fisher's exact test instead of chi-square to have two-sided results in which negative values indicate opposite to original.
12	SMS & Well-Being (Anderson et al., 2012)	Removed high and low socioeconomic study conditions	None
13	False Consensus 1 (Ross et al., 1977)	Original was likely paper-pencil	None
14	False Consensus 2 (Ross et al., 1977)	Original was likely paper-pencil	None
15	Position & Power (Giessner & Schubert, 2007)	Currency converted and adjusted for relevance to each sample	None
16	Framing (Tversky & Kahneman, 1981)	Original was likely paper-pencil; Replaced dollar amounts and consumer items to be appropriate for 2014; Currency converted and adjusted for relevance to each sample	Fisher's exact test instead of chi-square to have two-sided results in which negative values indicate opposite to original.
17	Trolley Dilemma 2 (Hauser et al., 2007)	Used just a subset of the scenarios	Fisher's exact test instead of chi-square to have two-sided results in which negative values indicate opposite to original.
18	Tempting Fate (Risen & Gilovich, 2008)	Original was likely paper-pencil; Removed "other person" conditions	None
19	Actions are Choices (Savani et al. 2010)	Original may have been paper-pencil; Adjusted analysis plan to estimate sample effect sizes	Asymptotic CIs reported rather than exact, non-central CIs reported in other studies.
20	Intuitive Reasoning (Norenzayan et al. 2002)	Categorized objects by selecting from multiple-choice list; Balanced random assignment to condition (original was 2/3, 1/3); Removed the practice trial	None
21	Less is Better (Hsee, 1998)	Original may have been paper-pencil; Currency converted and adjusted for relevance to each sample	None
22	Moral Typecasting (Gray & Wegner, 2009)	Original may have been paper-pencil	None
23	Moral Cleansing (Zhong & Liljenquist, 2006)	Original paper-pencil; Participants typed an adapted version of the story under guise of measuring personality and typing speed	None
24	Assimilation and Contrast (Schwarz et al., 1991)	Original paper-pencil in German	None
25	Choosing or Rejecting (Shafir, 1993)	Original paper-pencil. Original counterbalanced the order of parents, the replication did not.	Estimated effect size directly from the key Z test rather than estimating effect size with a logistic regression model.
26	Priming warmth (Zaval et al., 2014)	Excluded original question about current temperature at the start of the study with 10-minute delay to starting actual study	Excluding participants that made errors in sentence unscrambling was not preregistered, but decided a priori on recommendation of original authors.
27	Intentional Side Effects (Knobe, 2003)	Original may have been paper-pencil; Dependent variable changed from yes/no response to 7-point agreement scale	None
28	Direction and Similarity (Tversky & Gati, 1978)	Original was likely paper-pencil; Nations updated: Ceylon to Sri Lanka, West Germany to Germany, and U.S.S.R. to Russia	Additional mixed models reported in supplement

Notes: Additional description and supplementary analyses are available in Supplementary Notes (https://osf.io/4rbh9/). Full description of known differences from original study appear in the preregistered protocol such as notation of additional experimental conditions or outcome variables that were part of original study but not included in the replication (https://osf.io/ejcfw/). Unless noted otherwise, differences from original study were suggested by original authors or reviewed and approved during peer review. All studies differed in sample and setting of data collection from the original including the administration of studies sequentially in a slate. This is evaluated directly in the results section.

Table 2. Summary of effect sizes, confidence intervals, and significance test counts across samples for each of the 28 studies

	Original Study		Replication					
				Global effects		Significance Tests by Sample		
Effect	ES	95% CI	Median ES	ES	95% CI	Percentage <0 (p<.05)	Percentage ns	Percentage >0 (p<.05)
Cohen's q Effect Size								
Disgust & Homophobia (Inbar et al., 2009)	0.70	.05, 1.36	0.03	0.05	.01, .10	3.39	93.22	3.39
Assimilation & Contrast (Schwarz et al., 1991)	0.48	.07, .88	-0.06	-0.07	-.12, -.02	5.08	91.53	3.39
Cohen's d Effect Size								
Correspondence Bias (Miyamoto & Kitayama, 2002) - WEIRD	2.47	1.46, 3.49	1.78	1.81	1.75, 1.88	0.00	0.00	100.00
Correspondence Bias (Miyamoto & Kitayama, 2002) - less WEIRD	0.74	-.12, 1.59	1.86	1.84	1.74, 1.94	0.00	0.00	100.00
Intentional Side Effects (Knobe, 2003)	1.45	.79, 2.77	1.94	1.75	1.70, 1.80	0.00	5.08	94.92
Trolley Dilemma 1 (Hauser et al., 2007)	2.50	2.22, 2.86	1.42	1.35	1.28, 1.41	0.00	0.00	100.00
False Consensus 1 (Ross et al., 1977)	0.99	0.24, 2.29	1.08	1.18	1.13, 1.23	0.00	0.00	100.00
Moral Typecasting (Gray & Wegner, 2009)	0.80	.31, 1.29	1.04	0.95	.91, 1.00	0.00	5.00	95.00
False Consensus 2 (Ross et al., 1977)	0.80	0.22, 1.87	0.89	0.95	.90, 1.00	0.00	6.67	93.33
Intuitive Reasoning (Norenzayan et al. 2002) - WEIRD	0.00	-0.15, .15	0.95	0.95	.90, 1.00	0.00	2.33	97.67
Intuitive Reasoning (Norenzayan et al. 2002) - less WEIRD	0.69	.24, 1.13	0.50	0.56	.46, .65	0.00	42.86	57.14
Less is Better (Hsee, 1998)	0.69	.24, 1.13	0.86	0.78	.74, .83	0.00	10.53	89.47
Direction & SES (Huang et al., 2014) - WEIRD	0.83	.37, 1.28	0.66	0.55	.49, .61	4.35	30.43	65.22
Direction & SES (Huang et al., 2014) - less WEIRD	-0.59	-.99, -.19	-0.10	0.03	-.05, .13	5.56	83.33	11.11
Framing (Tversky & Kahneman, 1981)	1.08	.71, 1.45	0.38	0.40	.35, .45	0.00	54.55	45.45
Moral Foundations (Graham et al., 2009)	0.52	.40, .63	0.23	0.29	.25, .34	0.00	75.00	25.00
Trolley Dilemma 2 (Hauser et al., 2007)	0.34	.26, .42	0.22	0.25	.20, .30	0.00	81.67	18.33
Tempting Fate (Risen & Gilovich, 2008)	0.39	.03, .75	0.23	0.18	.14, .22	1.69	72.88	25.42
Priming consumerism (Bauer et al., 2012)	0.87	.41, 1.34	0.16	0.12	.07, .17	1.85	87.04	11.11
Incidental Anchors (Critcher & Gilovich, 2008)	0.30	.02, .58	0.00	0.04	-.01, .09	3.39	91.53	5.08
Position & Power (Giessner & Schubert, 2007)	0.55	.05, 1.05	0.01	0.03	-.01, .08	1.69	94.92	3.39
Direction & Similarity (Tversky & Gati, 1978)	0.48	.16, .80	0.03	0.01	-.02, .04	2.04	97.96	0.00
Moral Cleansing (Zhong & Liljenquist, 2006)	1.02	.39, 2.44	0.00	0.00	-.05, .04	0.00	94.23	5.77
Structure & Goal-pursuit (Kay et al., 2014)	0.49	0.001, .973	-0.02	-0.02	-.07, .03	0.00	100.00	0.00
Social Value Orientation (Van Lange et al., 1997)	0.19	<.001, .47	0.06	-0.03	-.08, .02	0.00	98.15	1.85
Priming warmth affects climate beliefs (Zaval et al., 2014)	0.31	.03, .59	0.00	-0.03	-.09, .03	5.36	89.29	5.36
Incidental Disfluency (Alter et al., 2007)	0.63	-.004, 1.25	-0.07	-0.03	-.08, .01	1.52	96.97	1.52
SMS & Well-Being (Anderson et al., 2012)	0.57	.20, .93	-0.05	-0.04	-.09, -.004	0.00	94.92	5.08
Choosing or Rejecting (Shafir, 1993)	0.35	-.04, .68	-0.04	-0.13	-.18, -.09	18.97	79.31	1.72
Affect & Risk (Rottenstreich & Hsee, 2001)	0.74	<.001, 1.74	-0.06	-0.08	-.13, -.03	3.33	95.00	1.67
Actions are Choices (Savani et al. 2010) - WEIRD	0.08	-.33, .50	-0.24	-0.21	-.23, -.18	46.51	53.49	0.00
Actions are Choices (Savani et al. 2010) - less WEIRD	-0.65	-1.01, -.30	-0.14	-0.12	-.16, -.08	28.57	71.43	0.00

Notes: All effect sizes (ES) presented in Cohen's d units except for Schwarz and Inbar for which Cohen's q is provided. 95% CIs for original effect sizes used cell sample sizes when available and assumed equal distribution across conditions when not available. For original studies that observed a difference between WEIRD and a particular less WEIRD sample, we present summary results for WEIRD and all less WEIRD samples separately to avoid potentially misrepresenting replication success within subsamples. Figure 2 plots WEIRD and less WEIRD distributions of effects across all studies.

Table 3. Heterogeneity tests for each of the 28 studies

		Heterogeneity tests															
		No moderators						WEIRD vs less WEIRD					Lab vs On-line				
Effect	ES	Tau	Q	df	p-value	I^2	I^2 95% CI	Tau	Q	p	I^2	I^2 95% CI	Tau	Q	p-value	I^2	I^2 95% CI
Cohen's q Effect Size																	
Disgust & Homophobia (Inbar et al., 2009)	0.05	0.00	55.80	58.00	0.56	3%	0%, 30%	0.00	2.89	0.09	0%	0%, 29%	0.00	0.18	0.67	5%	0%, 31%
Assimilation and Contrast (Schwarz et al., 1991)	-0.07	0.10	60.39	58.00	0.39	15%	0%, 33%	0.10	0.61	0.44	17%	0%, 35%	0.10	0.00	0.97	16%	0%, 34%
Cohen's d Effect Size																	
Correspondence Bias (Miyamoto & Kitayama, 2002)	1.82	0.00	235.65	57.00	<.001	65%	46%, 73%	0.00	1.47	0.23	64%	45%, 72%	0.00	2.83	0.09	64%	45%, 74%
Intentional Side Effects (Knobe, 2003)	1.75	0.14	631.72	58.00	<.001	93%	92%, 97%	*0.10*	*26.43*	*<.001*	*91%*	*87%, 95%*	0.14	2.55	0.11	93%	91%, 97%
Trolley Dilemma 1 (Hauser et al., 2007)	1.35	0.10	131.24	58.00	<.001	54%	32%, 66%	*0.10*	*4.80*	0.03	*51%*	*27%, 64%*	0.10	0.13	0.72	55%	32%, 67%
False Consensus 1 (Ross et al., 1977)	1.18	0.00	65.54	58.00	0.23	16%	0%, 41%	0.00	3.36	0.07	12%	0%, 38%	0.00	0.26	0.61	18%	0%, 43%
Moral Typecasting (Gray & Wegner, 2009)	0.95	0.10	203.30	59.00	<.001	73%	62%, 83%	0.10	6.02	0.01	71%	58%, 81%	0.10	0.52	0.47	71%	59%, 82%
False Consensus 2 (Ross et al., 1977)	0.95	0.00	100.19	57.00	<.001	43%	18%, 62%	0.00	0.00	0.97	44%	19%, 63%	0.00	0.17	0.68	46%	21%, 65%
Intuitive Reasoning (Norenzayan et al. 2002)	0.86	0.10	156.75	56.00	<.001	66%	54%, 81%	*0.10*	*20.58*	*<.001*	*55%*	*36%, 73%*	0.10	0.69	0.41	67%	55%, 81%
Less is Better (Hsee, 1998)	0.78	0.10	158.41	56.00	<.001	65%	49%, 77%	0.10	4.68	0.03	63%	46%, 75%	0.10	1.69	0.19	65%	49%, 77%
Framing (Tversky & Kahneman, 1981)	0.40	0.00	55.20	54.00	0.43	6%	0%, 36%	0.00	1.46	0.23	3%	0%, 37%	0.00	0.20	0.66	7%	0%, 38%
Direction & SES (Huang et al., 2014)	0.40	0.24	626.26	63.00	<.001	89%	84%, 92%	*0.22*	*13.01*	*<.001*	*87%*	*81%, 91%*	0.24	1.64	0.20	89%	84%, 92%
Moral Foundations (Graham et al., 2009)	0.29	0.09	175.26	59.00	<.001	64%	49%, 75%	0.09	0.25	0.62	65%	49%, 75%	0.09	1.26	0.26	65%	49%, 76%
Tempting Fate (Risen & Gilovich, 2008)	0.18	0.00	87.82	58.00	0.01	36%	6%, 54%	0.00	1.61	0.20	34%	3%, 53%	0.00	0.53	0.47	37%	7%, 55%
Trolley Dilemma 2 (Hauser et al., 2007)	0.25	0.00	60.40	59.00	0.42	12%	0%, 33%	0.00	0.90	0.34	10%	0%, 34%	0.00	0.14	0.71	11%	0%, 31%
Priming consumerism (Bauer et al., 2012)	0.12	0.00	63.78	53.00	0.15	12%	0%, 49%	0.00	0.04	0.85	14%	0%, 50%	0.00	0.30	0.58	15%	0%, 51%
Incidental Anchors (Critcher & Gilovich, 2008)	0.04	0.00	64.88	58.00	0.25	6%	0%, 43%	0.00	0.11	0.75	8%	0%, 44%	0.00	1.17	0.28	4%	0%, 41%
Social Value Orientation (Van Lange et al., 1997)	-0.03	0.00	103.56	53.00	<.001	50%	28%, 68%	0.00	1.15	0.28	50%	28%, 68%	0.00	1.15	0.28	49%	26%, 67%
Moral Cleansing (Zhong & Liljenquist, 2006)	0.00	0.00	65.59	51.00	0.08	22%	0%, 52%	0.00	1.17	0.28	21%	0%, 52%	*0.00*	*9.15*	*<.001*	*4%*	0%, 46%
Position & Power (Giessner & Schubert, 2007)	0.03	0.00	62.87	58.00	0.31	*3%*	0%, 42%	0.00	0.00	0.96	5%	0%, 43%	0.00	6.19	0.01	4%	0%, 35%
Direction and Similarity (Tversky & Gati, 1978)	0.01	0.00	15.33	48.00	0.99	0%	0%, 0%	0.00	0.42	0.52	0%	0%, 0%	0.00	0.12	0.73	0%	0%, 0%
SMS & Well-Being (Anderson et al., 2012)	-0.04	0.00	55.09	58.00	0.58	2%	0%, 30%	0.00	0.83	0.36	2%	0%, 30%	0.00	3.21	0.07	0%	0%, 16%
Priming warmth (Zaval et al., 2014)	-0.03	0.10	72.96	46.00	0.01	37%	8%, 63%	0.10	0.76	0.38	37%	8%, 63%	0.10	0.50	0.48	40%	11%, 64%
Structure and goal-pursuit (Kay et al., 2014)	-0.02	0.00	33.95	51.00	0.97	0%	0%, 2%	0.00	3.10	0.08	0%	0%, 0%	0.00	2.06	0.15	0%	0%, 0%
Incidental disfluency (Alter et al., 2007)	-0.03	0.00	59.46	65.00	0.67	0%	0%, 27%	0.00	1.38	0.24	0%	0%, 27%	0.00	0.91	0.34	0%	0%, 21%
Choosing or Rejecting (Shafir, 1993)	-0.13	0.00	51.67	40.00	0.10	26%	0%, 52%	0.00	0.55	0.46	26%	0%, 53%	0.00	0.14	0.71	25%	0%, 50%
Affect and Risk (Rottenstreich & Hsee, 2001)	-0.08	0.00	50.75	59.00	0.77	0%	0%, 21%	0.00	0.28	0.60	0%	0%, 22%	0.00	0.31	0.58	0%	0%, 25%
Actions are Choices (Savani et al. 2010)	-0.18	0.00	155.49	56.00	<.001	64%	47%, 76%	0.00	3.69	0.05	62%	43%, 74%	0.00	0.61	0.44	65%	48%, 77%

Notes: ES = Global Effect Size, repeated from Table 4 for comparison with Tau. All effect sizes based on Cohen's d units except for Schwarz and Inbar for which Cohen's q was used. Q(mod) have 1 df. Bonferroni correction for multiple comparisons suggests alpha=.004 (slate 1) and alpha = .003 (slate 2). Italics used for significant moderators. Random effects meta-analyses were conducted using the R package metafor (Viechtbauer, 2010). Between-study variance was estimated using REML.

Table 4. Global effect sizes and confidence intervals for each of the 28 studies compared with effect size when the study was administered first or last in its slate

	Global		First Position		Last Position	
Effect	ES	95% CI	ES	95% CI	ES	95% CI
Cohen's q Effect Size						
Disgust & Homophobia (Inbar et al., 2009)	0.05	.01, .10	0.01	-.16, .18	-0.06	-.23, .11
Assimilation and Contrast (Schwarz et al., 1991)	-0.07	-.12, -.02	-0.06	-.23, .12	-0.13	-.29, .03
Cohen's d Effect Size						
Correspondence Bias (Miyamoto & Kitayama, 2002)	1.82	1.76, 1.87	1.88	1.68, 2.07	1.63	1.43, 1.84
Intentional Side Effects (Knobe, 2003)	1.75	1.70, 1.80	1.47	1.27, 1.66	1.82	1.31, 2.03
Trolley Dilemma 1 (Hauser et al., 2007)	1.35	1.28, 1.41	1.57	1.33, 1.81	1.21	.98, 1.44
False Consensus 1 (Ross et al., 1977)	1.18	1.13, 1.23	1.22	1.05, 1.39	1.12	.93, 1.30
Moral Typecasting (Gray & Wegner, 2009)	0.95	.91, 1.00	1.07	.88, 1.26	1.20	1.01, 1.39
False Consensus 2 (Ross et al., 1977)	0.95	.90, 1.00	1.05	.88, 1.21	0.93	.75, 1.11
Intuitive Reasoning (Norenzayan et al. 2002)	0.86	.81, .91	0.69	.52, .87	0.71	.53, .89
Less is Better (Hsee, 1998)	0.78	.74, .83	0.75	.56, .93	0.85	.66, 1.03
Framing (Tversky & Kahneman, 1981)	0.40	.35, .45	0.47	.26, .68	0.41	.21, .62
Direction & SES (Huang et al., 2014)	0.40	.35, .45	0.31	.13, .49	0.35	.17, .52
Moral Foundations (Graham et al., 2009)	0.29	.25, .34	0.47	.30, .65	0.31	.14, .49
Tempting Fate (Risen & Gilovich, 2008)	0.18	.14, 22	0.12	-.05, .29	0.42	.25, .60
Trolley Dilemma 2 (Hauser et al., 2007)	0.25	.20, .30	0.20	.002, .41	0.24	.04, .44
Priming consumerism (Bauer et al., 2012)	0.12	.07, .17	0.03	-.16, .21	0.14	-.03, .32
Incidental Anchors (Critcher & Gilovich, 2008)	0.04	-.01, .09	0.09	-.08, .27	0.05	-.12, .22
Social Value Orientation (Van Lange et al., 1997)	-0.03	-.08, .02	-0.08	-.26, .10	-0.11	-.30, .08
Moral Cleansing (Zhong & Liljenquist, 2006)	0.00	-.05, .04	0.01	-.18, .20	0.17	-.02, .36
Position & Power (Giessner & Schubert, 2007)	0.03	-.01, .08	0.01	-.18, .19	-0.02	-.20, .15
Direction and Similarity (Tversky & Gati, 1978)	0.01	-.02, .04	0.13	-.01, .26	-0.01	-.14, .12
SMS & Well-Being (Anderson et al., 2012)	-0.04	-.09, .005	-0.08	-.26, .10	0.13	-.04, .30
Priming warmth (Zaval et al., 2014)	-0.03	-.09, .03	0.08	-.15, .30	-0.11	-.35, .14
Structure and goal-pursuit (Kay et al., 2014)	-0.02	-.07, .03	-0.01	-.18, .17	0.10	-.08, .27
Incidental disfluency (Alter et al., 2007)	-0.03	-.08, .01	-0.02	-.20, .16	-0.10	-.28, .07
Choosing or Rejecting (Shafir, 1993)	-0.13	-.18, -.09	-0.08	-.25, .09	-0.20	-.40, -.03
Affect and Risk (Rottenstreich & Hsee, 2001)	-0.08	-.13, -.02	-0.12	-.30, .07	-0.03	-.20, .14
Actions are Choices (Savani et al. 2010)	-0.18	-.21, -.16	-0.15	-.24, -.06	-0.20	-.29, -.11

Notes: ES = Effect Size in Cohen's d units except for Schwarz and Inbar for which Cohen's q is provided. Last position is 13 for Slate 1 effects and 15 for Slate 2 effects. Global column refers to overall effect size ignoring task position.

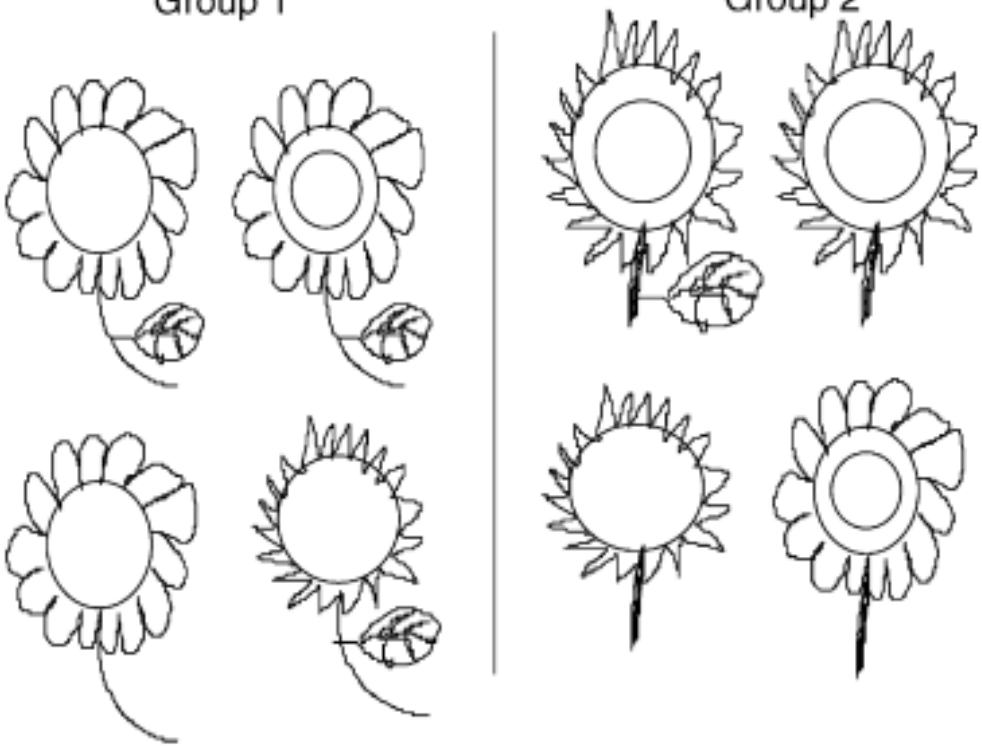
Table 5. Categorizing replication success (bold) or failure based on whether the replication global effect size is in the same direction as the original finding and statistically significant under different conditions

					Replication success criterion					
Effect	Original study sample size	Replication sample size	Replication global effect size	test	replication sample size, p < .05	replication sample size, p < .0001	Original study sample size, p < .05	2.5x original sample size, p < .05	50/group, p < .05	Minimum sample to detect w/ alpha=.05, power=.80
Correspondence Bias (Miyamoto & Kitayama, 2002)	107	7197	1.82	General Linear Model (main effect)	**< 1E-10**	**< 1E-10**	**4.65E-09**	**< 1E-10**	**< 1E-10**	12
Intentional Side Effects (Knobe, 2003)	78	7982	1.75	Welch Two Sample t-test	**< 1E-10**	**< 1E-10**	**< 1E-10**	**< 1E-10**	**< 1E-10**	14
Trolley Dilemma 1 (Hauser et al., 2007)	2646	6842	1.35	Two Sided Fisher's Exact Test	**< 1E-10**	**< 1E-10**	**< 1E-10**	**< 1E-10**	NA	20
False Consensus 1 (Ross et al., 1977)	80	7205	1.18	Welch Two Sample t-test	**< 1E-10**	**< 1E-10**	**6.98E-06**	**< 1E-10**	**5.18E-08**	26
Moral Typecasting (Gray & Wegner, 2009)	69	8002	0.95	Welch Two Sample t-test	**< 1E-10**	**< 1E-10**	**1.86E-04**	**3.06E-09**	**6.55E-06**	38
False Consensus 2 (Ross et al., 1977)	80	7827	0.95	Welch Two Sample t-test	**< 1E-10**	**< 1E-10**	**6.06E-05**	**2.04E-10**	**6.64E-06**	38
Intuitive Reasoning (Norenzayan et al. 2002)	157	7396	0.86	Welch Two Sample t-test	**< 1E-10**	**< 1E-10**	**< 1E-10**	**< 1E-10**	**3.92E-05**	46
Less is Better (Hsee, 1998)	83	7646	0.78	Welch Two Sample t-test	**< 1E-10**	**< 1E-10**	**6.18E-04**	**5.76E-08**	**1.69E-04**	54
Framing (Tversky & Kahneman, 1981)	181	7228	0.40	Two Sided Fisher's Exact Test	**< 1E-10**	**< 1E-10**	**0.031**	**6.29E-04**	NA	200
Direction & SES (Huang et al., 2014)	180	6591	0.40	Welch Two Sample t-test	**< 1E-10**	**< 1E-10**	0.080	**5.47E-03**	**0.0498**	200
Moral Foundations (Graham et al., 2009)	1209	6966	0.29	Fisher r-to-Z test (1 cor)	**< 1E-10**	**< 1E-10**	**4.12E-07**	**< 1E-10**	0.318	376
Trolley Dilemma 2 (Hauser et al., 2007)	2612	7923	0.25	Two Sided Fisher's Exact Test	**< 1E-10**	**< 1E-10**	**4.85E-08**	**< 1E-10**	NA	506
Tempting Fate (Risen & Gilovich, 2008)	120	8000	0.18	Two Sample t-test	**< 1E-10**	**< 1E-10**	0.325	0.119	0.369	972
Priming consumerism (Bauer et al., 2012)	77	6608	0.12	Two Sample t-test	**< 1E-10**	**< 1E-10**	0.594	0.399	0.546	2184
Disgust & Homophobia (Inbar et al., 2009)	44	7117	0.05	Fisher r-to-Z test (2 cor)	**0.024**	0.024	0.871	0.788	0.794	6283
Incidental Anchors (Critcher & Gilovich, 2008)	200	6826	0.04	Two Sample t-test	0.092	0.092	0.773	0.649	0.839	19626
Position & Power (Giessner & Schubert, 2007)	64	7890	0.03	Two Sample t-test	0.162	0.162	0.900	0.842	0.875	34886
Direction and Similarity (Tversky & Gati, 1978)	144	3549	0.01	One Sample t-test	0.550	0.550	0.973	0.983	0.953	313958
Moral Cleansing (Zhong & Liljenquist, 2006)	27	7001	0.00	Two Sample t-test	0.910	0.910	0.994	0.991	0.989	*NA*
Structure and goal-pursuit (Kay et al., 2014)	67	6506	-0.02	Welch Two Sample t-test	*0.347*	*0.347*	*0.924*	*0.880*	*0.907*	*NA*
Social Value Orientation (Van Lange et al., 1997)	536	6234	-0.03	Fisher r-to-Z test (1 cor)	*0.183*	*0.183*	*0.697*	*0.537*	*0.908*	*NA*
Incidental disfluency (Alter et al., 2007)	41	6935	-0.03	Two Sample t-test	*0.171*	*0.171*	*0.917*	*0.868*	*0.870*	*NA*
Priming warmth (Zaval et al., 2014)	192	4204	-0.03	Two Sample t-test	*0.274*	*0.274*	*0.816*	*0.712*	*0.866*	*NA*
SMS & Well-Being (Anderson et al., 2012)	116	6905	-0.04	Two Sample t-test	*0.079*	*0.079*	*0.820*	*0.719*	*0.833*	*NA*
Assimilation and Contrast (Schwarz et al., 1991)	100	7460	-0.07	Fisher r-to-Z test (2 cor)	*0.002*	*0.002*	*0.734*	*0.583*	*0.734*	*NA*
Affect and Risk (Rottenstreich & Hsee, 2001)	40	7218	-0.08	Two Sided Fisher's Exact Test	*0.002*	*0.002*	*0.831*	*0.735*	*NA*	*NA*
Choosing or Rejecting (Shafir, 1993)	170	7901	-0.13	One Sample Z-test	*5.47E-10*	*5.47E-10*	*0.186*	*0.079*	*0.314*	*NA*
Actions are Choices (Savani et al. 2010)	218	5882	-0.18	Generalized Linear Mixed Model with binomial (logit) link (main effect)	*8.04E-06*	*8.04E-06*	*NA*	*NA*	*NA*	*NA*
Successful Replications					15	14	11	12	8	
Unsuccessful Replications					13	14	16	15	16	
Success Rate					54%	50%	41%	44%	33%	

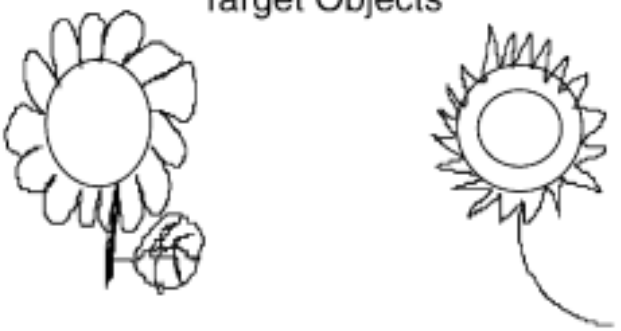
Notes: Findings are ordered by replication global effect size with negative values indicating effects in the opposite direction of the original WEIRD sample. Two effect sizes (Inbar et al., 2009; Schwarz et al., 1991) are on a different metric (Cohen's q) than the rest (Cohen's d). All p-values calculated based on the observed replication global effect size with different assumptions of sample size or alpha criterion (.05 or .0001). Replication success criteria p-values are bold if they meet criteria for successful replication. Replication success criteria p-values are in italics if the replication global effect size was in the opposite direction of the original WEIRD sample. Replication success could not be determined based on the original study sample size, 2.5x original study sample size, and 50/group for Savani et al. (2010) because this information could not be computed for the test used in this effect. Replication success could not be determined based on 50/group if the test was a Two Sided Fisher's Exact Test (4 findings), because this would require making strong assumptions on how the sample size per group is distributed in the 2x2 frequency table. Power analyses are conducted using the Cohen's d and Cohen's q values of the replication effect sizes. Note that if another effect size was used in the original study (e.g. correlation, odds ratio, proportion), these were transformed to Cohen's d values. The last column shows the sample size needed to detect a significant effect in the same direction of the original finding for the observed global effect size with alpha = .05 and power = .80. Cells are marked NA if the global effect size was in the opposite direction of the original finding.

Group 1

Group 2

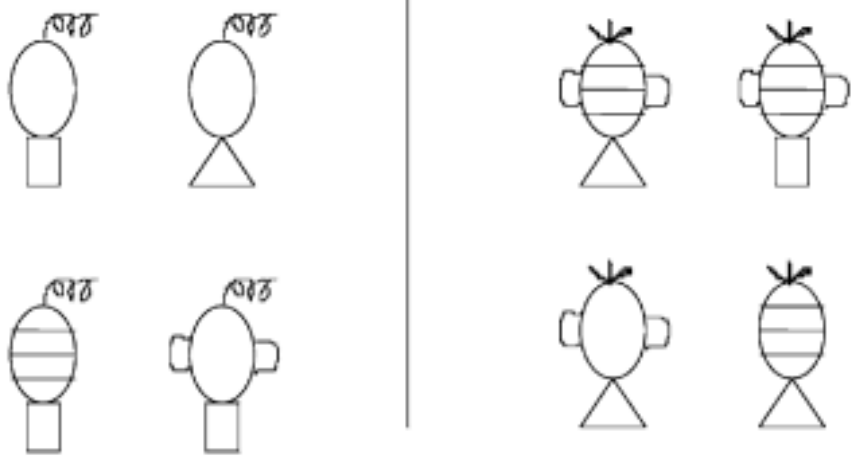


Target Objects



Group 1

Group 2



Target Objects

Original Effect Size

Cohen's q

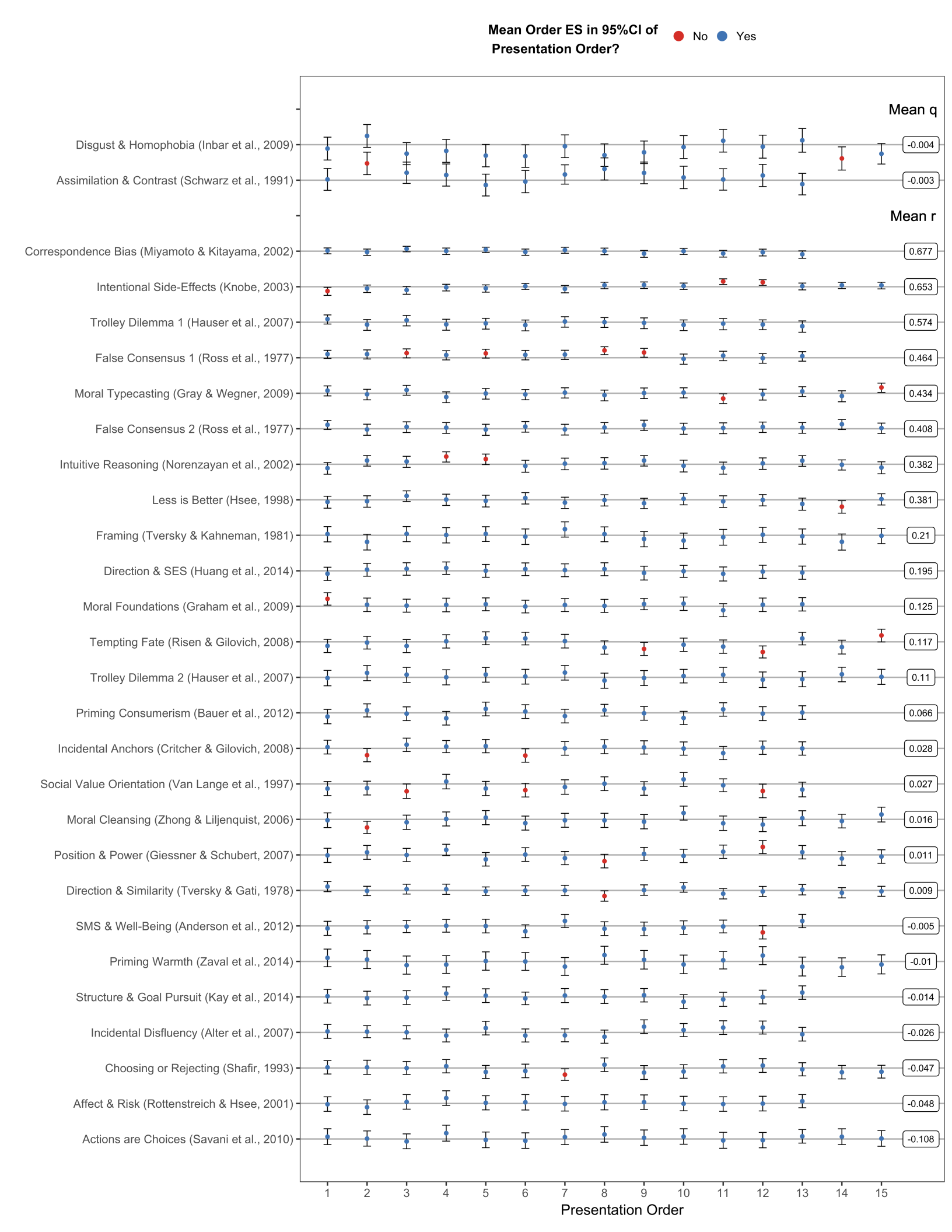
−3 −2 −1 0 1 2 3

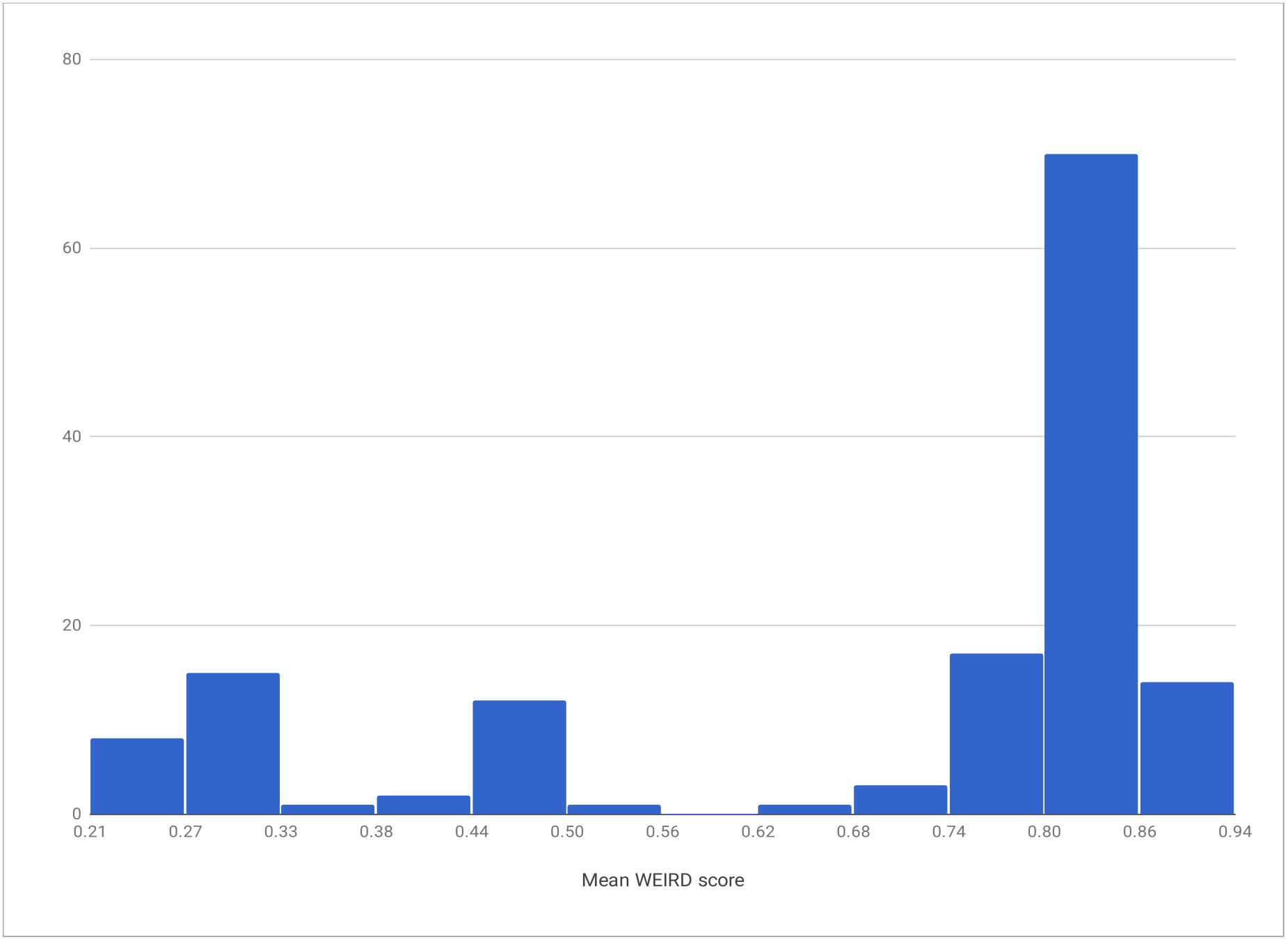
Disgust & Homophobia (Inbar et al., 2009)
Assimilation & Contrast (Schwarz et al., 1991)

Correspondence Bias (Miyamoto & Kitayama, 2002)
Intentional Side−Effects (Knobe, 2003)
Trolley Dilemma 1 (Hauser et al., 2007)
False Consensus 1 (Ross et al., 1977)
Moral Typecasting (Gray & Wegner, 2009)
False Consensus 2 (Ross et al., 1977)
Intuitive Reasoning (Norenzayan et al., 2002)
Less is Better (Hsee, 1998)
Framing (Tversky & Kahneman, 1981)
Direction & SES (Huang et al., 2014)
Moral Foundations (Graham et al., 2009)
Tempting Fate (Risen & Gilovich, 2008)
Trolley Dilemma 2 (Hauser et al., 2007)
Priming Consumerism (Bauer et al., 2012)
Incidental Anchors (Critcher & Gilovich, 2008)
Social Value Orientation (Van Lange et al., 1997)
Moral Cleansing (Zhong & Liljenquist, 2006)
Position & Power (Giessner & Schubert, 2007)
Direction & Similarity (Tversky & Gati, 1978)
SMS & Well−Being (Anderson et al., 2012)
Priming Warmth (Zaval et al., 2014)
Structure & Goal Pursuit (Kay et al., 2014)
Incidental Disfluency (Alter et al., 2007)
Choosing or Rejecting (Shafir, 1993)
Affect & Risk (Rottenstreich & Hsee, 2001)
Actions are Choices (Savani et al., 2010)

−1.0 −0.5 0.0 0.5 1.0

Effect Size r


Mean Order ES in 95%CI of Presentation Order?
No
Yes
Mean q
Disgust & Homophobia (Inbar et al., 2009)
-0.004
Assimilation & Contrast (Schwarz et al., 1991)
-0.003
Mean r
Correspondence Bias (Miyamoto & Kitayama, 2002)
0.677
Intentional Side-Effects (Knobe, 2003)
0.653
Trolley Dilemma 1 (Hauser et al., 2007)
0.574
False Consensus 1 (Ross et al., 1977)
0.464
Moral Typecasting (Gray & Wegner, 2009)
0.434
False Consensus 2 (Ross et al., 1977)
0.408
Intuitive Reasoning (Norenzayan et al., 2002)
0.382
Less is Better (Hsee, 1998)
0.381
Framing (Tversky & Kahneman, 1981)
0.21
Direction & SES (Huang et al., 2014)
0.195
Moral Foundations (Graham et al., 2009)
0.125
Tempting Fate (Risen & Gilovich, 2008)
0.117
Trolley Dilemma 2 (Hauser et al., 2007)
0.11
Priming Consumerism (Bauer et al., 2012)
0.066
Incidental Anchors (Critcher & Gilovich, 2008)
0.028
Social Value Orientation (Van Lange et al., 1997)
0.027
Moral Cleansing (Zhong & Liljenquist, 2006)
0.016
Position & Power (Giessner & Schubert, 2007)
0.011
Direction & Similarity (Tversky & Gati, 1978)
0.009
SMS & Well-Being (Anderson et al., 2012)
-0.005
Priming Warmth (Zaval et al., 2014)
-0.01
Structure & Goal Pursuit (Kay et al., 2014)
-0.014
Incidental Disfluency (Alter et al., 2007)
-0.026
Choosing or Rejecting (Shafir, 1993)
-0.047
Affect & Risk (Rottenstreich & Hsee, 2001)
-0.048
Actions are Choices (Savani et al., 2010)
-0.108
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Presentation Order


80
60
40
20
0
0.21
0.27
0.33
0.38
0.44
0.50
0.56
0.62
0.68
0.74
0.80
0.86
0.94
Mean WEIRD score