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

Richard A. Klein	University of Florida	raklein@ufl.edu
Michelangelo Vianello	University of Padua	michelangelo.vianello@unipd.it
Fred Hasselman	Radboud University Nijmegen	f.hasselman@bsi.ru.nl
Sinan Alper	Middle East Technical University	sin.alper@gmail.com
Mark Aveyard	American University of Sharjah	maveyard@aus.edu
Jordan R. Axt	University of Virginia	jaxt@virginia.edu
Štěpán Bahník	University of Würzburg	bahniks@seznam.cz
Mihály Berkics	Eötvös Loránd University, Budapest	berkics.mihaly@ppk.elte.hu
Michael J. Bernstein	Penn State University Abington	mjb70@psu.edu
Olga Bialobrzeska	University of Social Sciences and Humanities	obialobrzeska@swps.edu.pl
Konrad Bocian	University of Social Sciences and Humanities	kbocian1@swps.edu.pl
Mark J. Brandt	Tilburg University	M.j.brandt@tilburguniversity.edu
Katarzyna Cantarero	Institute of Psychology, Polish Academy of Sciences	kcantarero@psych.pan.pl
Zeynep Cemalcilar	Koç University	zcemalcilar@ku.edu.tr
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	armand.chatard@univ-poitiers.fr
Eva E. Chen	The Hong Kong University of Science and Technology	evaechen@ust.hk
Winnee Cheong	HELP University, Malaysia	cheongwn@help.edu.my
Sharon Coen	University of Salford	s.coen@salford.ac.uk
Brian Collisson	Marian University	bcollisson@marian.edu
John G. Conway	University of Florida	john.conway@ufl.edu
Katherine S. Corker	Kenyon College	corkerk@kenyon.edu
Paul G. Curran	Kenyon College	curranp@kenyon.edu
Fiery Cushman	Harvard University	cushman@wjh.harvard.edu
Anna Dalla Rosa	University of Padua	anna.dallarosa@studenti.unipd.it
William E. Davis	Texas A&M University	dbillium@gmail.com
Thierry Devos	San Diego State University	tdevos@mail.sdsu.edu
Canay Doğulu	Middle East Technical University	canaydogulu@gmail.com
Yarrow Dunham	Yale University	yarrow.dunham@yale.edu
Anja Eller	National Autonomous University of Mexico	eller@unam.mx
Carolyn Finck	Universidad de los Andes, Colombia	cfinck@uniandes.edu.co
Mike Friedman	Catholic University of Louvain	mike.friedman@uclouvain-mons.be
Steffen R. Giessner	Rotterdam School of Management, Erasmus University	sgiessner@rsm.nl
Timo Gnambs	Osnabrück University	timo.gnambs@uni-osnabrueck.de
Ángel Gómez	Universidad Nacional de Educación a Distancia	agomez@psi.uned.es
Jesse Graham	University of Southern California	jesse.graham@usc.edu
Jon E. Grahe	Pacific Lutheran University	graheje@plu.edu
Eva G. T. Green	University of Lausanne	eva.green@unil.ch
Matthew Haigh	Northumbria University	matthew.haigh@northumbria.ac.uk
Elizabeth L. Haines	William Paterson University	hainese@wpunj.edu
Marie E. Heffernan	University of Illinois at Urbana-Champaign	marieheffernan@gmail.com
Joshua A. Hicks	Texas A&M University	joshua.hicks@gmail.com
Petr Houdek	Jan Evangelista Purkyne University	petr.houdek@gmail.com
Jeffrey R. Huntsinger	Loyola University Chicago	jhuntsinger@luc.edu
Hans IJzerman	Tilburg University	h.ijzerman@gmail.com
Yoel Inbar	University of Toronto Scarborough	yi38@cornell.edu
Anna Kende	Eötvös Loránd University, Budapest	kende.anna@ppk.elte.hu
Åse H. Innes-Ker	Lund University	ase.innes-ker@psy.lu.se
Wiliam Jiménez-Leal	Universidad de los Andes, Colombia	w.jimenezleal@uniandes.edu.co
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

Victor N. Keller	University of Brasilia	vnfskeller@gmail.com
Nicolas Kervyn	Catholic University of Louvain	nicolas.o.kervyn@uclouvain.be
Lacy E. Krueger	Texas A&M University-Commerce	lacy.krueger@tamuc.edu
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
Samuel Lins	Pontifical Catholic University of Rio de Janeiro	samuel.bezerra.lins@gmail.com
Melissa-Sue John	Worcester Polytechnic Institute	mjohn@wpi.edu
Robyn K. Mallett	Loyola University Chicago	rmallett@luc.edu
Taciano L. Milfont	Victoria University of Wellington	taciano.milfont@vuw.ac.nz
Wendy L. Morris	McDaniel College	wmorris@mcdaniel.edu
Andriy Myachykov	Northumbria University	andriy.myachykov@northumbria.ac.uk
Nick Neave	Northumbria University	nick.neave@northumbria.ac.uk
Austin Lee Nichols	University of Navarra	anichols@unav.es
Susan L. O'Donnell	George Fox University	sodonnell@georgefox.edu
Gábor Orosz	Eötvös Loránd University, Budapest	orosz.gabor@ppk.elte.hu
Rolando Pérez-Sánchez	University of Costa Rica	rolarez@gmail.com
Boban Petrovic	Institute of Criminological and Sociological Research, Belgrade	bobanpetrovi@gmail.com
Ronaldo Pilati	University of Brasilia	rpilati@gmail.com
Monique M.H. Pollmann	Tilburg University	m.m.h.pollmann@tilburguniversity.edu
Erika Salomon	University of Illinois at Urbana-Champaign	salomon3@illinois.edu
Kathleen Schmidt	The University of Virginia's College at Wise	kes7z@virginia.edu
Maciej B. Sekerdej	Jagiellonian University	maciek@apple.phils.uj.edu.pl
Michael A. Smith	Northumbria University	michael4.smith@northumbria.ac.uk
Vanessa Smith-Castro	University of Costa Rica	vanessa.smith@ucr.ac.cr
Agata Sobkow	University of Social Sciences and Humanities, Faculty in Wroclaw	asobkow@swps.edu.pl
Jeroen Stouten	KULeuven	jeroen.stouten@kuleuven.be
Chris N. H. Street	University of British Columbia	c.street@psych.ubc.ca
Jakub Traczyk	University of Social Sciences and Humanities, Faculty in Wroclaw	jtraczyk@swps.edu.pl
David Torres	Universidad de Iberoamerica	datofez@gmail.com
Jordan Theriault	Boston College	jordan.theriault@bc.edu
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
Leigh Ann Vaughn	Ithaca College	lvaughn@ithaca.edu
Alexandra Vázquez	Universidad Nacional de Educación a Distancia	alx.vazquez@psi.uned.es
Catherine Verniers	Clermont Université	catherine.verniers@univ-bpclermont.fr
Mark Verschoor	Tilburg University	m.verschoor@tilburguniversity.edu
Marek A. Vranka	Charles University in Prague	marek.vranka@ff.cuni.cz
Aaron L. Wichman	Western Kentucky University	aaron.wichman@wku.edu
Lisa A. Williams	University of New South Wales	lwilliams@unsw.edu.au
Liane Young	Boston College	liane.young@bc.edu
John M. Zelenski	Carleton University	john_zelenski@carleton.ca
Brian A. Nosek	University of Virginia; Center for Open Science	nosek@virginia.edu

Authors' note: This proposal was supported by the Center for Open Science. The authors declare no conflict of interest with the proposed research. Correspondence concerning this paper should be addressed to Richard A. Klein, University of Florida, raklein@ufl.edu.

Abstract

This project is a unique investigation of a fundamental assumption in psychological science - that effects vary across person and situation. We gathered 28 classic and contemporary effects in psychological science and expect 58 unique samples for one slate of 13 effects, and 57 unique samples for a second slate of 15 effects. In total, the aggregate data will include 114 samples from more than two dozen countries or territories. We will examine the variation in replicability and effect magnitudes across sample and setting meta-analytically to test (a) the replicability of a variety of important effects in social and cognitive psychology, (b) which types of effects show evidence of context sensitivity across samples and setting, and (c) the extent to which that variation exceeds expectations from sampling error.

Word count = 128

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

Contributions of Authors

Coordinated project: Rick Klein, Brian Nosek, Michelangelo Vianello, Fred Hasselman

Designed the study: Štěpán Bahník, Zeynep Cemalcilar, Jesse Chandler, Katherine Corker, Fred Hasselman, Hans IJzerman, Rick 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, Rick 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, Rick 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: TBD

Analyzed data: TBD

Wrote report: TBD

Introduction

A central feature of psychological science is the appreciation of the variation in persons and situations in determining human behavior (Lewin, 1936; Ross & Nisbett, 1991). People differ in their interests, motivations, abilities, desires, and virtually everything else. Simultaneously, variation in situations promote and constrain behavior - sometimes minimizing differences across individuals, and sometimes having an interactive influence with those differences. This is so central to psychological research that investigating variation in behavior by person and situation is a reasonable working definition of social-personality psychology.

The lesson that behavior varies by person and situation is so ingrained that it can be highly accessible for explaining variation in results across investigations of the same phenomenon. Suppose a researcher, Josh Strohminger, conducts an experiment finding that experiencing threat reduces academic performance compared to a control condition. Another researcher, Nina Carp, conducts the same study at her institution and finds no effect. Two possible 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. Alternatively, an uninteresting, 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 positive, or Josh's result could be a false negative. The present proposal will contribute data points toward understanding the contribution of variation in sample and setting to the variation in observing psychological effects.

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

There is a voluminous history of research evidence for effects of variation by particular person characteristics, in particular situations, for particular experimental effects. 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 this effect is stronger in western than eastern cultures (Miyamoto & Kitayama, 2002) and when perceivers lack the cognitive resources to correct for this tendency (Gilbert, Pelham, & Krull, 1988). The standard model of investigating psychological processes is to identify an effect, and then investigate moderating influences that make the effect stronger or weaker.

Simultaneously, there is very little systematic evidence regarding the extent to which persons and situations--or samples and settings--influence the size of psychological effects *in general*. The default assumption is that psychological effects are awash in interactions among many variables. As such, when one confronts different outcomes from similar experiments, the readily available conclusion is that there is a moderating influence by sample or setting that accounts for the difference. That could be true, but it is not necessarily true. Without systematic investigation, it is difficult to know whether variation in effects is a function of persons, situations, or sampling error. If psychological effects vary less across sample and setting than assumed, then the readily available conclusion may be over-applied and the role of sampling error underestimated.

Evidence for the relative contribution of sample, setting, and sampling error to effect estimation is important for gauging the plausibility of Nina and Josh's discrepant findings in the general case. If effects are highly variable across sample and setting, then detecting any particular effect is difficult. Variation in effect sizes will routinely exceed what would be

expected if variation were a function only of 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 factors simultaneously influencing the effect, then it is difficult to isolate moderators to develop theory about the necessary conditions to obtain the effect. In this circumstance, the lack of consistency between Josh and Nina's results might produce collective indifference -- "there are just too many variables and variations to know why there was a difference."

Alternatively, variations in effect sizes may not exceed expected variation due to sampling error. In this case, the observed differences in effects are not indicating moderating influences of sample or setting, they are just a function of imprecision in effect estimation. 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. With small samples, Josh's positive result and Nina's null result will likely have confidence intervals that are overlapping each other leaving little to conclude other than "more data are needed".

It would be very useful to have systematic evidence about the extent to which variation in effect magnitudes exceed expectations of sampling error. If variation exceeding sampling error is pervasive, then it would support the conclusion that any given effect is awash with moderating influences by sample and setting. Alternatively, if variation exceeding expected sampling error occurs only occasionally, then--for any particular effect--assuming that moderating influences are the explanation is unjustified. More evidence would be needed before credibly attributing the differences between investigations to unseen moderators.

The difference between these interpretations is substantial, but there is very little *direct* evidence for which applies to phenomena in psychology. The present proposal examines this question following initial evidence from the "Many Labs" project (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 "Many Labs" project is that there was a small number of effects and there was no reason to presume them to vary substantially across sample and setting. They were just a collection of effects that could have done so. It is possible that the effects in the original "Many Labs" project are more robust and homogenous than the typical behavioral phenomena, or that the populations were more homogenous than initially expected.

The present research proposal is 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 samples, and (4) highly diverse samples. The selected effects are not random nor are they representative, but they do cover a wide range of topics in order to obtain preliminary evidence for the extent to which variation in effect magnitudes is attributable to sample and setting, or sampling error. At minimum, the present data can provide evidence that variation in effect magnitudes exceeds sampling error in (a) all effects investigated, (b) some effects, particularly those that have existing direct evidence for variation by sample, (c) some effects, but limited to those that have existing direct evidence, or, most surprisingly, (d) none of the effects investigated. Each of these outcomes would provide useful knowledge for fostering a broader

understanding of variation in estimating effects.

Research Questions

In Many Labs 2, we will employ an expanded version of the Many Labs paradigm to investigate a substantial number of new effects to learn more about variation in effect magnitudes across sample and setting. In particular, the study will include: (a) effects expected to vary in effect size, (b) effects that are thought to vary across cultural contexts and others that are thought to be invariant, (c) effects that are plausibly contingent on other features of the sample or setting, and (d) effects that have been observed in a variety of samples and settings, and others that are untested. A significant, additional interest is in increasing the precision of the effect size estimate for each of the 28 selected effects.

Overview of Research Design

Sampling Plan

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 must administer their assigned study procedure to at least 80 participants between August 15, 2014 and December 1, 2014.[1] Also, participating labs must meet the technology requirements: the study is to be administered via computers connected to the Internet for standardized administration and data being sent to a central database. Further, each team is required to create a video simulation of study administration to illustrate the features of the data collection setting. Finally, for samples that are not native English speakers, the lab must complete a translation and back translation of the study materials (cf. Brislin, 1970) and assess materials for content appropriateness for the national sample.

All contributors who meet these design and data collection requirements will receive authorship on the final report. Prior to registration of this protocol, 106 teams had volunteered to conduct a data collection, and 8 of those volunteered to administer both slates. This anticipates a minimum of 9,120 participants for the whole study if we recruit no additional data collection sites and each site does the minimum data collection. In the first Many Labs project, all labs completed the data collection, most labs exceeded the minimum sample, and some samples were substantially larger than the minimum. Labs decisions to stop data collection will be based on their access to participants and time constraints. None will have opportunity to observe the outcomes prior to conclusion of data collection.

Selection of Effects

The primary aim of the project is to estimate variability in effect magnitudes across sample and setting for the specified psychological effects. Our sampling plan establishes variability in sample and setting, and our research design includes a variety of psychological effects. After an intensive review we identified 28 effects for inclusion in this study. To obtain a candidate list of effects, we held a round of open nomination of effects and invited submissions for any effect that fit the defined criteria (see the organizing document). Those nominations were supplemented by ideas from the project team, and from direct queries for suggestions to

[1] In reality, review and implementation of this protocol was not finalized until mid-September, so the close of data collection was extended further into December.

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[2], (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 "targeted" for replication because of beliefs or concerns about the effect. Some research teams or sub areas of psychological research were identified as particularly effective at producing short, simple research procedures that tested interesting effects. This increased the likelihood of inclusion for those areas and researchers.

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 alternates. In only one instance did original authors react negatively to inclusion in the study. In this case, because we make no claim about the sample of studies being randomly selected or representative, we removed the effect from the study to avoid conflict.

Procedure

We pretested the amount of time required for each of the 28 effects, and created two slates of 13 and 15 effects that each required approximately 30 minutes to complete including demographics, instructions, and individual difference measures. We divided the studies across slate to be balanced on the criteria above and to avoid substantial overlap in topics. Participating labs will be assigned to complete one of the two slates, with eight labs volunteering to do both slates. As such, we expect that the effects in slates 1 and 2 will be examined by 57 samples each. Effects will be administered by a single experiment script that begins with informed consent, then presents the effects in that slate in a fully randomized order at the level of participants, then does the same for the individual difference measures, and then closes with demographics measures and debriefing. The studies will be conducted following approval of human subjects review boards. The Appendix shows the selected effects and a summary of the two slates on

[2] 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.

some of the selection criteria.

The original completed slates had 32 effects before peer review and pilot testing. One effect was removed during peer review at the request of the original authors. With the remaining 31 effects, we pilot tested both slates with participation among the collaborative team and 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.

Demographics

A few demographics will be included for characterizing each sample and as data for possible moderator investigations.

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

Sex. Participants can select "male" or "female" to indicate their biological sex.

Race/ethnicity. Participants indicate race/ethnicity by selecting from a drop-down menu populated with options determined by the replication lead for each site. Participants can also select "other" and write an open-response. Note that response items will not be standardized as some countries have very different definitions of race/ethnicity, and in some cases these terms have little meaning and the item will be omitted.

Cultural origins. Three items assessing cultural origins used a drop-down menu populated by a long list of countries or territories. Translated versions may just list the most probable responses as determined by the local researcher, in which case "other" will be included as a response option, and participants will be provided 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?

Home town. A single item "What is the name of your home town/city?" with an open response blank will be included as another potential variable of interest for the Huang et al., 2014 effect.

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

Political ideology. Participants rate 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 are 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 report 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 (Adler, Boyce, Chesney, Cohen, Folkman, Kahn, & Syme,

1994). Socio-economic status will be measured with the ladder technique (Adler et al., 1994). Participants are asked to indicate 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 (see here). 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 is also used in Effect 12 in Slate 1 (Anderson, Kraus, Galinsky & Keltner, 2012, Study 3). Participants in that slate will not receive the ladder item a second time. Use of the ladder item in that slate should account for its use as a dependent variable in that study.

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). To further research in this area, the current project will include two items at the end of the study, just prior to demographic items. The first item asks participants "In your honest opinion, should we use your data in our analyses in this study?" and has yes/no response options (Meade & Craig, 2012). The second item is an Instructional Manipulation Check (IMC; Oppenheimer, Meyvis, & Davidenko, 2009), also used in the first Many Labs project (Klein et al., 2014). The IMC will be modified to fit the format of the current project.

Individual Difference Measures

The following individual difference measures will be presented in a randomized order after all target effects are completed, and right before the demographics items. These measures will be particularly useful in tests for moderation of effect sizes.

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 use 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 will constrain the total time available to answer the three questions to 75 seconds. This will likely lower overall performance on average as it is somewhat faster than performance by some participants in pretesting.

Subjective well-being (Veenhoven, 2009). Subjective well-being is 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 is 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 .70 to .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 -- are measured with the Ten Item Personality Inventory (Gosling et al., 2003). Each trait is assessed with two items on seven point response scales from 1 = disagree strongly to 7 = agree strongly. The scale has been translated into several languages including, among others, German (Muck, Hell, & Gosling, 2007), Dutch (Hofmans, Kuppens, & Allik, 2008), Spanish and Catalaan (Renau et al., 2013), Japanese (Oshio, Abe, Cutrone, & Gosling, 2013) and many more (see Gosling, 2014). 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. The contamination subscale is included for Effect 10 in Slate 1. It will not appear in Slate 2.

Overall Analysis Plan

Each effect will be analyzed according to the analysis plan specified below, including decision rules for data exclusion. Descriptively, the primary effect of interest is the variability in effect size for each effect between the sample of samples. The analysis for this interest will follow closely with the procedure described in the first Many Labs paper and will produce a figure similar to Figure 1 of that article: for each effect, (1) an aggregate effect estimate across all samples, (2) a 99% confidence interval for the aggregate effect estimate, (3) display of the effect estimates for each individual sample, and (4) comparative display of the original effect estimate. The latter will probably have two data points in cases that only a subset of the samples or participants are anticipated a priori to be comparable to the original. In that case, the Figure will show the effect size for the restricted sample for the direct attempt to compare with the original effect size, and the effect size of the full sample.

Aggregate examination of the variability in effect estimates will use established meta-analytic statistics - tau2, Q and I2 - to determine if the amount of variability across samples exceeds that expected by random error. Because the study procedures are nearly identical (except for language translations), any variation exceeding random error is likely to be due to effects of sample or setting. This is the primary outcome of interest for each effect included in this study.

In the aggregate analysis, we expect that the effects *a priori* identified as ones that vary across samples and settings (i.e., ones with existing evidence for cultural variation) to show variability exceeding that which can be attributed to random error (i.e., show higher values of I2). Also, upon confirmation of the final study design, we will survey all project contributors for

their predictions of relative average effect magnitude and variation in effect size across samples and settings across the effects. These predictions will be compared with the actual effect sizes and variation in those effect sizes. Further, a collaborative team may conduct a prediction market for the effects included in Many Labs 2. That will be conducted independently of the main report for this project.

In addition to the focal research questions, the order of presentation is an obvious procedural factor that may moderate effect sizes. Across the 30 minute session, effects may weaken if participants tire or prior effects interfere with later effects. Although we did not observe this in the first Many Labs investigation, it is nonetheless a plausible moderator and cannot be ignored. Therefore, we will first examine whether each effect size differs as a function of their placement in the study procedure. To do so, we will select for each effect size the data at rank order K across all locations, where K varies from 1 (presented first) to 16 (presented last). For each value of K we will estimate each effect size, and both average effect size and variability represented as a function of K. With a moderator analysis we will examine if a linear or quadratic trend in order is present. This regression trend indicates whether the overall or individual effect sizes are sensitive to order. If no trend of order is observed in the moderation analysis then reporting the overall effect size is sufficient. Otherwise, we will report the range of effects across orders, with a focus on the effect size when the effect was administered first as the "purest" assessment of the effect magnitude without the influence of fatigue or any particular interference from one or more effects. Examination of specific interference effects is left to the follow-up commentaries on the main article (discussed next).

Additional Analysis via Commentaries

The amassed dataset will be very rich for exploring the individual effects, potential interactions between specific effects, and alternate ways to analyze the aggregate data. Our analysis plan focuses on the big picture and not, for example, exploring potential moderating influences on each of the individual effects. These are worthy analyses, but putting everything into one paper would be overwhelming.

Instead, we proposed an adaptation of the Registered Replication Reports format at *Perspectives on Psychological Science*, the intended outlet for this manuscript. The main paper will be authored by the entire community of researchers contributing to Many Labs 2. As such, micro-publications of each data collection will not be necessary, as is the present format. Instead, we proposed that the Editors solicit commentaries on the main paper that can include new data analysis. These commentaries would likely (though not necessarily) focus on a particular effect and use the rich dataset to examine its boundary conditions and moderating influences.

We believe that the extremely high-powered design of Many Labs 2 offers an opportunity to demonstrate the productive interplay of exploratory and confirmatory analysis strategies. That is, teams that would like to submit commentaries would be given access to one half of the dataset to analyze and write their commentary for editorial review. Those commentaries that are accepted would then be finalized, and the analysis would be subjected to a confirmatory phase. The identical analysis would be conducted on the other half of the data as a strong confirmatory test and reported in the commentary regardless of outcome. A change in the outcome between exploratory and confirmatory phases would not be a basis for changing the editorial decision.

We believe that this process could highlight the importance and interactivity of exploratory and confirmatory approaches to data analysis. Exploratory analysis provides an opportunity to learn from the data, and the subsequent confirmatory analysis provides a strong

test of the discoveries. Finally, we will make the full dataset (plus the two halves used for the exploratory/confirmatory commentaries) and all study materials available publicly at https://osf.io/8cd4r/ so that other teams can use it for their own investigations.

Selected Effects

Next, we describe the selected effects with a summary title of the effect with a citation, an abstract describing the main idea of the original research with the sample size, inferential test, and effect size that is the key result for replication. Then, we present the materials, procedure, and analysis plan. Many of the original studies were conducted with paper and pencil, but all of the replications will be conducted via computer. Also, many of the original studies were conducted in English, but all of the replications will be conducted in the dominant language of each setting of data collection. Any other known differences from the original study are noted in the method description.

The focus of this replication project is estimating the variability in effect magnitudes by sample and setting. As such, we aimed to identify or simplify original study designs that could be tested as simple, two-condition experiments or as correlational results. Some original studies had additional conditions that were relevant for the theoretical purposes of the investigation. In those cases, the replication designs identified the key conditions that are relevant for estimating the effect. Also, in some cases, multiple dependent variables were included in the original design. If the dependent variables could be administered quickly, they were usually retained in the replication designs. When multiple outcomes were included, because they are likely to be correlated outcomes, just one was identified as the primary object for replication and the others as secondary replications. All outcomes will be reported in the final text, but the primary outcome will be the focus for reporting purposes.

SLATE 1

1. 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), $F(1,176) = 20.39$, $MS_E = 5.63$, $p < .001$, $\eta_p^2 = 0.10$. US participants expected the high-SES person to live further north ($M = +0.98$, $SD = 1.85$) than the low-SES person ($M = -.69$, $SD = 2.19$), $t(78) = 3.69$, $p < .001$, $d = .82$, 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].

Materials. Participants will be randomly assigned to read a description of either a high or low SES person. The description of high-SES person will read: “Dr. Bennett lives in the city. He is a wealthy businessman who has travelled the world. He inherited a significant amount of money from a Great Aunt, and was educated at the best schools growing up. He enjoys fine

dining and going to the theater on weekends." The description of the low-SES person will read: "Mr. Bennett lives in the city. He is unemployed. He was born and raised in the city he now calls home. He struggles to pay the rent each month, and dropped out of high school before graduation. He enjoys a good hot dog and a six pack of beers when he can."

Then participants will view a map of a fictional city and will indicate where the person they read about might live by clicking on that location on the map. The map of the fictional city from the original study will be used (originally from Meier et al., 2011). Materials here: https://osf.io/exs7i/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_cNiQVmpTU8Xd1uB.

In the individual difference measures at the end of the slate, participants will report the name of their home town, and answer "Where do wealthier people live in your home town? with response options being north side, south side, or neither.

Analysis plan. The coordinates of the click on the map will be recorded (X, Y) from the top-left of the image. The mean difference between the high and low-SES conditions for north/south location of click (Y) will be compared with an independent samples *t*-test. All participants who indicate an area within the boundaries of the map will be included in the analysis.

The test for replicating the cultural difference observed in Huang et al. will be conducted on a subset of the participants that respond on the wealth in hometown individual difference item that wealthy people tend to live in the North (akin to original U.S. sample) versus wealthy people tend to live in the South (akin to original Hong Kong sample). The entire sample will be used for investigating variation in effects across sample and setting.

Known differences from original. Original participants were asked to guess the purpose of the study afterward, but none did and we will not be including that item.

The original was presented on pencil-and-paper and drew an "X" on the map, whereas the replication will be on a computer and participants will click to indicate the location on the map. With a monitor presentation, this also means the study will be completed on a vertical display as opposed to a horizontal paper. The original authors suggest this may be particularly important because associations between "up" and "good" or "down" and "bad" may interfere with any North/South associations. As such, at eight data collection sites, we will randomly assign participants to complete the slate on a regular monitor or on a Microsoft Surface tablets that is resting on the table like a paper-pencil administration. A focused examination of these sites will test whether this administration format matters.

Lastly, the original analysis emphasized *t*-tests against zero whereas the replication will focus specifically on the difference between conditions (e.g., independent samples *t*-test).

2. 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; $t(65) = 2.00$, $p = .05$, d = 0.50, 95% CI = [-.001, -.988]).

Materials and procedure. Participants will be asked to list their most important long-term goal. Afterwards, participants will be randomly assigned to the structured or random condition. Following the scenario script, participants will answer three questions regarding their willingness to pursue their listed goal and the subjective value of the goal. Materials here: osf.io/nkg5y . Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_dcMjcxf2UCsCd0N

Analysis plan. Following Kay et al. (2014), we will create an index of willingness to engage in goal pursuit for each participant by (1) regressing the mean of the two goal pursuit items on the centered mean of the goal subjective value item, (2) calculating the unstandardized residual for each participant, and (3) add to those the mean value for the self-regulation items measuring willingness to engage in goal pursuit. Then, the two conditions will be compared using an independent samples t-test.

Because of the analysis strategy, any participant with missing data on any one of the three items will not be included in analysis.

Known differences from the original. None known besides sampling and setting.

3. 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 manipulation of font was 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=.051, d=.64 [-.004, 1.28]).

Materials and procedure. Participants will be randomly assigned to complete syllogisms presented in easy- or hard-to-read font. Following Alter et al. (2007), the easy-to-read font will be *black Myriad Web 12-point* and the hard-to-read font will be *10% grey italicized Myriad Web 10-point*. Items will be presented on a single page in a fixed order: instructions, six syllogisms, and a mood item.

The original authors chose six syllogisms based on difficulty determined in previous research (Johnson-Laird & Bara, 1984; Zielinski, Goodwin, & Halford, 2006): two hard (20% correct), two moderate (50% correct), and two easy (85% correct). We will use the same syllogisms. Transient mood will be measured by asking "Please circle the number that best describes your current mood:" on a 7-point scale from " very unhappy" to "very happy". The mood item was included in the original study to evaluate whether disfluency could change mood that would then affect task performance. Additionally, a manipulation check will be added after the task to assess how difficult participants thought the text was to read. Materials here: https://osf.io/sizqu/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_bPWngKJWUb5DAVv.

Analysis plan. Similar to Alter et al. (2007), we will conduct an independent samples t-test to determine whether accuracy in solving moderately difficult syllogisms differ by font condition (fluent versus disfluent). The original study focused on the moderately difficult questions, on the basis that participants' performance could vary enough to detect changes in processing depth.

Our primary analysis strategy will be sensitive to potential differences across samples in

ability on syllogisms. We will first determine 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 will be the basis of computing mean syllogism performance for each participant.

For a direct comparison with the original effect size, only English in-lab samples will be used for two reasons: (1) we cannot adequately control for online participants "zooming in" on the page or otherwise making the font more readable, and (2) a different font may be used in some translated versions because the original font (Myriad Web) may not support all languages. All samples will be included in the investigation of cross-site variability in effect size.

As a secondary analysis, we will use the same two syllogisms from Alter et al (2007) for analysis regardless of performance to perfectly mirror the original analysis.

Known differences from original. The original was done with paper and pencil and it is unknown whether using a computer could affect fluency. However, other evidence suggests that this is not a limiting condition. A fluency effect was demonstrated in a previous study that administered words in a large or small font in an online study, and it was shown that the size of the font influenced participants' meta-memorial judgments (Kornell, Rhodes, Castel, & Tauber, 2011). Furthermore, in a computer-based experiment involving judging the validity of syllogisms, Morsanyi and Handley (2012; Experiment 2) found that font fluency affected participants bias to respond "yes," with those in normal font condition showing a higher rate of endorsing syllogisms as valid compared to the typical bias rate, whereas there was not a difference for those in the nonfluent font condition.

We will also take additional steps for in-lab replications to ensure the computer presentation is suitable. We will ask experimenters to maximize the browser window before participants arrive, and we will obtain information about the monitors used (e.g., display size, aspect ratio, model, resolution, and typical viewing distance) from researchers or automatically recorded meta-data. Researchers will also take two pictures of their monitors, one of the difficult-to-read condition and one of the easy-to-read condition so they are available for future review. Those data will be obtained for potential subsequent analyses to determine if those factors influenced the results.

The original authors hypothesize 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. An additional difference, as noted in the analysis plan, is that a different font from the original may be used in non-English samples when Myriad Web does not support the language.

4. 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 preferences and may also have different moral intuitions and principles. 1,532 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 (r = -.16, p < .0005, d = .32, 95%CI [.27, .38]) or fairness (r = -.21, p < .0005, d = .43, 95% CI [.38, .48]) were deemed more relevant for moral judgment by political liberals than conservatives ("individualizing" aggregate r= -.21, p < .0005, , d = .43, 95%CI [.38, .48]), whereas items that emphasized concerns for the ingroup (r = .12, p < .0005, d = .24, 95% CI [.19, .29]), authority (r = .21, p < .0005, d = .43, 95%CI [.38, 48]), or purity (r = .27, p < .0005, ,

d = .56, 95%CI [.51, .62]) were deemed more relevant for moral judgment by political conservatives than political liberals ("binding" aggregate r = .25, p < .0005, d = 0.52, 95% CI [.46, .57])

Materials and procedure. The moral relevance of five foundations will be measured by 3 items each for a total of 15. Participants will read the prompt, "When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking?". Them, they will rate the 15 moral relevance items in a randomized order on a 6-point scale from "not at all relevant" to "extremely relevant". At the end of the study package, participants will report their political ideology along with the other demographic measures. Materials here: https://osf.io/gdbp8/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_9T7LvQedxUzUFXn

Analysis plan. The 6 items for harm and fairness will be averaged to create a "individualizing" foundations moral relevance score and the 9 items for ingroup, authority, and purity will be averaged to create a "binding" foundations moral relevance score. The relationship between political ideology and the "binding" and "individualizing" aggregates will be calculated using zero order correlations. The primary target of replication is the relationship of political ideology with the "binding" foundations, and the relationship of political ideology with the "individualizing" foundations is a secondary replication. All participants who complete the corresponding measures will be included in analysis.

Known differences from original. We are conducting a simplified version of the original analyses with approval of the original authors (and updated analyses from the original study in line with our analysis for comparative purposes). Also, we altered the text of the political ideology item to more relevant to international samples that may not recognize "liberal" and "conservative" as having the same meaning as in the United States.

5. 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 ($50). In one condition, participants made the choice imagining a low probability (1%) of getting the outcome. In the other condition, participants imagined that the outcome was certain, they just needed to choose which one. When the outcome was unlikely 70% preferred the affectively attractive choice, when the outcome was certain 35% preferred the affectively attractive choice ($\chi^2(1,N=40) = 4.91$), p = .0267, Kramers φ = .35). This result supported the hypothesis that positive affect has greater influence on judgments made under conditions of uncertainty than judgements about definite outcomes.

Materials and procedure. Participants will be randomly assigned to make a choice with either a certain outcome or a 1% chance that the outcome will occur. The certain condition will read as follows: "Imagine that you have the opportunity to either **meet and kiss your favorite movie star** or **receive $50 in cash.**"

In the uncertain condition, participants will read: "Imagine that you have the opportunity to take part either in a lottery that offers a **1% chance to meet and kiss your favorite movie star** or a lottery that offers a **1% chance to receive $50 in cash.**" In both conditions, participants will choose one of the two options. Materials here: https://osf.io/pky9m/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_brBFoL7va7coZ0x.

Analysis plan. A two-way contingency table will be built with certainty condition (low-probability vs. certain) and choice (monetary reward vs. meeting favorite movie star) as factors.

The critical replication hypothesis will be given by a χ^2 test and the effect size by an odds ratio. All participants with valid data on the response will be included in the analysis.

Known differences from original. None known.

6. 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 towards others (1= *not at all*, 7 = *very much*) to conserve water ($M = 4.08$, $SD = 1.56$) compared to the control condition ($M = 5.33$, $SD = 1.30$), $t(76) = 3.86$, $p = .001$, $d = .88$, 95% CI [.41, 1.34]).

Materials. Participants are randomly assigned to the consumer or control condition. They will first read a description of a water shortage caused by a drought in which they are one of four people who share water from the same well. In the scenario, the four people sharing the well are referred to as either "Individuals" or "Consumers.", and receive information about past water usage indicating that they have been using more water than others. The passage reads:

You are [Individual/Consumer] A. You and three other [individuals/consumers] live in and share a water supply for a particular area. Typically, you use the most amount of water with your daily activities (e.g., washing clothes and dishes, flushing the toilet, watering the lawn, bathing, etc.). That is, you use about 160 gallons of water per day. [Individual/Consumer] B uses 140 gallons a day, [Individual/Consumer] uses about 120 gallons a day, and [Individual/Consumer] D uses about 100 gallons per day.

Unfortunately, this year a drought has depleted the normal supply of water in your area such that there is not enough water for you and the other three [individual/consumers] to draw on for your typical needs.

In order to maintain the water supply through this period of time, it is recommended that the total amount of water used should be cut by 25%. This is only a recommendation, so no water restriction has been explicitly imposed.

Below is a chart of each [individual's/consumer's] water usage and the specific activities that each individual would have to alter to cut their water usage.

Consumer	# of gallons per day	-25% of daily water use	Water Usage Alterations
A	160	120	Shorter showers, watering yard less frequently, cleaning dishes less often, flushing toilet only when necessary
B	140	105	Shorter showers, watering yard less frequently, cleaning dishes less often
C	120	90	Shorter showers, watering yard less frequently
D	100	75	Shorter showers

With a single-item from 1 = not at all to 7 = very much, participants answer "How much do you trust the other parties to use less water?" Materials here: https://osf.io/jv46k/ . Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_5cHQkTN7uGunuAt.

Analysis plan. We will compared the mean trust levels between conditions with an independent samples *t*-test. All participants with data will be included in analysis.

Known differences from original. The original experiment included four additional dependent variables: (1) responsibility for the crisis, (2) obligation to cut water usage, (3) how much they viewed others as partners, and (4) how much others should use less water. The central replication will be on the trust variable, while the other four dependent variables will be retained in the procedure but not analyzed for the focal replication.

7. 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. 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 ($M = 3.30$, $SD = 2.62$; $t(56) = 6.66$, $p < .001$, $d = 1.78$, 95% CI=[1.16, 2.39]). 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 ($M = 7.02$, $SD = 3.06$); $t(47) = 1.84$, $p = .069$, $d = .53$.

Materials and procedure. Participants will be randomly assigned to read a scenario in which a student wrote either a pro- or anti-capital punishment essay. Then, they will learn that the student had been assigned to take that position, reading, "Dr. Wallace is teaching a course on international politics at a midwestern university. In his class, students discuss a variety of topics and issues every week. Typically, Dr. Wallace solicits opinions about the topics from the students. In this week's class, the topic was capital punishment. Dr. Wallace asked Steve to write an essay [supporting/opposing] capital punishment. Steve agreed to do so and wrote the essay presented on the previous page." For the essay page and the constraint page, participants will not be allowed to move forwards until 10 seconds have passed.

After that, participants will answer three questions estimating the writer's true attitude, the attitude the writer would take if given the opportunity to speak freely, and the attitude of the average student at a Midwestern university (1 = against capital punishment, 15 = supports capital punishment). Participants will then answer the extent to which they believed the essay author had been constrained by the assignment using a 7-point scale (1 = strongly constrained, 7 = completely free). Finally, participants will indicate how persuasive they thought the essay was on a 7-point scale (1 = not at all persuasive, 7 = very persuasive). Materials here: https://osf.io/e426i/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_bEjoIILOJcahUuV.

Analysis plan. An ANCOVA will compare the mean estimates of the author's true

attitude across the two conditions, covarying for perceived constraint.

Known differences from original. Our focus is on comparing the two key conditions that elicited a cultural difference in the original research. The original experiment varied the diagnosticity of the essay. We have decided to focus on the low-diagnosticity conditions, which is where the significant cross-cultural difference in correspondence bias occurred.

Also, the original experiment included six outcome measures: 1) their estimate of the writer's true attitude, 2) the attitude the writer would express if free to choose, 3) the attitude of the average student at a Midwestern university (changed to the attitude of the average student in each respective country for the purpose of the current project), 4) their own attitude, 5) how much constraint the writer had while composing the essay, and 6) how persuasive they thought the essay was. We will focus on the estimate of the writer's true attitude as the primary target for replication. Item 2 will be examined as a secondary replication, items 3 and 6 as potential moderators, and item 5 as a covariate in the analysis.

The original authors suggested we alter the names and university location to be familiar for the national identity of each sample. Also, the original authors suggested that, for samples in countries without the death penalty, the prompt for the essay be changed from "I think that capital punishment should be abolished" to "I think that capital punishment should not be legalized." Finally, the original study was pen and paper. In adapting the task to an online version, the original authors suggested we make participants spend at least 10 seconds reading the essay and the explanation of constraint. We have done this by disabling the continue button for 10 seconds.

The original authors also suggested a few caveats that may produce cross-cultural variation in the effect observed from our implementation aside from differences in correspondence bias. Specifically, the authors suggested there may be differences in how participants read and interpret the situational constraint information (e.g., samples may vary in their familiarity with the seminar course format) and essays (e.g., samples may vary in their perception of the strength of the essays or familiarity with the death penalty). Finally, the authors suggested that presenting the study in a package of other studies could disrupt the effect. Our tests of order effects should address this. However, the original authors added that even when the study appears first, participants may respond differently because they are expecting to do more tasks afterwards. Our design does not address this possibility.

8. 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 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 = .41, $t(39)$ = 3.39, p = .002, r = .48). Disgust sensitivity was related to judgments of greater intentionality in the gay kissing condition, β = .79, $t(19)$ = 4.49, p = .0003, r = .72. and not the kissing condition, β = -.20, $t(19)$ = -.88, p = .38, r = .20. The correlation in gay kissing condition was stronger than the correlation in the kissing condition, z = 2.11, p = .03, d = .64, 95% CI=[.31, .96]. 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 intentional ratings is the target of direct replication.

Materials and Procedure. Participants will be randomly assigned to read one of the two versions of the following scenario: "A director was working on a music video. His assistant said: 'I took a look at the first cut of your video, and it looks to me like some of the images in it will encourage couples [homosexual men] to French kiss in public.' The director said: 'Look, I know that it will be encouraging couples [homosexual men] to French kiss in public, but I don't care at all about that. I just want to make a video that will increase sales of the album.'He included the images in the video. Sure enough, it encouraged couples [homosexual men] to French kiss in public."

Next, participants will be asked a condition matched set of three questions in a fixed order. First, "Did the director *intentionally* encourage couples [homosexual men] to French kiss in public?" answering on a 7-item scale from 1 = *not at all* to 7 = *definitely*. Second, "Is there anything wrong with couples [homosexual men] French kissing in public?" with a yes or no response. And, third, "Was it wrong of the director to make a video that he knew would encourage couples [homosexual men] to French kiss in public?" with a 7-item scale from 1 = *not at all* to 7 = *definitely*.

Finally, participants will complete the 25-item revised Disgust Sensitivity Scale (Olatunji, et al., 2007) among the individual difference measures of the procedure to separate it from the experimental manipulation. Materials here: https://osf.io/bfhp7/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_0HDz8AMWXFr9yvz.

Analysis plan. The five items of the contamination subscale of the Disgust Sensitivity Scale-Revised will be averaged to create a single index of disgust sensitivity. For the primary analysis, we will compute a correlation between disgust sensitivity and assessments of the director's intentionality in both the gay kissing and kissing conditions, and then compare the correlations with an *r*-to-*z* transformation. The other two outcome measures will be examined as secondary analyses following the same analysis strategy. All participants with relevant data will be included in analysis.

Known differences from original. The original study used 8-item short form of the disgust sensitivity scale. The original authors suggested using 25-item revised version because of its improved psychometric properties. In pretesting, the full 25-item scale was taking more time than desired. The original authors approved a revision using the 5-item contamination subscale.

9. 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%; $t(197.5) = 2.12$, $p = .03$, $d = 0.30$, 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.

Materials and procedure. Participants see a picture of a smartphone with either the model number "P17" or "P97" on the phone's display and read some background information about the smartphone. The text has been updated from the original to reflect more recent smartphones, and reads: "The Sony Ericsson P17 [P97] is an Android-powered smartphone with a 5.2" Full HD resolution display. The P17 [P97] is designed to be lightweight and compact, and features the most up-to-date processor and battery components. Purchasing a P17 [P97] allows

one to use Sony cloud storage app for a year for no additional charge. The Sony Ericsson P17 [P97] has also been made to be compatible with other Sony products, making transferring data from them to your Ericsson P17 [P97] as easy as a click of a button."

Participants learn that the phone will be introduced in the U.S. and Europe and then estimate the percentage of the smartphones that would be sold in the U.S. For participants in Asia, the phone will be planned to appear in Asia and the U.S. and they will estimate Asia sales. For participants in Europe, the phone will be planned to appear in Europe and the U.S. and they will estimate European sales. For participants in other regions, the regions will be the closest and second closest of U.S., Europe, and Asia and they will estimate the sales in the closest region. Materials here: https://osf.io/5j63p/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_2sJjfkN7AOd8xsp.

Analysis plan. The means for the P97 and P17 groups will be compared with an independent samples *t*-test. Participants whose answers will not fall between 0 and 100 will be excluded from analysis.

Known differences from original. The original study was conducted in the U.S. and had participants consider sales between U.S. and European markets. The replication will match markets with location as described above. The pictures and descriptions have been updated to reflect more modern smartphones. The original authors see no reason why the change to the phone image should affect the result, provided that the numbers remain salient, but it remains an untested assumption.

The authors avoided administering these studies on computer—and instead used only paper-and-pencil presentation—to avoid the possibility that the numeric keys on the keyboard might serve as numeric primes. That said, this methodological decision was based on speculation, and the authors report never having tested the influence of administration mode systematically. To test this factor, 10 sites will administer this as a paper-pencil task.

10. 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 larger families, resources have to be shared more frequently, facilitating cooperation and the development of a prosocial orientation (*sibling-prosocial 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; $F(2, 535) = 4.82$, $p < .01$, ds = .287 [.095, .478] and .210 [-.045, .465] respectively).

Materials and Procedure. Recent advances in measurement of SVO has introduced an alternative measure that has some psychometric advantages compared to the triple dominance measure. The SVO slider measure will be incorporated into the present replication (Murphy, Ackermann, & Handgraaf, 2011).

Participants will complete the SVO slider measure, then list how many older siblings, younger siblings, brothers, and sisters they have. The SVO slider measure consists of a series of six decomposed games in which participants select from a range of possible pairs of payoffs for themselves and a fictional other (Murphy et al., 2011). Materials here: https://osf.io/wkhit/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_4TxC1myNGotHyG9.

Analysis plan. The current replication focuses only on the observed direct positive correlation between greater prosocial orientation and number of siblings. Participants must respond to all 6 items of the SVO slider and have valid data for the two questions asking about older and younger siblings, to be included in the analysis. Total number of siblings will be obtained by adding the number of younger and older siblings. SVO slider scores will be scored in according to the procedure recommended by Murphy et al. (2011). The resulting SVO slider score will be correlated with the total number of siblings for the critical test.

Known differences from original. The original demonstration used a triple dominance measure of social value orientation with three categorical values. In discussion with the original author, the SVO slider was identified as a useful replacement to yield a continuous distribution of scores.

11. A DISSOCIATION BETWEEN MORAL JUDGMENTS AND JUSTIFICATIONS (Hauser, Cushman, Young, Kang-Xing & Mikhail, et al., 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 but the train kills one person instead of five. As a GREATER GOOD, a person pushes a fat man in front of a train, killing him, to save five people. While 89% of subjects judged the action in the foreseen side effect scenario as permissible (95% CI [.87, .91]), only 11% of subjects in the greater good scenario judged it as permissible (95% CI [.09, .13]). The difference between the proportions was significant. (χ^2 [1, N = 2646] = 1615.96, p < 0.001), w = 0.78, d = 2.51, 95% CI [2.22, 2.86], providing evidence for the principle of the double effect.

Materials and procedure. This study is replicated in Slate 1 and Slate 2 using different scenarios. Participants will be randomly assigned to one of two test scenarios. We will use two scenarios out of the original four described in Hauser et. al (2007). Participants in foreseen side effect condition read the following: "Denise is a passenger on a train whose driver has just shouted that the train's brakes have failed, and who then fainted of the shock. On the track ahead are five people; the banks are so steep that they will not be able to get off the track in time. The track has a side track leading off to the right, and Denise can turn the train onto it. Unfortunately there is one person on the right hand track. Denise can turn the train, killing the one; or she can refrain from turning the train, letting the five die." Then they will respond will a yes or no to the question, "Is it morally permissible for Denise to switch the train to the side track?"

Participants in the means to a greater good condition will respond to this scenario: "Frank is on a footbridge over the train tracks. He knows trains and can see that the one approaching the bridge is out of control. On the track under the bridge there are five people; the banks are so steep that they will not be able to get off the track in time. Frank knows that the only way to stop an out-of-control train is to drop a very heavy weight into its path. But the only available, sufficiently heavy weight is a large man wearing a backpack, also watching the train from the footbridge. Frank can shove the man with the backpack onto the track in the path of the train, killing him; or he can refrain from doing this, letting the five die." Then they will respond yes or no to the question, "Is it morally permissible for Frank to shove the man?".

After responding to the scenario, participants will be asked an additional question assessing any prior experience with the task. Materials here: https://osf.io/cnk7z/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_aXaDsDUF9HvCgEB.

Analysis plan. Subjects will be excluded from all analyses if they take fewer than four seconds to read and respond to either of the target scenarios. For the key confirmatory test comparing with the original effect size, we will include only participants that indicate having no prior experience with the task. The original authors suggested the effect may be weaker in participants with prior exposure. Prior exposure will be investigated as a moderator for the other analyses. A two-way contingency table will be built with Scenario (Means vs. Side-effect) and Response (Yes vs. No) as factors. The critical replication hypothesis will be given by a one tailed chi square test and the effect size by an odds ratio.

Known differences from original. The original research had 19 scenarios divided into four sets. Participants were randomly assigned to one of the sets which contained 4 scenarios, 3 moral dilemmas and one control scenario. We chose to present participants with only one target scenario to save time and since the analyses in the original were between-subjects. Also, the original study had a control condition at the beginning in which there were no people on the alternate track, so switching was obviously permissible. This condition is rarely present in research with this common scenario and is removed for the replication.

Given that this paradigm is widely known, the original authors suggested the effect may be weaker for participants who have previously been exposed to this sort of task. So, we have included the additional item assessing participants' prior knowledge of the task. The direct comparison with the original effect size will be on the subsample that is not familiar with the task. The investigation of variation across sample and setting will include all participants other than those who were excluded initially for responding in less than four seconds.

12. 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. 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 ($F(1,224) = 4.73$, $p = .03$) such that participants made to feel high in sociometric status had higher subjective well-being than those in the low sociometric status condition, $t(115) = 3.05$, $p = .003$, $d = .57$, 95% CI [.19, .94]. There were no differences between the two socioeconomic conditions, $t(109) = .06$, $p = .96$, $d = .01$.

Materials and Procedure. Participants will be randomly assigned to read a prompt to make them feel either high or low in sociometric status. In the high-sociometric condition, participants read, "Think of the ladder above as representing where people stand in the important groups to which they belong. For example, these can include their groups of friends, family, work group, etc... Now please compare yourself to the people at the very bottom rung of the ladder. These are people who have absolutely NO RESPECT, NO ADMIRATION, and NO INFLUENCE in ALL of their important social groups. In particular, we'd like you to COMPARE YOURSELF TO THESE PEOPLE in terms of your own respect, admiration, and influence in your important groups." In the low-sociometric condition, participants read, "Think of the ladder above as representing where people stand in the important groups to which they belong. For example, these can include their groups of friends, family, work group, etc... Now

please compare yourself to the people at the very top rung of the ladder. These are people who are the MOST RESPECTED, the MOST ADMIRED, and the MOST INFLUENTIAL in ALL of their important social groups. In particular, we'd like you to COMPARE YOURSELF TO THESE PEOPLE in terms of your own respect, admiration, and influence in your important groups."

Then, participants will write a short response to the following prompt: "Now imagine yourself in a getting acquainted interaction with one of these people. Think about how the SIMILARITIES AND DIFFERENCES BETWEEN YOU might impact what you would talk about, how the interaction is likely to go, and what you and the other person might say to each other. Please write a brief description about how you think this interaction would go." Then, participants will report which rung they occupied for the relevant status on a 10-rung ladder as a manipulation check. Finally, participants will complete three dependent measures in a fixed order: Satisfaction With Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985) and the Positive and Negative Affect Schedule (Watson, Clark, & Tellegen, 1988). All materials were provided by the original authors and are available here: https://osf.io/thcj9/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_b1lMXQGoBWcVLet.

Analysis plan. Following the original authors, the three dependent measures will be standardized and averaged into a single index of subjective well-being. The mean difference in subjective well-being between high and low-sociometric status conditions will be tested with an independent-samples *t*-test. All participants with data will be included in the analysis.

Known differences from original. We are using only the high- and low-sociometric status conditions and excluding the high- and low-socioeconomic status conditions that showed no differences in the original study.

13. 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%; $F(1,312) = 49.1$, $p < .001$, $d = .79$, 95% CI [.56, 1.02] for the main effect of experimental condition; meta-analysis (random effects model) of scenario effect sizes: $d = .66$). A later meta-analysis revealed that this effect is robust and moderate in size across a variety of paradigms ($r = .31$, Mullen et al., 1985).

Materials and Procedure. This study is replicated in Slate 1 and Slate 2 using different scenarios. In Slate 1, participants will be presented with the "supermarket" vignette.

Supermarket Story. "As you are leaving your neighborhood supermarket a man in a business suit asks you whether you like shopping in that store. You reply quite honestly that you do like shopping there and indicate that in addition to being close to your home the supermarket seems to have very good meats and produce at reasonably low prices. The man then reveals that a videotape crew has filmed your comments and asks you to sign a release allowing them to use the unedited film for a TV commercial that the supermarket chain is preparing."

Following the vignette, participants will be asked three questions: (1) What % of your peers do you estimate would sign the release?, (2) What % would refuse to sign it? [Total % should be 100%], and (3) Would you sign the release or refuse to sign it? . Materials here: osf.io/4my2z. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_025nTEnKVG4ul6t.

Analysis plan. An independent samples *t*-test will be conducted with participants' choice (sign release/refuse to sign) as the IV and participant estimate of % of peers who would sign the release as the DV. Note that participants self-select whether to sign or refuse to sign the release, so it is not random assignment to levels of the independent variable. Participants will be included in the analysis if they respond to all three questions and their estimate for the DV (e.g., "what percent of your peers would sign the release") falls between 0-100.

Known differences from original. Following the scenario estimates, Ross and colleagues also asked participants to predict the personality of the typical person who would choose each of the two alternatives on four dimensions (shyness, adventurousness, cooperativeness and trust). We are not including this secondary assessment. The personality prediction variables came after the variable of interest, thus the method of testing the effect of interest is effectively the same.

SLATE 2

14. 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.

Materials and Procedure. In Slate 2, participants will be presented with the "traffic ticket" vignette (the "supermarket" vignette will be administered to participants in Slate 1).

Traffic Ticket Story. "While driving through a rural area near your home you are stopped by a county police officer who informs you that you have been clocked (with radar) at 38 miles per hour in a 25-mph zone. You believe this information to be accurate. After the policeman leaves, you inspect your citation and find that the details on the summons regarding weather, visibility, time, and location of violation are highly inaccurate. The citation informs you that you may either pay a $80 fine by mail without appearing in court or you must appear in municipal court within the next two weeks to contest the charge."

Following the vignette, participants will be asked three questions: (1) What % of your peers do you estimate would pay the $80 fine by mail?, (2) What % would go to court to contest the charge? [Total % should be 100%], and (3) Would you pay the $80 fine by mail or appear in court? Materials here: osf.io/4my2z. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_0GJy9Iu0PYFJWkd.

Analysis plan. An independent samples *t*-test will be conducted with participants' choice (pay the fine/appear in court) as the IV and participant estimate of percent of peers who would pay the fine as the DV. Note that participants self-select whether to pay the fine or appear in court, so it is not random assignment to levels of the independent variable. Participants will be included in the analysis if they respond to all three questions and their estimate for the DV (e.g., "what percent of your peers would pay the fine by mail") falls between 0-100.

Known differences from original. The original traffic ticket scenario included a $20 fine. The fine has been adjusted to $80 to reflect inflation. Following the scenario estimates, Ross and colleagues also asked participants to predict the personality of the typical person who would choose each of the two alternatives on four dimensions (shyness, adventurousness,

cooperativeness and trust). We are not including this secondary assessment. The personality prediction variables came after the variable of interest, thus the method of testing the effect of interest is effectively the same.

15. 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 participants in the short line condition (M = 4.62, SD = 0.81; $t(62) = 2.20$, $p = .03$, d=.55, 95% CI [.049, 1.06]. This result was interpreted as showing that people associated vertical position with power, higher is more powerful.

Materials and Procedure. Participants will receive the following instructions: "In this next part you will be asked to evaluate the manager of a company based on very little information." On the next page, participants will read the following about the manager: "In the following you will see company A and a picture of a Manager A of this company. The average gross salary of the employees of company A is about 49,000 dollars. The company has 126 employees." Participants will then see a picture of the fictional manager and an organization chart with the manager at the top with a vertical line connecting to his team either 2 or 7 cm long. Then, they will respond to following items: (1) I think that Manager A is dominant, (2) I think that Manager A has a strong leader personality, (3) I think that Manager A is self-confident, (4) I think that Manager A has a lot of control in the company, and (5) I think that Manager A holds a very high status within the company on a 7 point scale from 1 = totally disagree to 7 = totally agree. Materials here: https://osf.io/79cjv/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_aVhzqs77Q05Ejqd.

Analysis plan. Responses to the five dependent measures will be averaged and an independent samples t-test will compare mean power rating between the 2 cm and 7 cm conditions. All participants who complete at least one item of the dependent measure will be included in the analysis.

Known differences from original. The original presented employee wages in Euros; that will be converted to other currencies as needed at the current exchange rate and adjusted when deemed necessary by the lead site researcher to maintain psychological equivalence (the rounded exchange rate for USD is presented above). In addition, the authors noted there may be an effect of presentation order. As with all effects, we will examine whether presentation order influenced effect size beyond what is expected by chance.

16. 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, $p = 7.4*10^{-7}$, 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.

Materials and procedure. Participants will receive one of two scenarios from the original with dollar amounts approximately adjusted for inflation and the consumer items being replaced with a ceramic vase and a wall hanging. Specifically, one condition will read: "Imagine that you are about to purchase a wall hanging for $250, and ceramic vase for $30. The salesman informs you that the ceramic vase you wish to buy is on sale for $20 at the other branch of the store, located 20 minutes drive away. Would you make the trip to the other store?"

The second condition will read: "Imagine that you are about to purchase a ceramic vase for $30, and a wall hanging for $250. The salesman informs you that the wall hanging you wish to buy is on sale for $240 at the other branch of the store, located 20 minutes drive away. Would you make the trip to the other store?" Participants will respond "Yes, I would go to the other branch" or "No, I would not go to the other branch." Materials here: https://osf.io/8t9ha/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_aW8rIyGPNQ2Wwh7.

Analysis plan. A two-way contingency table will be built with Price condition ($20 vs. $240) and Response (Yes vs. No) as factors. The critical replication hypothesis will be given by a χ^2 test and the effect size by an odds ratio. All participants with valid responses will be included in analysis.

Known differences from the original. In consultation with the original author, dollar amounts have been adjusted to be more appropriate for 2014. The stimuli were also replaced with consumer items that are relevant in 2014 and plausibly sold by a single salesperson. Further, for replications outside of the U.S., we will use amounts in the local currency that the replication team judges to be psychologically equivalent to these values. The default will be equivalence by exchange rate but may be adjusted further if there are substantial differences in wealth for the available sample.

17. 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 a different scenario. In Slate 2, participants will be 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 [.69, .74]), and 56% of subjects judged the action in the means to a greater good scenario as permissible (95% CI [.53, .59]). The difference between the proportions was significant. ($\chi^2[1, N = 2612] = 72.35$, $p < 0.001$), $w = 0.17$, $d = .34$, 95% CI [.26, .42].

Materials and Procedure. Participants will respond to one of two moral dilemmas. Unlike those described in Effect 11, these moral dilemma scenarios will be accompanied by an illustration of the situation.

In the greater good condition, participants will read the following: "Ned is taking his daily walks near the train tracks when he notices that the train that is approaching is out of control. Ned sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and they will not be able to get off the tracks in time. Fortunately, Ned is standing next to a switch, which he can throw, that will temporarily turn the train onto a side track. There is a heavy object on the side track. If the train

hits the object, the object will slow the train down, thereby giving the men time to escape. Unfortunately, the heavy object is a man, standing on the side track with his back turned. Ned can throw the switch, preventing the train from killing the men, but killing the man. Or he can refrain from doing this, letting the five die." Then they will answer, “"Is it morally permissible for Ned to throw the switch?".

In the foreseen side-effect condition (Scenario 4), participants will read: "Oscar is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Oscar sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Fortunately, Oscar is standing next to a switch, which he can throw, that will temporarily turn the train onto a side track. There is a heavy object on the side track. If the train hits the object, the object will slow the train down, thereby giving the men time to escape. Unfortunately, there is a man standing on the side track in front of the heavy object, with his back turned. Oscar can throw the switch, preventing the train from killing the men, but killing the man. Or he can refrain from doing this, letting the five die.". Then they will answer, "Is it morally permissible for Oscar to throw the switch?".

After responding to the scenario, participants will be asked an additional question assessing any prior experience with the task. Materials here: https://osf.io/ci864/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_5byLtAGiE0hAtjT.

Analysis plan. Participants will be excluded from all analyses if they take fewer than four seconds to read and respond to either of the target scenarios. For the key confirmatory test comparing with the original effect size, we will include only participants that indicate having no prior experience with the task. The original authors suggested the effect may be weaker in participants with prior exposure. Prior exposure will be investigated as a moderator for the other analyses. A two-way contingency table will be built with Scenario (Greater Good vs. Foreseen Side-effect) and Response (Yes vs. No) as factors. The critical replication hypothesis will be given by a one tailed chi square test and the effect size by an odds ratio.

Known differences from original. The original research had 19 scenarios divided into four sets. Participants were randomly assigned to one of the sets which contained 4 scenarios, 3 moral dilemmas and one control scenario. We chose to present participants with only one target scenario to save time and since the analyses in the original were between-subjects. Also, the original study had a control condition at the beginning in which there were no people on the alternate track, so switching was obviously permissible. This condition is rarely present in research with this common scenario and is removed for the replication. Given that this paradigm is widely known, the original authors suggested the effect may be weaker for participants who have previously been exposed to this sort of task. So, we have included the additional item assessing participants’ prior knowledge of the task. The direct comparison with the original effect size will be on the subsample that is not familiar with the task. The investigation of variation across sample and setting will include all participants other than those who were excluded initially for responding in less than four seconds.

18. 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 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 ($M = 2.53$, $SD = 2.24$; $t(116) = 2.15$, $p = .034$, $d = 0.39$).

Materials and procedure. Participants will be randomly assigned to read one of two scenarios. Both scenarios start the same: "Imagine that you are in a large lecture with a few hundred students and you are sitting in the middle section, a little more than half-way back in the room. The professor asks a question about the readings, but no one raises his or her hand to answer."

In the tempting fate version, the scenario continues with "You have not done the reading and feel confident that you would not be able to answer the question.", while in the control version it continues with "You have done the reading and feel confident that the professor would like your answer, but prefer not to volunteer answers in large classes." Both scenarios end with the final sentence: "The class sits in silence for two minutes before the professor explains that if no one volunteers, he will choose someone."

After reading the scenario, the belief that tempting fate is bad luck is measured with the question, "How likely do you believe it is that the professor will call on you?" on a 10-point scale ranging from 1 = "Not at all likely" to 10 = "Extremely Likely". Materials here: https://osf.io/3nkev/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_e8OVts4kSX6qSb3.

Analysis plan. The two groups will be compared with an independent samples *t*-test. All participants that answer the dependent measure will be included in analysis. The primary confirmatory test for comparing the original and replication effect size will be based on only the samples using undergraduate students. We will examine gender as a possible moderator of the effect in a supplemental, exploratory analysis.

Known differences from original. The original study design included self and other scenarios. No self-other differences were found. With the original author's approval, we limited the study to the two self conditions.

19. 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. The authors found no main effect of condition across cultures: $\beta = -0.13$, OR = 0.88, d=.10, $t(101) = 0.71$, $p = .48$. Among Americans, there was no difference between construing personal (M = .83, SD = .15) and interpersonal actions (M = .82, SD = .14) as choices, $t(88) = .39$, $p = .65$, $d = .04$. However, Indians were less likely to construe personal actions (M = .61, SD = .26) than interpersonal actions (M = .71, SD = .26) as choices, $t(126) = -3.69$, $p = 0.0002$, $d = .33$.

Materials and procedure. Participants were randomly assigned to personal- or interpersonal-choice conditions. In the personal-choice condition, participants had to recall eight actions that were mostly self-focused (e.g., participants were asked to recall the last time they

made a purchase for themselves). In the interpersonal-choice condition, participants had to recall eight matched actions that involved other people (e.g., participants were asked to recall the last time they made a purchase for someone else). Following the recollection of each action, participants indicated whether it constituted a choice. For each item participants also rated the importance of the action on a 7-point scale (Not at all important; Slightly important; Somewhat important; Moderately important; Quite important; Very important; Extremely important). Materials here: osf.io/pd4ac. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_dcXCkJscEmiTEc5.

Analysis plan. We will conduct a 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 will be 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.

Because some survey questions may be less suitable for non-student samples, we will only include university data collections in the primary confirmatory analysis to be compared with the original effect sizes. Data for all participants will be included to examine variability across sample and setting. However, participants must respond to all choice and importance of choice questions to be included in the analysis. The target effect size for replication will be the the results obtained for participants from labs in India, to compare to the effect found in the Indian sample in the original (Indian participants were more likely to construe interpersonal actions as choices than personal actions). Although we only have few labs from India, we are making extra efforts to recruit many participants in those labs. We anticipate that this effect will vary by sample, particularly in line with the original demonstration of cultural differences.

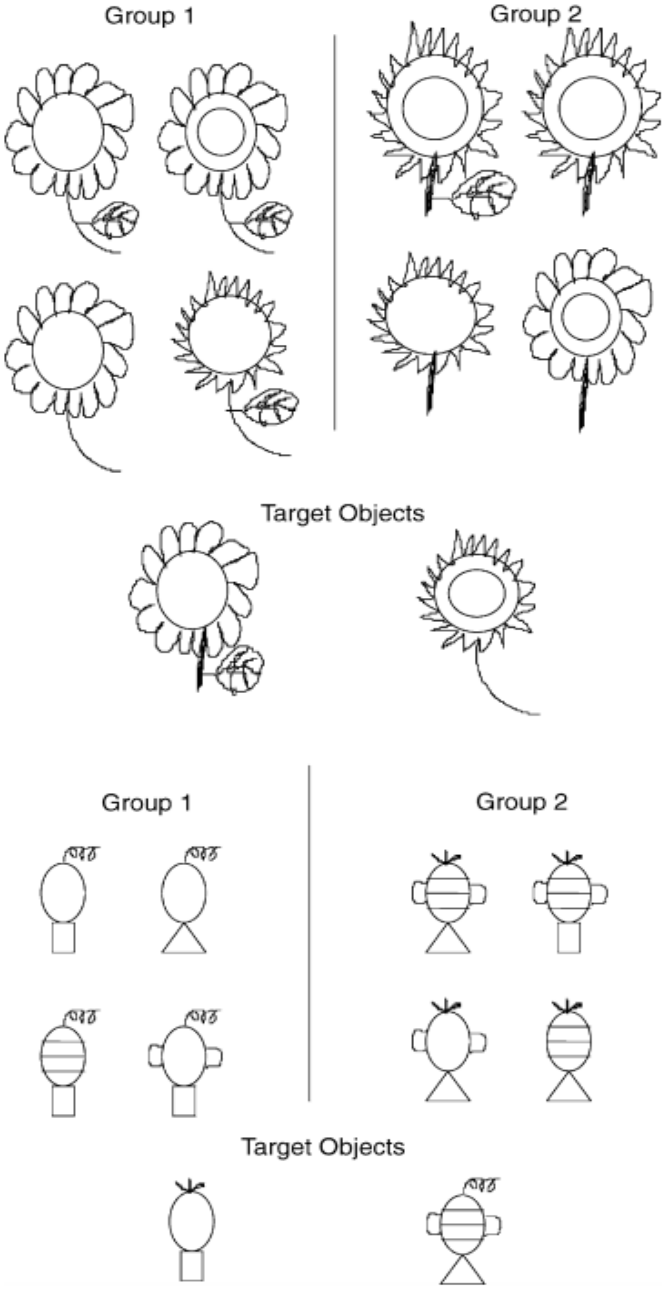
Known differences from original. Our analysis plan differs from the original in order to obtain an effect size within each of the dozens of samples.

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

Western thinking may be more rule based than East Asian thinking. 52 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"; ⅔ of sample) or similarity judgment (decide "which group the target object is most similar to"; ⅓ of sample) condition.

All participants categorized targets into two alternative groups of 4 exemplars. Both targets and group exemplars were defined according to 4 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.

Figure 1. Examples of Targets and Groups.



When asked "which Group the target object *belongs* to"participants across all three cultures preferred to classify based on rule (M = 67%) rather than on family resemblance (M = 33%; $F(1, 100) = 44.40$, $p < .001$, $r = .55$). When asked "which group the target object is more *similar* to", European Americans gave many more responses based on the unidimensional rule (M = 69%) than on family resemblance (M = 31%), $t(17) = 3.68$, $p = .002$, $d = 1.79$, 95% CI=[.64, 2.89]. On the contrary, East Asians gave fewer rule-based responses than family resemblance responses ($M_{rule} = 41\%$ vs. $M_{family} = 59\%$), $t(17) = 2.09$, $p = .05$, $d = 1.01$. Asian Americans were intermediate, having no preference for one rule over the other ($M_{rule} = 46\%$ vs. $M_{family} = 54\%$), $t < 1$.

Materials and Procedure. Participants will categorize target objects into one of two groups. In the belonging condition, participants will receive the instruction: "Which group does the target object belong to?" In the similarity condition, participants will receive the instruction: "Which group is the target object more similar to?" The instructions will end saying "Take your time while responding, but do not spend too much time on any single item."

Participants will judge 20 targets all with the same condition. All materials were provided by the original authors and are publicly available. In one group, all exemplars had 1 feature in common with each other and with the target. In the other group, there was no feature in common among all exemplars and the target, but 1 exemplar had 3 features in common with the target and 3 exemplars had two features in common with the target (see Figure 1). Materials here: osf.io/y3e7g. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_dhumIQeN4X1gVzD.

Analysis plan. We will compute for each subject the percentage of rule-based responses and test whether the mean of the two experimental groups (belong vs similar) on this DV is equal with a *t*-test for independent samples. The effect size will be given by a standardized mean difference. All participants with data will be included in analysis.

For additional analysis, a few items about cultural origins of participants and their parents are present in the individual differences assessment. These could be particularly useful for follow-up moderator analysis.

Known differences from original. The replication will use the same stimuli as the original but the implementation will be slightly different. In the original, participants assigned objects to categories by pressing a key on the keyboard, and the script then advanced automatically to the next trial. In our replication, participants will categorize the object by selecting from a multiple-choice list and will advance the page by clicking "Continue".

The original design had a ⅔ versus ⅓ split in assignment to condition. In consultation with a reviewer, we changed to equal weighted random assignment. Also, the original study had a practice trial, and the replication does not.

It is worth noting that another study in this slate also involves similarity judgements (Tversky & Gati, 1978) though they are dissimilar in content. It will be instructive to test whether the order between those two studies makes a difference for either one.

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. 83 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; $t(82) = 3.13$, $p = 0.002$, $d = .69$, 95% CI [.24, 1.14]), despite the gift being objectively less expensive.

Materials and Procedure. Participants will be asked to imagine that they were about to leave the country and had received a goodbye gift from a friend. Participants will be randomly assigned to the scarf or coat condition. The scarf scenario reads "It is a wool scarf, from a

nearby department store. The store carries a variety of wool scarves.The worst costs $10 and the best costs $100. The one your friend bought you costs $90." The coat scenario reads "It is a wool coat, from a nearby department store. The store carries a variety of wool coats. The worst costs $100 and the best costs $1,000. The one your friend bought you costs $110".

Following the scenario, participants will answer a question about the generosity of gift giver on a scale from 0 to 6, where 0 indicates "not generous at all" and 6 indicates "extremely generous". Materials here: https://osf.io/c4v8x/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_cI3khFhWnIMKE3b.

Analysis plan. The two conditions will be compared with an independent samples *t*-test with rated generosity of gift giver as the dependent variable. All participants with data will be included in the analysis.

Known differences from original. The original study included two additional questions for which statistics were not reported; one about their happiness about receiving the gift and one about the perceived expensiveness of the gift. The two additional questions were described in a footnote as showing the same effect as the generosity item. Dollar values will be approximately inflation adjusted to 2014 dollars. When necessary, dollars will be converted to the primary currency of the data collection site according to the exchange rate and adjusted as deemed necessary to maintain psychological equivalence (as determined by the local researcher).

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, $t(68) = 3.32$, $p = .001$, $d = .80$, 95% CI [.31, 1.30]). 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, $t(68) = 2.20$, $p = .03$, $d = .53$). 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, $t(68) = 3.49$, $p = .001$, $d = .85$).

Materials and procedure. Participants will be randomly assigned to read a scenario in which either a man or a baby commits an action that affects the other. For example, in the condition where the baby commits the action, the participant sees:

Sam

Roger

Imagine that Sam pushes a tray of glasses off a table. They shatter and one of the shards cuts into Roger's leg.

Then, participants will then complete 3 questions: (1) "How responsible is [person who committed action] for his behavior?", (2) "How intentional is [person who committed action]'s behavior?", and (3) "How much pain does [person who did not commit action] feel when he gets cut?"- responding to each on a scale from 1 ("Not at all responsible"/ "Completely unintentional"/ "No pain at all") to 7 ("Fully responsible"/ "Completely intentional"/ "Extreme pain"). Materials here: https://osf.io/szg3n/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_3ldCDrWXgDQsujz.

Analysis plan. We will compare the means on perceived responsibility between conditions with an independent samples *t*-test for the responsibility item. The intentionality item will be analyzed the same way as a secondary analysis. All participants with data will be included in analysis.

Known differences from original. As did the original study, we include all three dependent variables. However, for the aggregate analyses, we will use only the responsibility item and will report results for the other two items as secondary results.

23. 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; $F(1,25) = 6.99$, $p = .01$, $d = 1.06$, 95% CI [.20, 1.89]). 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 ($F(1,25) = 0.02$, $p = 0.89$, $d = 0.05$).

Materials and procedure. The opening instructions for all participants will read: "We are conducting a study on how people's typing skills may reflect certain aspects of their personality, and how that relationship may vary depending on the content of what they are typing. In other parts of the study, you complete some personality questionnaires that will allow us to explore this relationship. We would now like you to type the paragraph below. Although you should try to minimize errors, please don't type any slower or faster than you would under normal conditions—simply type at the speed you naturally would for a casual word-processing task."

Participants in the unethical condition will be asked to copy the following passage: "Two years ago, when I was a junior partner at a prestigious law firm, I was coming up for promotion against another junior partner, Chris. For several months, Chris had been working on a major case for the city that would make or break his career at the firm. However, he could not locate a key zoning document, without which, it was unlikely that he would have sufficient evidence to successfully argue his case. Late one evening, as I was rummaging through a corner filing cabinet, I happened to come across the zoning document that Chris was in desperate need of. I pulled it from the cabinet and walked over to the office shredder, knowing that my promotion would now be secured."

Participants in the ethical condition will be asked to copy this passage: "Two years ago, when I was a junior partner at a prestigious law firm, I was coming up for promotion against another junior partner, Chris. For several months, Chris had been working on a major case for the city that would make or break his career at the firm. However, he could not locate a key zoning document, without which, it was unlikely that he would have sufficient evidence to successfully argue his case. Late one evening, as I was rummaging through a corner filing cabinet, I happened to come across the zoning document that Chris was in desperate need of. I pulled it from the cabinet and placed it without a note on Chris' desk, knowing that he would be so relieved when he arrived to work the next morning."

Next, using a 7-point scale from 1 = not at all to 7 = very much, participants will answer "How much do you desire this product?" for five cleaning products (e.g., Dove shower soap, Crest toothpaste, Windex glass cleaner, Lysol countertop disinfectant, and Tide laundry detergent) and five control products (e.g., Post-it notes, Nantucket Nectars juice, Energizer batteries, Sony cd cases, and Snickers candy bars) presented in a randomized order. Materials here: https://osf.io/idgpt/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_8caXWyHBqIJK3lP.

Analysis plan. The key factor of interest is whether condition affects ratings of the cleaning products, so ratings of the five cleaning products will be averaged and compared between the two conditions (ethical or unethical) with an independent samples *t*-test. A second comparison is whether there is a condition difference between ratings of the other products. The theoretical expectation is that this difference will be weak or near zero. This will be examined as a secondary analysis as a 2 (ethical-unethical) x 2 (cleaning-other products) mixed model ANOVA, and a follow-up independent samples *t*-test comparing ratings of other products between the ethical and unethical conditions.[3]

Participants who copy less than half the target article will be excluded from analysis.

Known differences from original. The original was presented on pencil and paper and participants copied the text under the guise of a personality test. In the replication, the whole procedure will be administered on a computer and participants will type an adapted version of the original story under the guise of a study measuring personality and typing speed. This adaptation was recommended by the original authors.

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

456 participants answered a question about life satisfaction in a specific domain "How

[3] A reviewer noted a possible moderating influence that could be examined as a secondary factor. Lee and Schwarz (2010) observed that the effect is stronger for products that clean a dirty body part.

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 ($r = .67$, $p < .05$). When the specific question was asked second, the correlation between them was weaker ($r = .32$, $p < .05$). The difference between these correlations was significant, $z = 2.32$, $p < .01$, $d = 0.22$, 95% CI [.12, .31].

The authors suggest that the specific-first condition makes the relationship more accessible such that participants then 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.

Materials and procedure. Participants will rate their satisfaction on an 11-point scale from 1 = very dissatisfied to 11 = very satisfied in response to two questions: "How satisfied are you currently with your life-as-a-whole?" and "Please think about your relationship to your partner (spouse or date). How satisfied are you currently with your relationship?" The items will be presented on separate screens in a randomized order. Materials here: https://osf.io/m9iv4/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_ehNrfGOaZoxxURT.

Analysis plan. We will compute the correlation between responses to the general and specific question in each item order condition, and then compare the correlations using the Fisher *r*-to-*z* transformation. Participants with valid responses to both items will be included in the analysis.

Known differences from original. The original was administered in German. The original had nine conditions whereas the replication will use just two of those: one in which the question about life satisfaction is asked before the one about relationship satisfaction, and a second in which the order is reversed. In the original these conditions were dubbed the "general-specific" order and "one-specific-general" order. Here we use "specific-second" and "specific-first" to refer to these conditions. Further, in the original, the general-specific order included additional specific questions after the relationship item. Those additional specific questions are not relevant for the effect measured here so they are not retained. Finally, in the original procedure, no other measures preceded this task. 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 task order analysis will be particularly important for this effect, and the most direct comparison with the original is for the conditions in which this task is administered first.

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% ($z = 2.48$, $p < .02$, $d = 0.43$, 95% CI = [0.09, 0.77]). This finding was consistent with the

hypothesis that negative features are weighed more strongly when people are rejecting options, and positive features are weighed more strongly when people are selecting options (Shafir, 1993).

Materials and procedure. All participants will read the following prompt: "Imagine that you serve on the jury of an only-child sole-custody case following a relatively messy divorce. The facts of the case are complicated by ambiguous economic, social, and emotional considerations, and you decide to base your decision entirely on the following few observations." The last sentence will be randomized between participants to be either "To which parent would you award sole custody of the child?" or "Which parent would you deny sole custody of the child?" The parents' characteristics will be presented in a tabular format with five features each. Parent A (average) has average income, average health, average working hours, reasonable rapport with the child, and a relatively stable social life. Parent B (extreme) has above-average income, very close relationship with the child, extremely active social life, lots of work-related travel, and minor health problems.

After reading the prompt and characteristics, participants will choose which parent to award/deny custody. Materials here: https://osf.io/ek5gz/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_3DvOfaiJUOQ0WUZ.

Analysis plan. The proportion of participants awarding or denying custody for parent B will be summed from both groups and tested against 100% with a Z test. The effect size will be computed estimating a logistic regression model on the 2X2 table and then taking the exponentiation of the unstandardized beta parameter of the main effect of parent B, which can be interpreted as an odds ratio. All participants with data will be included in analysis.

Known differences from original. None known.

26. 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, $F(2, 288) = 3.88$, $p = .02$, and concern, $F(2, 288) = 4.74$, $p = .01$. 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; $t(191) = 2.08$, $p = .03$, $d = .30$, 95% CI [.02, .59]) or control conditions (M = x.xx, SD = x.xx; $t(xx) = x.xx$, $p = .02$, $d = .xx$). Likewise, participants in the heat-priming condition expressed greater concern (M = X.xx, SD = X.xx) about global warming than participants in the cold-priming (M = x.xx, SD = x.xx; $t(xx) = x.xx$, $p = .07$, $d = .xx$) or control conditions (M = x.xx, SD = x.xx; $t(xx) = x.xx$, $p = .03$, $d = .xx$). [Note: Some relevant statistics for clarifying the effect between experimental conditions are not available in the original article. We will follow-up with the original authors to obtain these values.]

Materials and procedure. First, participants will complete a 13-item scrambled sentences task, in which they will form a complete sentence by using four of five provided words. Participants will be randomly assigned to one of two conditions using different words. In one condition, 6 of the sentences will contain a word related to heat (e.g., boil, sunburn, hot). In the other condition, 6 of the sentences will contain a word related to cold (e.g., cold, frozen, shivers).

Next, participants will respond to two questions on a scale from 1 (not at all convinced/worried) to 4 (completely convinced/worried): (1) "How convinced are you that global warming is happening?", and (2) "How much do you personally worry about global warming?". Materials here: https://osf.io/a4sih/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_78PEVpJ0kcA62rj.

Analysis plan. Mean differences in belief and concern about global warming between heat and cold-priming conditions will be evaluated with an independent samples *t*-test. The scrambled sentence task could introduce unanticipated variation across translations. As such, the direct replication test will use only English language sites, and - like all other effects - all samples and settings with data will be included in analyses examining heterogeneity to see if factors, like translation, have an impact on effect estimates.

Known differences from original. The original experiment used an initial question about current temperature perception followed by a 10-minute delay of unrelated filler material. The initial question is not relevant to the direct replication so will not be included. Also, the original experiment had a control condition that will not be included. The primary target for replication is concern about global warming. Belief in global warming will be included as a secondary replication.

Translated versions will be excluded from the direct replication test, as noted above, due to concerns that direct translations of the scrambled sentences task may be impractical. Translators will be instructed to remain as true to the original as possible, but to emphasize the construct being manipulated instead of translating the words exactly.

27. INTENTIONAL ACTION AND SIDE EFFECTS IN ORDINARY LANGUAGE (Knobe, 2003, Study 1)

Consider 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. In the harm condition, 82% of ss said that the agent brought about the side-effect intentionally, whereas in the help condition, 77% said that the agent did not bring about the side-effect intentionally ($X^2(1, N = 78) = 27.2$, $p < .001$, $d = 1.46$). 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 $t(120) = 8.4$, $p < .001$, $d = 1.55$. The total amount of blame or praise attributed to the agent was associated with believing the agent brought about the side effect intentionally $r(120) = .53$, $p < .001$, $d = 1.25$, 95% CI [.79, 2.79].

Materials and Procedure. Participants will read a vignette where a company either harms or helps the environment: "The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm[help] the environment.' The chairman of the board answered, 'I don't care at all about harming[helping] the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed[helped]."

After reading the vignette, participants will be asked to indicate their agreement with the statement "The chairman harmed [helped] the environment intentionally," on a 7-point scale from 1=strongly disagree to 7=strongly agree. Participants judge how much blame the chairman

deserved in the harm condition, or how much praise the chairman deserved in the help condition, on a 7-point scale (1=no blame [praise]; 7=a lot of blame [praise]). Materials here: https://osf.io/dcbmw/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_b1qjejRBoOI1nwx.

Analysis plan. Ratings of intentionality in the harm and help conditions will be compared using an independent-samples *t*-test. This is the focal test for the direct replication. Blame and praise ratings will also be collected but are secondary analyses. All participants with data will be included in analysis.

Known differences from original. In the original study, participants indicated whether the chairman intentionally harmed or helped the environment using a yes/no question. Subsequent research examining this effect has used a 7-point agreement scale rather than the dichotomous response. We use the updated 7-point scale.

28. STUDIES OF SIMILARITY (Tversky & Gati, 1978, Study 2)

Tversky and Gati (1978) investigated the relationship between directionality and similarity. 77 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). For each pair, 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 presented the more prominent county was presented first "about an equal number of times". Results indicated that participant similarity ratings were higher when the less prominent country was displayed first compared to the more prominent country displayed first, $t(153) = 2.99$, $p = .001$, $d = .48$, 95% CI = [.16, .80], and that higher similarity ratings were given to the version of each pair that listed the more prominent country second, $t(20) = 2.92$, $p = .001$, $d = 1.31$, 95% CI [.33, 2.26].

Then, they did a follow-up study (N = 46) with the same design except that participants rated differences rather than similarities. Following the prior result, participant difference ratings were higher when the more prominent country was displayed first compared to the less prominent country displayed first, $t(45) = 2.24$, $p < .05$, $d = 0.67$, 95%CI [.34, .98] and higher difference ratings were given to the version of each pair that listed the more prominent country first, $t(20) = 2.72$, $p < .01$, $d = 1.22$, 95%CI [.64, 1.78].

Materials and Procedure. A list of country pairs was assembled in which one country was rated more prominent than the other. These items were taken from the original paper and updated as noted in the differences from original section. The 21 pairs of countries include: 1) U.S.A - Mexico, 2) Russia - Poland, 3) China - Albania, 4) U.S.A - Israel, 5) Japan - Philippines, 6) U.S.A - Canada, 7) Russia - Israel, 8) England - Ireland, 9) Germany - Austria, 10) Russia - France, 11) Belgium - Luxembourg, 12) U.S.A - Russia, 13) China - North Korea, 14) India - Sri Lanka, 15) U.S.A - France, 16) Russia - Cuba, 17) England - Jordan, 18) France - Israel, 19) U.S.A - Germany, 20) Russia - Syria, and 21) France - Algeria. We have created one list where 11 pairs have the prominent country first and another list with the opposite pairings (10 pairs that have prominent country first). In the original study an initial pilot was performed on 68 individuals who judged which of each pair was more prominent. In the list of the pairs of countries reported here, the first country was judged as the more prominent country by at least ⅔ of the participants.

Participants will be randomly assigned to one of the two order conditions above, and will be randomly assigned to rate similarities or differences between the two countries. The similarities condition reads: "You will be presented with a number of country pairs and be asked

to judge the similarity of one country to another. Please use the scale ranging from 1 = no similarity to 20 = maximal similarity." Participants will then rate the similarity of 21 country pairs with one country being more prominent than the other. Participants will rate each pair with a single-item from 1 = *no similarity* to 20 = *maximal similarity*. The differences condition will read virtually the same instructions except with "difference" in place of "similarity", and they will rate the countries on a list of 1 = *minimal difference* to 20 = *maximal difference*.

Materials here: https://osf.io/a2jwk/. Test the study here: https://ufl.qualtrics.com/SE/?SID=SV_39Q9WUKfgyMRk1f.

Analysis plan. We will perform three analyses on the data. For the primary analysis, we will analyze the data through a general linear mixed model with a random effect for the item pair nested within subject, and a fixed factor 'order' representing the order of the pair (prominent first vs. prominent second). Fitting this model will allow evaluation of both effects. If the intercept is significantly greater than 0, this would confirm the finding that at the participant level, if there is an effect for the factor 'order' the pairs where the prominent country appeared second will be rated as more similar than when the prominent country appeared first. We will convert the Beta provided by this intercept term into a Cohen's *d* effect size.

Second, we will recreate the original analysis used to get a participant-level effect of making similarity judgments where either the more or less prominent country comes first. We will compute an asymmetry score for each subject, calculated as the average similarity for comparisons where the prominent country appears second minus the average for the comparisons where the prominent country appears first. Using a one-sample *t*-test, we will test this difference score against zero (original d=.48). Third, using a matched-pairs *t*-test, we will compare the average score for each pair when it was prominent-first compared to prominent-second (original d= 1.31).

Because these latter two analyses do not account for the fact that the variance in ratings is crossed between participants and pairs, they will be secondary and only used as a comparison for the original analysis. All participants with data will be included in the analysis.

These analyses will be repeated for the differences conditions and reported as a separate study. Because of the random assignment to similarity or difference conditions, each site will have half as much data for its critical test as the other effects. This will likely increase the standard error of its estimates by comparison.

Known differences from the original. The original study was likely completed in a paper and pencil format. Also, in the original, the similarity and difference conditions were separate studies. A reviewer suggested including both for the present design.

In the current 21 pairs, Ceylon is changed into Sri Lanka, West Germany changed to Germany, and U.S.S.R changed to Russia. Because this test will be performed in many different countries, the country in each pair that is considered most prominent might differ, depending on the sample (e.g., participants in Israel might judge Israel to be more prominent than France). It is also worth noting *a priori* that another study in this slate involves similarity judgements (Norenzayan et al., 2002) and it may be relevant to analyze order effects between these two effects in particular.

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

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63*, 596-612.

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

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

Carnelley, K. B. & Janoff-Bulman, R. (1992). Optimism about love relationships: general vs specific lessons from one's personal experiences. *Journal of Social and Personal Relationships*, *9*(1), 5-20. doi:10.1177/0265407592091001

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

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

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

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

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

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

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

Gable, P. A., & Harmon-Jones, E. (2008). Approach-motivated positive affect reduces breadth of attention. *Psychological Science*, *19*, 476-482. doi: 10.1111/j.1467-9280.2008.02112.x

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*(1), 21-38.

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology, 54*, 733-740.

Gnambs, T. (2014). A meta-analysis of dependability coefficients (test-retest reliabilities) for measures of the Big Five. *Journal of Research in Personality*. Advance online publication. 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. (2014, June 17). *GOZ Lab: Scales we've developed*. Retrieved from http://homepage.psy.utexas.edu/HomePage/Faculty/Gosling/scales_we.htm

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

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.

Hazan, C., & Shaver, P. (1987). Romantic love conceptualized as an attachment process. *Journal of personality and social psychology*, *52*(3), 511.

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

Hofmans, J., Kuppens, P., & Allik, J. (2008). Is short in length short in content? An examination of the domain representation of the Ten Item Personality Inventory scales in Dutch language. *Personality and Individual Differences*, *45*, 750-755. doi: 10.1016/j.paid.2008.08.004

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.

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2014). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*. Advance online publication, April 2014. 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

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

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

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

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

Keysar, B. (1994). The illusory transparency of intention: linguistic perspective taking in text. *Cognitive Psychology, 26*, 165–208. doi: 10.1006/cogp.1994.1006

Kim, H. S., & Sherman, D. K. (2007). "Express yourself": Culture and the effect of self-expression on choice. *Journal of Personality and Social Psychology*, *92*, 1-11. doi: 10.1037/0022-3514.92.1.1

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, S., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong,

W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M., Joy-Gaba, J. A., Kappes, H. B., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van 't Veer, A. E., Vaughn, L. A., Vranka, M., Wichman, A. L., Woodzicka, J. A., & 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

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

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

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

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

Muck, P. M., Hell, B., & Gosling, S. D. (2007). Construct validation of a short Five-Factor model instrument. A self-peer study on the German adaptation of the Ten-Item Personality Inventory (TIPI-G). *European Journal of Psychological Assessment, 23*, 166-175. doi: 10.1027/1015-5759.23.3.166^

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

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*(8), 771-781.

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.

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

Oshio, A., Abe, S., Cutrone, P., & Gosling, S. D. (2013). Big Five content representation of the Japanese version of the Ten-Item Personality Inventory. *Psychology, 4*, 924. doi:10.4236/psych.2013.412133

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

Putrevu, S. (2014). Effects of mood and elaboration on processing and evaluation of goal□ framed appeals. *Psychology & Marketing, 31*, 134-146. doi: 10.1002/mar.20682

Renau, V., Oberst, U., Gosling, S., Rusiñol, J., & Chamarro, A. (2013). Translation and validation of the Ten-Item-Personality Inventory into Spanish and Catalan. *Aloma: Revista de Psicologia, Ciències de l'Educació i de l'Esport, 31*(2).

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*. 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

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

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

Snyder, M. (1974). The self-monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30,* 526-537. doi: 10.1037/h0037039

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

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.

Uleman, J. S., Rhee, E., Bardoliwalla, N., Semin, G., & Toyama, M. (2000). The relational self: Closeness to ingroups depends on who they are, culture, and the type of closeness. *Asian Journal of Social Psychology, 3*, 1-17, doi: 10.1111/1467-839X.00052

Uskul, A. K., Hynie, M., & Lalonde, R. N. (2004). Interdependence as a mediator between culture and interpersonal closeness for Euro-Canadians and Turks. *Journal of Cross-Cultural Psychology, 35,* 174-191, doi: 10.1177/0022022103262243

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,* 4, 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

Watson, R. E., Pritzker, L., & Madison, P. (1955). Hostility in neurotics and normals. *The Journal of Abnormal and Social Psychology*, *50*(1): 36-40. doi: 10.1037/h0041326

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. *Manuscript submitted for publication*.

Appendix. Individual Difference Measures and Original Articles of Included Effects. Citation counts from Google Scholar on September 14, 2014.

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.	2798	
	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.	35	
	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.	905	
	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.	1851	
	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.	303	
	Data quality: Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. Psychological Methods, 17, 437–455.	63	
	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.	13	
	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.	94	
	Disgust Sensitivity, Contaminiation 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.	190	

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.	0	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.	2	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.	295	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.	606	1
5	Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. Psychological Science, 12, 185-190.	462	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.	28	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.	91	1
8	Inbar, Y., Pizarro, D., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. Emotion, 9, 435-439.	191	1
9	Critcher, C. R., & Gilovich, T. (2008). Incidental environmental anchors. Journal of Behavioral Decision Making, 21, 241-251.	56	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.	629	3
11	Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. Mind & Language, 22, 1-21.	353	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.	37	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.	1715	1.1

Effect #	***Slate 2***	# citations	Study #
14	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.	1715	1.2
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.	107	1a
16	Tversky, A., Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211, 453-458.	10786	10
17	Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. Mind & Language, 22, 1-21.	353	1.2
18	Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. Journal of Personality and Social Psychology, 95, 293.	50	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.	43	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.	276	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.	215	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.	87	1a
23	Zhong, C. B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. Science, 313, 1451–1452.	433	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.	324	1
25	Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. Memory & Cognition, 21, 546-556.	396	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.	8	3a
27	Knobe, J. (2003). Intentional action and side effects in ordinary language. Analysis, 63, 190-193.	481	1
28	Tversky, A., & Gati, I. (1978). Studies of similarity. Cognition and categorization, 1, 79-98.	500	2