Psychological Bulletin 2020, Vol. 146, No. 5, 451–479
ISSN: 0033-2909 http://dx.doi.org/10.1037/bul0000220

Crowdsourcing hypothesis tests: Making transparent how design choices shape research results

Justin F. Landy[1], Miaolei (Liam) Jia[2], Isabel L. Ding[3], Domenico Viganola[4], Warren Tierney[5], Anna Dreber[6,7], Magnus Johannesson[6], Thomas Pfeiffer[8], Charles R. Ebersole[9], Quentin F. Gronau[10], Alexander Ly[10,11], Don van den Bergh[10], Maarten Marsman[10], Koen Derks[12], Eric-Jan Wagenmakers[10], Andrew Proctor[6], Daniel M. Bartels[13], Christopher W. Bauman[14], William J. Brady[15], Felix Cheung[16], Andrei Cimpian[15], Simone Dohle[17], M. Brent Donnellan[18], Adam Hahn[17], Michael P. Hall[19], William Jiménez-Leal[20], David J. Johnson[21], Richard E. Lucas[18], Benoît Monin[22], Andres Montealegre[20,23], Elizabeth Mullen[24], Jun Pang[25], Jennifer Ray[15], Diego A. Reinero[15], Jesse Reynolds[22], Walter Sowden[19,26], Daniel Storage[27], Runkun Su[3], Christina M. Tworek[28], Jay J. Van Bavel[15], Daniel Walco[29], Julian Wills[15], Xiaobing Xu[30], Kai Chi Yam[3], Xiaoyu Yang[31], William A. Cunningham[32], Martin Schweinsberg[33], Molly Urwitz[6], The Crowdsourcing Hypothesis Tests Collaboration, Eric L. Uhlmann[34]

[1]Nova Southeastern University, [2]Warwick Business School, University of Warwick, [3]National University of Singapore, [4]George Mason University, [5]Kemmy Business School, University of Limerick, [6]Stockholm School of Economics, [7]University of Innsbruck, [8]Massey University, [9]University of Virginia, [10]University of Amsterdam, [11]Centrum Wiskunde & Informatica, [12]Nyenrode Business University, [13]University of Chicago, [14]University of California, Irvine, [15]New York University, [16]University of Hong Kong, [17]University of Cologne, [18]Michigan State University, [19]University of Michigan, Ann Arbor, [20]Universidad de los Andes, [21]University of Maryland at College Park, [22]Stanford University, [23]Cornell University, [24]San José State University, [25]Renmin University of China, [26]Tripler Army Medical Center, Hawaii, [27]University of Denver, [28]HarrisX, [29]New York Yankees, [30]Hainan University, [31]Tsinghua University, [32]University of Toronto, [33]ESMT Berlin, [34]INSEAD

To what extent are research results influenced by subjective decisions that scientists make as they design studies? Fifteen research teams independently designed studies to answer five original research questions related to moral judgments, negotiations, and implicit cognition. Participants from two separate large samples (total $N > 15{,}000$) were then randomly assigned to complete one version of each study. Effect sizes varied dramatically across different sets of materials designed to test the same hypothesis: materials from different teams rendered statistically significant effects in opposite directions for four out of five hypotheses, with the narrowest range in estimates being $d = -0.37$ to $+0.26$. Meta-analysis and a Bayesian perspective on the results revealed overall support for two hypotheses, and a lack of support for three hypotheses. Overall, practically none of the variability in effect sizes was attributable to the skill of the research team in designing materials, while considerable variability was attributable to the hypothesis being tested. In a forecasting survey, predictions of other scientists were significantly correlated with study results, both across and within hypotheses. Crowdsourced testing of research hypotheses helps reveal the true consistency of empirical support for a scientific claim.

Keywords: Crowdsourcing, scientific transparency, stimulus sampling, forecasting, conceptual replications, research robustness

Scientific theories are meant to be generalizable. They organize findings, ideas, and observations into systems of knowledge that can make predictions across situations and contexts. Theories are more useful when they can explain a wider variety of phenomena. Understanding a theory's scope is critical to successfully applying it. In order to be generalizable, theories often make use of abstract concepts or conceptual variables to organize their hypothesized relationships. For instance, cognitive dissonance theory, one of the most influential theories in social psychology, states that when individuals have inconsistent cognitions, they will experience psychological distress or discomfort that motivates them to reduce the inconsistency (Festinger, 1957). This theory makes use of conceptual variables to describe its relationships of interest. In particular, "cognitions" refer to any of several types of mental constructs, including attitudes, beliefs, self-concepts, and knowledge that one has engaged in a certain behavior. Reducing inconsistency can take many forms, such as altering one or both of the cognitions to become consistent,

Contact Authors: jlandy@nova.edu, miaolei.jia@wbs.ac.uk, isabel.ding@u.nus.edu, domenico.viganola@phdstudent.hhs.se, eric.luis.uhlmann@gmail.com

or adding new cognitions that resolve the discrepancy. These conceptual variables allow researchers to use the theory to make predictions about many different situations in which people experience inconsistency.

Researchers must operationalize abstract and conceptual variables into concrete terms for empirical testing. For example, to study cognitive dissonance, a researcher might identify two cognitions that could reasonably be brought into conflict with one another (the independent variable). Then, the psychologist could identify a way of resolving the conflict to provide to participants (the dependent variable). Indeed, psychologists have studied cognitive dissonance by measuring attitudes toward a boring task after inducing some participants to lie to the next participant and say the task was exciting (Festinger & Carlsmith, 1957), by measuring preferences toward appliances after obliging participants to choose between two attractive options to receive as a gift (Brehm, 1956), or by assessing interest in a study group after undergoing an uncomfortable initiation (Aronson & Mills, 1959). Each of these concrete operationalizations widens the understood boundaries of the conceptual variables involved in an effect and thus the generalizability of the effect itself (Schmidt, 2009; Stroebe & Strack, 2014).

Although generalizability is a critical goal of scientific research, the standard model of conducting research creates many challenges for establishing robust generalizability of an effect across contexts. Researchers and/or labs often work in isolation or in small groups, generating their own hypotheses, measures, and operationalizations. These operationalizations represent a small subset of the possible, theoretically justifiable methods that they could have used to test their hypotheses (Baribault et al., 2018; Judd, Westfall, & Kenny, 2012; Monin & Oppenheimer, 2014; Monin, Pizarro, & Beer, 2007; Wells & Windschitl, 1999; Westfall, Judd, & Kenny, 2015). In particular, scientists may use methods that are likely to confirm their preconceptions (McGuire, 1973, 1983; Nickerson, 1998). For example, researchers who theorize that moral judgments are intuitive tend to use simple and emotionally evocative scenarios, whereas researchers who theorize that moral judgments are rooted in reasoning tend to use complex stimuli that pit different values against each other and stimulate deliberation (Monin et al., 2007). Such assumptions may guide which operationalizations are used to test hypotheses and theories, and divergence across operationalizations may then affect which theory is empirically supported.

After one or a few operationalizations and stimulus sets are tested, researchers choose which observations to report to the broader scientific community in academic journals. There is substantial evidence that scientific publishing is biased in favor of positive or statistically significant findings, leaving negative and null results underreported (Greenwald, 1975; Ioannidis, 2005; Ioannidis & Trikalinos 2007; Pfeiffer, Bertram, & Ioannidis, 2011; Rosenthal, 1979; Schimmack, 2012; Simonsohn, Nelson, & Simmons, 2014). Null results are important for understanding generalizability because they provide insights about where the boundaries of a theory lie; nonetheless, the scientific community may be left largely unaware of them due to biases in publishing (LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018; Zwaan, Etz, Lucas, & Donnellan, 2018).

After initial observations are reported, other researchers may conduct follow-up research. These follow-ups have the potential to increase understanding of generalizability by inspiring new operationalizations and instantiations of effects and theories. Still, scientific culture and professional advancement often privilege novelty over increased certainty and incremental refinement (Everett & Earp, 2015; Giner-Sorolla, 2012; Nosek, Spies, & Motyl, 2012), which may disincentivize researchers from conducting tests of previously published ideas in favor of pursuing new ideas and theories (Makel, Plucker, & Hegarty, 2012). Although scientific culture has been changing with respect to valuations of replications, particularly in psychology, these changes have been more focused on direct replications (testing the same idea with the same materials and methodology; Alogna et al., 2014; Ebersole et al., 2016; Klein et al., 2014; Open Science Collaboration, 2015; Pashler & Harris, 2012; Simons, 2014) than on conceptual replications (testing established ideas with a new approach; Crandall & Sherman, 2016; Finkel, Eastwick, & Reis, 2015, 2017). Furthermore, failed conceptual replications are far more susceptible to alternative explanations based on methodological differences than are direct replications, and as a consequence may be left unpublished or dismissed by original researchers and other scientists (Baribault et al., 2018; Doyen, Klein, Simons, & Cleeremans, 2014; Earp, in press; Hendrick, 1990; Schmidt, 2009; Simons, 2014). Taken together, these forces within the standard model of conducting psychological research may impede tests of generalizability of scientific theories and phenomena. The standard model may thus stunt theory development by limiting contributions to the literature to ones based on a

relatively small subset of operationalizations, and to unrealistically positive results.

The Current Research

To address these challenges, we introduce a crowdsourced approach to hypothesis testing. In the crowdsourcing initiative reported here, up to 13 research teams (out of a total of 15 teams) independently created stimuli to address the same five research questions, while fully blind to one another's approaches, and to the original methods and the direction of the original results. The original hypotheses, which were all unpublished at the time the project began, dealt with topics including moral judgment, negotiations, and implicit cognition. Large samples of research participants were then randomly assigned to different teams' versions of the same study, with a commitment to publish the results from all study designs as a fundamental component of the project. The analyses were also pre-registered, which has been argued to reduce bias (Nosek, Ebersole, DeHaven, & Mellor, 2018; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Wicherts, Veldkamp, Augusteijn, Bakker, Van Aert, & Van Assen, 2016), although a causal effect remains to be empirically demonstrated. Comparisons of the estimated effect sizes associated with the same hypothesis across the studies created by the different teams reveal the extent to which the empirical results are contingent on the decisions scientists make as they design their study. Aggregating results across teams via meta-analysis, taking into account both average effects and variability across teams, provides a systematic assessment of the relative strength of support for each hypothesis.

There are a number of potential benefits to a crowdsourcing approach to hypothesis testing. Crowdsourcing the operationalization of research ideas makes transparent the true consistency of support for an empirical prediction, and provides a more stringent test of robustness than employing a narrow set of stimuli (Monin & Oppenheimer, 2014), directly replicating multiple independent and dependent variables that have been used previously (Caruso, Shapira, & Landy, 2017), or even the innovative approach of radically randomizing features of the same basic experimental design (e.g., symbols, colors, and presentation speeds in a cognitive priming paradigm; Baribault et al., 2018). Rather than varying features of the same basic design to address concerns about stimulus sampling (Baribault et al., 2018), we had different researchers design distinct studies to test the same research questions, providing an arguably wider-ranging test of the conceptual robustness of each original finding. The extent to which divergent approaches produce different results is further revealed. Uniquely, the conceptual replications are developed by independent research teams with no prior knowledge of the original authors' method or results to bias them (Silberzahn et al., 2018; Silberzahn & Uhlmann, 2015), unlike in the usual practice of science, in which conceptual replications are conducted after the dissemination of the original results. Materials designers also did not know the direction of the original hypotheses and results, but were rather provided with a non-directional version of each research question (see below). This was to prevent materials designers from constructing materials aimed at confirming a directional hypothesis, while not giving alternative directional hypotheses a chance (Monin et al., 2007). In other words, we believe that we were more concerned with answering the research questions that drove the five original, unpublished studies than we were with confirming their results.

Because all crowdsourced conceptual replications were pre-registered and reported, this approach is also free of reporting bias, unlike traditional conceptual replications where null effects may be attributed to departures from the original methodology and therefore left unpublished. Moreover, because participants from the same large sample are randomly assigned to different conceptual replications, discrepant results, including "failed" replications, cannot be attributed to differences in the populations being sampled (McShane, Tackett, Böckenholt, & Gelman, 2019; Tiokhin, Hackman, Munira, Jemsin & Hruschka, 2019; Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). Heterogeneity in results above-and-beyond what would be expected based on sampling error can confidently be attributed to design choices.

In the present initiative, we also recruited a second large sample and repeated our initial studies with the same methodologies and materials. This effort is, to our knowledge, the first time an entire crowdsourced set of studies has itself been directly replicated. Doing so allowed us to simultaneously take into account both conceptual and direct replications when assessing the strength of evidence for each finding. Altogether, we provide a new framework for determining the generalizability and context-dependency of new findings, with the goal of identifying more deeply robust phenomena, which we believe may hold utility for select research questions in the future. In the Discussion, we elaborate at greater length on when crowdsourcing hypothesis tests is likely to prove most (and least) useful.

We additionally examine whether scientists are able to predict *a priori* how design choices impact research results. Prior work has demonstrated that researchers can anticipate whether a published result will independently replicate based on the research report alone (Camerer et al., 2016; Dreber et al., 2015), and predict the effects of performance interventions starting only from a few benchmark effects and the materials for the additional treatments (DellaVigna & Pope, 2018a, 2018b). Other forecasting studies with scientists have returned more mixed results (Coffman & Niehaus, 2014; Dunaway, Edmonds, & Manley, 2013; Groh, Krishnan, McKenzie, & Vishwanath, 2016; Sanders, Mitchell, & Chonaire, 2015). We therefore conducted a forecasting survey asking an independent crowd of scientists to attempt to predict the results of each study based solely on its sample size, methodology, and materials. Notably, all prior work has examined forecasting accuracy *across* different hypotheses that vary in their truth value and alignment with empirical reality. In contrast, we assessed whether scientists are accurate in their beliefs about the outcomes of *different* experiments designed to test the *same* research question. Scientists' intuitions about the impact of researcher choices may or may not map onto the actual downstream consequences.

Method

Main Studies and Replication Studies

In two separate data collection efforts (the initial investigations ["Main Studies"] and direct replications ["Replication Studies"]), we randomly assigned participants to different sets of study materials designed independently by up to 13 teams of researchers to test the same five research questions. There were 15 teams of materials designers in total, from which up to 13 teams designed materials for each research question (i.e., not all teams made materials to test all five original hypotheses). The five research questions were gathered by emailing colleagues conducting research in the area of moral judgment and asking if they had initial evidence for an effect that they would like to volunteer for crowdsourced testing by other research groups. In three cases (Hypotheses 1, 3, and 4), project coordinators volunteered an effect from their research program that fit these criteria, and in two cases, members of other teams volunteered an effect (Hypotheses 2 and 5). In the present research, we examined the overall degree of support for each hypothesis, and also quantified the heterogeneity across different sets of study materials. To our knowledge, this instance is the first time a large-scale meta-scientific project has itself been directly replicated in full with a new sample. All materials, data, and analysis scripts from this project are publicly available at https://osf.io/9jzy4/.

Target hypotheses. We identified five directional hypotheses in the areas of moral judgment, negotiation, and implicit cognition, each of which had been supported by one then-unpublished study.[1] Table 1 shows the directional hypotheses, as well as the nondirectional forms in which they were presented to materials designers. Below we elaborate briefly on the theoretical basis for each research question.

Hypothesis 1: Awareness of automatic prejudice. Influential dual-process theories of intergroup attitudes propose that individuals have both explicit, consciously endorsed attitudes towards negatively stereotyped groups, and also implicit ones that they may not endorse (Dovidio, Kawakami, & Gaertner, 2002; Fazio, Jackson, Dunton, & Williams, 1995; Gawronski & Bodenhausen, 2006; Greenwald & Banaji, 1995; Wilson, Lindsey, & Schooler, 2000). Rather than in propositional logic, these implicit attitudes are based in simple associations (e.g., Black-Criminal, Female-Weak), that are conditioned by the cultural environment (Gawronski & Bodenhausen, 2006; Uhlmann, Poehlman, & Nosek, 2012). As a result, even consciously egalitarian individuals often harbor prejudiced associations that may "leak out" and affect their judgments and behaviors without them realizing it. Low correspondence between self-reported and implicit measures of intergroup attitudes has been interpreted as indicating a lack of introspective access into the latter (Banaji, 2001; Greenwald & Banaji, 1995). Nonetheless, people could potentially be aware of their spontaneous affective reactions without endorsing them. Indeed, Hahn and colleagues (Hahn, Judd, Hirsh, & Blair, 2014; Hahn & Gawronski, 2019) provide evidence that people can accurately predict their performance on Implicit Association Tests of associations with social groups (Greenwald, McGhee, & Schwartz, 1998). Uhlmann and Cunningham (2000) constructed questionnaire items examining whether individuals directly self-report negative gut feelings towards minorities. Representative

[1] The original study supporting Hypothesis 4 has since been published as a supplemental study in Landy, Walco, and Bartels (2017).

items include "Although I don't necessarily agree with them, I sometimes have prejudiced feelings (like gut reactions or spontaneous thoughts) that I don't feel I can prevent" and "At times stereotypical thoughts about minorities coming into my head without my necessarily intending them to." In the original research, approximately three-quarters of undergraduates agreed with such statements, and overall endorsement was confirmed by mean responses statistically significantly above the neutral scale midpoint of four (*1= strongly disagree, 4= neutral, 7= strongly agree*). As the Uhlmann and Cunningham (2000) investigations were never published, the present initiative crowdsourced the question of whether people self-report automatic intergroup prejudices, assigning a dozen independent research teams to create their own awareness measures. Specifically, we examined whether the majority of people, without further prompting or consciousness-raising, agree on questionnaire measures that they harbor such automatic biases towards stigmatized groups.

Hypothesis 2: Extreme offers reduce trust. Negotiators are routinely advised to make extreme first offers to benefit from the anchoring effect (Tversky & Kahneman, 1974). When sellers make extreme first offers, final prices tend to be high; in contrast, when buyers make extreme first offers, final prices tend to be low (Ames & Mason, 2015; Galinsky, Leonardelli, Okhuysen, & Mussweiler, 2005; Galinsky & Mussweiler, 2001). Evidence suggests this effect is robust across cultures, issues, and power positions (Gunia, Swaab, Sivanathan, & Galinsky, 2013). Yet, more recent research has examined the conditions under which this advice might not be accurate (Loschelder, Swaab, Trötschel, & Galinsky, 2014; Loschelder, Trötschel, Swaab, Friese, & Galinsky, 2016; Maaravi & Levy, 2017). The present Hypothesis 2 explores one mechanism for why extreme first offers might backfire in negotiations with multiple issues. Specifically, extreme first offers may interfere with value creation processes such as trust building and information exchange. Building on previous research that showed that extreme first offers can cause offense and even impasses (Schweinsberg, Ku, Wang, & Pillutla, 2012), Schweinsberg (2013) examined the specific hypothesis that extreme first offers lower trust in the counterpart. Ultimately, this line of research may show that extreme first offers can help negotiators claim a larger percentage of the bargaining zone for themselves, but that extreme first offers also shrink the overall size of the bargaining zone by reducing information exchange and trust. Thus, extreme first offers might help negotiators claim a larger percentage of a smaller bargaining zone, making them ultimately worse off. Negotiators might be blind to this extreme first offer disadvantage because their salient comparison is between value they claimed versus value claimed by their counterpart, and not the more relevant but counterfactual comparison between value they claimed from an extreme offer versus value they could have claimed from a more moderate first offer. The present research focuses on just one part of this argument, providing crowdsourced tests of the prediction that "negotiators who make extreme first offers are trusted less, relative to negotiators who make moderate first offers."

Hypothesis 3: Moral praise for needless work. It is easy to find anecdotal examples in which individuals received moral praise for continuing to work despite coming into sudden wealth and no longer needing to earn a salary (Belsie, 2011). In scenario studies based on such real life cases, Americans positively evaluate the moral character of individuals with working class occupations (e.g., potato peeler in a restaurant kitchen) who continue their employment after winning the lottery (Poehlman, 2007; Uhlmann, Poehlman, & Bargh, 2009). A number of sources for such moral intuitions are plausible, among these a tendency to value work contributions that parallels general disapproval of shirkers and non-contributors (Jordan, Hoffman, Bloom, & Rand, 2016), use of work behavior as a signal of underlying traits (Uhlmann, Pizarro, & Diermeier, 2015), the influence of the Protestant work ethic in some cultures (Uhlmann & Sanchez-Burks, 2014), and post-materialist value systems in which work is pursued for meaning and fulfillment rather than as an economic necessity (Inglehart, 1997; Inglehart & Welzel, 2005). A separate project to this one examines the extent to which these and other work morality effects directly replicate across different national cultures (Tierney et al., 2019a). Of interest to the present initiative is how conceptually robust the findings are to alternative study designs. We therefore crowdsourced the research question of whether working in the absence of material need elicits moral praise, limiting our samples to U.S.-based participants, the group originally theorized to exhibit these effects.

Hypothesis 4: Proximal authorities drive legitimacy of performance enhancers. People in the United States express widespread normative opposition to the use of Performance-Enhancing Drugs (PEDs), especially among competitive athletes, but it is not clear what underpins these judgments. While most studies of opposition to PEDs have examined perceptions of fairness (e.g., Dodge,

Williams, Marzell, & Turrisi, 2012; Fitz, Nadler, Manogaran, Chong, & Reiner, 2014; Scheske & Schnall, 2012), some research also suggests that the sheer fact that PEDs are prohibited also contributes to opposition toward them (Sattler, Forlini, Racine, & Sauer, 2013). This distinction between fairness concerns and explicit rules roughly parallels the insight from Social Domain Theory (Turiel, 1983; 2002) that acts can be "wrong" in at least two qualitatively different ways: moral offenses violate universal moral standards like fairness, whereas "conventional" offenses violate consensually accepted norms or the dictates of legitimate authorities. Landy, Walco, and Bartels (2017) investigated whether opposition to PED use exhibits properties of conventional offenses by manipulating whether or not an athlete's use of PEDs "violates the law and the rules of his [competition] circuit" (Study 2 of the original report), and found that this manipulation significantly affected people's judgments of how wrong it was for the athlete to use PEDs. A follow-up study (Supplemental Study 1 of the original report) found that PED use was considered more wrong when it violated a dictate of a legitimate proximal authority (the competition circuit) than when it violated the law. An additional study replicated this finding (Study 12 of the original report), but a further study did not (Study 13 of the original report), so it is unclear whether proximal authority or legal authority contributes more to opposition to PED use. Since all of these studies were unpublished at the beginning of this project, we applied our crowdsourcing methodology to obtain a more definitive answer to this question.

Hypothesis 5: The tendency to make deontological judgments is positively correlated with happiness. In order to bridge the normative-descriptive divide between the fields of philosophical ethics (how should people morally behave) and moral psychology (how and why do people morally behave) cognitive science must map out how variation in moral cognitions are systematically related to variances in outcomes related to human flourishing. The goal of this original research was to contribute to this endeavor by examining how the tendency to make utilitarian versus deontological moral judgments (Bentham 1970/1823; Kahane, 2015; Kant, 1993/1785; Mill, 1861) relates to personal happiness and well-being (Kahneman, Diener, & Schwarz, 1999; Ryff, 1989; Waterman, 1993). The idea that happiness and morality are tightly intertwined has a long history in philosophy (see, e.g., Annas, 1993; Aristotle, 340 BCE/2002; Foot, 2001; Kraut, 1979), and recent empirical work suggests that people consider moral goodness to be an element of what "happiness" consists of (Phillips, Freitas, Mott, Gruber, & Knobe, 2017; Phillips, Nyholm, & Liao, 2014). However, prior work has not examined the relationship (if any) between specific *moral orientations* and happiness.

Hypothesis 5 posits that people who are more inclined to base their moral judgments on the violation of rules, duties, and obligations (deontological judgments) versus material outcomes (utilitarian judgments) are also more likely to experience happiness in their lives. This prediction is based on philosophical and scientific evidence that has demonstrated shared psychological and neurological mechanisms between these dimensions (e.g., Everett, Pizarro, & Crockett, 2016; Greene, 2013; Lieberman, 2013; Phillips et al., 2017; Singer, 2005). To test this hypothesis, Sowden and Hall (2015) asked participants to judge several morally questionable behaviors that pitted utilitarian and deontological considerations against one another (Greene et al. 2001) and compared an index of those judgments to how they responded to measures of subjective well-being (Diener et al., 1985; Watson et al., 1988) and eudaimonic happiness (Waterman et al., 2010). The crowdsourced project posed the research question to independent researchers, who separately designed studies relating moral judgments to individual happiness.

[INSERT TABLE 1 ABOUT HERE]

Method

Materials. A subset of the project coordinators (Landy, Jia, Ding, Uhlmann) recruited 15 teams of researchers through their professional networks to independently design materials to test each hypothesis. Of these 15 teams, four included the researchers who developed the original materials for at least one of the five hypotheses. Teams ranged in size from one researcher to five, and members ranged in experience from graduate students to full professors. We opted not to standardize team size because research teams vary greatly in size in the natural practice of science (Wuchty, Jones, & Uzzi, 2007a, 2007b). All studies were required to be designed to be administered in an online survey. Note that recruiting through our professional networks would, if anything, be expected to bias our results towards homogeneity and consistency between materials designers. Likewise, the restriction to using only brief, online questionnaires rather than behavioral measures, video stimuli, or elaborate laboratory experiments with a cover story and research confederates, also artificially constrains variability in study

designs. Yet, as we detail below, we still observed remarkable heterogeneity in results.

To avoid biasing their designs, materials designers were provided with the non-directional versions of the five hypotheses presented in Table 1, and developed materials to test each hypothesis independently of the other teams. The team of Xu and Yang designed materials only for Hypotheses 1, 2, and 5, and the team of Cimpian, Tworek, and Storage designed materials only for Hypotheses 3 and 4. We also included the original materials from the unpublished studies that initially supported each hypothesis and conducted direct replications with them; the teams of Uhlmann, Schweinsberg, and Uhlmann and Cunningham only contributed these original materials. The original materials for Hypothesis 5 were developed by the team of Sowden and Hall, but were much longer than any other materials set, so this team also developed a shorter set of materials for Hypothesis 5 and data were collected using both versions. In all, 64 sets of materials, including the five sets of original materials, were created through this crowdsourced process. The materials and analyses for both studies were pre-registered at https://osf.io/9jzy4/ (see also Supplement 1, as well as Supplement 2 for deviations from the pre-registered analyses).

Participants. In total, 8,080 participants located in the United States began the Main Studies on Amazon Mechanical Turk (MTurk; Chandler, Mueller, & Paolacci, 2014, Chandler, Paolacci, & Mueller, 2013); of these, 7,500 completed the entire study. In accordance with our pre-registered stopping rule (see https://osf.io/avnuc/), we ceased data collection after $N = 7{,}500$ participants finished the survey. In the Replication Studies, 7,500 English-speaking adult participants located in the United States were recruited via PureProfile, a survey firm – we employed this different sampling method for the Replication because, in the Main Studies, we had already essentially exhausted the number of Mechanical Turk participants that a typical lab samples (see Stewart et al., 2015). In both data collection efforts, responses from participants who completed their assigned materials for one or more hypotheses but did not complete the all assigned materials in their entirety were retained, resulting in slightly different sample sizes across the five hypotheses (Main Studies: Hypothesis 1 $N = 7{,}175$; Hypothesis 2 $N = 7{,}160$; Hypothesis 3 $N = 7{,}146$; Hypothesis 4 $N = 7{,}158$; Hypothesis 5 $N = 7{,}758$; Replication Studies: Hypothesis 1 $N = 7{,}586$; Hypothesis 2 $N = 7{,}631$; Hypothesis 3 $N = 7{,}568$; Hypothesis 4 $N = 7{,}576$; Hypothesis 5 $N = 8{,}231$). On a per-cell basis, there were approximately 300 participants for Hypotheses 1–4, and 600 participants for Hypothesis 5 (which was tested using a Pearson correlation, rather than a comparison between experimental groups).

Procedure. In the Main Studies, participants were randomly assigned to one set of materials for each of the five hypotheses, and, for designs with multiple conditions, one condition per hypothesis. The order in which the five sets of materials were presented was randomized for each participant. After responding to all five sets of materials, participants completed a demographics questionnaire including questions about their age, gender, and other characteristics. Additionally, a separate subsample of participants was randomly assigned to only complete the full original materials for Hypothesis 5, due to their length. The materials designed by the team of Jiménez-Leal and Montealegre to test Hypothesis 4 were run separately approximately two months after the rest of Main Studies were run, because we discovered that, due to a coding error, one of the two conditions from these materials was not presented to participants in the original run (new data were therefore collected for both conditions of this design). The procedure for the Replication Studies was essentially identical to that of the Main Studies; the only modifications were fixing the aforementioned condition missing from Hypothesis 4, and pre-registering some exploratory analyses conducted on the data from the Main Studies (see Supplement 2), this time as confirmatory tests (see https://osf.io/8s69w/).

Forecasting Study

The online Forecasting Study was open to any scientist, and had two purposes. First, it tested the extent to which researchers ($N = 141$) were able to predict the results of the Main Studies and Replication Studies, in terms of the standardized effect size that would be obtained from each set of materials, and also with regard to statistical significance (the likelihood that a p-value below .05 would be found). Second, it determined how independent reviewers evaluate each set of materials based on whether it provides an adequate test of the original hypothesis. Variability across different study versions is far more meaningful if they provide valid tests of the original research idea. We placed half of the forecasters at random into a monetarily incentivized version of the survey; potential payoff ranged between $0 and $60, meaning financial incentives were present in the treatment condition but not strong. Further methodological details for the forecasting survey can be found in Supplements 3 and 5,

and the pre-registration can be found at https://osf.io/9jzy4/.

Results

Main Studies and Replication Studies

Given our key theoretical question regarding heterogeneity in estimates, as well as large sample sizes that might render even small and theoretically uninteresting differences statistically significant, our primary focus is on dispersion in effect sizes across different study designs. Yet, since the $p < .05$ threshold is widely used as the lower bound criterion for concluding the presence of an effect, we likewise examined patterns of statistical significance, both at the level of individual designs and aggregated across them. This reliance on both effect sizes and statistical significance levels to quantify the project results was pre-registered in advance. Because of the potential issues associated with relying on statistical significance to draw conclusions, we report the results of null hypothesis significance tests in Supplement 9, and focus here on the analyses of effect sizes.

Meta-analytic statistics. To examine the support for each hypothesis, as well as the variation across study designs for each of them, we computed effect size estimates for the results from each of the 64 sets of materials. The diversity in effect size estimates from different study designs created to test the same theoretical ideas constitute the primary output of this project. For Hypotheses 1-4, the effect sizes were independent-groups Cohen's *d*s, and for Hypothesis 5, they were Pearson *r*s. Effect size estimates and sampling variances were calculated via bootstrapping, using the *bootES* package for R (Kirby & Gerlanc, 2013)[2], then combined in random-effects meta-analyses using the *metafor* package (Viechtbauer, 2010), to obtain an overall estimate for the size of each hypothesized effect.[3] This model treats each observed effect size y_i as a function of the average true effect size μ, between-study variability, $u_i \sim N(0, \tau^2)$, and sampling error, $e_i \sim N(0, v_i)$ (see Viechtbauer, 2010)[4]:

$$y_i = \mu + u_i + e_i$$

The heterogeneity among effect sizes (τ^2) was estimated using Restricted Maximum Likelihood Estimation. Positive effect sizes indicate results consistent with the original, unpublished findings, whereas negative effect sizes indicate results in the opposite direction. Figures 1a–1e present forest plots of the observed effect sizes in these analyses. For ease of comparison across the five figures, the Pearson *r* effect sizes for Hypothesis 5 have been converted to Cohen's *d*s (Rosenthal & DiMatteo, 2001). The top panel of each figure presents observed effect sizes and the estimated mean effect size from the Main Studies, and the middle panel presents observed effect sizes and the estimated mean effect size from the Replication Studies. Beneath these panels, the estimated mean effect size for each hypothesis, computed by combining all individual effect sizes in the Main Studies and Replication Studies ($k = 26$ for Hypotheses 1, 2, 3, and 5; $k = 24$ for Hypothesis 4) is presented. The bottom panel presents effect sizes computed by meta-analytically combining the Main Studies' and Replication Studies' effect sizes for each set of materials (i.e., this panel

[2] Materials designed by the team of Donnellan, Lucas, Cheung, and Johnson for Hypotheses 2, 3, and 4 employed within-subjects designs, whereas the other materials for these hypotheses employed one-sample or between-subjects designs. To ensure that all effect sizes were comparable in the meta-analyses, the repeated-measures *d*s for the within-subjects designs were converted to independent-groups *d*s (see Morris & DeShon, 2002). bootES does not have a feature to convert between effect size metrics, so custom bootstrapping code was used (see https://osf.io/avnuc/). This custom code returns the same effect size estimates and variance terms for the repeated-measures *d*s as bootES, and converts the repeated-measures *d*s to independent-groups *d*s according to Equation 11 in Morris and DeShon (2002).

[3] Fixed-effects models showed similar estimated mean effect sizes. In the Main Study, the point estimate was not statistically significant for Hypothesis 1, $p = .093$, and was statistically significant for Hypothesis 4, $p < .001$. In the Replication, the estimated effect sizes were again similar when fixed-effects models were used, but the point estimates for Hypotheses 1 and 4 were statistically significant, $p < .001$. Yet, fixed-effects models are not generally recommended when meta-analyzing studies with different methods (Borenstein, Hedges, Higgins, & Rothstein, 2010; Hunter & Schmidt, 2000), so we focus on the random-effects models.

[4] This analytic approach is not ideal, because it ignores the multivariate nature of the data: each hypothesis can be thought of as a separate outcome variable. It also ignores the multilevel nature of the data (designs are nested within hypotheses), and individual-level correlations across designs resulting from the fact that each participant completed up to five different study designs. We therefore also ran a one-stage multivariate meta-analysis on our individual participant data to model these aspects of the data. The results are very similar to the reported univariate meta-analyses, and this approach has its own disadvantages, particularly that analysis of heterogeneity jointly across outcomes jointly is complicated by the non-nested participant design (see Supplement 8). Therefore, we focus here on the more familiar analytic approach.

presents the results of 12 or 13 meta-analyses, each with k = 2 studies).[5]

[INSERT FIGURES 1A-E ABOUT HERE]

In the Main Studies, these analyses showed a statistically significant aggregated effect in the expected direction for Hypotheses 2, 3, and 5 (estimated mean effect sizes: d = 1.04, 95% *CI* [0.61, 1.47], $p < .001$; d = 0.33, 95% *CI* [0.17, 0.50], $p < .001$; r = .06, 95% *CI* [0.01, 0.11], p = .010), and no statistically significant aggregated effect as expected under Hypotheses 1 and 4 (d = 0.07, 95% *CI* [-0.22, 0.37], p = .623; d = 0.07, 95% *CI* [-0.05, 0.20], p = .269). Note that in the case of Hypothesis 5, the aggregated estimate was very small, and the threshold for statistical significance may only have been crossed due to the large sample and the resultant high power to detect even trivially small effects. In the Replication Studies, the patterns of results were similar, though the estimated mean effect sizes tended to be somewhat smaller, overall. Hypotheses 2 and 3 (d = 0.60, 95% *CI* [0.32, 0.88], $p < .001$; d = 0.24, 95% *CI* [0.11, 0.38], $p < .001$) were associated with a statistically significant effect in the expected direction. Hypothesis 5 did not receive statistically significant overall support in the Replication Studies (r = .03, 95% *CI* [-.04, .09], p = .417), though the estimated mean effect size was not meaningfully different than in the Main Studies. Consistent with the Main Studies, Hypotheses 1 and 4 were again not supported (d = -0.07, 95% *CI* [-0.33, 0.19], p = .588; d = 0.03, 95% *CI* [-0.05, 0.19], p = .465). Overall, then, the meta-analytic results were largely consistent across the Main Studies and the Replication Studies, reflecting overall support for Hypotheses 2 and 3, and an overall lack of support for Hypotheses 1, 4, and 5. Similar results were found when relying on null hypothesis significance testing (see Supplement 9).

Just as importantly, inspection of the forest plots suggests substantial variation among effect sizes, even within the same hypothesis. We assessed this more formally by examining the Q, I^2, and τ^2 statistics in each meta-analysis (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006). The Q statistic is a test for heterogeneity - a significant Q statistic means that heterogeneity in true effects can be expected. Because all participants in each study were drawn from the same large online sample and randomly assigned to conditions, it is unlikely that heterogeneity can be attributed to hidden moderators (e.g., different populations being sampled, different study environments, etc., see Van Bavel et al., 2016), and thus is likely due to differences in the materials. The I^2 statistic quantifies the percentage of variance among effect sizes attributable to heterogeneity, rather than sampling variance. By convention, I^2 values of 25%, 50%, and 75% indicate low, moderate, and high levels of unexplained heterogeneity, respectively (Higgins, Thompson, Deeks, & Altman, 2003). Yet, Q and I^2 are also sensitive to sample size; large samples tend to produce large and significant Q statistics and large I^2 values. Therefore, we also report the τ^2 statistic as an absolute measure of the amount of heterogeneity in our data. The τ^2 statistic is an estimate of the variance of true effect sizes (Borenstein, Hedges, Higgins, & Rothstein, 2009). All five hypotheses showed statistically significant and high levels of heterogeneity in the Main Study and the Replication (see Table 2). In the Main Study, only about 1%, 2%, 6%, 12%, and 24% of the variance across the effect sizes for Hypotheses 1, 2, 3, 4, and 5, respectively, can be attributed to chance variation. Similarly, in the Replication, we would only expect to observe about 1%, 3%, 9%, 22%, and 14% of the variance across the effect sizes for Hypotheses 1, 2, 3, 4, and 5, respectively, by chance. The *vast majority* of observed variance across effect sizes in both studies is unexplained heterogeneity. Moreover, the τ^2 statistics are rather large, relative to the estimated mean effect sizes, suggesting that these large I^2 values are not simply due to our large effect sizes resulting in low sampling variance - there are meaningful levels of absolute heterogeneity in our data. One can also see this pattern simply by visually inspecting the forest plots (Figures 1a-1e), which show considerable dispersion among effect sizes.

[INSERT TABLE 2 ABOUT HERE]

Explaining heterogeneity in effect sizes. We therefore sought to explain this observed heterogeneity. First, we computed intraclass correlation coefficients (ICCs) predicting observed effect sizes from the hypothesis they tested, and the team that designed the materials (see Klein et al., 2014, for a similar analysis). In order to compare across all observed effect sizes, the Pearson rs from Hypothesis 5 were converted into Cohen's ds (Rosenthal & DiMatteo, 2001), as above. In the Main

[5] When meta-analytically combining the Main Studies' and Replication Studies' effect sizes for each individual set of materials, we employed fixed-effects models, unlike in the rest of our meta-analytic models. This is because the two effect sizes being combined come from studies with *identical* materials and methods, so they should, in principle, be measuring the same true population effect size.

Studies, the hypothesis being tested was moderately predictive of observed effect sizes, *ICC* = .40, 95% *CI* [.15, .86], whereas team did not explain statistically significant variance, *ICC* = -.13, 95% *CI* [-.23, .09]. The negative ICC for team indicates that between-team variance is lower than within-team variance. This means that which team designed a set of materials had no predictive relationship with the observed effect size (see Bartko, 1976). In other words, some teams were not "better" than others at designing study materials that produced large effect sizes across hypotheses. We followed up this analysis with a random-effects meta-regression, predicting effect sizes from hypothesis and team, with the median hypothesis (Hypothesis 5, in the Main Study) and the median team (Sowden & Hall) as the reference levels. Hypothesis 2 produced statistically significantly larger effect sizes than the median hypothesis, $\beta = 0.85$, 95% *CI* [0.47, 1.23], $p < .001$, but, consistent with the analysis above, no team produced statistically significantly larger or smaller effect sizes than the median team, *p*s > .086. Moreover, after accounting for both hypothesis and team, there was still substantial and statistically significant residual heterogeneity across effect sizes, $Q(44) = 1291.64$, $p < .001$, $I^2 = 97.39\%$, 95% *CI* [96.22, 98.40], $\tau^2 = 0.24$, 95% *CI* [0.16, 0.38]. In the present research, the subjective choices that researchers make in stimulus design have a substantial impact on observed effect sizes, but if a research team produces a large effect size for one research question, it does not necessarily mean that they will produce a large effect size for another question. This pattern fails to support the hypothesis that some researchers have a "flair" for obtaining large and statistically significant results (see, e.g., Baumeister, 2016). Still, more research is needed on this point, since other research topics (e.g., stereotype threat, motivated reasoning), or more finely parsed subtopics, may yet yield evidence for expertise effects in conducting conceptual replications.[6]

As might be expected, independent ratings of the quality of each study design (assessed in the Forecasting Study) were positively correlated with the obtained results. Higher quality sets of materials yielded larger observed effect sizes in the direction predicted by each original hypothesis (Cohen's *d*s), $r(62) = .31$, $p = .012$. Thus, it is possible that the inclusion of low-quality materials biases our analyses against finding support for hypotheses that are in fact true, when properly tested. We therefore repeated all of the meta-analytic analyses above, excluding 18 sets of materials that were rated as below 5 on a scale of 0 (not at all informative) to 10 (extremely informative) by independent raters in the Forecasting Study. As described in greater detail in Supplement 6, the results were substantively quite similar for all five hypotheses.

It is also possible that rather than artificially reducing the degree of observed support for a given hypothesis, lower quality materials introduce psychometric artifacts such as poor reliability and validity which bias effects toward zero. We therefore further examined whether quality ratings predict larger effect size estimates in absolute terms, in other words larger estimates either consistent or inconsistent with the original hypothesis. Independent ratings of the quality of each study design were directionally positively correlated with the absolute value of the effect size estimates, but this relationship was not statistically significant, $r(62) = .20$, $p = .12$. Overall, the results suggest that the observed variability in effect sizes was not driven by a subset of lower quality study designs.

Aggregating results of the Main Studies and Replication Studies. Leveraging the combined samples of the Main Studies and Replication Studies allowed for more precise effect size estimates from each study version, as well as higher-powered estimates of the overall degree of support for each of the five original hypotheses. Aggregating all of the effect sizes across the two studies in random-effects meta-analyses ($k = 26$ for Hypotheses 1, 2, 3, and 5; $k = 24$ for Hypothesis 4) produced similar results to the separate meta-analyses above. Hypotheses 2 and 3 were supported ($d = 0.82$, 95% *CI* [0.55, 1.08], $p < .001$; $d = 0.29$, 95% *CI* [0.18, 0.39], $p < .001$). Hypothesis 5 was also associated with a statistically significant estimate in the expected direction, though, as above, the effect was negligible in size ($r = .04$, 95% *CI* [.01, .08], $p = .026$), leading to the conclusion that H5 was not empirically supported by the crowdsourced initiative. Later we report a Bayesian analysis casting further doubt on Hypothesis 5. Even under the null hypothesis significance testing framework, Hypotheses 1 and 4 were not supported (d =

[6] We also re-ran these analyses, restricting the data to Hypotheses 3, 4, and 5, which are clearly within the same general area of research, moral psychology, to see if we could find support for the flair hypothesis within a particular area of research. Once again, however, we found no evidence that observed effect sizes are predicted by the identity of the researchers that designed the materials (see Supplement 9 for details).

0.00, 95% *CI* [-0.19, 0.19], $p = .997$, and $d = 0.05$, 95% *CI* [-0.02, 0.13], $p = .179$). We repeated these analyses selecting only study versions rated as 5 or above in informativeness by the independent raters, (see Supplement 6), and nesting study (Main Studies versus Replication Studies) within each hypothesis (see Supplement 9). Both of these additional analyses produced qualitatively similar results to the results above.

Comparing the results of the Main Studies and Replication Studies. As there is no single approach to determining whether an effect directly replicated or not (Brandt et al., 2014; Open Science Collaboration, 2015), we pre-registered a number of criteria for whether the results of the Main Studies held up in the Replication Studies. These included correlating the Main Studies' and Replication Studies' effect sizes, comparing the statistical significance levels and direction of effects, and testing for statistically significant differences between the effect sizes from the Main Studies and the corresponding effect sizes from the Replication Studies. We further examined whether the effect was statistically significant after meta-analyzing across both the Main Studies and Replication Studies (see Figures 1a-1e), and we report a Bayesian analysis of differences in the Main Study and Replication results in Supplement 7.

Each of these criteria is an imperfect and incomplete measure of replication. For instance, a near perfect correlation in effect sizes could emerge even if replication effect sizes were dramatically smaller, so long as the rank ordering of effects remained consistent. Given such a pattern, it would be unreasonable to conclude the effects were robust and replicable. When it comes to comparing whether the replication effect is statistically significantly different from the original effect or not, this method is low in informational value when an original study has a statistically significant *p*-value close to .05 with a lower bound of the confidence interval close to zero. With this *p*-value, it is highly unlikely to find a statistically significant difference from the original result unless the replication point estimate is in the opposite direction of the original finding.

With these caveats in mind, we turn to comparing the results from the Main Studies and Replication Studies. In 51 out of 64 cases (80%), the Replication Studies' effect was directionally consistent with the effect from the Main Studies'. In 36 of those 51 cases (71%), when new participants were run using the same study design, statistically significant results were again statistically significant in the same direction, and non-significant effects were again non-significant. Further, 13 of 44 (30%) statistically significant findings from the Main Studies were not statistically significant in the Replication Studies. At the same time, 6 of 20 (30%) non-significant findings from the Main Studies were statistically significant in the Replication Studies.

We next examined whether effect sizes were significantly different in size between the two studies. We conducted *z*-tests comparing each team-by-hypothesis combination across the two studies (e.g., Team 5's materials for Hypothesis 1 from the Main Studies, versus Team 5's materials for Hypothesis 1 from the Replication Studies). Replication Studies' effect sizes were statistically significantly smaller than the corresponding effect in the Main Studies, according to *z*-tests, in 21 out of 64 cases, and statistically significantly larger in just one case, with no significant difference in 42 out of 64 cases. This pattern agrees with the qualitative observation above that effect sizes tended to be somewhat smaller in the Replication Studies than in the Main Studies. This was quite unexpected – if anything, we anticipated that Mechanical Turk, as the less expensive, more expedient data source, might potentially yield smaller effect sizes. We can only speculate that the general decline effect across the two samples resulted from the slightly different populations of online respondents that were sampled, but the precise difference between the two samples that drove this result is unclear.

When directly replicated, a substantial minority of individual effect sizes reversed direction, changed significance levels across the $p < .05$ threshold, or were statistically significantly different from the initial result. At the same time, correlating the 64 effect sizes obtained in the Main Studies with the 64 effect sizes from the Replication Studies revealed very high correspondence between them in the aggregate, $r(62) = .92$, 95% *CI* [.88, .95], $p < .001$ (see Figure 2). Moreover, descriptively, the major overall findings from our Main Studies emerged in the Replication Studies as well. Effect sizes were again radically dispersed, with statistically significant effects in opposing directions obtained from different sets of materials designed to test three of the five research questions. Meta-analyzing across study versions, Hypotheses 2 and 3 were again supported, and Hypotheses 1 and 4 were not. The directional and statistically significant, but very small estimate for Hypothesis 5 in the Main Studies was not statistically significant in the Replication Studies, yet also not meaningfully different in size (Gelman & Stern, 2006). Variability in effect sizes

was again far more attributable to whether the hypothesis itself enjoyed overall support than to the skill of particular research teams at designing studies that returned large effects (see Supplement 9).

[INSERT FIGURE 2 ABOUT HERE]

Publication bias analyses. We present funnel plots and the results of Egger's test (Egger, Smith, Schneider, & Minder, 1997) for all of our meta-analytic results in Supplement 9. Because *all* of the study designs are reported in this article, there is, by definition, no publication bias in the results we have reported. Yet, we *did* find evidence of funnel plot asymmetries for Hypotheses 1, 2, and 5. As we discuss in greater detail in Supplement 9, these must reflect "sample size effects" that are idiosyncratic to the designs tested in this research. This result highlights one further advantage of crowdsourcing in comparison to the traditional practice of science: In a traditional meta-analysis of multiple studies conducted at different times, one cannot be certain whether funnel plot asymmetries reflect publication bias or some other sample size effect (see, e.g., Deeks, Macaskill, & Irwig, 2005), whereas in a crowdsourced project like this one, there is, by the very nature of the design, no publication bias.

Bayesian perspective on the results. Supplement 7 provides an extended report of Bayesian analyses of the overall project results (the pre-registered analysis plan is available at https://osf.io/9jzy4/). To summarize briefly, the Bayesian analyses find compelling evidence in favor of Hypotheses 2 and 3, moderate evidence against Hypothesis 1 and 4, and strong evidence against Hypothesis 5. Overall, two of five original hypotheses were confirmed aggregating across the different study designs. This pattern is generally consistent with the frequentist analyses reported above, with the exception that the frequentist approach suggests a very small but statistically significant ($p < .05$) effect in the direction predicted by Hypothesis 5 after aggregating across the different study designs, whereas the Bayesian analyses find strong evidence *against* this prediction. The project coordinators, original authors who initially proposed Hypothesis 5, as well as further authors on this article concur with the Bayesian analyses that the effect is not empirically supported by the crowdsourcing hypotheses tests project, due to the small estimate of the effect, and heterogeneity across designs. Regarding the main meta-scientific focus of this initiative, namely variability in results due to researcher choices, for all five hypotheses strong evidence of heterogeneity across different study designs emerged in the Bayesian analyses.

Forecasting Survey

We set up the forecasting survey to test if scientists' predictions about the effect sizes and statistical significance levels (whether $p < .05$ or not) associated with the different sets of study materials would be positively correlated with the realized outcomes. Note that in asking forecasters to predict statistical significance levels, we are not endorsing the idea that something magical happens at $p = .05$, or the binary assumption of there being a result if $p < .05$ and none if $p > .05$ (Greenland, 2017). Yet, given that in many fields and journals this criterion is used to indicate the minimum support required to claim an effect (see McShane & Gelman, 2017), we find that it is interesting to see whether a crowd of researchers can predict this binary outcome.

In addition, we tested whether monetary incentives or individual characteristics of the forecasters increased the accuracy of the predictions. The planned analyses for the forecasting study are detailed at https://osf.io/9jzy4/. Standard errors are clustered at two non-nested levels in all the regressions employing individual-level data: individual level and team-hypothesis version level (i.e., the level of a single study). Double clustering renders estimates robust to potential violations of independence among forecasts generated by the same individual over different versions of the study materials, and among predictions about the same set of study materials generated by different researchers.

Overall accuracy. To test our primary hypotheses regarding the accuracy of scientists' predictions, we examined whether there existed positive correlations between scientists' forecasts and the estimated effect sizes and statistical significance levels ($p < .05$ or not) from the different study versions in the Main Studies, at the team-hypothesis version level. In addition, we performed paired *t*-tests on aggregated prediction data and observed effect sizes to test whether scientists generally underestimated or overestimated the strength of each finding. As hypothesized, we observed a positive correlation between scientists' forecasts and the results being statistically significant in the predicted direction, $r(62) = 0.59$, 95% *CI* [0.40, 0.73], $p < .001$. The correlation between scientists' predictions and the observed effect sizes was likewise statistically significant: $r(62) = 0.71$, 95% *CI* [0.56 ,0.81], $p < .001$.

[INSERT FIGURES 3A AND 3B ABOUT HERE]

We tested whether scientists underestimated or overestimated the realized outcomes by employing paired *t*-tests between the vector collecting the average forecasts and the vectors collecting the effect sizes and directional statistical significance of each study version.

Descriptively, for both effect sizes and directional statistical significance, predictions and outcomes were fairly aligned, with no differences reaching statistical significance. For directional statistical significance in terms of $p < .05$, the mean of the observed outcomes is $M = 0.58$ ($SD = 0.50$) and the mean of the forecasted outcomes is $M = 0.48$ ($SD = 0.09$), $t(63) = -1.78$, 95% *CI* of the difference of the means [-0.21, 0.01], $p = .080$. For effect sizes, the mean of the observed outcomes is $M = 0.31$ ($SD = 0.56$) and the mean of the forecasted outcomes is $M = 0.25$ ($SD = 0.10$), $t(63) = -1.02$, 95% *CI* of the difference of the means [-0.19, 0.06], $p = .311$. Evidence from the analysis of the forecasting survey supports the hypothesis that scientists' predictions are positively correlated with the realized outcomes, both in terms of effect sizes and in terms of whether the result is statistically significant or not for the different sets of study materials. Moreover, the analysis shows no evidence of systematic underestimation or overestimation of the realized outcomes.

Sensitivity to design choices. To test if forecasters were sensitive to how different versions of the materials designed to test the same hypotheses affect research outcomes, we ran individual level regressions. These analyses tested whether scientists could predict results *within* each hypothesis, rather than only across them. The outcome (realized statistical significance in terms of $p < .05$, observed effect size) was the dependent variable and the individual prediction was the independent variable. As for all other individual level regressions, the standard errors were clustered at two non-nested levels: individual level (to account for the fact that each individual made several forecasts) and team-hypothesis version level (to account for the fact that the forecasts about the same set of materials might possibly be correlated). The model was estimated with either hypothesis fixed effects (exploiting only the variation in predictions across teams, as shown in equations (1*a*) and (1*b*)) or team fixed effects (exploiting only the variation in predictions across hypotheses, as shown in (2a) and (2b)).

$$(1a) \qquad SS_{th} = \beta_0 + \beta_1 x_{ith} + Hyp_h + \varepsilon_{ith}$$

$$(1b) \qquad EE_{th} = \beta_0 + \beta_1 \hat{x}_{ith} + Hyp_h + \varepsilon_{ith}$$

The dependent variables SS_{th} and EE_{th} are the realized outcomes, the dummy variable being positive if the study is statistically significant in (1*a*), and realized effect size in (1*b*), respectively. The independent variables are the individuals' forecasts, x_{ith} for the predictions regarding statistical significance in terms of $p < .05$ and $\hat{x}_{ith}$ for the predictions regarding effect size. Hyp_h identify the hypothesis fixed effects, and $Team_t$ are the team fixed effects.

$$(2a) \qquad SS_{th} = \beta_0 + \beta_1 x_{ith} + Team_t + \varepsilon_{ith}$$

$$(2b) \qquad EE_{th} = \beta_0 + \beta_1 \hat{x}_{ith} + Team_t + \varepsilon_{ith}$$

Separately including only hypothesis or only team fixed effects allows us to test if the forecasts are associated with the realized outcomes using only the variation in forecasts within hypotheses (making predictions for the different teams within hypotheses) or only the variation in forecasts within teams (making predictions for the different hypotheses within teams).

The individual prediction coefficient was statistically significant in the expected direction in both the regressions with only hypothesis fixed effects, and in the regressions with only team fixed effects. This holds for predicting both statistical significance levels ($\beta_1 = .148$, $t(9018) = 4.07$, $p < .001$ controlling for hypotheses, $\beta_1 = .255$, $t(9007) = 4.38$, $p < .001$ controlling for teams), and effect sizes ($\beta_1 = 0.097$, $t(9018) = 2.16$, $p = .031$ controlling for hypotheses, $\beta_1 = 0.228$, $t(9007) = 2.68$, $p = .007$ controlling for teams), and shows that forecasters were able to anticipate results from different teams of materials designers within each hypothesis, as well as different hypotheses within each team of materials designers. For completeness, we also estimated the results without any fixed effects ($\beta_1 = 0.309$, $t(9022) = 43.04$ for predictions on whether the result is statistically significant ($p < .05$) or not, $\beta_1 = 0.309$, $t(9022) = 2.38$ for predictions regarding effect size) and with both team and hypotheses fixed effects ($\beta_1 = 0.089$, $t(9003) = 2.78$ for predictions whether the result is statistically significant ($p < .05$) or not, $\beta_1 = 0.091$, $t(9003) = 2.77$ for predictions regarding effect size), and the individual prediction coefficient is statistically significant in these models as well (see Tables S5.4a and S5.5 in Supplement 5).[7] Furthermore, we estimated equations (*1a*) and (*2a*) as

[7] In all four models, there was a statistically significant association between individual forecasts and outcomes. This is true for both the predictions regarding whether the study will find a statistically significant effect in the hypothesized direction and the predictions regarding the realized effect size. Note however that, as the independent variable (i.e., the individual forecasts) are likely to be measured with error, the estimated coefficients reported in this paragraph are potentially biased downwards. Measurement error would artificially reduce the correspondence between forecasts and outcomes, leading to a conservative test of forecaster accuracy.

a probit model (see Table S5.4b in Supplement 5), obtaining similar results as those obtained using the linear probability model. In short, scientists were able to predict not only which hypotheses would receive empirical support (see Figure 3a) but also variability in results for the same hypothesis based on the design choices made by different research teams (see Figure 3b).

We report several further analyses of the Forecasting Study in Supplement 5, for the interested reader. In particular, we examine whether monetary incentives increase the accuracy of forecasts (they do not, at least with the relatively small incentives on offer in this study), whether characteristics of the forecaster, such as job rank and confidence in their forecasts, predict accuracy (they do not consistently do so), and repeat our primary analyses for the data from the Replication Studies and aggregating across the Main Studies and Replication Studies (the results are similar to those reported here).

Discussion

How contingent is support for scientific hypotheses on the subjective choices that researchers make when designing studies? Concerns about the potential dependency of findings on the stimuli used to capture them have been raised repeatedly (e.g., Baribault et al., 2018; Campbell & Fiske, 1959; Judd et al., 2012; Monin & Oppenheimer, 2014; Monin et al., 2007; Wells & Windschitl, 1999). In contrast, the extent to which this problem presents a challenge to conducting research investigations and interpreting research findings has never been directly examined. In this crowdsourced project, when up to 13 independent research teams designed their own studies to test five original research questions, variability in observed effect sizes proved dramatic, with the Bayesian analyses confirming overwhelming evidence of heterogeneity for four of five hypotheses and compelling evidence in the fifth case (see Supplement 7). Descriptively, different research teams designed studies that returned statistically significant effects in opposing directions for the same research question for four out of five hypotheses in the Main Studies, and three out of five hypotheses in the Replication Studies (see Supplement 9). In other words, even when some or most teams created studies that substantiated a theoretical prediction, at least one other team's design found the opposite. Even the most consistently supported original hypotheses still exhibited a wide range of effect sizes, with the smallest range being d = -0.37 to d = 0.26 (Hypothesis 4, Replication Studies). While the hypothesis being tested explained substantial variability in effect sizes (i.e., some hypotheses received more consistent support than others), there remained substantial unexplained heterogeneity after accounting for the hypothesis being tested, implying that idiosyncratic choices in stimulus design have a very large effect on observed results, over and above the overall support (or lack thereof) for the hypothesis in question.

Crowdsourcing makes more transparent the true consistency of support for a scientific prediction, and provides the opportunity to leverage the collective experience and perspectives of a crowd of scientists via aggregation (Bates & Granger, 1969; Galton, 1907; Lorenz, Rauhut, Schweitzer, & Helbing, 2011; Silberzahn et al., 2018; Surowiecki, 2004). Meta-analytically combining effect sizes across the various conceptual replications yielded overall support for two of five of the original predictions, and a Bayesian analysis likewise supported two of five hypotheses. Crowdsourcing hypothesis tests can confirm and disconfirm predictions in a convincing way, by providing converging evidence across independent investigators who are unbiased by each other's approaches or knowledge of the original finding.

Contrary to the "flair" hypothesis (Baumeister, 2016) that some researchers are more adept at obtaining empirical support for their predictions, none of the 15 different teams involved in this project designed studies associated with more consistent support for the original ideas. This non-effect occurred despite variable seniority of team leaders, who ranged from doctoral students to chaired full professors, with citation counts ranging from zero into the tens of thousands. The present findings further suggest that replication results are more attributable to the robustness and generalizability of the original finding than the skill of the scientist carrying out the replication (whether a direct or conceptual replication; Bench et al., 2017; Open Science Collaboration, 2015). Although replicating some studies certainly requires specialized technical knowledge (e.g., of neuroimaging technology), evidence that disappointing reproducibility rates for published research (e.g., Dewald, Thursby, & Anderson, 1986; Klein et al., 2014; LeBel, 2015; Open Science Collaboration, 2015) are due to a dearth of replicator competence remains lacking. That said, further meta-scientific work is needed on the role of expertise in replication results (Tierney et al., 2019b).

A substantial degree of variability in the results was accounted for by the original hypotheses themselves, which — as noted earlier — differed in their overall empirical support (see Figure 1). Although the original effects all replicated using the original materials (when

combining the results of the Main Studies and Replication Studies), three effects were unsupported overall in the alternative study designs, in some cases returning estimates in the opposite direction than predicted.

As confirmed in the Bayesian analyses of the project results (Supplement 7), all five original hypotheses exhibited wide variability in support across different study designs. Although the present project was able to parse the two, in typical research contexts this heterogeneity in results due to study design choices co-exists and potentially interacts with heterogeneity in results due to population differences (McShane et al., 2019; Tiokhin et al., 2018). Discrepant results and variability in research findings (Open Science Collaboration, 2015; Schweinsberg et al., 2016) are perhaps unavoidable, and might best be embraced as a normal aspect of the scientific process. In terms of building solid theory, it may be necessary to vary stimuli and study designs (Baribault et al., 2018; Caruso et al., 2017; the present initiative), employ a variety of statistical specifications (Silberzahn et al., in 2018; Simonsohn, Simmons, & Nelson, 2016; Steegen et al., 2016), and replicate findings across more geographic locations and populations (Henrich, Heine, & Norenzayan, 2010), before drawing definitive conclusions. With regard to communicating findings both within and outside the scientific community, more conservative messaging regarding new research conducted in a single population or relying heavily on a specific experimental paradigm seems warranted.

Implications for the Five Original Hypotheses

The primary goal of this initiative was to examine effect size dispersion when independent investigators design studies to address the same research questions. A secondary purpose was to evaluate the evidence for the five original effects targeted in the crowdsourced conceptual replications. Below we assess current support and potential future directions for Hypothesis 1-5, in consultation with the original team that volunteered each research idea for the initiative.

Hypothesis 1: Awareness of automatic prejudice. This effect directly replicated using the original Uhlmann and Cunningham (2000) questionnaire items, with participants in both the Main Study and Replication expressing overall agreement to the items "Although I don't necessarily agree with them, I sometimes have prejudiced feelings (like gut reactions or spontaneous thoughts) that I don't feel I can prevent", and "At times stereotypical thoughts about minorities coming into my head without my necessarily intending them to." As in the original data collections by Uhlmann and Cunningham (2000), mean responses to these items were significantly above the neutral scale midpoint of four (*1= strongly disagree, 4= neutral, 7= strongly agree*). At the same time, conceptual replications by different research teams employing alternative questions failed to confirm the hypothesis that participants so openly self-report automatic prejudices. Aggregating across the different study designs via meta-analysis reveals no statistically significant effect in the expected direction, and a Bayesian analysis found moderate evidence against H1. On reflection, the double-barreled nature of the original items, invoking both lack of intentions and prejudiced reactions, as well as the use of qualifiers ("I sometimes", "At times") might have biased participants' responses towards agreement. Further shortcomings of the original study design are the lack of a relative comparison group (e.g., non-minorities and members of dominant groups such as White men), and the absence of any probe items regarding positive or favorable thoughts.

In sum, the present initiative to crowdsource hypothesis tests casts serious doubt on whether overall endorsement of self-perceived automatic prejudice is generally as high as initially reported by Uhlmann and Cunningham (2000). Yet, it does not call into question evidence that different measures of beliefs are correlated at an individual level with scores on implicit measures of attitude (Hahn et al., 2014) and that awareness of automatic associations can be experimentally increased (Hahn & Gawronski, 2019). As of yet there are no systematic reviews or meta-analyses on the empirical relationships between awareness indices and automatic associations. From the present crowdsourced project, we cannot conclude that everyday people believe themselves to be as biased as implicit and indirect measures of automatic associations suggest they are. Indeed, the present results, relying on a wide array of study designs, suggest they do not generally see themselves as implicitly prejudiced. Opportunities to improve validated self-report measures of beliefs about one's automatic prejudices towards various social groups, and to use them as predictors and outcome measures in future investigations, remain open.

Hypothesis 2: Extreme offers reduce trust. This crowdsourcing initiative found consistent evidence for Hypothesis 2 across the range of conceptual replications, as well as in the direct replications using the original materials. Both frequentist and Bayesian analyses supported this particular prediction, with the Bayesian analyses confirming compelling evidence for this hypothesis despite heterogeneity in estimates across

different study designs. This result is consistent with a recent meta-analysis (Huffmeier, 2014), which found that "hardline" negotiation tactics (of which extreme first offers are one example) are associated with more negative "socioemotional" outcomes in negotiations (i.e., perceptions that the hardline negotiator is unreasonable and uncooperative). However, this meta-analysis did not *specifically* examine extreme first offers or trust. Although our findings provide initial support for the idea that extreme first offers indeed reduce trust on the part of the recipient, that this reduced trust consequently diminishes information exchange, and value creation remains to be demonstrated. It also remains unclear to what extent such effects generalize across cultures. Given that negotiators in some cultural settings may be more accustomed to receiving extreme first offers than negotiators in other cultural settings, this effect may indeed be culturally moderated. This possibility is currently being examined in an ongoing international replication project (Schweinsberg et al., 2019) that will assess the cultural boundary conditions of this effect.

Hypothesis 3: Moral praise for needless work. Earlier findings that Americans morally praise individuals who continue at their job after coming into sudden wealth were likewise confirmed by the crowdsourced initiative. Aggregating via meta-analysis across distinct studies independently created by different research teams, both the frequentist and Bayesian analyses find compelling evidence in favor of the needless work hypothesis. Although the robustness of the effect to different operationalizations is now confirmed in two large U.S. samples via the present host of conceptual replications, the original hypothesis of cross-cultural variability has yet to be put to a rigorous empirical test. The original research predicted that praise for those who work in the absence of any material need is steeped in the Protestant work ethic, and hence should be strongest among those with greater degrees of exposure to U.S. culture (Poehlman, 2007; Uhlmann et al., 2009).

As there is no systematic literature review or meta-analysis on this topic, an ongoing crowdsourced project by Tierney et al. (2019a) will attempt to directly replicate this and other original findings regarding work morality across four countries (the United States, the United Kingdom, Australia, and India). Relying on a "creative destruction" approach to replication, the Tierney et al. (2019a) initiative will pit the original prediction that moral praise for needless work only characterizes U.S. culture against theories positing the general moralization of work across cultures, regional differences within the United States (i.e., New England vs. other regions; Fisher, 1989), and valorization of work as a means of personal fulfillment in post-materialist societies (Inglehart, 1997; Inglehart & Welzel, 2005). Thus, further facets of the robustness, generalizability, and potential cultural boundedness of this effect remain to be explored in future research. For now, we conclude that aggregating across the crowdsourced study designs, the needless work hypothesis is supported for U.S. participants, but the originally hypothesized moderation by culture (Poehlman, 2007; Uhlmann et al., 2009) remains to be demonstrated.

Hypothesis 4: Proximal authorities drive legitimacy of performance enhancers. The original finding that the dictates of proximal authorities (e.g., the league, the competitive circuit) have a larger impact on judgments of the acceptability of using performance enhancing drugs (PEDs) than the law was not supported in this crowdsourced initiative. Although the finding directly replicated using the original materials, across a dozen different, independently-developed study designs, people were not more opposed to the use of PEDs when they are banned by a proximal authority than when they are illegal, and the Bayesian analysis found moderate evidence against this hypothesis. This result concurs with follow-up studies done by the research team who contributed this hypothesis (Landy, Walco, & Bartels, 2017), which were conducted after this project began. These subsequent studies find that both types of authority contribute to normative judgments of PED use, to similar degrees. There is currently no systematic review or meta-analysis of judgments of PED use, but Landy, Walco, and Bartels (2017) employed an exploratory, "deep-dive" methodology, in which they tested 11 different potential explanations for opposition to the use of these substances. They concluded that PED use is opposed for three primary reasons: it violates moral norms of fairness, it poses a risk of harm to the user, and it tends to violate legitimate conventional rules. The present results help to clarify this last reason, by showing that the precise source of those rules - the law or a more proximal authority - does not affect levels of opposition.

Hypothesis 5: Deontological judgments predict happiness. Although the original pattern of results once again directly replicated using the original materials, the hypothesis that individuals who tend to make deontological (vs. utilitarian) judgments report different levels of personal happiness was not supported overall by the crowdsourced conceptual replications. Although a statistically significant directional effect in support of H5

was reported in the Main Studies, the aggregated estimate was close to zero, and the effect did not reach statistical significance in the Replication Studies. Overall, the Bayesian analysis found strong evidence *against* this original prediction. There has not previously been a systematic review or meta-analysis of the relationship between moral stance and happiness, though prior research has linked both processes to emotional and intuitive responding (e.g., Everett et al., 2016; Greene, 2013; Lieberman, 2013; Phillips et al., 2017; Singer, 2005). These results fail to find support for an association between deontological moral judgments and hedonic happiness that has been suggested – although not empirically confirmed – by this prior work. Although laypeople appear to believe that part of what brings happiness is living a moral life (Phillips et al., 2017; Phillips et al., 2014), adherence to deontological vs. utilitarian ethical principles does not seem to relate to one's overall happiness.

Forecasting Findings

Scientists can predict whether a published finding will replicate from the research reports (Camerer et al., 2016; Dreber et al., 2015) and benchmark findings plus the materials for further experimental conditions (DellaVigna & Pope, 2018a, 2018b). We find that examination of the materials for an unpublished study is sufficient for scientists to successfully anticipate the outcome. In our forecasting survey, predictions by independent scientists were significantly correlated with both effect sizes and whether the observed results were statistically significant in the hypothesized direction, and the average predictions were similar to the observed outcomes. Monetary incentives failed to improve forecasters' predictive performance. Although speculative, it is possible that scientists who opted into and completed an extensive survey about predicting research findings were sufficiently intrinsically motivated to be accurate, so external incentives did not further increase their motivation (see Lakhani & Wolf, 2005, and Lakhani, Jeppesen, Lohse, & Panetta, 2007, regarding the tendency for crowdsourced initiatives to leverage intrinsic motivations). Another potential explanation is that the financial incentives (up to $60) were not sufficiently strong to affect accuracy.

Comparatively more senior academics (in terms of job rank) were more accurate at forecasting statistical significance levels (i.e., whether the study's outcome would be $p < .05$ in the predicted direction or not), but not effect sizes (see Supplement 5). Other indices of scientific eminence, such as number of peer reviewed publications, were unrelated to forecasting accuracy. In a separate investigation, DellaVigna and Pope (2018a) found that more senior academics (in terms of job rank and citations), if anything, underperformed junior academics at predicting how different incentives would influence the effort and performance of experimental subjects. Moreover, academics in general did no better than lay people (undergraduates, MBA students, and MTurk workers) at rank-ordering the effectiveness of different experimental treatments (DellaVigna & Pope, 2018a). More research is needed on whether traditional indices of scientific eminence (Sternberg, 2016; Vazire, 2017) are associated with any advantage in designing or predicting the results of scientific studies.

Unique to the present study, we show that independent scientists are not only able to predict study results with some success by merely examining the materials, but are also sensitive to how design choices influence the degree of empirical support for a specific claim. Forecasters predicted research results with significant accuracy not just across but also within each of the five hypotheses. This suggests some fine-grained sensitivity to how different operationalizations of the same hypothesis can impact results. More forecasting surveys and other tools aggregating beliefs such as prediction markets are needed to determine the accuracy of scientists' intuitions about how contextual factors affect research outcomes— for instance, whether scientists are able to anticipate cultural differences in effects, and whether specializing in research on culture confers any special advantage. Ongoing projects from our group examine whether academics can predict the heterogeneity statistics in replication results for prime-to-behavior effects (Tierney et al., 2019b), differences in replication effect sizes when the same experiment is run in multiple laboratories (Schweinsberg et al., 2019), and whether findings from the field of strategic management generalize to other time periods and geographies (Delios et al., 2019).

Limitations and Future Directions

This project represents an early foray into the crowdsourcing of stimulus selection and study designs (see also Baribault et al., 2018), with important limitations that should be addressed in future initiatives. The primary meta-scientific purpose of this initiative was to examine the impact of scientists' design choices on effect size estimates. Still, a number of aspects of our approach may have led to artificial homogeneity in study designs. In particular, materials designers were restricted to creating simple experiments with a self-reported dependent measure that could be run online in five minutes or less.

Further, the key statistical test of the hypothesis had to be a simple comparison between two conditions (for Hypotheses 1-4), or a Pearson correlation (for Hypothesis 5). Full thirty-minute- to hour-long-laboratory paradigms with factorial designs, research confederates, and more complex manipulations and outcome measures (e.g., behavioral measures) contain far more researcher choice points and may be associated with even greater heterogeneity in research results. In addition, the project coordinators recruited the materials designers from their own social networks, potentially biasing the project towards demographic and intellectual homogeneity (Ibarra, 1995, 1997). Future initiatives should recruit materials designers more broadly, to better represent the diversity of perspectives within a field or subfield (McGuire, 1973; Monin et al., 2007; Duarte, Crawford, Stern, Haidt, Jussim, & Tetlock, 2015).

Another limitation is that our participants all participated in tests of multiple research questions. In meta-analysis, it is typically assumed that all samples are independent of one another, but this assumption is violated in our data. This assumption is not problematic in the univariate meta-analyses that we present in the main text, but it does complicate the multivariate meta-analysis we present in Supplement 8. Future crowdsourced initiatives could perhaps assign each participant to only *one* research design, or focus exclusively on a single research question, to avoid these participant-level correlations across hypotheses, which are not accounted for in our primary analyses. This would allow for a straightforward multivariate meta-analytic approach, in which participants are nested within designs, which are nested within hypotheses. Each research team was also free to develop their own dependent measures, which meant that we could not directly compare raw results across different designs, but could only compare standardized effect sizes. Future projects in this vein might constrain dependent measures to allow clean, straightforward comparisons of the effects of multiple, independently developed experimental manipulations.

This initiative to crowdsource hypothesis tests also targeted only five original hypotheses, leaving us unable to identify which features of a research idea might be associated with more or less heterogeneity in study designs and outcomes. Some research ideas may naturally feature a greater latitude of construal (Beck, McCauley, Segal, & Hershey, 1988; Dunning, Meyerowitz, & Holzberg, 1989), leading different teams to create more varied experimental paradigms in order to test them. In the extreme, hypotheses that are theoretically underspecified (unlike the present H1-H5) may result in a chaos of operationalizations as the materials designers impose their own priors and assumptions on the idea. Thus, one way to reduce the role of subjective researcher choices in research outcomes may be to more fully flesh out the underlying theory at the outset (Dijksterhuis, 2014; McGuire, 1973; Stroebe & Strack, 2014).

Our limited number of target hypotheses also means one cannot generalize the present results to all hypotheses in all subfields. We cannot conclude that only 40% of research ideas that directly replicate will be supported in conceptual replications, or that for the majority of research questions different designs will return statistically significant effects in opposing directions. Those are the results of this project only, and further initiatives to crowdsource hypothesis tests are needed before drawing definitive conclusions about the impact of subjective researcher choices on empirical outcomes.

Perhaps the most concrete methodological limitation of the present project was the modest sample of forecasters (N = 141), which reduced the statistical power of the relevant analyses. Our sample size was comparable to those for prior surveys examining the forecasting abilities of academics. For example, DellaVigna and Pope (2018a, 2018b) recruited 208 academics for their forecasting research, Dreber et al. (2015) had 47 and 45 active traders in their two prediction markets for replications, Camerer et al.'s (2016) prediction market had 97 participants, Forsell et al. (in press) included 78 participants, and Camerer et al. (2018) featured two conditions with 114 and 92 participants in each treatment. For the present project, we recruited the largest sample we could, given our forecasters' massive task of reviewing, making quality assessments, and predicting the results from 64 distinct sets of experimental materials. Still, the relatively small group of forecasters in our survey indeed limits our conclusions. Furthermore, there may be overlap between the samples of forecasters included in this and other studies, as they were recruited by similar methods. Further research is needed using higher-powered designs, especially with regards to the potential role of forecaster characteristics in moderating predictive accuracy.

Finally, the crowdsourcing hypothesis tests approach shares certain costs and benefits with other crowd approaches to scientific research (Uhlmann et al., in press). In comparison to the standard approach of relying on a small team, recruiting a crowd of collaborators enables big science, democratizes access to projects, and more

effectively assesses the robustness of the findings. Yet at the same time, crowdsourcing study designs is inefficient, in that for the same effort and expense, initial evidence for a far greater number of interesting ideas could have been obtained using a small team or solo investigator approach. In future work, the return on investment from crowdsourcing hypothesis tests may be greatest for theoretically important findings that are well established with a specific paradigm, and whose robustness to alternative methodological approaches is of general interest.

Conclusions

The present crowdsourced project illustrates the dramatic consequences of researcher design choices for scientific results. This initiative also provides a roadmap for future crowdsourced approaches to testing the generality of scientific theories. If a scientific prediction is theoretically important enough, or has practically significant policy and societal implications, future investigations could assign it to multiple laboratories to independently operationalize and carry out empirical tests. The extent to which the results converge (and diverge) across investigations can then be used to inform discussion and debate, revise theory, and formulate policy.

Scientists craft theories with the ambitious goal of unifying potentially disparate findings into coherent, generalizable structures of knowledge. This process is often arduous and lengthy and may be impeded by features of the standard approach to scientific inquiry. Nonetheless, this process can be streamlined through collective action. As the present investigation demonstrates, bringing many perspectives and operationalizations to bear on hypotheses provides a richer account of phenomena than would occur if researchers and teams worked in isolation. Moreover, we also showed that independent researchers are able to identify not only the hypotheses that are more likely to be supported by an empirical investigation but also the research designs that, within a specific hypothesis, are more likely to lead to significant effect. This suggests that researchers can determine the features of the hypotheses, of the methods, and of the research designs that are systematically associated with the effect size and the statistical significance of a research question. Through crowdsourced collaborations such as this one, researchers can craft theories with more confidence and better understand just how far they extend.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., … Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556-578.

Ames, D. R., & Mason, M. F. (2015). Tandem anchoring: Informational and politeness effects of range offers in social exchange. *Journal of Personality and Social Psychology*, *108*(2), 254–274.

Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology*, *59*, 177–181.

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger III, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 117–150). Washington, DC: American Psychological Association.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., ... & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607-2612.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, *83*, 762-765.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, *20*, 451-468.

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, *66*, 153-158.

Beck, L., McCauley, C., Segal, M., & Hershey, L. (1988). Individual differences in prototypicality judgments about trait categories. *Journal of Personality and Social Psychology*, *55*, 286-292.

Belsie, L., (2011). Powerball numbers: Why do lottery winners keep working? *The Christian Science Monitor*. Available at: https://www.csmonitor.com/Business/new-economy/2011/0602/Powerball-numbers-Why-do-lottery-winners-keep-working

Bench, S. W., Rivera, G. N., Schlegel, R. J., Hicks, J. A., & Lench, H. C. (2017). Does expertise matter in replication? An examination of the Reproducibility Project: Psychology. *Journal of Experimental Social Psychology*, *68*, 181-184.

Bentham, J. (1970). An introduction to the principles of morals and legislation. London: Althone Press (Original work published 1823).

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Chichester (UK): Wiley.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97-111.

Brandt, M. J., Ijzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R., … van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50,* 217-224.

Brehm, J. W. (1956). Post decision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology, 52(3),* 384-389.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability

of laboratory experiments in economics. *Science, 351,* 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E-J., & Wu, H. (2018). Evaluating replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*, 637-644.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56(2),* 81-105.

Caruso, E. M., Shapira, O., & Landy, J. F. (2017). Show me the money: A systematic exploration of manipulations, moderators, and mechanisms of priming effects. *Psychological Science, 28,* 1148-1159.

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

Chandler, J., Paolacci, G., & Mueller, P. (2013). Risks and rewards of crowdsourcing marketplaces. In P. Michelucci (Ed.) *Handbook of Human Computation*. New York, NY: Sage.

Coffman, L., & Niehaus, P. (2014). *Pathways of persuasion*. Working paper.

Crandall, C.S., & Sherman, J.W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology, 66*, 93-99.

Deeks, J. J., Macaskill, P., & Irwig, L. (2005). The performance tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology, 58,* 882-893.

Delios, A., Tan, H., Wu, T., Wang, Y., Viganola, D., Gordon, M., Dreber, A., Johannesson, M., Pfeiffer, T., & Uhlmann, E. (2019). *Can you step into the same river twice? Examining the context sensitivity of research findings from archival data*. Project in progress.

DellaVigna, S., & Pope, D.G. (2018a). Predicting experimental results: Who knows what? *Journal of Political Economy, 126,* 2410-2456.

DellaVigna, S., & Pope, D. (2018b). What motivates effort? Evidence and expert forecasts. *The Review of Economic Studies, 85*(2), 1029-1069.

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in empirical economics: The journal of money, credit and banking project. *American Economic Review, 76*, 587-603.

Diener, E., Emmons, R.A., Larson, R.J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71-75.

Dijksterhuis, A. (2014). Welcome back theory! *Perspectives on Psychological Science, 9*(1), 72-75.

Dodge, T., Williams, K. J., Marzell, M., & Turrisi, R. (2012). Judging cheaters: Is substance misuse viewed similarly in the athletic and academic domains? *Psychology of Addictive Behaviors, 26*(3), 678-682.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*, 62–68.

Doyen, S., Klein, O., Simons, D. J., & Cleeremans, A. (2014). On the other side of the mirror:
Priming in cognitive and social psychology. *Social Cognition, 32*, 12–32.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y. Nosek B.A., &
Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of
scientific research. *Proceedings of the National Academy of Sciences*, *112*, 15343-15347.

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. (2015). Political
diversity will improve social and personality psychological science. *Behavioral and Brain Sciences, 38,* 1-13.

Dunaway, B., Edmonds, A., & Manley, D. (2013). The folk probably do think what you think they think. *Australasian Journal of Philosophy, 91(3),* 421-441.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The
role of idiosyncratic trait definitions in self- serving assessments of ability. *Journal of Personality and Social Psychology, 57,* 1082-1090.

Earp, B. D. (in press). Falsification: How does it relate to reproducibility? In J.-F. Morin, C. Olsson, & E. O. Atikcan (Eds.), *Key Concepts in Research Methods.* Oxford: Oxford University Press.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68-82.

Egger, M., Smith, G. D., Schneider, M. & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315,* 629-634.

Everett, J.A.C., & Earp, B.D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology, 6*(1152), 1-4.

Everett, J.A.C., Pizarro, D.A., & Crockett, M.J. (2016). Inference of trustworthiness from
intuitive moral judgments. *Journal of Experimental Psychology: General, 145*(6), 772-787.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic
activation as an unobtrusive measure of racial attitudes: A bona-fide pipeline? *Journal of Personality and Social Psychology, 69*, 1013–1027.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford University Press.

Festinger, L. & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58*, 203-210.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology:
Illustrating epistemological and pragmatic considerations with the case of relationship
science. *Journal of Personality and Social Psychology, 108*, 275-297.

Finkel, E. J., Eastwick, P. E., & Reis, H. T. (2017). Replicability and other features of a high-quality science: Toward a

balanced and empirical approach. *Journal of Personality and Social Psychology, 113*, 244-253.

Fisher, D. H. (1989). *Albion's seed: Four British folkways in America.* New York, NY: Oxford University Press.

Fitz, N. S., Nadler, R., Manogaran, P., Chong, E. W. J., & Reiner, P. B. (2014). Public attitudes toward cognitive enhancement. *Neuroethics, 7*, 173-188.

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B.N., Johannesson, M. & Dreber, A. (in press). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology.*

Galinsky, A. D., Leonardelli, G. J., Okhuysen, G. A., & Mussweiler, T. (2005). Regulatory focus at the bargaining table: Promoting distributive and integrative success. *Personality and Social Psychology Bulletin*, *31*(8), 1087–1098.

Galinsky, A. D., & Mussweiler, T. (2001). First offers as anchors: The role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology*, *81*(4), 657–669.

Galton, F. (1907). Vox Populi. *Nature*, *75*, 450–451.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*, 692–731.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328-331.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the

publication bottleneck but undermine science. *Perspectives on Psychological Science, 7*, 562–571.

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them.* New York, NY: Penguin Press.

Greenland, S. (2017). Invited commentary: The need for cognitive science in methodology.

American Journal of Epidemiology 186(6), 639-646.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological*

Bulletin, *82*, 1-20.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and

stereotypes. *Psychological Review, 102*, 4–27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

Groh, M., Krishnan, N., McKenzie, D., & Vishwanath, T. (2016). The impact of soft skill training on female youth employment: Evidence from a randomized experiment in Jordan. *IZA Journal of Labor and Development, 5*(9), 1-23.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2017). Informed Bayesian *t*-tests. *Manuscript submitted for publication*. Retrieved from https://arxiv.org/abs/1704.02479

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ... Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology, 81*, 80–97. Retrieved from https://doi.org/10.1016/j.jmp.2017.09.005

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). bridgesampling: An R package for estimating normalizing constants. *Manuscript submitted for publication and uploaded to arXiv*. Retrieved from https://arxiv.org/abs/1710.08162

Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology, 2*, 123–138.

Gunia, B. C., Swaab, R. I., Sivanathan, N., & Galinsky, A. D. (2013). The remarkable robustness

of the first-offer effect: Across culture, power, and issues. *Personality and Social Psychology Bulletin*, *39*(12), 1547–1558.

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to

acknowledgement. *Journal of Personality and Social Psychology*, *116*, 769-794.

Hahn, A., Judd, C.M., Hirsh, H.K., & Blair, I.V. (2014). Awareness of implicit attitudes.

Journal of Experimental Psychology: General, *143*, 1369–1392.

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: Are they important? *Journal of Social Behavior and Personality*, *5*(4), 41–49.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?

Behavioral and Brain Sciences, 33, 61-83.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal, 327*, 557-560.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods, 11*, 193-206.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*, 275-292.

Ibarra, H. (1995). Race, opportunity and diversity of social circles in managerial managers'

networks. *Academy of Management Journal, 38*(3), 673-703.

Ibarra, H. (1997). Paving an alternate route: Gender differences in network strategies for

career development. *Social Psychology Quarterly, 60*(1), 91-102.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and*

political change in 43 societies. Princeton, NJ: Princeton University Press.

Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change, and democracy: The*

human development sequence. Cambridge, MA: Cambridge University Press.

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

http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0020124

Ioannidis, J. P. A., & Trikalinos T. A. (2007). An exploratory test for an excess of significant

findings. *Clinical Trials, 4*, 245-253.

JASP Team. (2018). *JASP (Version 0.9.2.0)[Computer software]*. Retrieved from https://jasp-stats.org/

Jordan, J.J., Hoffman, M., Bloom, P., & Rand, D.G. (2016). Third-party punishment as a costly

signal of trustworthiness. *Nature, 530*, 473-476.
Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social
psychology: A new and comprehensive solution to a pervasive but largely ignored
problem. *Journal of Personality and Social Psychology, 103*, 54-69.
Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience, 10*(5), 551-560.
Kahneman, D., Diener, E., & Schwarz, N., eds. (1999). *Well-being: The foundations of hedonic psychology.* New York, NY: Russell Sage Foundation.
Kant, I. (1993). *Grounding of the metaphysics of morals*, 3rd ed. Trans. J.W. Ellington. Indianapolis: Hackett (Original work published 1775).
Kirby, K. N., & Gerlanc, D. (2013). BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods, 45*, 905-927.
Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45(3)*, 142–152.
Lakhani, K. R. & Wolf, R. G. (2005). Why hackers do what they do: Understanding motivation
and effort in free/open source software projects. In J. Feller, B. Fitzgerald, S. Hissam, & K. R. Lakhani (Eds.), *Perspectives on free and open source software*, pp. 3-21. Cambridge, MA: MIT Press.
Lakhani, K. R., Jeppesen, L. B., Lohse, P. A., & Panetta, J. A. (2007). *The value of openness in*
scientific problem solving. Division of Research, Harvard Business School.
Landy, J. F., Walco, D. K., & Bartels, D. M. (2017). What's wrong with using steroids? Exploring whether and why people oppose the use of performance-enhancing drugs. *Journal of Personality and Social Psychology, 113*, 377-392.
LeBel, E. P. (2015, October 13). A list of successful and unsuccessful high-powered direct
replications of social psychology findings.
https://proveyourselfwrong.wordpress.com/2015/10/13/a-list-of-successful-and-unsucces
sful-high-powered-direct-replications-of-social-psychology-findings/
LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified
framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science, 1*(3), 389-402.
Lieberman, M.D. (2013). *Social: Why our brains are wired to connect.* New York, NY: Crown
Publishers.
Loschelder, D. D., Swaab, R. I., Trötschel, R., & Galinsky, A. D. (2014). The first-mover
disadvantage: The folly of revealing compatible preferences. *Psychological Science, 25*(4), 954–962.
Loschelder, D. D., Trötschel, R., Swaab, R. I., Friese, M., & Galinsky, A. D. (2016). The
information-anchoring model of first offers: When moving first helps versus hurts negotiators. *Journal of Applied Psychology, 101*(7), 995–1012.
Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can
undermine the wisdom of crowd effect. *Proceedings of the National Academy of*
Sciences, 108, 9020–9025.
Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*, 1122-1135.
Maaravi, Y., & Levy, A. (2017). When your anchor sinks your boat: Information asymmetry in
distributive negotiations and the disadvantage of making the first offer. *Judgment and Decision Making, 12*(5), 420–429.
Makel, M.C., Plucker, J.A., & Hegarty, B. (2012). Replications in psychology research how
often do they really occur? *Perspectives on Psychological Science, 7*(6), 537–542.
McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology, 26*(3), 446-456.
McGuire, W.J. (1983). A contextualist theory of knowledge: Its implications for innovations and
reform in psychological research. In L. Berkowitz (Ed.), *Advances in Experimental*
Social Psychology (Vol. 16, pp. 1-47). New York, NY: Academic Press.
McShane, B.B., & Gelman, A. (2017). Abandon statistical significance. *Nature, 551*(7682), 558.
McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large scale replication
projects in contemporary psychological research. *The American Statistician, 73*, 99-105.
Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica, 6*, 831–860.
Mill, J. S. (1861/2004). *Utilitarianism and other essays*. London, UK: Penguin Books.
Monin, B., & Oppenheimer, D.M. (2014). The limits of direct replications and the virtues of
stimulus sampling [Commentary on Klein et al., 2014]. *Social Psychology, 45*, 299-300.
Monin, B., Pizarro, D., & Beer, J. (2007). Deciding vs. reacting: Conceptions of moral judgment
and the reason-affect debate. *Review of General Psychology, 11*(2), 99-111.
Morey, R. D., & Rouder, J. N. (2015). *BayesFactor 0.9.11*. Comprehensive R Archive Network. Retrieved from
http://cran.r-project.org/web/packages/BayesFactor/index.html
Morris, S. B. & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-group designs. *Psychological Methods, 7,* 105-125.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.
Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration
revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600-2606.
Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and

practices to promote truth over publishability. *Perspectives on Psychological Science, 7,*
615-631.
Open Science Collaboration (2015). Estimating the reproducibility of psychological science.
Science, 349(6251).
Pashler, H. & Harris, C. (2012). Is the replicability crisis overblown? Three arguments
examined. *Perspectives on Psychological Science, 7,* 531-536.
Pfeiffer, T., Bertram, L., & Ioannidis, J. (2011). Quantifying selective reporting and the Proteus phenomenon for multiple datasets with similar bias. *PLOS ONE, 6,* e18362.
Phillips, J., De Freitas, J., Mott, C., Gruber, J., & Knobe, J. (2017). True happiness: The role of morality in the folk concept of happiness. *Journal of Experimental Psychology: General, 146*(2), 165-181.
Poehlman, T.A. (2007). *Ideological inheritance: Implicit Puritanism in American moral cognition.* Doctoral dissertation, Yale University.
R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
Riley, R. D., Price, M. J., Jackson, D., Wardle, M., Gueyffier, F., Wang, J., … White, I. R. (2015). Multivariate meta-analysis using individual participant data. *Research Synthesis Methods*, *6*(2), 157–174. doi:10.1002/jrsm.1129
Rosenthal, R. (1979). The "file drawer problem" and the tolerance for null results. *Psychological*
Bulletin, 86(3), 638-641.
Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59-82.
Ryff, C. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology, 57*(6), 1069-1081.
Sanders, M., Mitchell, F., & Chonaire, A.N. (2015). *Just common sense? How well do experts*
and lay-people do at predicting the findings of behavioural science experiments. Working
paper.
Sattler, S., Forlini, C., Racine, E., & Sauer, C. (2013). Impact of contextual factors and substance
characteristics on perspectives toward cognitive enhancement. *PLOS ONE, 8*(8), e71542.
Scheibehenne, B., Gronau, Q. F., Jamil, T., & Wagenmakers, E.-J. (2017). Fixed or random? A resolution through model-averaging. Reply to Carlsson, Schimmack, Williams, and Burkner. *Psychological Science, 28,* 1698–1701.
Scheske, C. & Schnall, S. (2012). The ethics of "smart drugs": Moral judgments about healthy
people's use of cognitive-enhancing drugs. *Basic and Applied Social Psychology, 34,* 508-515.
Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple study articles. *Psychological Methods, 17(4),* 551-566.
Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected
in the social sciences. *Review of General Psychology, 13*(2), 90-100.
Schweinsberg, M. (2013). *Starting high shrinks the pie.* Unpublished raw data.
Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., … Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55-67.
Schweinsberg, M., Ku, G., Wang, C. S., & Pillutla, M. M. (2012). Starting high and ending with nothing: The role of anchors and power in negotiations. *Journal of Experimental Social Psychology*, *48*(1), 226–231.
Schweinsberg, M., Viganola, D., Prasad, V., Dreber, A., Johannesson, M., Pfeiffer, T., Tierney,
W.T., Eitan, O. … & Uhlmann, E.L. (2019). *The pipeline project 2: Opening pre-publication independent replication to the world.* Project in progress.
Silberzahn, R., & Uhlmann, E.L. (2015). Many hands make tight work: Crowdsourcing research
can balance discussions, validate findings and better inform policy. *Nature, 526,* 189-191.
Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... & Carlsson, R. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337-356.
Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science,*
9(1), 76-80.
Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-Curve: A key to the file drawer.
Journal of Experimental Psychology: General, 143, 534-547.
Simonsohn, U., Simmons, J., & Nelson, L. (2016). *Specification curve: Descriptive and*
inferential statistics for all plausible specifications. Unpublished manuscript.
Singer, P. (2005). Ethics and intuitions. *Journal of Ethics, 9,* 331-352.
Sowden, W. & Hall, M. (2015). *Exploring the relationship between morality and happiness.* Unpublished raw data.
Stan Development Team. (2018). *RStan: the R interface to Stan.* Retrieved from http://mc-stan.org/ (R package version 2.17.3)
Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency
through a multiverse analysis. *Perspectives on Psychological Science, 11(5)*, 702 –712.
Sternberg, R. J. (2016). "Am I famous yet?" Judging scholarly merit in psychological
science an introduction. *Perspectives on Psychological Science, 11*(6), 877-881.
Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk Workers. *Judgment and Decision Making, 10,* 479-491.
Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication.
Perspectives in Psychological Science, 9, 59–71.
Surowiecki, J. (2004). *The wisdom of crowds*. New York, NY: Random House.
Tierney, W., Ebersole, C., Hardy, J., Chapman, H., Gantman, A., Vanaman, M., DeMarree, K.,

Wylie, J., Storbeck J., & Uhlmann, E. L. (2019a). *A creative destruction approach to replication.* Registered Report proposal under review.

Tierney, W.T., Viganola, Ebersole, C., Hardy, J., D., Gordon, M., Dreber, A., Johannesson, M., Pfeiffer, T., Molden, D., Grossman, I., Bauman, C., DeMarree, K., Devos, T., Huynh, Q., Bozo, J., Diermeier, D., Heinze, J., & Uhlmann, E. L. (2019b). *Replication ring for priming effects on judgments and behaviors.* Registered Report proposal in preparation.

Tiokhin, L., Hackman, J., Munira, S., Jesmin, K., Hruschka, D. (2019). Generalizability is not optional: insights from a cross-cultural study of social discounting. *Royal Society Open Science*, *6*(2), 181386.

Turiel, E. (1983). *The development of social knowledge: Morality and convention.* Cambridge, UK: Cambridge University Press.

Turiel, E. (2002). *The culture of morality.* Cambridge, UK: Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Uhlmann, E. L., & Cunningham, W.A. (2000). *Awareness of automatic prejudice.* Unpublished

raw data.

Uhlmann, E.L., Ebersole, C., Chartier, C., Errington, T., Kidwell, M., Lai, C.K., McCarthy, R.,

Riegelman, A., Silberzahn, R., & Nosek, B.A. (in press). Scientific Utopia III: Crowdsourcing Science. *Perspectives on Psychological Science.*

Uhlmann, E. L., Poehlman, T. A., & Nosek, B. A. (2012). Automatic associations: Personal

attitudes or cultural knowledge? In J. Hanson (Ed.), *Ideology, psychology, and law* (pp. 228–260). Oxford, UK: Oxford University Press.

Uhlmann, E.L., Pizarro, D., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10,* 72-81.

Uhlmann, E.L., Poehlman, T.A., & Bargh, J.A. (2009). American moral exceptionalism. In J.T.

Jost, A.C. Kay, & H. Thorisdottir (Eds.) *Social and Psychological Bases of Ideology and System Justification.* (pp. 27-52). New York, NY: Oxford University Press.

Uhlmann, E.L., & Sanchez-Burks, J. (2014). The implicit legacy of American Protestantism.

Journal of Cross-Cultural Psychology, 45, 991-1005.

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

van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E., Derks, K., ... Wagenmakers, E.-J. (2019). How to interpret the output of a Bayesian ANOVA in JASP. *In preparation*.

van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *Psychological Bulletin* from 1990–2013. *Journal of Open Psychology Data*, *5*(1), 4. Retrieved from http://doi.org/10.5334/jopd.33

Vazire, S. (2017). Our obsession with eminence warps research. *Nature, 547,* 7.

Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a

replication attempt. *Journal of Experimental Psychology: General, 143,* 1457-1475.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1-48.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. doi: https://doi.org/10.3758/s13423-017-1323-7

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. doi:

https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H.L.J. & Kievit, R.A. (2012).

An agenda for purely confirmatory research. *Perspectives on Psychological Science,*

7(6), 632-638.

Waterman, A.S. (1993). Two conceptions of happiness: contrasts of personal expressiveness (eudaimonia) and hedonic enjoyment. *Journal of Personality and Social Psychology, 64,* 678–91.

Waterman, A.S.., Schwartz, S.J., Zamboanga, B.L., Ravert, R.D. Williams, M.K., Agocha, V.B., Kim, S.Y., & Donnellan, B. (2010). The questionnaire for eudaimonic well-being: Psychometric properties, demographic comparisons, and evidence of validity. *The Journal of Positive Psychology, 5*(1), 41-61.

Watson, D., & Clark, LA. (1988). Development and validation of brief measures of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling in social psychological

experimentation. *Personality and Social Psychology Bulletin, 25,* 1115-1125.

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science, 10*(3), 390-399.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7,* 1832.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological*

Review, 107, 101–126.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007a). The increasing dominance of teams in the

production of knowledge. *Science, 316,* 1036–1038.

Wuchty, S., Jones, B., & Uzzi, B. (2007b). Why do team authored papers get cited more? *Science, 317,* 1496-1497.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, e120.

Table 1. Directional and nondirectional formulations of the five hypotheses.

Hypothesis 1
Directional: People explicitly self-report an awareness of harboring negative automatic associations with members of negatively stereotyped social groups.
Nondirectional: When directly asked, do people explicitly self-report an awareness of harboring negative automatic associations with members of negatively stereotyped social groups?
Hypothesis 2
Directional: Negotiators who make extreme first offers are trusted less, relative to negotiators who make moderate first offers.
Nondirectional: Are negotiators who make extreme first offers trusted more, less, or the same relative to negotiators who make moderate first offers?
Hypothesis 3
Directional: A person continuing to work despite having no material/financial need to work has beneficial effects on moral judgments of that individual.
Nondirectional: What are the effects of continuing to work despite having no material/financial need to work on moral judgments of that individual — beneficial, detrimental, or no effect?
Hypothesis 4
Directional: Part of why people are opposed to the use of performance enhancing drugs in sports is because they are “against the rules”. But, whether the performance enhancer is against the rules established by a proximal authority (e.g., the league) contributes more to this judgment than whether it is against the law.
Nondirectional: Part of why people are opposed to the use of performance enhancing drugs in sports is because they are "against the rules". But which contributes more to this judgment — whether the performance enhancer is against the law, or whether it is against the rules established by a more proximal authority (e.g., the league)?
Hypothesis 5
Directional: The tendency to make deontological (as opposed to utilitarian) judgments is positively related to personal happiness.
Nondirectional: Is a utilitarian vs. deontological moral orientation related to personal happiness?

Table 2. Effect sizes and Q, I^2, and τ^2 statistics from meta-analyses of Main Studies and Replication Studies.

Main Studies

Hypothesis	Description	*k*	Effect Size [95% *CI*]	*Q*	I^2 [95% *CI*]	τ^2 [95% *CI*]
1	Awareness of automatic prejudice	13	d = 0.07 [-0.22, 0.37]	$Q(12)$ = 897.51***	99.08% [98.20, 99.67]	0.28 [0.14, 0.81]
2	Extreme offers reduce trust	13	d = 1.04 [0.61, 1.47]	$Q(12)$ = 568.36***	98.25% [96.58, 99.36]	0.61 [0.31, 1.70]
3	Moral praise for needless work	13	d = 0.33 [0.17, 0.50]	$Q(12)$ = 152.45***	93.55% [87.39, 97.68]	0.09 [0.04, 0.26]
4	Proximal authorities drive legitimacy of performance enhancers	12	d = 0.07 [-0.05, 0.20]	$Q(11)$ = 89.72***	87.94% [75.82, 95.85]	0.04 [0.02, 0.13]
5	Deontological judgments predict happiness	13	r = 0.06 [0.01, 0.11]	$Q(12)$ = 52.91***	75.65% [52.68, 90.62]	0.01 [0.00, 0.02]

Replication Studies

Hypothesis	Description	*k*	Effect Size [95% *CI*]	*Q*	I^2 [95% *CI*]	τ^2 [95% *CI*]
1	Awareness of automatic prejudice	13	d = -0.07 [-0.33, 0.19]	$Q(12)$ = 773.19***	98.88% [97.81, 99.60]	0.23 [0.12, 0.64]
2	Extreme offers reduce trust	13	d = 0.61 [0.32, 0.88]	$Q(12)$ = 372.40***	97.09% [94.34, 98.98]	0.26 [0.13, 0.73]
3	Moral praise for needless work	13	d = 0.24 [0.11, 0.38]	$Q(12)$ = 129.49***	91.26% [82.81, 96.85]	0.05 [0.03, 0.16]
4	Proximal authorities drive legitimacy of performance enhancers	12	d = 0.03 [-0.06, 0.12]	$Q(11)$ = 47.45***	78.06% [55.84, 92.65]	0.02 [0.01, 0.07]
5	Deontological judgments predict happiness	13	r = 0.03 [-0.04, 0.09]	Q(12) = 90.93***	86.39% [73.53, 94.97]	0.01 [0.00, 0.03]

Note. ***p < .001.

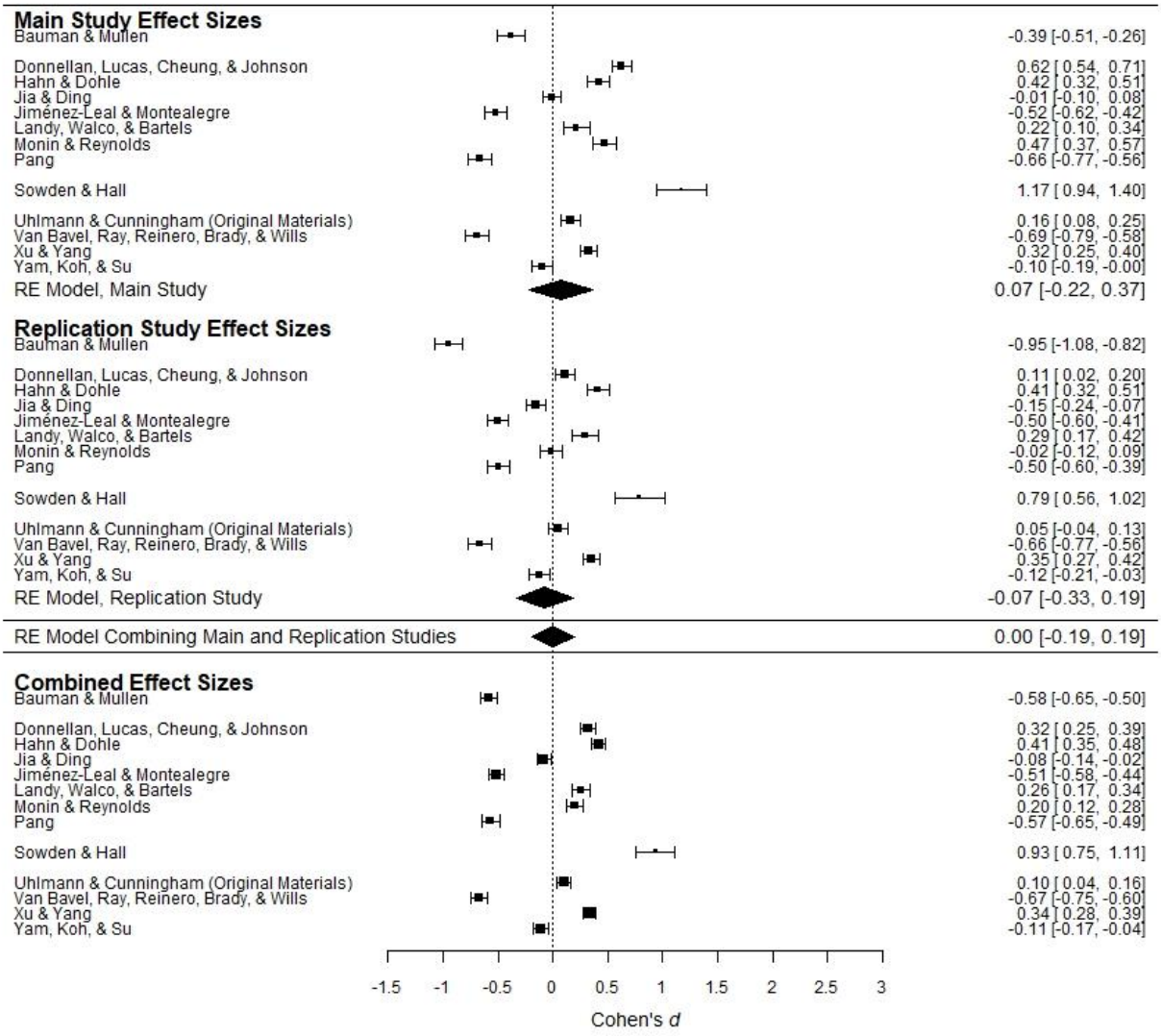

Figure 1a. Forest plot of observed effect sizes (independent-groups Cohen's *d*s) for Hypothesis 1. The research question was "When directly asked, do people explicitly self-report an awareness of harboring negative automatic associations with members of negatively stereotyped social groups?"

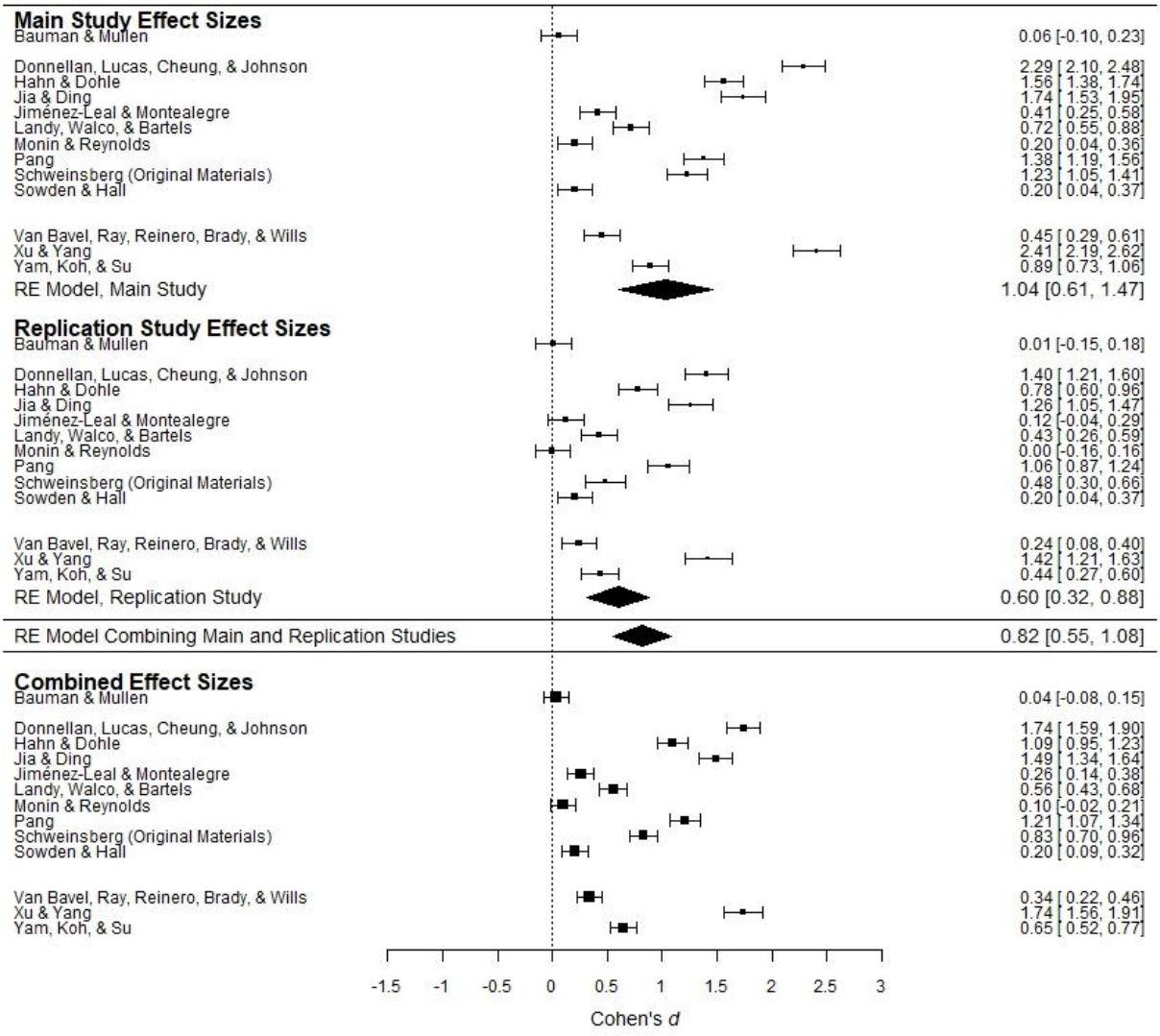

Figure 1b. Forest plot of observed effect sizes (independent-groups Cohen's *d*s) for Hypothesis 2. The research question was "Are negotiators who make extreme first offers trusted more, less, or the same relative to negotiators who make moderate first offers?"

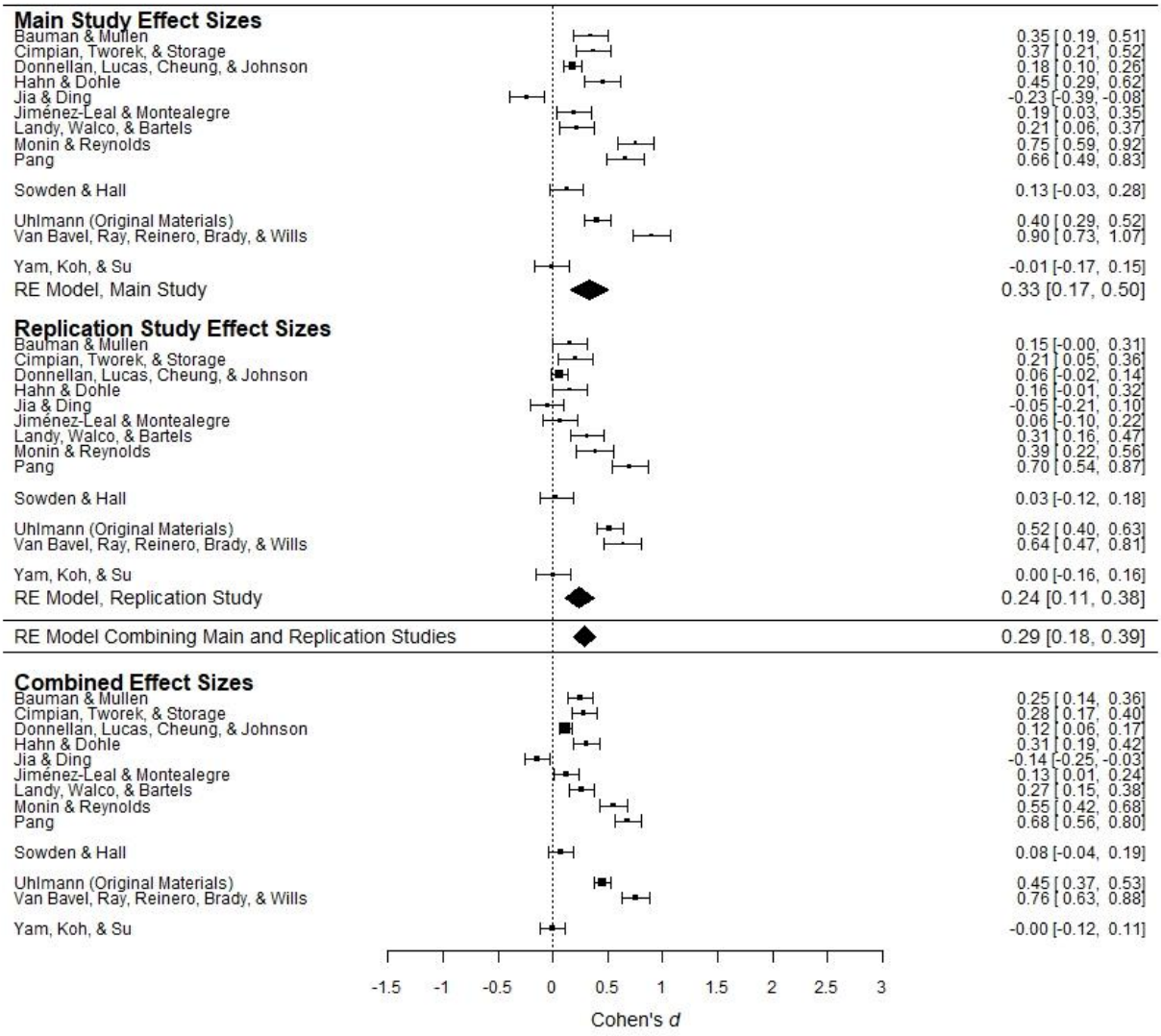

Figure 1c. Forest plot of observed effect sizes (independent-groups Cohen's *d*s) for Hypothesis 3. The research question was "What are the effects of continuing to work despite having no material/financial need to work on moral judgments of that individual - beneficial, detrimental, or no effect?"

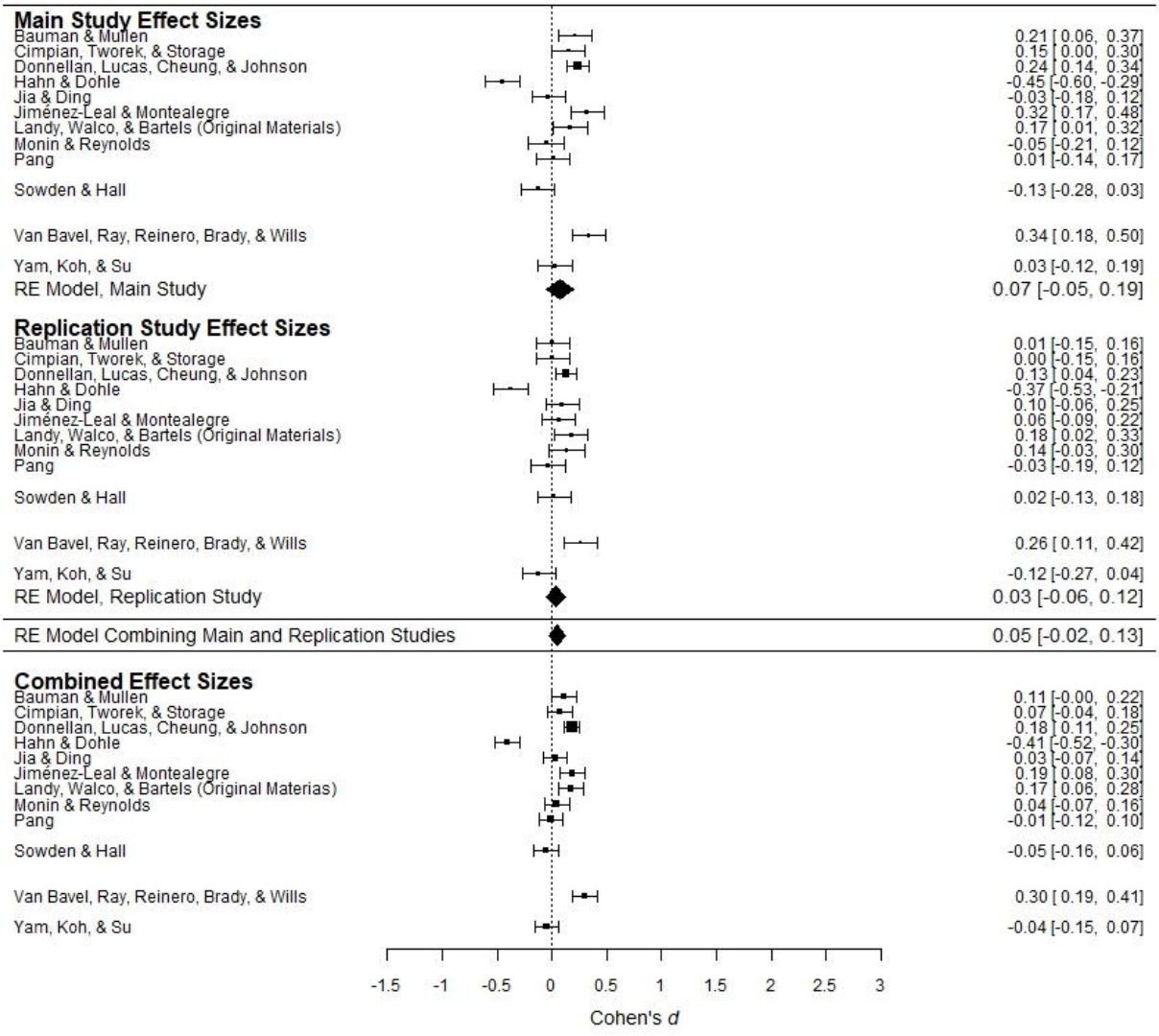

Figure 1d. Forest plot of observed effect sizes (independent-groups Cohen's *d*s) for Hypothesis 4. The research question was "Part of why people are opposed to the use of performance enhancing drugs in sports is because they are 'against the rules'. But which contributes more to this judgment - whether the performance enhancer is against the law, or whether it is against the rules established by a more proximal authority (e.g., the league)?"

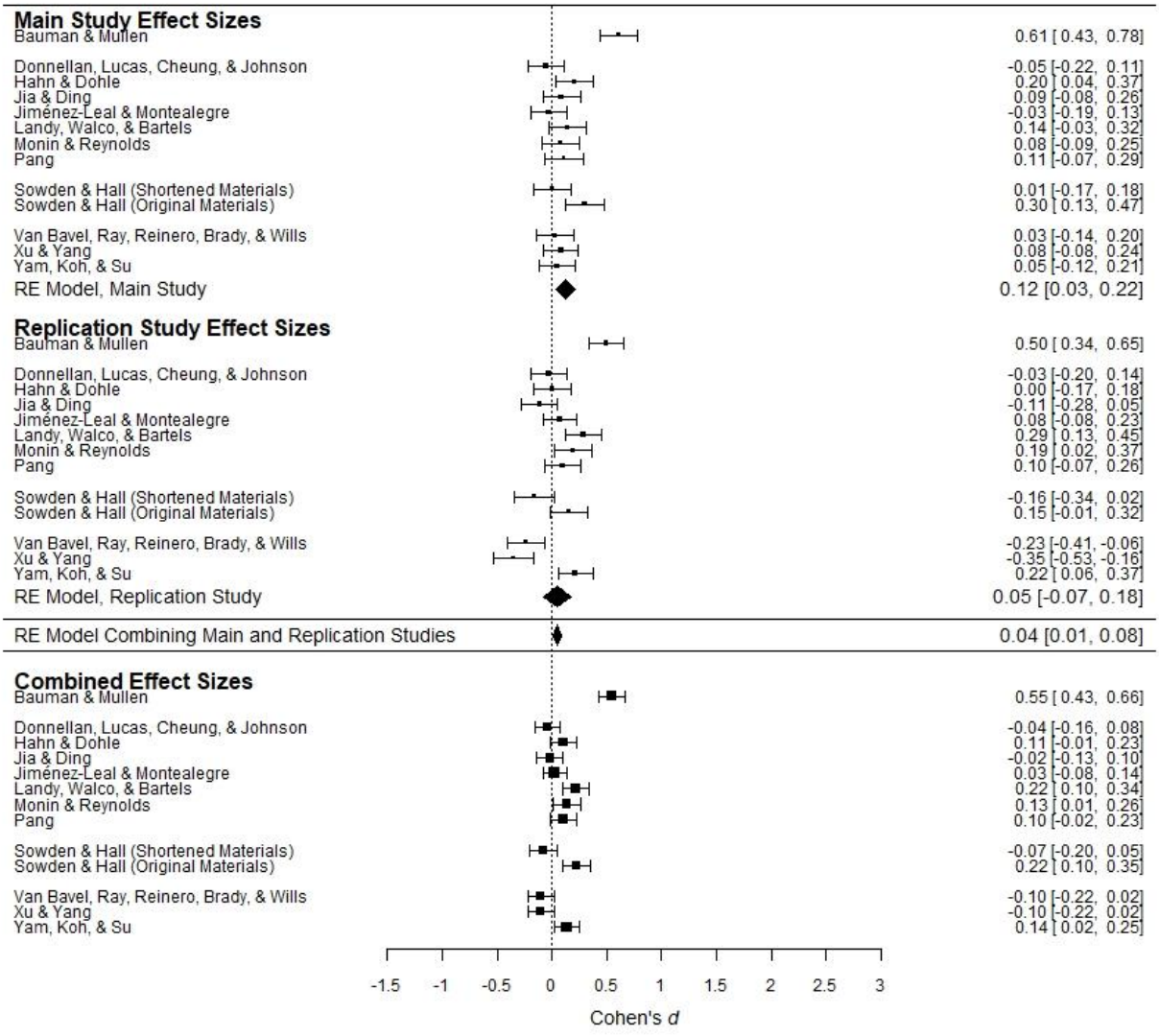

Figure 1e. Forest plot of observed effect sizes (converted to Cohen's *d*s, for comparison to other hypotheses) for Hypothesis 5. The research question was "Is a utilitarian vs. deontological moral orientation related to personal happiness?"

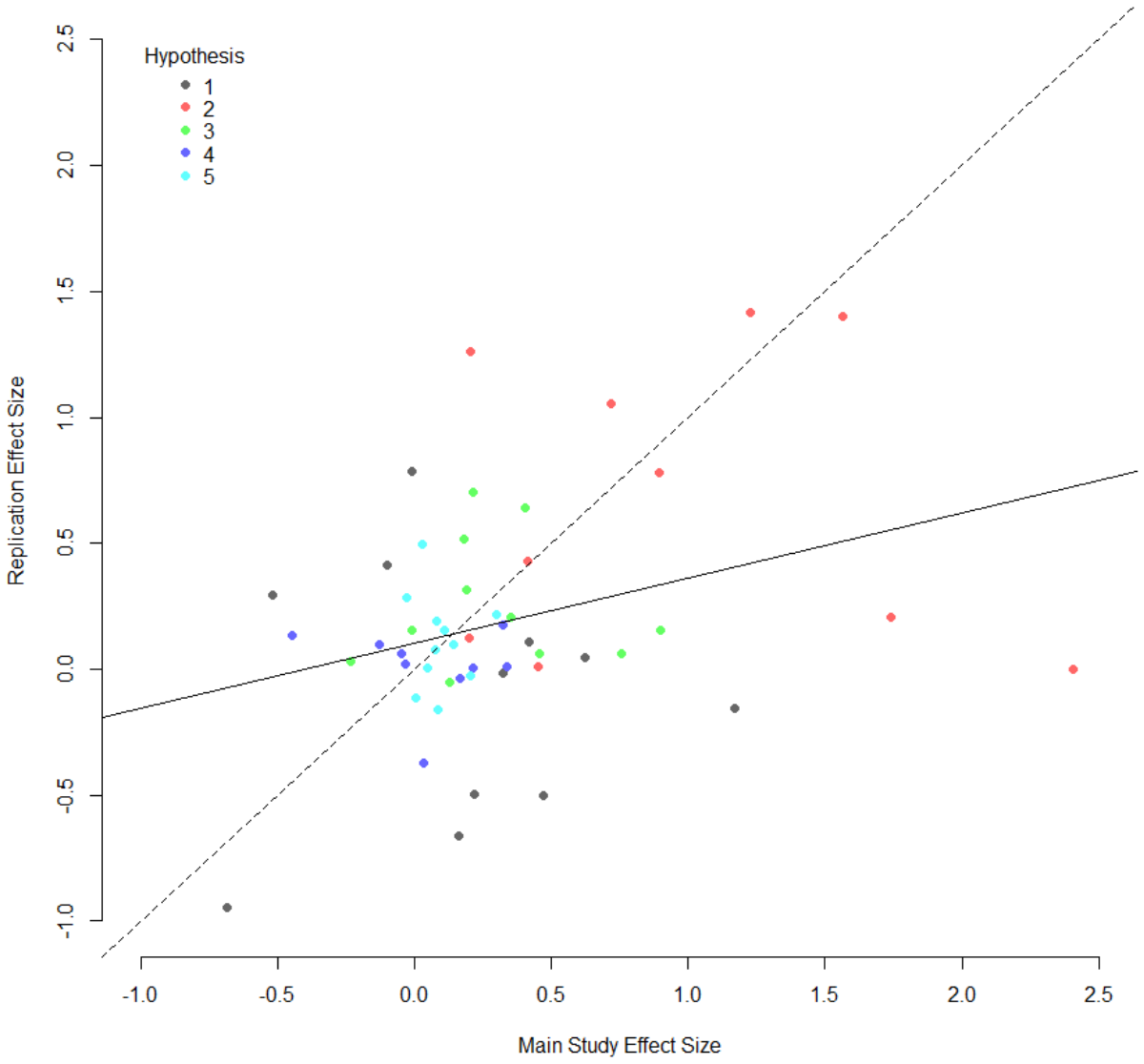

Figure 2. Scatter plot comparing Main Study and Replication effect sizes (Cohen's *d*s). Each point in the scatter plot consists of one of 64 study designs. The continuous segment represents the fitted line; the dashed segment represents the 45-degree line. H1: Awareness of automatic prejudice, H2: Extreme offers reduce trust, H3: Moral praise for needless work, H4: Proximal authorities drive legitimacy of performance enhancers, H5: Deontological judgments predict happiness.

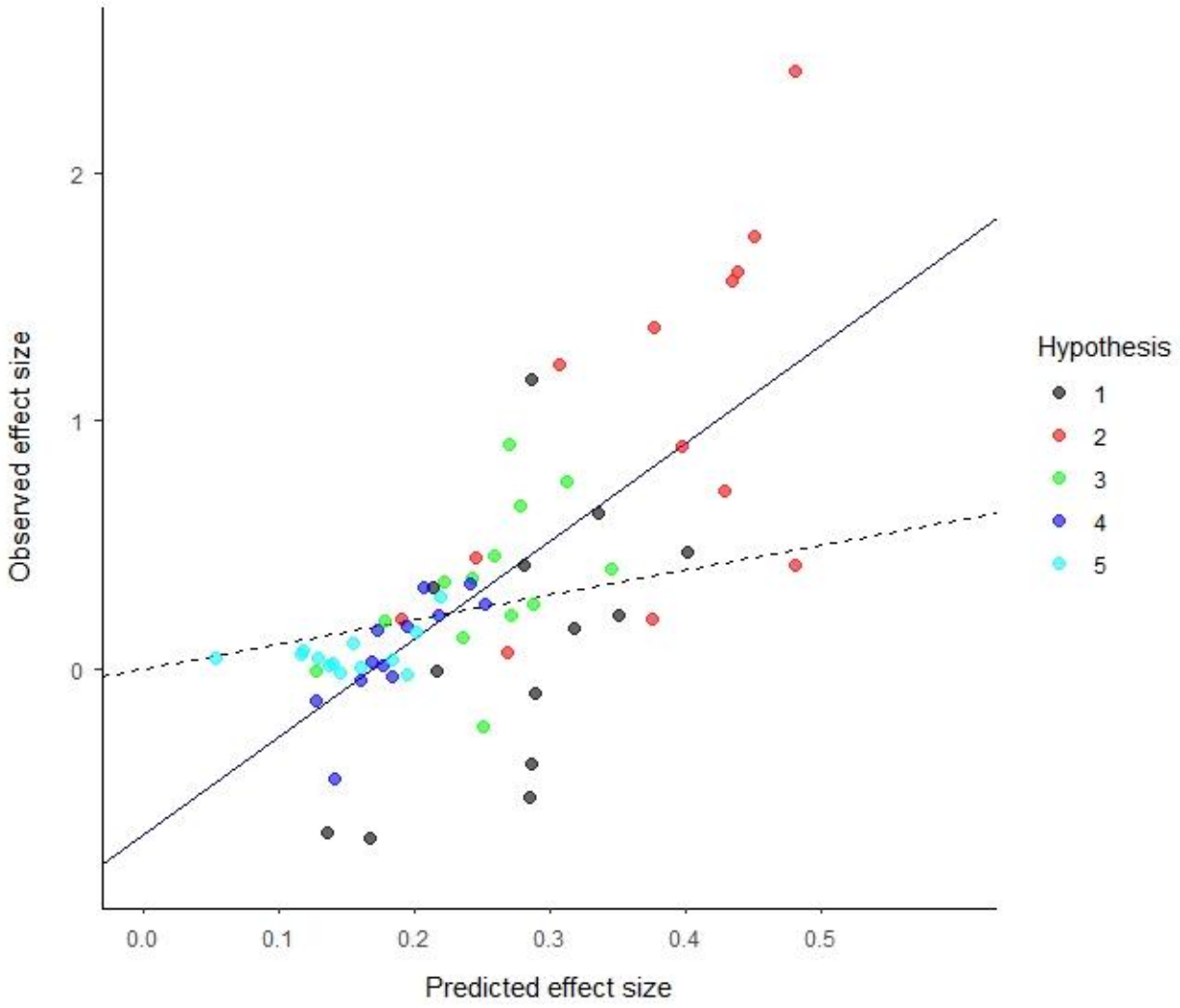

Figure 3a. Correlation between average predicted effect size and observed effect size for each study design. The continuous segment represents the fitted line; the dashed segment represents $y = x$. H1: Awareness of automatic prejudice, H2: Extreme offers reduce trust, H3: Moral praise for needless work, H4: Proximal authorities drive legitimacy of performance enhancers, H5: Deontological judgments predict happiness.

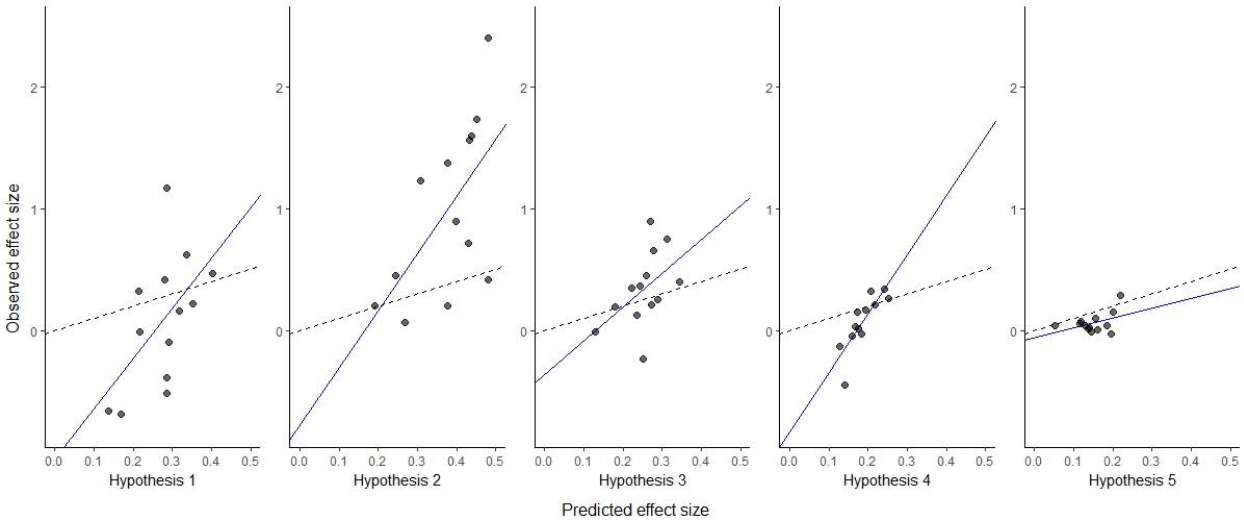

Figure 3b. Correlation between average predicted effect size and observed effect size for each version of the study materials, separately for each of the five hypotheses. Continuous segments represent fitted lines; dashed segments represent $y = x$. H1: Awareness of automatic prejudice, H2: Extreme offers reduce trust, H3: Moral praise for needless work, H4: Proximal authorities drive legitimacy of performance enhancers, H5: Deontological judgments predict happiness.

Appendix: The Crowdsourcing Hypothesis Tests Collaboration

Matúš Adamkovič[1], Ravin Alaei[2], Casper J. Albers[3], Aurélien Allard[4], Ian A. Anderson[5], Michael R. Andreychik[6], Peter Babinčák[7], Bradley J. Baker[8], Gabriel Baník[7], Ernest Baskin[9], Jozef Bavolar[10], Ruud M. W. J. Berkers[11], Michał Białek[12], Joel Blanke[13], Johannes Breuer[14], Ambra Brizi[15], Stephanie E. V. Brown[16], Florian Brühlmann[17], Hendrik Bruns[18], Leigh Caldwell[19], Jean-François Campourcy[20], Eugene Y. Chan[21], Yen-Ping Chang[22], Benjamin Y. Cheung[23], Alycia Chin[24]*, Kit W. Cho[25], Simon Columbus[26], Paul Conway[27], Conrad A. Corretti[28], Adam W. Craig[29], Paul G. Curran[30], Alexander F. Danvers[31], Ian G. J. Dawson[32], Martin V. Day[33], Erik Dietl[34], Johannes T. Doerflinger[35], Alice Dominici[36], Vilius Dranseika[37,38], Peter A. Edelsbrunner[39], John E. Edlund[40], Matthew Fisher[41], Anna Fung[42], Oliver Genschow[43], Timo Gnambs[44,45], Matthew H. Goldberg[46], Lorenz Graf-Vlachy[47], Andrew C. Hafenbrack[42], Sebastian Hafenbrädl[48], Andree Hartanto[49], Patrick R. Heck[50], Joseph P. Heffner[51], Joseph Hilgard[52], Felix Holzmeister[53], Oleksandr V. Horchak[54], Tina S.-T. Huang[55], Joachim Hüffmeier[56], Sean Hughes[57], Ian Hussey[57], Roland Imhoff[58], Bastian Jaeger[59], Konrad Jamro[60], Samuel G. B. Johnson[61], Andrew Jones[62], Lucas Keller[35], Olga Kombeiz[34], Lacy E. Krueger[63], Anthony Lantian[64], Justin P. Laplante[65], Ljiljana B. Lazarevic[66], Jonathan Leclerc[67], Nicole Legate[68], James M. Leonhardt[69], Desmond W. Leung[70,71], Carmel A. Levitan[72], Hause Lin[2], Qinglan Liu[73], Marco Tullio Liuzza[74], Kenneth D. Locke[75], Albert L. Ly[76], Melanie MacEacheron[77], Christopher R. Madan[78], Harry Manley[79], Silvia Mari[80], Marcel Martončik[7], Scott L. McLean[81], Jonathon McPhetres[82,83], Brett G. Mercier[84], Corinna Michels[43], Michael C. Mullarkey[85], Erica D. Musser[86], Ladislas Nalborczyk[87,57], Gustav Nilsonne[88,89], Nicholas G. Otis[90], Sarah M. G. Otner[91], Philipp E. Otto[92], Oscar Oviedo-Trespalacios[93,94], Mariola Paruzel-Czachura[95], Francesco Pellegrini[96], Vitor M. D. Pereira[97], Hannah Perfecto[98], Gerit Pfuhl[99], Mark H. Phillips[100], Ori Plonsky[101], Maura Pozzi[102], Danka B. Purić[66], Brett Raymond-Barker[103], David E. Redman[104], Caleb J. Reynolds[27], Ivan Ropovik[7], Lukas Röseler[105,106], Janna K. Ruessmann[43], William H. Ryan[90], Nika Sablaturova[107], Kurt J. Schuepfer[108], Astrid Schütz[106], Miroslav Sirota[109], Matthias Stefan[53], Eric L. Stocks[110], Garrett L. Strosser[111], Jordan W. Suchow[112], Anna Szabelska[113], Kian Siong Tey[5], Leonid Tiokhin[114], Jais Troian[115], Till Utesch[116], Alejandro Vásquez-Echeverría[117], Leigh Ann Vaughn[118], Mark Verschoor[3], Bettina von Helversen[119], Pascal Wallisch[120], Sophia C. Weissgerber[121], Aaron L. Wichman[122], Jan K. Woike[123,124], Iris Žeželj[66], Janis H. Zickfeld[125,126], Yeonsin Ahn[5], Philippe F. Blaettchen[5], Xi Kang[5], Yoo Jin Lee[5], Philip M. Parker[5], Paul A. Parker[5], Jamie S. Song[5], May-Anne Very[5], Lynn Wong[5]

[1]University of Presov, [2]University of Toronto, [3]University of Groningen, [4]University of Paris VIII, [5]INSEAD, [6]Fairfield University, [7]University of Prešov, [8]University of Massachusetts, [9]Saint Joseph's University, [10]Pavol Josef Šafárik University in Košice, [11]Max Planck Institute for Human Cognitive & Brain Sciences, [12]Kozminski University, [13]Stockholm School of Economics, [14]GESIS – Leibniz Institute for the Social Sciences, [15]Sapienza University of Rome, [16]Texas A&M University, [17]University of Basel, [18]University of Hamburg, [19]Irrational Agency, [20]Université Clermont Auvergne, [21]Monash University, [22] Institute of Sociology, Academia Sinica, Taiwan, [23]University of British Columbia, [24]Public Company Accounting Oversight Board, [25]University of Houston–Downtown, [26]Vrije Universiteit Amsterdam, [27]Florida State University, [28]The University of Texas at Dallas, [29]University of Kentucky, [30]Grand Valley State University, [31]University of Arizona, [32]University of Southampton, [33]Memorial University of Newfoundland, [34]Loughborough University, [35]University of Konstanz, [36]European University Institute, [37]Kaunas University of Technology, [38]Vilnius University, [39]ETH Zurich, [40]Rochester Institute of Technology, [41]Southern Methodist University, [42]University of Washington, [43]University of Cologne, [44]Leibniz Institute for Educational Trajectories, [45]Johannes Kepler University Linz, [46]Yale University, [47]University of Passau, [48]IESE Business School, [49]Singapore Management University, [50]Geisinger Health System, [51]Brown University, [52]Illinois State University, [53]University of Innsbruck, [54]Instituto Universitário de Lisboa (ISCTE-IUL), CIS-IUL, [55]University College London, [56]TU Dortmund University, [57]Department of Experimental Clinical and Health Psychology, Ghent University, [58]Johannes Gutenberg University Mainz, [59]Tilburg University, [60]University of Massachusetts Dartmouth, [61]University of Bath, [62]University of Liverpool, [63]Texas A&M University-Commerce, [64]Université Paris Nanterre, [65]Clark University, [66]University of Belgrade, [67]John Molson School of Business,

Concordia University, [68]Illinois Institute of Technology, [69]University of Nevada, Reno, [70]Baruch College, City University of New York, [71]The Graduate Center, City University of New York, [72]Occidental College, [73]Hubei University, [74]Magna Græcia University of Catanzaro, [75]University of Idaho, [76]Loma Linda University, [77]University of Western Ontario, [78]University of Nottingham, [79]Chulalongkorn University, [80]University of Milano - Bicocca, [81]Walden University, [82]Massachusetts Institute of Technology, [83]University of Regina, [84]University of California, Irvine, [85]University of Texas-Austin, [86]Florida International University, [87]Univ. Grenoble Alpes, CNRS, LPNC, 38000, Grenoble, France, [88]Karolinska Institutet, [89]Stockholm University, [90]University of California, Berkeley, [91]Imperial College Business School, [92]European University Viadrina, [93]Queensland University of Technology (QUT), [94]Universidad del Norte, [95]University of Silesia, [96]Università degli Studi di Padova, [97]LanCog, CFUL, Faculdade de Letras, Universidade de Lisboa, Alameda da Universidade, [98]Washington University in St. Louis, [99]UiT The Arctic University of Norway, [100]Abilene Christian University, [101]Duke University, [102]Università Cattolica del Sacro Cuore, [103]University of Roehampton, [104]Pacific Lutheran University, [105]Harz University of Applied Sciences, [106]University of Bamberg, [107]Masaryk University, [108]Miami University, [109]University of Essex, [110]University of Texas at Tyler, [111]Southern Utah University, [112]Stevens Institute of Technology, [113]Queen's University Belfast, [114]Eindhoven University of Technology, [115]Istanbul Bilgi University, [116]University of Münster, [117]University of the Republic, [118]Ithaca College, [119]University of Bremen, [120]New York University, [121]University of Kassel, [122]Western Kentucky University, [123]Max Planck Institute for Human Development, [124]DIW, Berlin, Germany, [125]University of Oslo, [126]University of Mannheim, [127]McGill University

The first through 135th members of the Crowdsourcing Hypothesis Tests Collaboration lent their expertise as evaluators of study quality and forecasters. The 136th through 144th members of the Crowdsourcing Hypothesis Tests Collaboration lent their expertise as evaluators of study quality and forecasters in a pilot version of the Forecasting Study. In addition, all members of the Crowdsourcing Hypothesis Tests Collaboration, along with the authors whose full names are listed on the first page, revised and approved the final version of this article.

*The Public Company Accounting Oversight Board, as a matter of policy, disclaims responsibility for any private publication or statement by any of its economic research fellows, consultants, or employees.