Reexamining the effect of gustatory disgust on moral judgment: A multi-lab direct replication of Eskine, Kacinik, and Prinz (2011)

Eric Ghelfi[1], Cody D. Christopherson[2], Heather L. Urry[3], Richie L. Lenne[4], Nicole Legate[5], Mary Ann Fischer[6], Fieke M. A. Wagemans[7], Brady Wiggins[8], Tamara Barrett[9], Michelle Bornstein[10], Bianca de Haan[11], Joshua Guberman[12], Nada Issa[13], Joan Kim[14], Elim Na[15], Justin O'Brien[16], Aidan Paulk[17], Tayler Peck[18], Marissa Sashihara[19], Karen Sheelar[20], Justin Song[21], Hannah Steinberg[22], & Dasan Sullivan[23]

[1] Brigham Young University

[2] Southern Oregon University

[3] Tufts University

[4] University of Minnesota

[5] Illinois Institute of Technology

[6] Indiana University Northwest

[7] Tilburg University and University of Duisburg-Essen

[8] Brigham Young University-Idaho

[9] Utah State University

[10] Tufts University

[11] Brunel University

[12] University of Michigan

[13] Indiana University Northwest

[14] Massachusetts General Hospital

[15] Boston University School of Medicine

[16] Brunel University

[17] Oregon College of Oriental Medicine

[18] Southern Oregon University

[19] Tufts University

[20] Southern Oregon University

[21] Tufts University

[22] Pacific Graduate School of Psychology/Stanford University Doctor of Psychology Consortium

[23] Southern Oregon University

Author Note

Correspondence concerning this article should be addressed to Eric Ghelfi, Department of Psychology, Brigham Young University, 1001 SWKT, Provo, UT 84602. E-mail: ericghelfi1@gmail.com

Abstract

Eskine, Kacinik, and Prinz's (2011) influential experiment demonstrated that gustatory disgust triggers a heightened sense of moral wrongness. We report a large-scale multi-site direct replication of this study conducted by participants in the Collaborative Replications and Education Project. Participants in each sample were randomly assigned to one of three beverage conditions: bitter/disgusting, control, or sweet. Then, participants made a series of judgments indicating the moral wrongness of the behavior depicted in each of six vignettes. In the original study ($N = 57$), drinking the bitter beverage led to higher ratings of moral wrongness than drinking the control and sweet beverages; a beverage contrast was significant among conservative ($N = 19$) but not liberal ($N = 25$) participants. In this report, random effects meta-analyses across all participants ($N = 1{,}137$ in $k = 11$ studies), conservative participants ($N = 142$, $k = 5$), and liberal participants ($N = 635$, $k = 9$) revealed standardized effect sizes that were smaller than reported in the original study. Some were in the opposite of the predicted direction, all had 95% confidence intervals containing zero, and most were smaller than the effect size the original authors could meaningfully detect. In linear mixed-effects regressions, drinking the bitter beverage led to higher ratings of moral wrongness than drinking the control beverage but not the sweet beverage. Bayes Factor tests reveal greater relative support for the null hypothesis. The overall pattern provides little to no support for the theory that physical disgust via taste perception harshens judgments of moral wrongness.

Keywords: disgust, moral judgment

Reexamining the effect of gustatory disgust on moral judgment: A multi-lab direct replication of Eskine, Kacinik, and Prinz (2011)

This paper presents results from a multi-lab replication of an experiment suggesting that gustatory disgust (i.e., taste perception) can make moral judgments harsher (Eskine, Kacinik, & Prinz, 2011). Previous studies had revealed a similar link between sensory disgust induced by other means, such as olfactory stimulation (Schnall et al., 2008b), and judgments of moral wrongness. Moreover, since Eskine and colleagues' (2011) study, empirical and conceptual work has grown around the idea that inducing perceptions of gustatory disgust is an especially effective way of increasing the severity with which people make moral judgments (Hellmann, Thoben, & Echterhoff, 2013; Schnall, Haidt, Clore, & Jordan, 2015). However, a recent meta-analysis failed to replicate the association between disgust and moral judgment (Landy & Goodwin, 2015a). To our knowledge, there are no attempts to replicate the relation between physical disgust via taste perception and moral wrongness. The purpose of this project is therefore to precisely estimate the effect of gustatory disgust on moral judgment by replicating the methods used by Eskine et al. (2011).

In the 1990's, the body of research that would eventually coalesce under the banner of embodied cognition began taking shape (Wilson, 2002). This research proposes that real-world thinking and problem-solving are deeply situated within sensoriperceptual processes and are not merely cognitive processes (Anderson, 2003; Ionescu & Vasc, 2014). For example, it has been theorized that while disgust likely evolved to steer organisms away from pathogens, it also evolved to guide humans in their decision-making in the domains of mate selection and morality (Tybur, Lieberman, Kurzban, & DeScioli, 2013). Other researchers agree that the cognitive computation systems involved in pathogenic ("core") disgust likely overlap with the computational systems involved in perceptions of moral disgust (Curtis & Biran, 2001; Inbar & Pizarro, 2014). Indeed, the insula plays a role in perceptions of both pathogenic and moral disgust (Vicario, Rafal, Martino, & Avenanti,

2017), and pathogen and moral disgust both cause activation in the levator labii muscle region, defined as the raising of the upper lip and wrinkling of the nose (Cannon, Schnall, & White, 2011; Chapman, 2018). Moreover, some argue that the two constructs are so closely related that inducing incidental disgust (e.g., the smell of garbage or the taste of a bitter drink) can amplify moral disgust and so harshen moral judgment (Inbar & Pizarro, 2014).

The idea that similar parts of the brain are activated by both physical and moral disgust has prompted other researchers to examine how inducing a sensoriperceptual experience of disgust can affect moral judgment more specifically (Cameron, Payne, & Doris, 2013; Case, Oaten, & Stevenson, 2012; Gill & Nichols, 2008; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Landy & Goodwin, 2015a). Several studies suggest that olfactory (e.g., Inbar, Pizarro, & Bloom, 2012) and gustatory (e.g., Chapman, 2018; Hellmann et al., 2013) input, for example, can affect moral judgment. In particular, these studies find that perceptions of sensory disgust lead to harsher moral judgment across several means of inducing disgust (Cameron et al., 2013; Eskine et al., 2011; Schnall et al., 2008b; Wheatley & Haidt, 2005). However, a recent meta-analytic review estimated a negligible effect size of disgust on moral judgment ($d = .11$ for the studies included in the meta-analysis and $d =$ -.01 accounting for publication bias; Landy & Goodwin, 2015a). These effect sizes were calculated across multiple means of inducing disgust. There is no available meta-analysis of studies using gustatory disgust inductions and no discoverable replications of the effect of gustatory disgust on moral judgment, such as the one reported by Eskine and colleagues (2011). As we explain in the next section, the effect sizes Eskine and colleagues found were much larger than effect sizes found in other studies investigating similar effects. This difference—combined with the small sample size of the study—made it an important candidate to replicate.

Original Study

Eskine and colleagues (2011) found that participants who drank a bitter beverage before reading six moral vignettes judged characters' actions more harshly than participants who drank a sweet beverage or water. Eskine and colleagues selected a bitter drink to induce disgust because they hypothesized that a strong bitter taste would be reliably experienced as disgusting. A manipulation check confirmed that their sample experienced the bitter drink (Swedish Bitters) as disgusting. In addition to the main effect of gustatory disgust on moral judgment, politically conservative participants showed more sensitivity to the effect, suggesting that political conservatives were more likely to be influenced by incidental sensoriperceptual cues (i.e., gustatory disgust) when making moral judgments.

The original research ($N = 57$, each randomly assigned to one of three conditions; $N = 54$ after exclusions) yielded large, statistically significant effects ($d = 1.09$ and $p = .001$ for the disgust group vs. the neutral control, and $d = 1.22$ and $p = .003$ for the disgust group vs. the sweet control). Although the authors did not report the F test for the interaction between beverage conditions and political orientation, they conducted two contrast analyses within politically conservative and liberal subgroups. In the conservative subgroup, the contrast between the disgust condition and the other conditions was large and statistically significant, $t(16) = 4.473$, $p < .001$, $d = 2.21$. The same contrast in the liberal subgroup was medium-large but not statistically significant, $t(22) = 1.703$, $p = .103$, $d = 0.74$.[1] It is unclear whether the difference between these two contrasts is significant. These findings have informed subsequent theoretical and empirical work in the psychology of morality and

[1] The original authors reported only the t value and degrees of freedom. We used that information in R to calculate the p value using the userfriendlyscience::convert.t.to.p function and the effect size using the BSDA::tes function. Based on the reported df, there were $N = 24$ liberal participants in this analysis; the authors did not report how many of them were in the disgust and nondisgust groups. We, thus, assumed that 1/3 of them ($n = 8$) were randomized to the disgust group, and 2/3 ($n = 16$) to the nondigust group (control or sweet) when calculating the effect size.

politics (e.g., Chapman, 2018; Chapman & Anderson, 2012; Haidt, 2012; Vicario et al., 2017), yet no attempts have been made to more precisely estimate the size of the effect. Given the wide reaching impact of this study, it is important to replicate it to gauge the generalizability and reliability of the effect.

Current Study

This study is a multi-lab effort to replicate the methods used by Eskine et al. (2011). We test (1) whether a bitter beverage indeed prompts harsher moral judgments than both a sweet beverage and water, and (2) whether these effects are stronger in politically conservative populations than in politically liberal ones. This study's overarching goal is to provide high-powered estimates of the effect observed by Eskine, Kacinik, and Prinz (2011) using a crowdsourcing approach (see Hagger et al., 2016; Moshontz et al., 2018). That is, instead of evaluating the "replication success" of the original study based solely on the results of a single replication, we meta-analyze many distinct replication attempts using frequentist and Bayesian statistical approaches.

In sum, our goal in this report is to estimate both the main effect of beverage type on moral judgments and to test political orientation as a moderator. We examined two contrasts for the effect of beverage type: bitter versus control and bitter versus sweet. With respect to moderation by political orientation, we focused primarily on the comparison between conservative and liberal participants.

Disclosures

Preregistration. Before data collection, each lab preregistered their materials, protocol, and analysis plans on the Open Science Framework (OSF). In addition, we registered our analysis plan for this manuscript on the OSF prior to conducting the analyses reported herein; see our project at https://osf.io/5ygsp. Our preregistration represents a partial preregistration in that each author knew the results of at least one particular

replication before we preregistered the meta-analysis plan.

Data, materials, and online resources. After labs completed their study, they uploaded raw data, analyses (including syntax), and a clear explanation of results to their own page at the OSF. The data and all other relevant documentation of each of the labs included in this meta-analysis can be found at https://osf.io/4hkjv/forks/. We uploaded the data and analysis code for this manuscript to our project page at https://osf.io/5ygsp.

Reporting. This manuscript is based on analysis of existing data rather than new data collection. We report how we determined the sample size for this manuscript, and all data exclusions, manipulations, and measures.

Ethical approval. Each individual site secured approval from an institutional review board or similar committee prior to data collection, and carried out the study in accordance with the provisions of the World Medical Association Declaration of Helsinki.

Method

The replication studies reported herein were conducted as part of the Collaborative Replications and Education Project (CREP; see our in-detail text box describing CREP in the appendix). All participants provided informed consent prior to participating. Eleven replication studies took place at five universities in the United States and three universities in Europe (United Kingdom, Germany, and the Netherlands). Table A1 in the appendix lists labs by study number, site, and mentor.

Design

The design was a 3×3 between-subjects factorial design. The first factor was the experimental manipulation, a drink intended to produce either a neutral (control), sweet, or bitter taste sensation. The drink conditions were water, fruit punch, and Swedish Bitters (an herbal digestive aid). The second factor was political orientation. The political orientation

conditions were conservative, liberal and other.[2]

Target Sample Size

The minimum target for the overall sample for meta-analysis was 2.5 times the original sample of 57, or 143 participants, which would be sufficient for a "small telescopes" analysis (Simonsohn, 2015). Individual labs were permitted to continue to submit qualified replication samples until December 2017. The CREP recommended each lab collect data from at least 57 participants (matching the original sample), but not all labs hit this target. We included labs who did not hit the target N in our analysis because we were primarily concerned with having adequate power to accurately estimate the overall effect size in the weighted meta-analysis across all samples (and we were not concerned about the power of an individual sample).

Deviations from Original Study

The CREP team contacted the original authors, who provided the original moral dilemma vignettes, manipulation check, and informed consent forms. We could not obtain the original "imageability distractor task" or "cover story" so we created our own based on the description in the paper. Our distractor task is available here: https://osf.io/ju7nq/. Our cover story consisted of the following paragraph:

"In this study you will be asked to read several vignettes and make judgments about the characters in them. Your job will be to judge the actions of the characters. During this task, you will be asked to drink a beverage. The purpose of this study is to determine whether motor movements involved with drinking influence your judgments while reading about others. In order to successfully attain this, please drink each dose in a single swift

[2] The original authors also assessed these three political orientations, but exclusively focused on conservative and liberal categories when computing their ANOVA and their follow-up contrast analyses. Their design was, thus, effectively 3×2 between-subjects factorial.

motion, as if you were drinking a shot."

Two problems arose with using an international sample to replicate: First, the terms "liberal" and "conservative" as used in the United States do not necessarily translate to major political axes within other countries. The European labs adapted this question to capture similar concepts in their respective countries. The two authors responsible for coding political orientation in the pooled sample communicated with faculty supervisors of non-American replications to better understand what participants meant in responding to the political orientation question. For example, in a British sample, "right wing" and "center-right" were coded as conservative, "center-left" and "left wing" were coded as liberal, and responses such as "center" or "none" were coded as other.

Second, the drinks differ slightly from country to country. For example, Minute Maid berry punch, the drink used in the original study, is not widely available outside of the US, and Swedish Bitters can refer to something other than a bitter herb mix. In these cases we suggested researchers substitute a sweet drink with similar sugar content and a bitter but inert drink, respectively. We also gave individual labs discretion to substitute any brand of Swedish Bitters given that the brand was not specified in the original paper. We used the manipulation check to ensure that participants found the sweet drink sweet but not neutral or disgusting and the bitter drink disgusting but not sweet or neutral.

A third deviation from the original study was that three of the studies (7, 9 and 10) presented the moral vignettes via Qualtrics surveys rather than by paper and pencil. Subjects in these three studies used a slider scale that yielded a number rather than making a mark on a line on paper to indicate their moral judgments.

A fourth deviation from the original study was our decision to compute a moral wrongness composite score for all participants who rated at least 3 out of 6 vignettes.[3] We

[3] Eskine et al. (2011) indicated that, "Three of the 57 participants correctly guessed our hypothesis and were

preregistered this approach to maximize statistical power and minimize selection biases. There were very few exclusions based on this preregistered criterion; three of the participants tested rated fewer than 3 vignettes. See *Preliminary Analyses* of the moral wrongness composite for further details.

Protocol Fidelity

Labs qualified to participate in this CREP project by submitting an application, preparing an OSF page with all materials and procedures, and submitting a video of a mock experimental procedure. Volunteer psychology professors working with CREP reviewed this material for fidelity to the CREP protocol. The CREP protocol required the use of exact stimuli such as the moral judgment task, use of approximate stimuli where necessary (for example, Minute Maid Berry Punch is not easily available outside of the United States), and appropriate data reporting. See this link for explicit CREP instructions and materials: https://osf.io/4hkjv/. Labs were required to incorporate reviewer feedback on both written procedures and the video to increase precise compliance. Each lab included at least one student researcher and one supervising faculty sponsor. Labs that successfully complied with CREP protocol, passed review, and submitted the target N (57) were eligible to receive a CREP certificate and a monetary reward (monetary rewards ended in July 2017 as funding ended).

---

therefore excluded from all analyses. An overall moral-judgment score was obtained for each of the remaining 54 participants (bitter condition: $n = 15$; sweet condition: $n = 18$; control condition: $n = 21$) by averaging his or her ratings of the six vignettes" (p. 296). They did not indicate whether any participants had scores for fewer than six of the vignettes, or if they set a minimum number of vignettes for which scores were required to compute the moral wrongness composite. Without an explicit reference to the contrary, we assume from this that all participants provided scores for all six vignettes and, thus, that our decision to compute a moral wrongness composite score for participants who rated at least 3 out of 6 vignettes represents a departure from the original research.

Participants

We analyzed data from a total of 1,137 participants in $k = 11$ studies. The per-study N ranged from 24 to 439 (median = 65). An additional four participants declined to confirm consent to use of their data, three participants were under age 18 years, three participants rated fewer than 3 vignettes, and beverage condition was unknown for an additional two participants in one study.

In our preregistration, we indicated that if linear mixed-effects analyses indicate that the effects of the beverage contrasts or their interactions with political orientation vary across levels of participant knowledge of the hypothesis, we would exclude participants demonstrating knowledge of the hypothesis. However, in part because we could not obtain the open-ended responses necessary to code level of knowledge of the hypothesis for 3 out of 11 studies, we retained all available participants in our analyses. See *Data analysis and inference criteria*.

With regard to gender, 671 (59%) identified as female, 392 (34%) as male, and 6 (1%) as nonbinary; there was no gender information available for 68 (6%) participants.

Materials

Materials were provided on the OSF website including demographic information and two versions of the moral vignette packet (identical but each in a different order).

As with the original study, labs used six vignettes first presented by Wheatley and Haidt (2005). The six vignettes focus on the following main characters (and situations): Bob (has a sexual relationship with his second cousin), Frank (cooks and eats his dead dog), George (lawyer who seeks clients at the hospital emergency room), Arnold (politician who condemns corruption but accepts bribes himself), Robert (shoplifts clothing), and Tim (takes books and other items out of the library without checking them out).

Moral judgments consisted of reading each vignette and answering the question "How morally wrong is this?" by making a mark on a line or selecting a number on a scale with *not at all wrong*, *moderately wrong*, and *extremely wrong* anchors. Ratings were scaled to a range of 0-100 consistent with Eskine et al.

A distractor task about "imageability" was also created based on the description in the original study (i.e., ". . . they described their language background and rated sentences for their imageability.") The intention of the distractor task is to accompany the cover story in disguising the purpose of the study from participants.

Participants rated how much they enjoyed their beverage, and how sweet, bitter, neutral, or disgusting they found their beverage using a 7-point scale ranging from 1, *Not at all*, to 4, *Neutral*, to 7, *Very much*. This portion of the questionnaire served as a manipulation check to ensure that each type of beverage elicited the intended response in the participants.

Labs were required to have access to drinking water, fruit punch, any brand of Swedish Bitters, cups, a private space to fill out the experimental forms, a booklet or survey containing the moral vignettes, the forms or survey used for a manipulation check, demographics including political orientation, the distractor task, and the procedures including the cover story. The specific questions used to collect demographic information differed somewhat between labs. For example, some labs used an open answer box to collect political orientation data, while others gave participants specific options to choose from.

The cover story was based on the description in the original paper: "They were told that we were exploring the effects of motor interference (specifically arm-hand movements) on cognitive processing, and we therefore directed them to drink a beverage during a moral-judgment task to instantiate this movement in a natural way" (Eskine et al., 2011, p. 296).

Procedure

Prior to the arrival of participants, experimenters consulted a pre-printed list of random numbers or the online survey used for random assignment to determine the appropriate drink preparation for the respective participant. Participants were run individually in physically separate spaces.

After providing informed consent, experimenters explained the study was an examination of the influence of motor interference on cognitive processing. On the informed consent form, experimenters provided a list of ingredients for the appropriate beverage condition to avoid exposing participants to allergens unwittingly. Researchers called attention to the ingredient list in case participants had not read the consent form closely. (Participants at the Tufts site verbally confirmed they were not allergic to any of the ingredients during the consent process.) Experimenters provided the beverage and told participants to drink it in a swift motion "as if drinking a shot." Participants then completed the first half of the moral judgment task. Experimenters administered a second serving of the beverage and then instructed participants to complete the second half of the moral judgment task.

Participants then completed the distractor task, beverage ratings, and demographic information including political orientation. Finally, they responded to the prompt "What do you think this study is about? Please provide a few details to explain your answer." Upon completion, participants were debriefed verbally and in writing. The complete protocol, including a script for researchers, is available on the OSF page for this study.

Two authors (EG and MAF) who were blind to beverage assignment coded political orientation responses into one of three categories, conservative, liberal, or other. The "other" category included participants who declined to provide information. The two coders exhibited excellent agreement, Cohen's kappa = 1.00, 95% CI[0.99, 1] (weighted kappa =

0.99). We used the first coder's responses in subsequent analyses. Across studies, 648 (57%) were coded as liberal, 162 (14%) as conservative, and 327 (29%) as other.

The same two authors coded the open-ended responses in which participants explained what they thought the study was about into one of three categories capturing levels of knowledge about the hypothesis. The three categories were naive (i.e., no insight into the hypothesis; e.g., "How the mechanical act of drinking would affect an individual's moral judgment."), partially suspicious (i.e., insight into part of the hypothesis; e.g., "This study may be about how much we liked or disliked the beverage and whether or not that influenced our answers."), and fully suspicious (i.e., clear and accurate guess of hypothesis; e.g., "This study is about how arm movement/taste can determine how you feel/think about something. I had a bad taste in my mouth so I saw everyone as bad in the short stories."). The two coders exhibited very good agreement, Cohen's kappa = 0.76, 95% CI[0.71, 0.80] (weighted kappa = 0.83). When the two coders disagreed, we selected the value representing greater knowledge of the hypothesis. We could not obtain participant responses to this question for Studies 1, 2, and 3 and thus could not code level of knowledge of the hypothesis for any of the participants in these three studies.

Note that because the knowledge/suspicion question was presented after the beverage rating manipulation check, it is possible that thinking about the beverages during the ratings task triggered participants' subsequent guesses about the purpose of the study. As a result, some participants we classified as partially or fully suspicious may have actually been naïve when they made wrongness judgments but were "clued in" to the hypothesis by the time they were asked to guess the purpose of the study. Thus the partially and fully suspicious groups may include subjects who had been naïve when making judgments; we can however, be fairly certain that subjects who responded with naiveté even with the benefit of exposure to the beverage rating hints, were in fact naïve when they made their judgments.

Data analysis and inference criteria

We present the rationale for and details of all analyses where applicable in the Results section. Broadly speaking, in preliminary analyses, we computed descriptive statistics for moral wrongness and beverage ratings, and internal consistency reliability for the moral wrongness composite. In confirmatory analyses, we conducted random effects meta-analyses for two effects of interest, bitter versus control and bitter versus sweet. We also conducted one-sided tests to determine whether observed effects were smaller than the effect size the original study could have detected with 33% power and/or equivalent to zero. To complement these approaches, we conducted linear mixed-effects regression (LMER) modeling of individual participant judgments of moral wrongness. Finally, we conducted a series of four Bayes Factors (BF) tests described by Verhagen and Wagenmakers (2014) for each replication study.

For analyses based on null hypothesis significance testing, we set an alpha criterion of .05 (two-tailed). For BF tests, we considered BF greater than 3 to provide nonanecdotal evidence for the replication hypothesis and BF less than 1/3 to be nonanecdotal evidence against the replication hypothesis.

We wrote this manuscript as an R Markdown document in RStudio 1.2.1335 (RStudio Team, 2015), and analyzed our data using R (Version 3.5.2; R Core Team, 2017) and the R-packages *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2015), *bindrcpp* (Version 0.2.2; Müller, 2018), *boot* (Version 1.3.20; Davison & Hinkley, 1997), *BSDA* (Version 1.2.0; Arnholt & Evans, 2017), *coda* (Version 0.19.2; Plummer, Best, Cowles, & Vines, 2006), *compute.es* (Version 0.2.4; Re, 2013), *dplyr* (Version 0.8.0.1; Wickham et al., 2017), *effsize* (Version 0.7.4; Torchiano, 2017), *emmeans* (Version 1.3.4; Lenth, 2018), *ggplot2* (Version 3.1.1; Wickham, 2009), *gridExtra* (Version 2.3; Auguie, 2017), *lattice* (Version 0.20.38; Sarkar, 2008), *lme4* (Version 1.1.21; Bates, Maechler, Bolker, & Walker, 2015), *lmerTest* (Version 3.1.0; Kuznetsova, Brockhoff, & Christensen, 2017), *MASS* (Version 7.3.51.1;

Venables & Ripley, 2002), *Matrix* (Version 1.2.15; Bates & Maechler, 2017), *MBESS* (Version 4.4.3; Kelley, 2017), *MCMCpack* (Version 1.4.4; Martin, Quinn, & Park, 2011), *metafor* (Version 2.0.0; Viechtbauer, 2010), *pacman* (Version 0.5.0; Rinker & Kurkiewicz, 2017), *papaja* (Version 0.1.0.9842; Aust & Barth, 2017), *polspline* (Version 1.1.13; Kooperberg, 2015), *psych* (Version 1.8.12; Revelle, 2017), *pwr* (Version 1.2.2; Champely, 2017), *R2WinBUGS* (Version 2.1.21; Sturtz, Ligges, & Gelman, 2005), *SDMTools* (Version 1.1.221; VanDerWal, Falconi, Januchowski, Shoo, & Storlie, 2014), *sjlabelled* (Version 1.0.16; Lüdecke, 2018a), *sjPlot* (Version 2.6.2; Lüdecke, 2018b), *sjstats* (Version 0.17.3; Lüdecke, 2018c), *stargazer* (Version 5.2.2; Hlavac, 2018), *stringr* (Version 1.4.0; Wickham, 2019), *tidyr* (Version 0.8.3; Wickham & Henry, 2017), *TOSTER* (Version 0.3.4; Lakens, 2017), *viridis* (Version 0.5.1; Garnier, 2018a, 2018b), and *viridisLite* (Version 0.3.0; Garnier, 2018b).

When analyzing the data, we discovered that some studies contributed no or very few observations to one or more cells of the study design. We made the *post hoc* decision to only include studies for which there were at least $n = 2$ participants in each cell for random effect meta-analyses, which allowed us to compute both a mean and standard deviation. This left $k = 5$ studies for comparisons between the beverage conditions among only conservative participants and $k = 9$ studies for comparisons among only liberal participants.[4] We included all available observations in our LMER models.

[4] The original study had $N = 19$ conservative participants and $N = 25$ liberal participants. Assuming an even distribution of participants to beverage conditions, they had a minimum of 6 participants per cell. Results are similar if we instead require at least $n = 6$ in each cell to be consistent with the original study, although doing so reduces the number of studies to $k = 3$ for comparisons among conservatives and $k = 7$ for comparisons among liberals. See version 2 (dated January 14, 2019) of our preprint posted on PsyArXiv at https://psyarxiv.com/349pk/.

Results

Preliminary Analyses

Moral Wrongness Composite.

Table 1

Descriptives and internal consistency reliability for the moral wrongness composite in each replication study.

	Number of Participants				Mean (Standard Deviation)			Internal Consistency	
study	bitter	control	sweet	total	bitter	control	sweet	alpha [95% CI]	omega [95% CI]
All participants									
1	20	19	20	59	75.28 (10.92)	65.37 (17.06)	67.94 (15.64)	0.52 [0.3, 0.69]	0.64 [0, 0.74]
2	22	22	21	65	67.08 (11.12)	67.52 (12.49)	69.8 (8.59)	0.28 [0, 0.53]	0.37 [0, 0.59]
3	19	19	18	56	62.26 (17.09)	64.25 (14.59)	66.96 (13.82)	0.53 [0.27, 0.7]	0.46 [0.09, 0.61]
4	40	24	36	100	68.99 (11.93)	70.7 (15.7)	71.31 (12.32)	0.5 [0.3, 0.64]	0.38 [0.02, 0.48]
5	17	22	20	59	76.53 (13.97)	76.65 (11.06)	73.43 (10.74)	0.26 [0, 0.51]	0.44 [0.08, 1]
6	9	9	6	24	78.22 (11.96)	64.05 (20.92)	60.68 (22.09)	0.69 [0.31, 0.87]	0.7 [0.2, 0.85]
7	24	29	28	81	67.03 (14.27)	65.94 (12.96)	69.05 (15.24)	0.45 [0.19, 0.62]	0.4 [0.21, 0.53]
8	23	22	23	68	70.2 (11.27)	72.55 (12.63)	70.8 (15.9)	0.53 [0.33, 0.67]	0.55 [0.19, 0.71]
9	51	55	54	160	70.2 (12.36)	71.7 (14.36)	69.77 (12.49)	0.57 [0.44, 0.71]	0.58 [0.43, 0.7]
10	147	142	150	439	71.79 (14.65)	68.31 (13.78)	73.69 (14.24)	0.53 [0.45, 0.59]	0.54 [0.47, 0.6]
11	9	8	9	26	70.2 (17.26)	64.76 (8.88)	70.59 (19.38)	0.62 [0.28, 0.76]	0.58 [0.01, 0.71]
Conservatives									
1	2	4	2	8	78.58 (2.57)	63.42 (8.64)	93.18 (9.64)	0.7 [0, 0.89]	0.58 [0.44, 0.64]
2	16	13	17	46	66.14 (11.48)	61.9 (10.31)	71.05 (7.9)	0.19 [0, 0.52]	0.37 [0, 1]
4	8	6	9	23	63.97 (6.27)	72.74 (6.17)	74.65 (9.82)	NA [NA, NA]	0.34 [0.01, 1]

5	3	7	2	12	78.1 (17.51)	75.73 (13.99)	82.8 (0.76)	0.48 [0, 0.72]	NA [NA, NA]
10	17	16	20	53	76.67 (12.78)	70.25 (10.81)	72.23 (13.9)	0.4 [0.08, 0.63]	0.47 [0.19, 0.63]
Liberals									
1	7	5	9	21	75.32 (12.2)	50.88 (10.46)	65.28 (13.94)	0.64 [0.26, 0.84]	0.75 [0, 0.85]
3	12	7	7	26	58.72 (14.3)	55.03 (13.64)	65.9 (16.01)	0.48 [0.03, 0.74]	NA [NA, NA]
4	21	11	13	45	70.38 (13.59)	69.05 (16.07)	70.39 (15.49)	0.62 [0.42, 0.77]	0.67 [0.42, 1]
5	7	8	7	22	71.74 (12.14)	74.48 (12.41)	76.29 (5.57)	NA [NA, NA]	0.11 [0, 1]
6	3	3	2	8	78.42 (17.77)	73.95 (15.98)	61.36 (1.4)	0.54 [0, 0.82]	NA [NA, NA]
7	20	20	23	63	65.89 (15.08)	65.8 (12.92)	68.19 (14.72)	0.42 [0.12, 0.64]	0.41 [0.13, 0.57]
8	10	10	8	28	65.53 (6.15)	78.8 (13.42)	70.04 (10.58)	0.4 [0, 0.68]	NA [NA, NA]
9	30	31	29	90	69.91 (12.82)	72.33 (12.72)	67.25 (12.3)	0.57 [0.4, 0.7]	0.59 [0.37, 0.74]
10	108	111	113	332	70.43 (14.99)	68.3 (13.74)	73.72 (14.55)	0.54 [0.46, 0.61]	0.55 [0.45, 0.62]

Note. Internal consistency was calculated across beverage groups. 95% confidence intervals (CI) were determined via 2,000 bias-corrected and accelerated bootstrap replications. Results were reported as NA for alpha when negative correlations between some items and the total score led to a negative value (conservatives in Study 4 and liberals in Study 5). Results were reported as NA for omega when the model failed to converge (conservatives in Study 5 and liberals in Study 3, 6, and 8).

We computed a moral wrongness composite for participants who provided moral wrongness judgments for at least 3 of 6 vignettes. All participants did so for at least 4 vignettes; 99.12% had a value for 5 or 6 vignettes (7.04% had a value for 5 vignettes, 92.08% had a value for 6 vignettes).

Table 1 shows descriptive statistics for the moral wrongness composite in each beverage group for each replication study. We will provide descriptive statistics across studies in the *Confirmatory Analyses* section of the Results.[5]

Table 1 also shows internal consistency reliability of the moral wrongness composite across beverage groups in each replication study. Cronbach's alpha calculated across all political orientations within each study ranged from 0.26 to 0.69 (median = 0.53), and omega ranged from 0.37 to 0.70 (median = 0.54). Internal consistency by study therefore ranged from poor to acceptable, with most falling below .70. Across participants in all studies, Cronbach's alpha was 0.50, 95% CI[0.45, 0.54], and omega was 0.49, 95% CI[0.43, 0.54].

Beverage Ratings. As a manipulation check, we assessed the extent to which the three beverages had the intended effect on subjective ratings (bitter, disgusting, neutral, and sweet) for each study. We computed descriptive statistics and linear regression models that compared ratings for the bitter group to ratings made by each of the other two groups across all participants and for conservative and liberal subgroups for each study (see Table A3 in the appendix). In addition, to summarize across studies, we examined four LMER models assessing the fixed effects of beverage type and political orientation on each rating with a random intercept for studies (see Table A4 in the appendix). The beverage contrasts were consistently significant in expected directions for all three political orientations.

Based on estimated marginal means from the LMER models, the bitter group

[5] See Table A2 in the appendix for linear regression models that compared moral wrongness judgments for the bitter group to ratings made by each of the other two groups across all participants and for conservative and liberal subgroups for each study.

perceived their beverage to be quite bitter, $M = 6.18$, 95% CI [5.98, 6.38], and disgusting, $M = 5.84$, 95% CI [5.63, 6.05], and not very sweet, $M = 1.44$, 95% CI [1.27, 1.61], or neutral, $M = 1.48$, 95% CI [1.27, 1.68] (7-point scale). The control group perceived their beverage to be quite neutral, $M = 5.65$, 95% CI [5.45, 5.85], and not very bitter, $M = 1.67$, 95% CI [1.48, 1.87], disgusting, $M = 1.36$, 95% CI [1.15, 1.57], or sweet, $M = 1.74$, 95% CI [1.57, 1.91]. The sweet group perceived their beverage to be quite sweet, $M = 5.55$, 95% CI [5.39, 5.71], and not very bitter, $M = 1.98$, 95% CI [1.79, 2.17], disgusting, $M = 1.95$, 95% CI [1.74, 2.15], or neutral, $M = 2.28$, 95% CI [2.08, 2.47].

Level of Knowledge of the Hypothesis. Of the 933 participants for whom we were able to code knowledge of the hypothesis, $N = 543$ (58.20%) were naive, $N = 336$ (36.01%) were partially suspicious, and $N = 54$ (5.79%) were fully suspicious. The original authors did not report how many participants were partially suspicious, but indicated that 3 out of 57 (5%) "correctly guessed our hypothesis" (Eskine et al., 2011, p. 296).

Based on predicted probabilities obtained in a generalized linear mixed-effects logistic regression assessing the fixed effects of beverage type and political orientation with a random intercept for studies on level of knowledge (0 = naive, 1 = partially or fully suspicious), more participants in the bitter (60.65%) and sweet (52.37%) groups were partially or fully suspicious relative to the participants in the control group (38.85%). The difference was statistically significant for bitter versus control, $b_{logodds} = 0.65$, $SE = 0.26$, $p = .014$, but not for sweet versus control, $b_{logodds} = 0.37$, $SE = 0.26$, $p = .145$ (see Figure A1 in the appendix).

Confirmatory Analyses

Random Effects Meta-analyses and One-sided Tests. In our first set of analyses, we estimated the standardized effects of the two beverage contrasts on moral wrongness within and across political orientations and within and across replication studies by conducting random-effect meta-analyses. These meta-analyses enable us to estimate to what extent drinking a disgusting, bitter beverage harshens moral judgments in standardized

units relative to both water (control) and juice (sweet). We would conclude there is support for the hypotheses if the two overall effects (bitter versus control and bitter versus sweet) were significantly greater than zero in the positive direction, perhaps only among conservative participants as in the original study. We used a random effects approach to determine to what extent effect sizes vary from one study to the next. We excluded the original study from these analyses as is typical for registered replication reports (e.g., Wagenmakers et al., 2016); this yields estimates that are based only on unpublished studies that were registered in advance and, thus, are unbiased.

In addition to random effects meta-analyses, we conducted one-sided tests examining whether the standardized meta-analytic effect sizes for the two contrasts of interest were significantly smaller than the effect size the original authors had 33% power to detect ($d_{33\%}$; small-telescopes approach; Simonsohn (2015)). The small-telescopes approach tells us whether the replication study effect size is large enough to have been detectable in the original study. If the replication study were well-powered and the replication effect significantly smaller than $d_{33\%}$, we would conclude that the original study was unable to draw meaningful conclusions about the studied effect. Across all participants in the original study, $d_{33\%} = 0.53$ for the bitter versus control contrast, and $d_{33\%} = 0.55$ for the bitter versus sweet contrast. Among conservative participants, $d_{33\%} = 0.94$ for both contrasts. Among liberal participants, $d_{33\%} = 0.80$ for both contrasts.[6]

Finally, we also conducted one-sided tests examining whether the replication effect was effectively equivalent to zero. We preregistered two sets of standardized effect size

[6] The original report did not specify the number of participants in each beverage group for conservative (N = 19) and liberal (N = 25) subgroups. We therefore assumed an even distribution and estimated 19/3 = 6.33 conservative participants in each beverage group and 25/3 = 8.33 liberal participants in each beverage group when calculating $d_{33\%}$. For results of one-sided tests of the moral wrongness composite in each replication study asking whether the observed effect size was smaller than $d_{33\%}$ or equivalent to 0 +/- $d_{33\%}$, see Tables A5 and A6 in the appendix, respectively.

equivalence bounds around 0, specifically 0 +/- $d_{33\%}$ and 0 +/- 0.30 (Lakens, Scheel, & Isager, 2018). The equivalence bound of +/- $d_{33\%}$ reflects the smallest effect size the original study could have detected. The equivalence bound of +/- 0.30 reflects the smallest effect size of interest to us in light of possible inadequate power to detect equivalence to zero of effects smaller than that. To simplify presentation of results, we deviate from our preregistration and only report results for the more stringent of the two, 0 +/- 0.30. If the meta-analytic effect sizes were equivalent to 0 +/- 0.30, this would suggest that the effects are unlikely to be larger than +/- 0.30.

Figures 1 and 2 display the standardized effect sizes and 95% confidence interval for each contrast within and across studies across all participants and for conservative and liberal subgroups. Figures A2 and A3 in the appendix display the parallel information for raw mean differences instead of standardized mean differences; the pattern is quite similar. In the text below we summarize results for the two contrasts of interest, bitter versus control and bitter versus sweet across all participants as well as separately within conservative and liberal subgroups.

All Participants.

Across all participants in $k = 11$ studies, mean moral wrongness across studies, weighted by number of participants in each group for each study, was 70.65 ($SD = 3.52$) in the bitter group, 68.94 ($SD = 3.41$) in the control group, and 71.29 ($SD = 2.86$) in the sweet group.

The overall effect for bitter versus control was negligible and in the predicted direction. This effect was significantly smaller than $d_{33\%} = 0.53$, $Z = -5.22$, $p = < .001$. It was also equivalent to 0 +/- 0.30, $Z = -2.42$, $p = .008$. Heterogeneity was low.

The overall effect for bitter versus sweet was negligible but in the opposite of the predicted direction. This effect was significantly smaller than $d_{33\%} = 0.55$, $Z = -8.31$, $p <$

.001. It was also equivalent to 0 +/- 0.30, $Z = 3.51$, $p < .001$. Heterogeneity was low.

Conservative Participants.

Across conservative participants in $k = 5$ studies, mean moral wrongness across studies, weighted by number of participants in each group for each study, was 70.98 ($SD = 6.99$) in the bitter group, 68.45 ($SD = 5.88$) in the control group, and 73.53 ($SD = 5.64$) in the sweet group.

The overall effect for bitter versus control was small and in the predicted direction. This effect was significantly smaller than $d_{33\%} = 0.94$, $Z = -1.80$, $p = .036$. It was not equivalent to 0 +/- 0.30, $Z = -0.22$, $p = .412$. Heterogeneity was substantial.

The overall effect for bitter versus sweet was medium and in the opposite of the predicted direction. This effect was significantly smaller than $d_{33\%} = 0.94$, $Z = -4.46$, $p < .001$. However, it was not equivalent to 0 +/- 0.30, $Z = -0.52$, $p = .697$. Heterogeneity was moderate.

Liberal Participants.

Across liberal participants in $k = 9$ studies, mean moral wrongness across studies, weighted by number of participants in each group for each study, was 69.38 ($SD = 3.97$) in the bitter group, 68.66 ($SD = 5.94$) in the control group, and 71.23 ($SD = 4.01$) in the sweet group.

The overall effect for bitter versus control was near zero. This effect was significantly smaller than $d_{33\%} = 0.80$, $Z = -4.95$, $p < .001$, and equivalent to 0 +/- 0.30, $Z = -1.78$, $p = .038$. Heterogeneity was moderate.

The overall effect for bitter versus sweet was negligible and in the opposite of the predicted direction. This effect was significantly smaller than $d_{33\%} = 0.80$, $Z = -8.97$, $p < .001$. It was also equivalent to 0 +/- 0.30, $Z = 1.91$, $p = .028$. Heterogeneity was low.

Linear Mixed-Effects Regression Models. The random effects meta-analyses and one-sided tests above enabled us to examine the standardized effect of the beverage manipulation on moral wrongness within and across political orientation groups. These analyses also addressed the extent to which beverage effects vary across replication studies. They did not, however, reveal whether political orientation and/or knowledge of the hypothesis formally moderate the effects of beverage condition on moral wrongness. They also did not account for potential random variation in moral wrongness across participants and vignettes. Thus, we conducted linear mixed-effects regressions (LMER) of the individual participant replication data to address these gaps. Apportioning all relevant sources of variation simultaneously in LMER may increase sensitivity to detect predicted effects. For the same reasons expressed for the random effects meta-analyses, we excluded the original study from these analyses.

LMER Model Specification.

We report three LMER models.[7] All models included fixed effects reflecting beverage type with two contrasts, bitter [.5] versus control [-.5] (bvc; sweet [0]) and bitter [.5] versus sweet [-.5] (bvs; control [0]), and political orientation also with two contrasts, conservative [.5] versus liberal [-.5] (cvl; other [0]), and conservative [.5] versus other [-.5] (cvo), and their interactions.

In Model 1, we analyzed moral wrongness as a composite averaged across vignettes with a random intercept to allow variation in average moral wrongness across studies. Because we used effect coding, the intercept reflects mean moral wrongness across all participants. There were $N = 1137$ observations (participants) for this analysis.

[7] We originally planned to report models that included random slopes for beverage type and political orientation contrasts across studies, levels of knowledge of the hypothesis, and/or vignettes. However, with only three levels of knowledge, the plan to examine random variation as a function of this variable was ill-advised. Also, we excluded models with random slopes (convergence issues), and treated level of knowledge of the hypothesis as an additional fixed effect.

In Model 2, we added level of knowledge of the hypothesis as a fixed effect that could interact with both beverage type and political orientation.[8] We recoded the knowledge variable to 0, naive, or 1, partially or fully suspicious[9]; the intercept therefore reflects mean moral wrongness among naive participants. We could not retrieve the open-ended responses to the question about what the study was about for three studies. As such, we could not code level of knowledge of the hypothesis for any of the participants in those studies. With these study-level exclusions, there were $N = 933$ observations (participants) for this analysis.

In Model 3, we analyzed moral wrongness for each vignette with the fixed effects from Model 2 and with random intercepts that allowed variation in mean moral wrongness across studies, participants, and vignettes. The intercept reflects mean moral wrongness among naive participants. There were $N = 5502$ observations reflecting 933 participants each with moral wrongness ratings for up to 6 vignettes. A total of 96 moral wrongness ratings were missing (1.71%) across 3 studies. We did not impute missing values.

LMER Model Results.

Confirmatory.

Table 2 summarizes results of the three LMER models. See the appendix for figures depicting moral wrongness in all design cells after accounting for fixed and random sources of variation in Model 1 or 2.

Consistent with the hypothesis that physical disgust induces moral wrongness, participants who drank a bitter beverage made significantly harsher moral judgments than those who drank water (see BvC term in Models 1, 2, and 3, respectively). However, contrary to the hypothesis, participants who drank a bitter beverage made numerically

[8] As shown earlier in this manuscript, level of knowledge of the hypothesis varied as a function of beverage type, thus these two predictors are not orthogonal.

[9] There were no conservative participants who were coded as suspicious, thus we could not retain all three levels of knowledge of the hypothesis.

milder moral judgments than those who drank the sweet juice; this difference was not statistically significant (see BvS term in Models 1, 2, and 3, respectively).

The bitter minus control effect did not vary significantly by conservative or liberal political orientation (see BvC × CvL and BvC × CvL × Knowledge terms in Models 1, 2, and 3, respectively). Similarly, the bitter minus sweet effect did not vary significantly by conservative or liberal political orientation (see BvS × CvL and BvS × CvL × Knowledge terms in Models 1, 2, and 3, respectively). See bottom panel of Figure A4.

The bitter minus control effect did, however, vary significantly by level of knowledge of the hypothesis (see BvC × Knowledge terms in Models 2 and 3). Based on estimated marginal means from Model 2, among naive participants, the bitter beverage, $M = 72.42$ ($SE = 1.56$), resulted in harsher moral judgments than water, $M = 67.90$ ($SE = 1.35$); this simple effect was statistically significant, $t(45.35) = 2.46$, $p = .018$. Among suspicious participants, the bitter beverage, $M = 70.43$ ($SE = 1.70$), resulted in milder moral judgments than water, $M = 74.02$ ($SE = 2.03$); this simple effect was not statistically significant, $t(87.03) = -1.41$, $p = .161$.

The bitter minus sweet effect did not vary significantly by level of knowledge of the hypothesis (see BvS × Knowledge terms in Models 2 and 3). Based on estimated marginal means from Model 2, among naive participants, the bitter beverage, $M = 72.42$ ($SE = 1.56$), resulted in similar moral judgments compared to the sweet juice, $M = 72.15$ ($SE = 1.48$). Among suspicious participants, the bitter beverage, $M = 70.43$ ($SE = 1.70$), also resulted in similar moral judgments compared to the sweet juice, $M = 71.31$ ($SE = 1.60$).

The proportion of variance in moral wrongness due to studies was 0.03 and 0.01, respectively, based on intraclass correlation coefficients (ICC) computed based on Models 1 and 2. The proportion of variance across studies was lower in Model 3, ICC = 0.00, likely due to variance accounted for by participants, ICC = 0.14, and items, ICC = 0.05.

Table 2

Linear mixed-effects models examining moral wrongness

	(1)	(2)	(3)
(Intercept)	69.937*** (0.874)	70.826*** (0.956)	70.782*** (2.706)
bitter - control (BvC)	3.276* (1.369)	5.849** (2.045)	5.794** (2.045)
bitter - sweet (BvS)	−2.030 (1.339)	−2.657 (2.135)	−2.526 (2.136)
conservative - liberal (CvL)	2.388 (1.233)	1.042 (1.760)	1.118 (1.762)
conservative - other (CvO)	−1.864 (1.373)	1.467 (2.141)	1.370 (2.141)
Level of Knowledge (0 = naive)		1.093 (1.255)	1.083 (1.257)
BvC × CvL	2.823 (3.271)	5.905 (4.720)	6.061 (4.721)
BvC × CvO	−0.702 (3.717)	−5.826 (5.601)	−5.636 (5.598)
BvS × CvL	1.484 (3.218)	5.981 (4.906)	6.199 (4.909)
BvS × CvO	−6.109 (3.666)	−7.857 (5.997)	−7.784 (5.990)
BvC × Knowledge		−10.050** (3.622)	−9.883** (3.631)
BvS × Knowledge		3.878 (3.377)	3.754 (3.378)
CvL × Knowledge		3.534 (2.837)	3.663 (2.840)
CvO × Knowledge		−3.718 (3.235)	−3.625 (3.235)
BvC × CvL × Knowledge		−10.074 (8.267)	−10.139 (8.284)
BvC × CvO × Knowledge		−0.757 (9.421)	−0.794 (9.429)
BvS × CvL × Knowledge		−8.651 (7.796)	−8.802 (7.797)
BvS × CvO × Knowledge		9.230 (8.904)	9.114 (8.898)
Observations	1,137	933	5,502
Log Likelihood	−4,590.005	−3,735.196	−25,537.830
Akaike Inf. Crit.	9,202.011	7,510.392	51,119.670
Bayesian Inf. Crit.	9,257.409	7,607.160	51,265.150

Note: *p<0.05; **p<0.01; ***p<0.001

Exploratory.

Confirmatory LMER analyses as a whole did not reveal support for the hypothesis that drinking a bitter, disgusting beverage would promote a heightened sense of moral wrongness relative to drinking both water or sweet juice, either in all participants or among conservative participants. As such, we conducted some exploratory LMER analyses.

1. LMER Results for Participants Rating All Six Vignettes

We included all participants in our analyses who provided ratings of at least three of the six vignettes. It is possible that hypothesized effects might emerge only among participants who rated all six. Table A7 in the appendix presents the three original LMER models again, this time after excluding participants who rated fewer than six vignettes. Estimates in these models are all rather similar in magnitude and direction to the models that included participants who rated only three or more vignettes; none of them reveal hypothesized elevations of moral wrongness for the bitter group relative to both the control and sweet groups across everyone or as a function of political orientation.

2. LMER Results by Vignette

It is possible that hypothesized effects might emerge in a subset of vignettes. For example, researchers have argued in the past that moral judgments of purity violations might be particularly susceptible to manipulations of physical disgust (Horberg, Oveis, Keltner, & Cohen, 2009). Table A8 in the appendix presents six linear mixed-effects models which are structured the same as Model 1 but with moral wrongness for an individual vignette as the criterion variable in each. Table A9 presents two linear mixed-effects models which are structured the same as Models 1 and 2 but with moral wrongness across the two purity violation vignettes as the criterion variable. None of these models reveal elevations of moral wrongness for the bitter group relative to both the control and sweet groups across everyone or as a function of political orientation. These results fail to support a causal effect of

physical disgust on moral wrongness judgments.

3. Association between Moral Wrongness and Disgust Ratings

Eskine et al. (2011) reported a positive between-subjects association between disgust ratings and moral wrongness judgments across participants, $\beta = .53$, $t(52) = 4.45$, $p < .001$. In the present replication studies, zero-order correlations between these two variables by study ranged from r = -0.11 to 0.42. A bivariate LMER assessing the fixed effect of disgust ratings (standardized) on moral wrongness (standardized) revealed a positive association, β = 0.07, p = .014. In two additional LMERs, bitter ratings were not associated with moral wrongness, $\beta = 0.01$, p = .808, and bitter and disgust ratings were strongly positively associated with one another, $\beta = 0.77$, $p = <$ .001. An LMER assessing the fixed effects of all four beverage ratings yielded a positive association between disgust and moral wrongness controlling for the other ratings, $\beta = 0.15$, p = .003. Across studies, each one-unit increase in disgust specifically (7-point scale) was associated with a $b = 0.87$-unit increase in moral wrongness (101-point scale). By contrast, there was a negative association between bitterness and moral wrongness controlling for the other ratings, β = -0.12, p = .019, and near-zero associations between neutral, β = -0.04, p = .363, and sweet, $\beta = 0.02$, p = .675, ratings and moral wrongness. These results establish a small, correlational link between physical disgust and moral wrongness judgments that likely does not reflect general unpleasantness since controlling for bitterness does not mitigate the association.

Bayes Factor Tests. Finally, the random effects meta-analyses and the LMER analyses above all rely on a frequentist statistical perspective with presentation of standardized effect sizes, confidence intervals, and p values. In our last set of analyses, we conducted a set of Bayes Factor (BF) tests. Unlike frequentist methods, BF tests quantify relative levels of evidence for the alternative hypotheses (i.e., that the bitter/disgusting beverage would harshen moral judgments compared to the control and sweet beverages) versus the null hypotheses (i.e., that there would be no effect of the beverage contrasts). Like

Wagenmakers et al. (2016), these tests focus on each of the replication studies individually. Three out of four of them explicitly incorporate the original study, as described below.

Bayes Factor Model Specification.

Following the logic and code given by Verhagen and Wagenmakers (2014), below we report four complementary BF tests for each replication study. In each case, the BF represents a comparison between two models; it captures the extent of evidence for one model relative to the other. First, the Jeffreys-Zellner-Siow (JZS) BF, which is independent from the original finding, determines the relative evidence for the effect being present versus absent in the replication by setting a standard two-sided Cauchy(0, 1) distribution as the prior, ignoring the original study. Second, the replication BF test sets the prior based on the posterior distribution from the original study. It determines the relative evidence for the original effect versus a null effect. Third, the equality-of-effect-size BF test determines whether the effect size in the replication study was equal to the effect size in the original study by determining whether variance τ^2 of the effect sizes is zero relative to nonzero. Fourth, the fixed-effect meta-analysis BF test pools the original and replication study data, and uses the two-sided Cauchy(0,1) distribution as the prior (similar to the JZS BF test).[10]

Bayes Factor Results.

Figure 3 summarizes Bayes Factor results across all participants in all studies. (See Table A10 and accompanying text in the appendix for Bayes Factor test results within conservative and liberal subgroups.)

---

[10] For the meta-analysis BF approach, we could have preregistered a meta-analysis BF across the original and all replication studies. Instead, we preregistered this series of separate two-study meta-analyses (across the original and each individual replication study) because doing so parallels what is possible with the other three types of Bayes Factors. Moreover, it is consistent with the approach taken in the registered replication report by Wagenmakers et al. (2016), who reported BFs for individual replication studies in addition to a random effects meta-analysis across all replication studies, which we did as well.

There was relatively consistent evidence against the replication hypothesis (bitter > control and bitter > sweet) across most studies for the equality-of-effect-size, JZS, and replication BF tests. By contrast, about half the studies provided evidence for the replication hypothesis based on the fixed-effect meta-analysis BF (MetaBayes) test. That said, because the MetaBayes BF tests pool the original and replication effects, the very large original effect size likely has a strong influence on the outcome. Also noteworthy, of the four replication studies for which MetaBayes provided more than nonanecdotal support for the replication hypothesis for bitter versus control, only one had a large enough sample size to inspire confidence in precision of estimation (Study 10 $N = 439$); the other three studies were on the smaller side (Study 1 $N = 59$, Study 6 $N = 24$, and Study 11 $N = 26$, respectively).

Discussion

Several studies have shown that experimental manipulations of disgust can amplify moral judgments (e.g., Harlé & Sanfey, 2010; Moretti & Pellegrino, 2010; Schnall et al., 2008a, 2008b; Van Dillen, Wal, & Bos, 2012; Wheatley & Haidt, 2005). Although more recent research has cast some doubt on the existence of this relationship, some studies have proposed that the effect might be more robust for a specific type of manipulation: gustatory and olfactory disgust inductions (see Landy & Goodwin, 2015a, also see 2015b; Schnall et al., 2015). As this notion was largely based on one low-powered study with a particularly large effect size, we were interested in obtaining an accurate estimate of its effect size by conducting a high-powered meta-analysis of 11 preregistered direct replication studies. Overall, we found little to no support for Eskine and colleagues' (2011) conclusions.

We adopted a multifaceted analysis strategy in an effort to fairly examine the research question (i.e., does gustatory disgust induced via a bitter drink amplify moral wrongness judgments?) from slightly different angles. In part, we employed a frequentist approach, running random effects meta-analyses, one-sided tests, and linear mixed-effects regression models to make inferences across all replication studies. We observed a very small

meta-analytic effect of the bitter drink inducing harsher moral judgments than water. This effect was small and statistically significant in linear mixed-effects regressions. However, in the random effects meta-analyses, effect size confidence intervals encompassed zero in all but one replication, resulting in an overall confidence interval that included zero (see Figure 1). In addition, standardized effects from those meta-analyses were significantly smaller than the smallest effect the original authors could have found, and unlikely to be larger than +/- 0.30 based on equivalence tests. Moreover, across frequentist tests, we found no effect of the bitter drink relative to the sweet one; if anything, the bitter drink led to more lenient moral judgments than the sweet drink. This latter finding is important in that it undermines the notion that disgust is responsible for any effects on moral judgments; instead, any effect on moral judgments might be explained by the act of drinking something with flavor. Finally, we found no evidence that political conservatives were especially harsh in their judgments after consuming a bitter drink. It should be noted, though, that this latter finding is based on a relatively small sample of conservative participants (a limitation discussed more elaborately below).

We also used a Bayesian approach, which quantifies relative strength of the evidence in favor of versus against the replication hypothesis. We computed four different Bayes Factor (BF) tests, of which three (i.e., Jeffreys-Zellner-Siow, replication, and equality-of-effect-size BF tests) showed evidence against the replication hypothesis across most studies. The fixed-effect meta-analytic BF showed more support for the idea that disgust amplifies the harshness of moral judgments. However, this is not surprising, as this approach pools each of the replication studies with the original study, which had a particularly large effect. Even in this case, though, only four of the eleven replication studies (three of which had smaller sample sizes) favored support for the replication hypothesis.

Based on results from these various tests, we conclude there is little to no support for the notion that gustatory disgust can increase ratings of moral wrongness. With that, our

study adds to the growing number of studies that fail to find support for a relationship between incidental manipulations of disgust and moral judgments (e.g., Johnson, Cheung, & Donnellan, 2014; Case et al., 2012; Johnson et al., 2016; Landy & Goodwin, 2015a). Our results not only cast doubt on Eskine and colleagues' (2011) conclusion that a disgusting taste impacts moral judgments, they also fail to support Schnall and colleagues' (2015) proposal of a special potency for gustatory inductions of disgust to influence moral condemnation.

It is possible that the effect of incidental disgust on moral judgments depends heavily on moderator variables, such as individual difference measures. For example, it has been suggested that the relationship is stronger for individuals who are generally sensitive to bodily sensations (i.e., high on Private Body Consciousness, Schnall et al., 2008b, also see 2015). We did not measure Private Body Consciousness (PBC), or any other individual difference measure, consistently across our studies so we cannot directly test this hypothesis. However, an earlier replication study by Johnson and colleagues (2016) tested the importance of PBC to the disgust-moral judgment link, but did not find evidence indicating it has a moderating effect (also see Johnson et al., 2014).

One moderator variable that we identified is participant knowledge of the hypothesis. A total of 6% of our participants demonstrated full knowledge of the hypothesis that drinking a bitter beverage would harshen moral judgments; an additional 36% partially guessed that hypothesis. We could not separate out effects by partial versus full suspicion due to absence of observations in one cell of the factorial design. However, when combining the fully and partially suspicious groups in our linear mixed-effects regression models, we found that the bitter versus water effect was larger in the predicted direction among naive participants than among fully or partially suspicious participants. This result is in line with the idea that induced disgust has a stronger effect on individuals who are unaware of a link between the taste of the drink and the moral judgments (Schnall et al., 2015). Contrary to this idea,

however, knowledge of the hypothesis did not moderate the bitter versus sweet effect.

It is unclear how closely the coding scheme we used to assess knowledge of the hypothesis maps onto the scheme used by the original authors. They reported that three of 57 participants (5%), who were excluded from analyses, correctly guessed the hypothesis. Based on that percentage, which is close to the 6% who demonstrated full knowledge in the present work, these participants may have had full knowledge, not just partial. Regardless, we can conclude that knowledge of the beverage hypothesis moderates the effect of beverage type on moral judgments to some extent. Future studies would benefit from refinement of deception procedures to minimize such knowledge. On this point, it is noteworthy that more participants in the bitter and sweet groups were partially or fully suspicious relative to the participants in the water control group; deception procedures would ideally yield similar rates of suspicion in all conditions so that level of knowledge of the hypothesis is orthogonal to the beverage manipulation.

There are some limitations of the current investigation that may have impacted our findings. One potential limitation is the low reliability of the moral vignettes across most of our replication samples. This may have influenced our ability to detect effects, especially if our average reliability was substantially lower than what was observed in the original study. Unfortunately, information about reliability was not reported in Eskine et al. (2011), or in Wheatley and Haidt (2005), the developers of these vignettes. However, linear mixed-effects analyses of the individual vignettes reveal that the bitter drink does not induce harsher moral judgments for any of the vignettes when compared to the sweet drink or water. It therefore seems unlikely that the observed low reliability of moral vignettes explain our mostly null findings.

Another limitation is that we had few samples with a sufficient number of conservative participants when testing the moderating effect of political ideology. Only five out of 11 replications had at least two conservative participants in each of the three beverage

conditions. Among those, only three had more than six conservative participants in each of the three beverage conditions (i.e., the estimated number of conservatives per condition in the original study). This likely decreased our ability to detect a potential moderation of political ideology. It should be noted, though, that we had substantially more conservative participants across replication samples ($N = 162$) than the original study ($N = 19$).

Conclusion

This work joins the growing number of crowdsourced replication projects combining multiple laboratories' efforts to replicate a single study (e.g., Moshontz et al., 2018; Hagger et al., 2016; Schweinsberg et al., 2016). Multi-lab replications have an advantage over single-lab replications in that they can more accurately estimate effect sizes given the possibility of a much larger sample size. It also joins the growing movement of conducting replications in the classroom (see also Leighton, Legate, LePine, Anderson, & Grahe, 2018; Wagge et al., 2019). Replications conducted by students represent more than a valuable pedagogical tool; they have been shown to be a promising means of carrying out high-quality replications (Frank & Saxe, 2012; Hawkins et al., 2018). Here we demonstrate that pedagogical replications can provide valuable evidence about the robustness of an effect that has not yet been submitted to independent replication. We believe pedagogical replications could be implemented widely to advance psychological science.

Author Contributions

We have collectively identified a set of major and minor contributors to this work. Major contributors are listed on the byline from 1-8. Minor contributors, in alphabetical order, are listed on the byline from 9-23. Contributions according to the CreDiT taxonomy (https://casrai.org/credit) were as follows: Conceptualization, all contributors; Data curation, E.G., F.M.A.W., and M.A.F.; Formal analysis, H.L.U. and R.L.L.; Funding acquisition, H.L.U. (for project at Tufts); Investigation, all minor contributors; Methodology, all contributors; Project administration, E.G. and C.D.C; Resources, all contributors;

Supervision, all major contributors; Validation, H.L.U. and R.L.L.; Visualization, H.L.U. and R.L.L.; Writing - original draft, all major contributors; Writing - review & editing, all contributors.

Conflicts of Interest

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Acknowledgements

The authors wish to acknowledge Jon Grahe and Hanz IJzerman for early guidance and reviews, and Jordan Wagge, Mark Brandt, and all the volunteer reviewers for their work with the Collaborative Replication & Education Project: https://osf.io/wfc6u/wiki/home/.

Funding

A Tufts University Faculty Research Award to H.L.U. covered the costs of the replication at the Tufts site.

Supplemental Material

Supplementary material referenced in this manuscript may be found at *link goes here*.

Prior Versions

We uploaded prior versions of this manuscript to the PsyArXiv preprint server at https://psyarxiv.com/349pk/.

References

Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence*, *149*(1), 91–130. doi:10.1016/S0004-3702(03)00054-7

Arnholt, A. T., & Evans, B. (2017). *BSDA: Basic statistics and data analysis*. Retrieved from https://CRAN.R-project.org/package=BSDA

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics*. Retrieved from https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bates, D., & Maechler, M. (2017). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Cameron, C. D., Payne, B. K., & Doris, J. M. (2013). Morality in high definition: Emotion differentiation calibrates the influence of incidental disgust on moral judgments. *Journal of Experimental Social Psychology*, *49*(4), 719–725.

Cannon, P. R., Schnall, S., & White, M. (2011). Transgressions and expressions: Affective facial muscle activity predicts moral judgments. *Social Psychological and Personality Science*, *2*(3), 325–331.

Case, T. I., Oaten, M. J., & Stevenson, R. J. (2012). Disgust and moral judgment. *Emotions, Imagination, and Moral Reasoning*, 195–218.

Champely, S. (2017). *Pwr: Basic functions for power analysis*. Retrieved from

https://CRAN.R-project.org/package=pwr

Chapman, H. A. (2018). A component process model of disgust, anger, and moral judgment. *Atlas of Moral Psychology*, 70.

Chapman, H. A., & Anderson, A. K. (2012). Understanding disgust. *Annals of the New York Academy of Sciences*, *1251*(1), 62–76.

Curtis, V., & Biran, A. (2001). Dirt, disgust, and disease: Is hygiene in our genes? *Perspectives in Biology and Medicine*, *44*(1), 17–31.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press. Retrieved from http://statwww.epfl.ch/davison/BMA/

Eskine, K. J., Kacinik, N. A., & Prinz, J. J. (2011). A bad taste in the mouth: Gustatory disgust influences moral judgment. *Psychological Science*, *22*(3), 295–299.

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*(6), 600–604.

Garnier, S. (2018a). *Viridis: Default color maps from 'matplotlib'*. Retrieved from https://CRAN.R-project.org/package=viridis

Garnier, S. (2018b). *ViridisLite: Default color maps from 'matplotlib' (lite version)*. Retrieved from https://CRAN.R-project.org/package=viridisLite

Gill, M. B., & Nichols, S. (2008). Sentimentalist pluralism: Moral psychology and philosophical ethics. *Philosophical Issues*, *18*(1), 143–163.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537),

2105–2108.

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

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion.* Vintage.

Harlé, K. M., & Sanfey, A. G. (2010). Effects of approach and withdrawal motivation on interactive economic decisions. *Cognition and Emotion*, *24*(8), 1456–1465.

Hawkins, R. X., Smith, E. N., Au, C., Arias, J. M., Catapano, R., Hermann, E., ... Reynolds, J. (2018). Improving the replicability of psychological science through pedagogy. *Advances in Methods and Practices in Psychological Science*, *1*(1), 7–18.

Hellmann, J. H., Thoben, D. F., & Echterhoff, G. (2013). The sweet taste of revenge: Gustatory experience induces metaphor-consistent judgments of a harmful act. *Social Cognition*, *31*(5), 531–542.

Hlavac, M. (2018). *Stargazer: Well-formatted regression and summary statistics tables.* Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). Retrieved from https://CRAN.R-project.org/package=stargazer

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology*, *97*(6), 963.

Inbar, Y., & Pizarro, D. A. (2014). Pollution and purity in moral and political judgment. *Advances in Experimental Moral Psychology: Affect, Character, and Commitments*, 111–29.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2012). Disgusting smells cause decreased liking of

gay men. *Emotion*, *12*(1), 23.

Ionescu, T., & Vasc, D. (2014). Embodied cognition: Challenges for psychology and education. *Procedia-Social and Behavioral Sciences*, *128*, 275–280.

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments? *Social Psychology*, *45*(3), 209–215. doi:10.1027/1864-9335/a000186

Johnson, D. J., Wortman, J., Cheung, F., Hein, M., Lucas, R. E., Donnellan, M. B., ... Narr, R. K. (2016). The effects of disgust on moral judgments: Testing moderators. *Social Psychological and Personality Science*, *7*(7), 640–647.

Kelley, K. (2017). *MBESS: The mbess r package*. Retrieved from https://CRAN.R-project.org/package=MBESS

Kooperberg, C. (2015). *Polspline: Polynomial spline routines*. Retrieved from https://CRAN.R-project.org/package=polspline

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi:10.18637/jss.v082.i13

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *1*, 1–8. doi:10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269.

Landy, J. F., & Goodwin, G. P. (2015a). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science*,

10(4), 518–536. doi:10.1177/1745691615583128

Landy, J. F., & Goodwin, G. P. (2015b). Our conclusions were tentative, but appropriate: A reply to schnall et al. (2015). *Perspectives on Psychological Science*, *10*(4), 539–540. doi:10.1177/1745691615590570

Leighton, D. C., Legate, N., LePine, S., Anderson, S. F., & Grahe, J. (2018). Self-esteem, self-disclosure, self-expression, and connection on facebook: A collaborative replication meta-analysis. *Psi Chi Journal of Psychological Research*, *23*(2).

Lenth, R. (2018). *Emmeans: Estimated marginal means, aka least-squares means*. Retrieved from https://CRAN.R-project.org/package=emmeans

Lüdecke, D. (2018a). *Sjlabelled: Labelled data utility functions (version 1.0.11)*. doi:10.5281/zenodo.1249215

Lüdecke, D. (2018b). *SjPlot: Data visualization for statistics in social science*. Retrieved from https://CRAN.R-project.org/package=sjPlot

Lüdecke, D. (2018c). *Sjstats: Statistical functions for regression models*. Retrieved from https://CRAN.R-project.org/package=sjstats

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, *42*(9), 22. Retrieved from http://www.jstatsoft.org/v42/i09/

Moretti, L., & Pellegrino, G. di. (2010). Disgust selectively modulates reciprocal fairness in economic interactions. *Emotion*, *10*(2), 169–180. doi:10.1037/a0017826

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., ... Chartier, C. R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *PsyArXiv*. doi:10.31234/osf.io/785qu

Müller, K. (2018). *Bindrcpp: An 'rcpp' interface to active bindings*. Retrieved from https://CRAN.R-project.org/package=bindrcpp

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from https://journal.r-project.org/archive/

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Re, A. C. D. (2013). *Compute.es: Compute effect sizes. R Package*. Retrieved from http://cran.r-project.org/web/packages/compute.es

Revelle, W. (2017). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Rinker, T. W., & Kurkiewicz, D. (2017). *pacman: Package management for R*. Buffalo, New York: University at Buffalo/SUNY. Retrieved from http://github.com/trinker/pacman

RStudio Team. (2015). *RStudio: Integrated development for r*. Boston, MA: RStudio, Inc. Retrieved from http://www.rstudio.com/

Sarkar, D. (2008). *Lattice: Multivariate data visualization with r*. New York: Springer. Retrieved from http://lmdvr.r-forge.r-project.org

Schnall, S., Benton, J., & Harvey, S. (2008a). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science*, *19*(12), 1219–1222.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008b). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, *34*(8), 1096–1109. doi:10.1177/0146167208317771

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2015). Landy and goodwin confirmed most of our findings then drew the wrong conclusions.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., ... others. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. doi:10.1177/0956797614567341

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running winbugs from r. *Journal of Statistical Software*, *12*(3), 1–16. Retrieved from http://www.jstatsoft.org

Torchiano, M. (2017). *Effsize: Efficient effect size computation*. Retrieved from https://CRAN.R-project.org/package=effsize

Tybur, J. M., Lieberman, D., Kurzban, R., & DeScioli, P. (2013). Disgust: Evolved function and structure. *Psychological Review*, *120*(1), 65.

VanDerWal, J., Falconi, L., Januchowski, S., Shoo, L., & Storlie, C. (2014). *SDMTools: Species distribution modelling tools: Tools for processing data associated with species distribution modelling exercises*. Retrieved from

https://CRAN.R-project.org/package=SDMTools

Van Dillen, L. F., Wal, R. C. van der, & Bos, K. van den. (2012). On the role of attention and emotion in morality: Attentional control modulates unrelated disgust in moral judgments. *Personality and Social Psychology Bulletin*, *38*(9), 1222–1231.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475.

Vicario, C. M., Rafal, R. D., Martino, D., & Avenanti, A. (2017). Core, social and moral disgust are bounded: A review on behavioral and neural bases of repugnance in clinical disorders. *Neuroscience & Biobehavioral Reviews*, *80*, 185–200.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R., ... others. (2016). Registered replication report: Strack, martin, & stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928.

Wagge, J. R., Brandt, M. J., Lazarevic, L. B., Legate, N., Christopherson, C., Wiggins, B., & Grahe, J. E. (2019). Publishing research with undergraduate students via replication work: The collaborative replications and education project. *Frontiers in Psychology*, *10*, 247.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, *16*(10), 780–784.

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., Francois, R., Henry, L., & Mueller, K. (2017). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2017). *Tidyr: Easily tidy data with 'spread()' and 'gather()' functions*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636. doi:10.3758/BF03196322

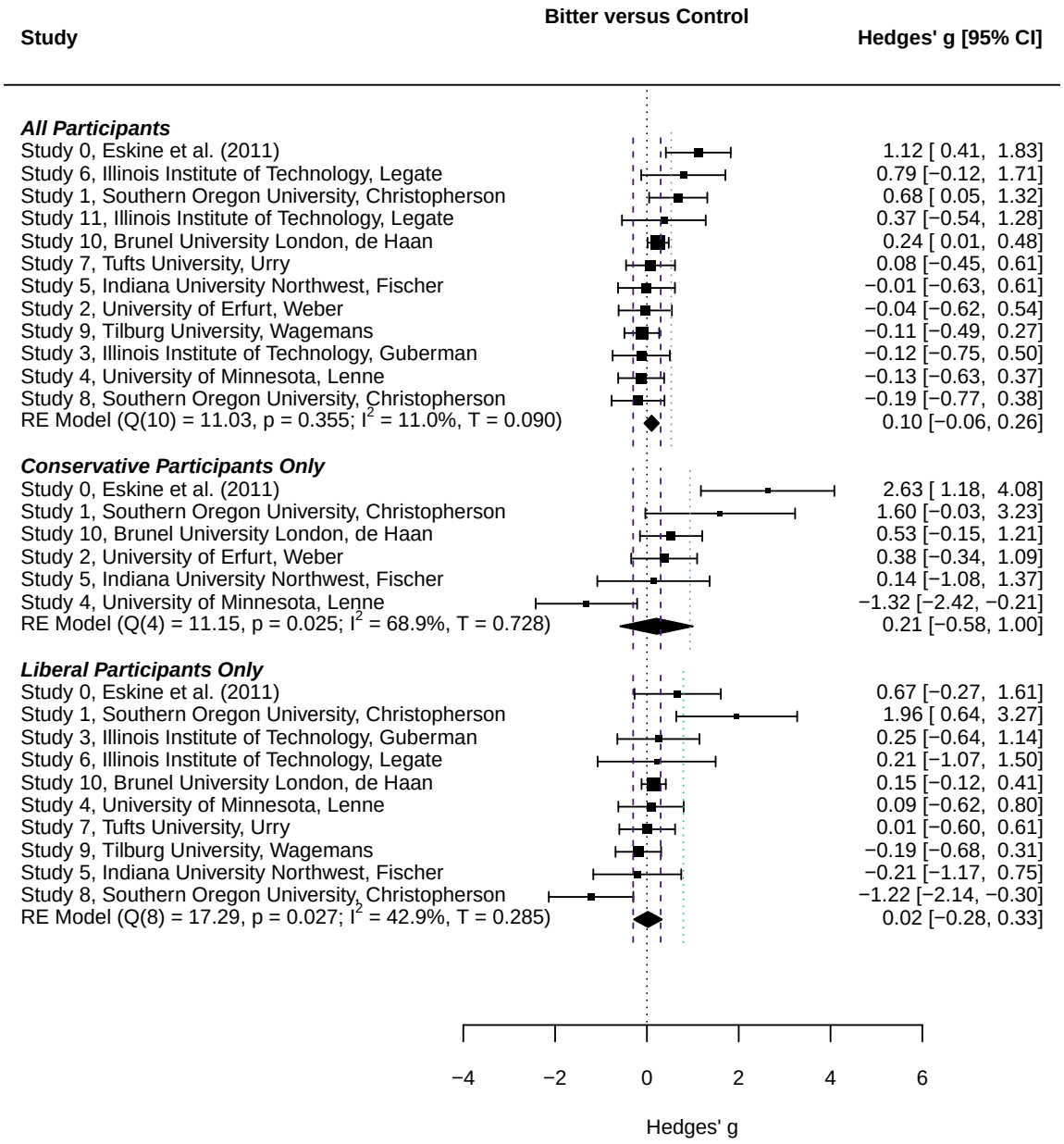

Figure 1. Effect sizes (standardized) for the original Eskine et al. (2011) study (Bitter versus Control) and multiple replications across all participants, and for conservative and liberal subgroups. Within each subgroup, we present studies in descending order by point estimate. Error bars represent 95% confidence intervals. RE Model refers to the overall estimate across replication studies based on a random effects model; it excludes the original effect size. Purple dashed vertical lines indicate +/- 0.30 equivalence bounds. The green dotted vertical line indicates d33%.

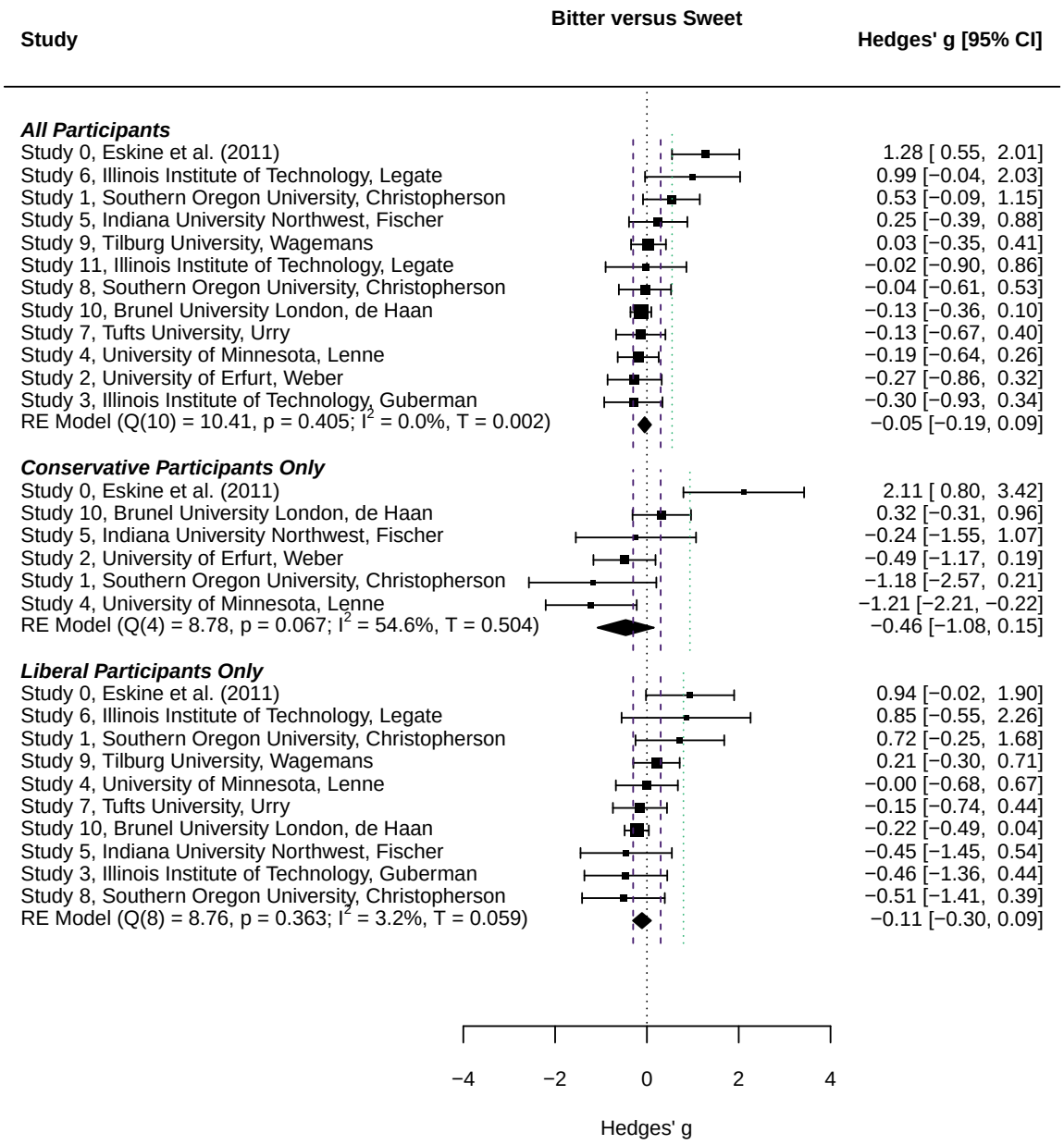

Figure 2. Effect sizes (standardized) for the original Eskine et al. (2011) study (Bitter versus Sweet) and multiple replications across all participants, and for conservative and liberal subgroups. Within each subgroup, we present studies in descending order by point estimate. Error bars represent 95% confidence intervals. RE Model refers to the overall estimate across replication studies based on a random effects model; it excludes the original effect size. Purple dashed vertical lines indicate +/- 0.30 equivalence bounds. The green dotted vertical line indicates d33%.

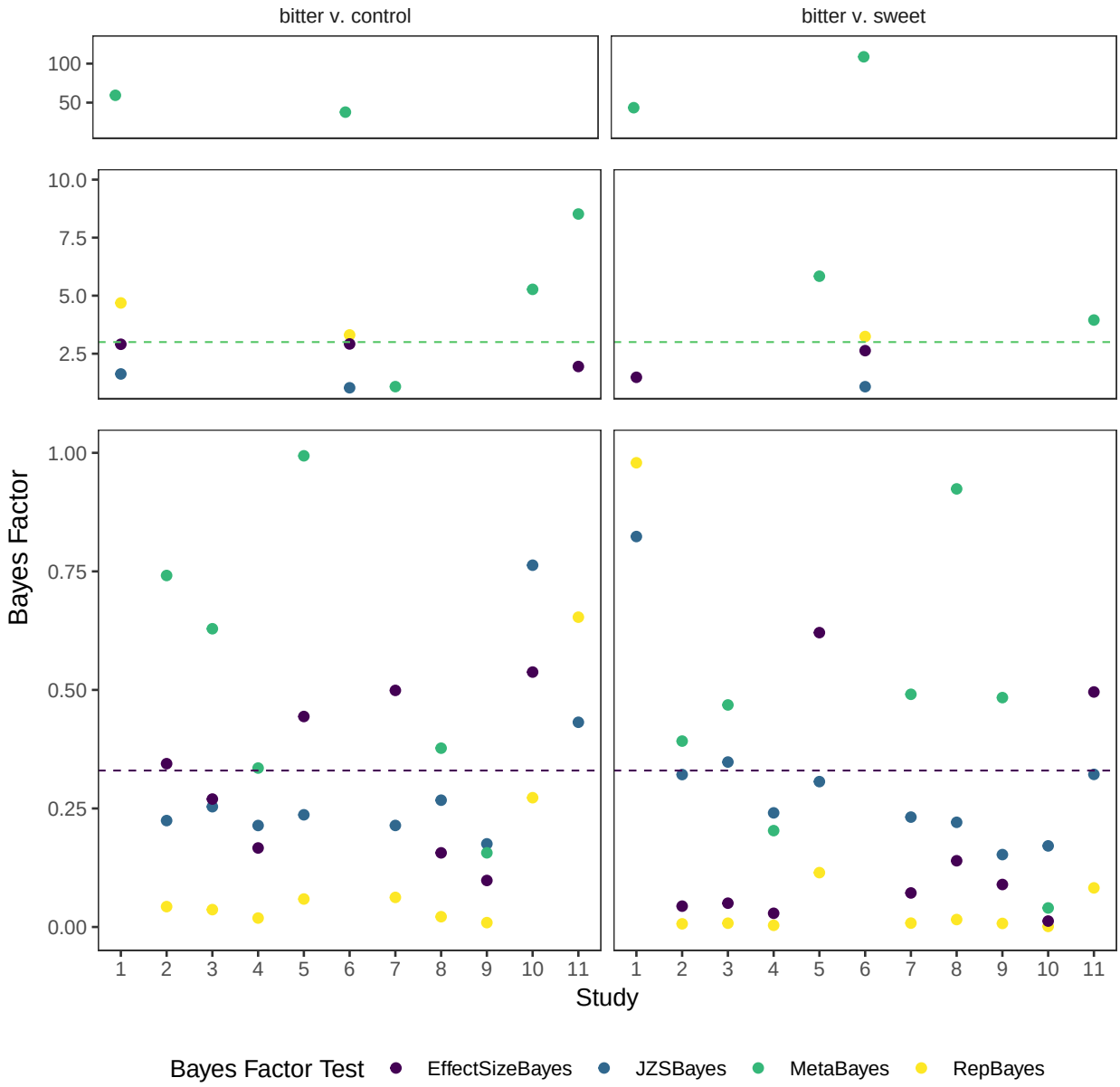

Figure 3. Results of Bayes Factors tests across all participants in each replication study for each contrast, bitter versus control (left) and bitter versus sweet (right). Each color represents one of the four Bayes Factor tests. Horizontal lines mark BF = 3 and 0.33, thresholds above and below which provide nonanecdotal evidence for and against the replication hypothesis, respectively. JZSBayes = Jeffreys-Zellner-Siow (JZS) Bayes Factor; RepBayes = Replication test Bayes Factor; EffectSizeBayes = equality-of-effect-size Bayes Factor; MetaBayes = fixed-effect meta-analysis Bayes Factor. For JZSBayes, RepBayes, and MetaBayes, we report BF_{10}. For EffectSizeBayes, however, we report BF_{01} because replication success is supported by evidence favoring zero variance, the null hypothesis. In all four cases, a larger value constitutes greater relative support for the replication hypothesis.

Appendix

In this appendix, we provide the supplementary material referenced in the main manuscript. At the end of this document, we provide the R session information at the time we generated this document to support analytic reproducibility.

In-Detail Text Box: The Collaborative Replications and Education Project

Each replication presented in this manuscript was conducted as a part of the Collaborative Replications and Education Project (CREP; https://osf.io/wfc6u/), an initiative to involve undergraduate methods and capstone students under faculty supervision in the replication of psychological research. The CREP, like other pedagogical replication projects (e.g., Hawkins et al., 2018), aims to train students on best practices while contributing high-quality replication data to the field. There are a number of quality checks built into the CREP to ensure high-fidelity replications, such as a procedure and materials review by faculty experts before students are approved to begin data collection.

The CREP process incorporates best replication/open practices recommended by scholars (Brandt et al., 2014) and open science organizations (Center for Open Science). Students joining the CREP first request a study from our list of available studies to replicate (selected by the CREP team for impact, feasibility, and potential student interest). On approval, students then build a study page on the OSF, uploading their materials, IRB approval, and a video of their procedure to demonstrate their fidelity to the original procedures. Next, the CREP team (consisting of two to three PhD-level researchers trained on a particular study) reviews the projects on that study.

Following a successful review, students preregister their project and collect the required sample size (usually equivalent to the original study's sample size). Finally, students post a summary of their results on their study page, along with their raw data and codebook. At this point, the CREP team again reviews the project, deciding whether to award a CREP certificate of completion based on attaining the required minimum sample size and carrying out procedures as planned. Additionally, students are encouraged to present their findings at conferences, and be involved with manuscripts for the meta-analytic findings of CREP studies.

Study numbers, sites, and mentors

Table A1 lists labs by study number, site, and mentor.

Descriptive Statistics and Regressions Within Studies: Moral Wrongness Judgments

Table A2 depicts descriptive statistics for the moral wrongness composite across all participants and within the conservative and liberal subgroups in each study. This table also depicts results of linear regression models comparing moral wrongness judgments made by the bitter group to moral wrongness judgments made by each of the other two groups.

Table A1

Study number, site, and mentor for each replication study.

Study	Site and Mentor
1	Southern Oregon University, Christopherson
2	University of Erfurt, Weber
3	Illinois Institute of Technology, Guberman
4	University of Minnesota, Lenne
5	Indiana University Northwest, Fischer
6	Illinois Institute of Technology, Legate
7	Tufts University, Urry
8	Southern Oregon University, Christopherson
9	Tilburg University, Wagemans
10	Brunel University London, de Haan
11	Illinois Institute of Technology, Legate

Note. Some sites conducted more than one replication in separate semesters.

Table A2

Mean (standard deviation) for the moral wrongness composite in each replication study along with results of multiple regressions examining the effect of the bitter versus control and bitter versus sweet contrasts.

	Descriptives			Regression Results	
study	bitter	control	sweet	Bitter versus Control predictor	Bitter versus Sweet predictor
All participants					
1	75.28 (10.92)	65.37 (17.06)	67.94 (15.64)	$t(56) = 1.52$, $p = .134$	$t(56) = 0.59$, $p = .559$
2	67.08 (11.12)	67.52 (12.49)	69.80 (8.59)	$t(62) = 0.32$, $p = .747$	$t(62) = -0.87$, $p = .390$
3	62.26 (17.09)	64.25 (14.59)	66.96 (13.82)	$t(53) = 0.08$, $p = .935$	$t(53) = -0.85$, $p = .399$
4	68.99 (11.93)	70.70 (15.70)	71.31 (12.32)	$t(97) = -0.18$, $p = .857$	$t(97) = -0.53$, $p = .596$
5	76.53 (13.97)	76.65 (11.06)	73.43 (10.74)	$t(56) = -0.52$, $p = .605$	$t(56) = 0.96$, $p = .340$
6	78.22 (11.96)	64.05 (20.92)	60.68 (22.09)	$t(21) = 0.69$, $p = .496$	$t(21) = 1.21$, $p = .241$
7	67.03 (14.27)	65.94 (12.96)	69.05 (15.24)	$t(78) = 0.64$, $p = .525$	$t(78) = -0.77$, $p = .441$
8	70.20 (11.27)	72.55 (12.63)	70.80 (15.90)	$t(65) = -0.59$, $p = .558$	$t(65) = 0.17$, $p = .867$
9	70.20 (12.36)	71.70 (14.36)	69.77 (12.49)	$t(157) = -0.79$, $p = .433$	$t(157) = 0.54$, $p = .590$
10	71.79 (14.65)	68.31 (13.78)	73.69 (14.24)	$t(436) = 3.06$, $p = .002$	$t(436) = -2.54$, $p = .011$
11	70.20 (17.26)	64.76 (8.88)	70.59 (19.38)	$t(23) = 0.82$, $p = .418$	$t(23) = -0.47$, $p = .644$
Conservatives					
1	78.58 (2.57)	63.42 (8.64)	93.18 (9.64)	$t(5) = 3.95$, $p = .011$	$t(5) = -3.33$, $p = .021$
2	66.14 (11.48)	61.90 (10.31)	71.05 (7.90)	$t(43) = 2.06$, $p = .046$	$t(43) = -2.31$, $p = .026$

4	63.97 (6.27)	72.74 (6.17)	74.65 (9.82)	$t(20) = -0.92$, $p = .370$	$t(20) = -1.87$, $p = .077$
5	78.10 (17.51)	75.73 (13.99)	82.80 (0.76)	$t(9) = 0.56$, $p = .586$	$t(9) = -0.53$, $p = .609$
10	76.67 (12.78)	70.25 (10.81)	72.23 (13.90)	$t(50) = 1.11$, $p = .274$	$t(50) = 0.34$, $p = .734$
Liberals					
1	75.32 (12.20)	50.88 (10.46)	65.28 (13.94)	$t(18) = 2.99$, $p = .008$	$t(18) = -0.39$, $p = .703$
3	58.72 (14.30)	55.03 (13.64)	65.90 (16.01)	$t(23) = 1.12$, $p = .276$	$t(23) = -1.38$, $p = .180$
4	70.38 (13.59)	69.05 (16.07)	70.39 (15.49)	$t(42) = 0.26$, $p = .797$	$t(42) = -0.14$, $p = .891$
5	71.74 (12.14)	74.48 (12.41)	76.29 (5.57)	$t(19) = -0.10$, $p = .923$	$t(19) = -0.65$, $p = .521$
6	78.42 (17.77)	73.95 (15.98)	61.36 (1.40)	$t(5) = -0.36$, $p = .730$	$t(5) = 1.20$, $p = .284$
7	65.89 (15.08)	65.80 (12.92)	68.19 (14.72)	$t(60) = 0.32$, $p = .750$	$t(60) = -0.63$, $p = .533$
8	65.53 (6.15)	78.80 (13.42)	70.04 (10.58)	$t(25) = -2.66$, $p = .013$	$t(25) = 0.49$, $p = .631$
9	69.91 (12.82)	72.33 (12.72)	67.25 (12.30)	$t(87) = -1.34$, $p = .184$	$t(87) = 1.36$, $p = .178$
10	70.43 (14.99)	68.30 (13.74)	73.72 (14.55)	$t(329) = 2.25$, $p = .025$	$t(329) = -2.60$, $p = .010$

Descriptive Statistics and Regressions Within Studies: Beverage Ratings

Table A3 depicts descriptive statistics for ratings of the extent to which the beverages were perceived to be bitter, disgusting, neutral, and sweet across all participants and within the conservative and liberal subgroups in each study. This table also depicts results of linear regression models comparing ratings for the bitter group to ratings made by each of the other two groups.

Table A3

Means (standard deviations) for beverage ratings in each replication study.

	Descriptives			Regression Results	
study	bitter	control	sweet	Bitter versus Control predictor	Bitter versus Sweet predictor
All participants: bitter					
1	6.10 (1.45)	1.74 (1.52)	1.90 (1.25)	$t(56) = 5.76$, $p < .001$	$t(56) = 5.21$, $p < .001$
2	6.32 (0.72)	1.27 (0.88)	1.48 (1.08)	$t(62) = 11.11$, $p < .001$	$t(62) = 9.70$, $p < .001$
3	6.53 (0.61)	1.39 (0.83)	1.93 (1.46)	$t(53) = 9.85$, $p < .001$	$t(53) = 6.97$, $p < .001$
4	6.20 (1.16)	2.17 (1.63)	2.06 (1.53)	$t(97) = 5.90$, $p < .001$	$t(97) = 7.11$, $p < .001$
5	6.94 (0.24)	1.36 (1.18)	2.15 (1.69)	$t(56) = 9.61$, $p < .001$	$t(56) = 5.91$, $p < .001$
6	6.78 (0.44)	2.00 (2.00)	1.17 (0.41)	$t(21) = 3.63$, $p = .002$	$t(21) = 5.34$, $p < .001$
7	6.74 (0.45)	1.18 (0.77)	1.56 (1.15)	$t(75) = 14.64$, $p < .001$	$t(75) = 11.75$, $p < .001$
8	6.13 (1.18)	1.59 (1.22)	1.74 (1.10)	$t(65) = 7.76$, $p < .001$	$t(65) = 7.10$, $p < .001$
9	6.20 (1.04)	1.65 (1.34)	1.52 (1.09)	$t(157) = 11.33$, $p < .001$	$t(157) = 12.33$, $p < .001$
10	6.10 (1.46)	1.77 (1.42)	2.61 (1.97)	$t(434) = 15.41$, $p < .001$	$t(434) = 8.01$, $p < .001$
11	6.00 (1.22)	2.50 (2.14)	1.89 (1.05)	$t(23) = 2.24$, $p = .035$	$t(23) = 3.77$, $p = .001$
All participants: disgust					
1	6.10 (1.17)	1.21 (0.92)	2.45 (1.93)	$t(56) = 7.78$, $p < .001$	$t(56) = 3.10$, $p = .003$
2	4.73 (1.64)	1.27 (0.77)	2.29 (1.23)	$t(62) = 6.74$, $p < .001$	$t(62) = 2.13$, $p = .037$
3	5.74 (1.33)	1.21 (0.71)	1.54 (0.98)	$t(53) = 8.29$, $p < .001$	$t(53) = 6.49$, $p < .001$

4	6.10 (1.24)	1.75 (1.39)	1.67 (1.04)	$t(97) = 7.53$, $p < .001$	$t(97) = 8.85$, $p < .001$
5	6.76 (0.56)	1.00 (0.00)	1.80 (1.44)	$t(56) = 13.70$, $p < .001$	$t(56) = 8.49$, $p < .001$
6	5.89 (1.05)	1.67 (1.41)	2.50 (2.07)	$t(21) = 4.00$, $p = .001$	$t(21) = 1.82$, $p = .082$
7	5.70 (1.82)	1.07 (0.26)	1.70 (1.23)	$t(76) = 9.18$, $p < .001$	$t(76) = 5.76$, $p < .001$
8	6.39 (0.89)	1.18 (0.66)	1.87 (1.49)	$t(65) = 10.56$, $p < .001$	$t(65) = 6.95$, $p < .001$
9	5.76 (1.46)	1.24 (0.69)	1.81 (1.27)	$t(157) = 12.98$, $p < .001$	$t(157) = 8.53$, $p < .001$
10	6.13 (1.51)	1.74 (1.50)	2.29 (1.93)	$t(434) = 14.55$, $p < .001$	$t(434) = 9.80$, $p < .001$
11	5.22 (1.99)	1.38 (0.52)	1.78 (0.97)	$t(23) = 3.75$, $p = .001$	$t(23) = 2.76$, $p = .011$
All participants: neutral					
1	1.40 (1.19)	6.37 (1.30)	2.50 (1.91)	$t(56) = -10.56$, $p < .001$	$t(56) = 3.35$, $p = .001$
2	1.27 (0.63)	5.91 (1.48)	1.76 (0.94)	$t(62) = -15.55$, $p < .001$	$t(62) = 6.40$, $p < .001$
3	1.74 (0.87)	5.42 (2.27)	3.17 (1.64)	$t(53) = -6.21$, $p < .001$	$t(53) = 0.85$, $p = .399$
4	1.62 (1.23)	5.54 (1.79)	2.14 (1.07)	$t(97) = -11.68$, $p < .001$	$t(97) = 5.12$, $p < .001$
5	1.71 (1.49)	5.45 (1.77)	3.15 (2.21)	$t(56) = -6.04$, $p < .001$	$t(56) = 0.84$, $p = .405$
6	1.11 (0.33)	4.78 (1.99)	2.17 (1.17)	$t(21) = -5.40$, $p < .001$	$t(21) = 1.21$, $p = .241$
7	1.50 (1.14)	6.64 (1.04)	2.15 (1.19)	$t(70) = -17.32$, $p < .001$	$t(70) = 6.96$, $p < .001$
8	1.43 (0.79)	5.77 (1.72)	2.48 (1.70)	$t(65) = -10.06$, $p < .001$	$t(65) = 3.00$, $p = .004$
9	1.24 (0.59)	6.05 (1.54)	1.67 (1.06)	$t(157) = -24.14$, $p < .001$	$t(157) = 10.33$, $p < .001$
10	1.47 (1.10)	5.42 (1.89)	2.16 (1.47)	$t(432) = -23.24$, $p < .001$	$t(432) = 8.39$, $p < .001$
11	1.78 (1.09)	5.25 (1.83)	2.44 (1.59)	$t(23) = -4.85$, $p < .001$	$t(23) = 1.70$, $p = .102$
All participants: sweet					

1	1.40 (1.10)	1.84 (1.34)	5.85 (1.04)	$t(56) = 5.50$, $p < .001$	$t(56) = -13.21$, $p < .001$
2	1.18 (0.39)	1.45 (1.06)	6.19 (0.75)	$t(62) = 10.88$, $p < .001$	$t(62) = -23.48$, $p < .001$
3	1.32 (0.82)	1.50 (1.19)	5.29 (1.47)	$t(53) = 5.39$, $p < .001$	$t(53) = -11.46$, $p < .001$
4	1.15 (0.36)	2.04 (1.49)	6.03 (0.77)	$t(97) = 7.41$, $p < .001$	$t(97) = -23.59$, $p < .001$
5	1.00 (0.00)	1.14 (0.64)	5.45 (1.67)	$t(56) = 7.39$, $p < .001$	$t(56) = -15.16$, $p < .001$
6	1.00 (0.00)	2.56 (1.74)	6.00 (1.10)	$t(21) = 1.85$, $p = .078$	$t(21) = -7.47$, $p < .001$
7	1.14 (0.35)	1.45 (0.78)	5.93 (0.62)	$t(75) = 14.12$, $p < .001$	$t(75) = -30.88$, $p < .001$
8	1.09 (0.29)	1.91 (1.57)	5.91 (1.04)	$t(65) = 5.62$, $p < .001$	$t(65) = -15.78$, $p < .001$
9	1.47 (0.90)	1.75 (1.27)	5.81 (0.99)	$t(157) = 10.66$, $p < .001$	$t(157) = -23.54$, $p < .001$
10	1.86 (1.50)	1.96 (1.45)	4.97 (1.76)	$t(435) = 9.03$, $p < .001$	$t(435) = -19.21$, $p < .001$
11	1.11 (0.33)	1.62 (1.06)	5.78 (0.67)	$t(23) = 5.85$, $p < .001$	$t(23) = -14.61$, $p < .001$
Conservatives: bitter					
1	6.50 (0.71)	1.75 (1.50)	1.50 (0.71)	$t(5) = 2.56$, $p = .051$	$t(5) = 2.54$, $p = .052$
2	6.38 (0.72)	1.23 (0.83)	1.53 (1.18)	$t(43) = 8.81$, $p < .001$	$t(43) = 7.87$, $p < .001$
4	5.50 (1.41)	2.33 (2.16)	2.11 (1.27)	$t(20) = 1.96$, $p = .065$	$t(20) = 2.66$, $p = .015$
5	7.00 (0.00)	1.57 (1.51)	1.00 (0.00)	$t(9) = 3.32$, $p = .009$	$t(9) = 3.38$, $p = .008$
10	5.94 (1.52)	2.12 (1.63)	2.30 (1.84)	$t(50) = 3.97$, $p < .001$	$t(50) = 3.64$, $p = .001$
Conservatives: disgust					
1	7.00 (0.00)	1.00 (0.00)	4.50 (0.71)	$t(5) = 21.24$, $p < .001$	$t(5) = -1.91$, $p = .115$
2	4.69 (1.66)	1.23 (0.60)	2.18 (1.24)	$t(43) = 5.26$, $p < .001$	$t(43) = 2.00$, $p = .052$
4	6.62 (0.52)	1.83 (1.33)	1.67 (0.87)	$t(20) = 5.33$, $p < .001$	$t(20) = 6.54$, $p < .001$

5	7.00 (0.00)	1.00 (0.00)	1.00 (0.00)	undefined	undefined
10	6.12 (1.27)	1.88 (1.67)	1.50 (1.10)	$t(50) = 4.80$, $p < .001$	$t(50) = 6.55$, $p < .001$
Conservatives: neutral					
1	2.00 (1.41)	6.50 (1.00)	2.50 (2.12)	$t(5) = -4.36$, $p = .007$	$t(5) = 1.53$, $p = .186$
2	1.31 (0.70)	5.62 (1.56)	1.65 (0.79)	$t(43) = -12.16$, $p < .001$	$t(43) = 5.71$, $p < .001$
4	1.88 (1.81)	4.67 (2.25)	2.00 (0.71)	$t(20) = -3.56$, $p = .002$	$t(20) = 1.83$, $p = .082$
5	1.00 (0.00)	4.43 (2.07)	4.00 (4.24)	$t(9) = -1.48$, $p = .174$	$t(9) = -0.74$, $p = .477$
10	1.41 (0.71)	5.50 (2.03)	2.32 (1.34)	$t(49) = -8.37$, $p < .001$	$t(49) = 2.74$, $p = .009$
Conservatives: sweet					
1	1.00 (0.00)	2.50 (1.73)	6.50 (0.71)	$t(5) = 1.28$, $p = .256$	$t(5) = -4.16$, $p = .009$
2	1.25 (0.45)	1.46 (1.13)	6.24 (0.75)	$t(43) = 8.74$, $p < .001$	$t(43) = -20.01$, $p < .001$
4	1.12 (0.35)	1.83 (1.33)	5.67 (0.71)	$t(20) = 3.97$, $p = .001$	$t(20) = -11.79$, $p < .001$
5	1.00 (0.00)	1.00 (0.00)	3.50 (3.54)	$t(9) = 1.79$, $p = .107$	$t(9) = -2.70$, $p = .025$
10	1.76 (1.20)	1.81 (1.33)	5.50 (0.95)	$t(50) = 5.27$, $p < .001$	$t(50) = -11.35$, $p < .001$
Liberals: bitter					
1	5.86 (2.04)	1.60 (1.34)	1.44 (1.01)	$t(18) = 2.67$, $p = .016$	$t(18) = 3.44$, $p = .003$
3	6.42 (0.67)	1.36 (0.48)	1.50 (0.50)	$t(23) = 10.02$, $p < .001$	$t(23) = 9.19$, $p < .001$
4	6.57 (0.75)	1.91 (1.45)	2.31 (1.89)	$t(42) = 5.43$, $p < .001$	$t(42) = 4.34$, $p < .001$
5	7.00 (0.00)	1.50 (1.41)	1.57 (1.13)	$t(19) = 5.88$, $p < .001$	$t(19) = 5.47$, $p < .001$
6	7.00 (0.00)	1.00 (0.00)	1.00 (0.00)	undefined	undefined
7	6.70 (0.47)	1.25 (0.91)	1.65 (1.23)	$t(60) = 11.48$, $p < .001$	$t(60) = 9.43$, $p < .001$

8	6.30 (0.82)	1.40 (1.26)	1.50 (0.76)	$t(25) = 6.39$, $p < .001$	$t(25) = 5.67$, $p < .001$
9	6.17 (1.02)	1.71 (1.47)	1.59 (1.35)	$t(87) = 7.55$, $p < .001$	$t(87) = 8.06$, $p < .001$
10	6.17 (1.38)	1.72 (1.40)	2.64 (1.99)	$t(328) = 14.28$, $p < .001$	$t(328) = 6.94$, $p < .001$
Liberals: disgust					
1	6.29 (0.49)	1.00 (0.00)	2.22 (2.05)	$t(18) = 4.55$, $p < .001$	$t(18) = 2.30$, $p = .034$
3	5.75 (1.29)	1.57 (1.13)	1.29 (0.49)	$t(23) = 3.99$, $p = .001$	$t(23) = 4.87$, $p < .001$
4	6.10 (1.41)	1.36 (0.92)	1.92 (1.32)	$t(42) = 5.90$, $p < .001$	$t(42) = 4.21$, $p < .001$
5	6.57 (0.79)	1.00 (0.00)	1.86 (1.46)	$t(19) = 7.77$, $p < .001$	$t(19) = 4.51$, $p < .001$
6	5.33 (1.53)	2.33 (2.31)	1.50 (0.71)	$t(5) = 0.83$, $p = .446$	$t(5) = 1.61$, $p = .169$
7	5.85 (1.60)	1.10 (0.31)	1.86 (1.32)	$t(59) = 8.36$, $p < .001$	$t(59) = 5.00$, $p < .001$
8	6.50 (0.53)	1.00 (0.00)	2.38 (2.00)	$t(25) = 7.89$, $p < .001$	$t(25) = 2.98$, $p = .006$
9	5.70 (1.64)	1.32 (0.75)	1.66 (1.04)	$t(87) = 8.83$, $p < .001$	$t(87) = 6.85$, $p < .001$
10	6.25 (1.35)	1.77 (1.54)	2.30 (1.89)	$t(328) = 13.36$, $p < .001$	$t(328) = 9.13$, $p < .001$
Liberals: neutral					
1	1.71 (1.89)	7.00 (0.00)	1.67 (1.00)	$t(18) = -8.09$, $p < .001$	$t(18) = 4.74$, $p < .001$
3	1.50 (0.67)	5.43 (2.07)	2.50 (1.71)	$t(23) = -5.30$, $p < .001$	$t(23) = 1.49$, $p = .149$
4	1.67 (1.28)	6.18 (1.60)	2.31 (0.95)	$t(42) = -9.36$, $p < .001$	$t(42) = 3.77$, $p < .001$
5	2.14 (2.04)	6.12 (1.36)	2.86 (2.19)	$t(19) = -4.37$, $p < .001$	$t(19) = 1.49$, $p = .153$
6	1.33 (0.58)	3.67 (2.08)	1.00 (0.00)	$t(5) = -2.49$, $p = .055$	$t(5) = 1.34$, $p = .237$
7	1.58 (1.22)	6.44 (1.26)	2.10 (1.26)	$t(53) = -12.47$, $p < .001$	$t(53) = 5.55$, $p < .001$
8	1.40 (0.84)	5.50 (2.01)	2.88 (1.96)	$t(25) = -5.09$, $p < .001$	$t(25) = 0.82$, $p = .418$

9	1.30 (0.70)	5.90 (1.87)	1.86 (1.16)	$t(87) = -14.53$, $p < .001$	$t(87) = 5.75$, $p < .001$
10	1.47 (1.09)	5.42 (1.88)	2.11 (1.51)	$t(327) = -20.27$, $p < .001$	$t(327) = 7.55$, $p < .001$
Liberals: sweet					
1	1.14 (0.38)	1.60 (1.34)	6.33 (0.71)	$t(18) = 5.09$, $p < .001$	$t(18) = -13.67$, $p < .001$
3	1.42 (1.00)	1.50 (1.32)	5.93 (0.84)	$t(23) = 4.61$, $p < .001$	$t(23) = -9.48$, $p < .001$
4	1.19 (0.40)	1.91 (1.64)	6.31 (0.75)	$t(42) = 5.62$, $p < .001$	$t(42) = -15.19$, $p < .001$
5	1.00 (0.00)	1.38 (1.06)	5.86 (1.21)	$t(19) = 4.94$, $p < .001$	$t(19) = -10.86$, $p < .001$
6	1.00 (0.00)	3.67 (2.31)	7.00 (0.00)	$t(5) = 0.31$, $p = .769$	$t(5) = -3.91$, $p = .011$
7	1.16 (0.37)	1.40 (0.68)	6.00 (0.62)	$t(58) = 13.83$, $p < .001$	$t(58) = -30.67$, $p < .001$
8	1.10 (0.32)	1.60 (1.26)	5.88 (1.36)	$t(25) = 4.50$, $p < .001$	$t(25) = -10.19$, $p < .001$
9	1.43 (0.86)	1.77 (1.33)	5.62 (1.15)	$t(87) = 6.98$, $p < .001$	$t(87) = -15.72$, $p < .001$
10	1.82 (1.48)	1.96 (1.48)	4.90 (1.82)	$t(329) = 7.51$, $p < .001$	$t(329) = -16.22$, $p < .001$

Note. T-test results are reported as "undefined" when there was no variability in ratings for all three beverage groups.

Table A4 shows the result of linear mixed-effect regressions examining beverage ratings across studies.

Knowledge of the Hypothesis

Figure A1 shows predicted probabilities of being partially or fully suspicious as a function of both beverage type and political orientation obtained in a generalized linear mixed-effect logistic regression.

Random Effects Meta-Analyses of Raw Mean Differences in Moral Wrongness

Figures A2 and A3 display raw mean differences and 95% confidence intervals for each contrast within and across studies across all participants and for conservative and liberal subgroups. These figures also report estimates of heterogeneity (Cochrane's Q, I^2, and Tau). The pattern is similar to that shown with standardized effect sizes. For both contrasts and all subgroups, overall raw mean differences are small and confidence intervals include zero.

Figures Depicting Moral Wrongness in all Design Cells

Figure A4 plots mean moral wrongness across vignettes for each participant by beverage type (top panel) and by beverage type and political orientation (bottom panel). The color of each data point represents study.

Figure A5 plots mean moral wrongness across vignettes for each participant by beverage type, political orientation, and level of knowledge of the hypothesis.

In both figures, the black vertical lines represent +/- 1 SE around the predicted estimate (black diamond) after accounting for fixed and random sources of variation in LMER Model 1 or 2, respectively.

Table A4

Linear mixed-effects models examining beverage ratings

	bitter	disgust	neutral	sweet
	(1)	(2)	(3)	(4)
(Intercept)	3.276*** (0.065)	3.048*** (0.076)	3.134*** (0.068)	2.910*** (0.045)
bitter - control (BvC)	5.808*** (0.139)	5.584*** (0.139)	−3.316*** (0.143)	−2.941*** (0.128)
sweet - control (SvC)	−2.600*** (0.136)	−2.204*** (0.136)	−1.718*** (0.140)	5.283*** (0.125)
conservative - liberal (CvL)	−0.085 (0.122)	−0.164 (0.124)	−0.030 (0.126)	−0.034 (0.107)
conservative - other (CvO)	−0.003 (0.137)	0.106 (0.139)	−0.233 (0.141)	−0.031 (0.122)
BvC × CvL	−0.087 (0.331)	−0.296 (0.332)	−0.215 (0.341)	−0.294 (0.305)
BvC × CvO	−0.421 (0.327)	0.234 (0.327)	0.503 (0.337)	0.674* (0.301)
SvC × CvL	0.092 (0.378)	0.348 (0.379)	0.657 (0.389)	0.157 (0.348)
SvC × CvO	−0.153 (0.372)	−0.439 (0.373)	−0.042 (0.383)	−0.173 (0.342)
Observations	1,132	1,133	1,125	1,133
Log Likelihood	−1,996.329	−2,002.325	−2,015.673	−1,904.810
Akaike Inf. Crit.	4,014.658	4,026.649	4,053.347	3,831.621
Bayesian Inf. Crit.	4,070.008	4,082.008	4,108.628	3,886.980

Note: *p<0.05; **p<0.01; ***p<0.001

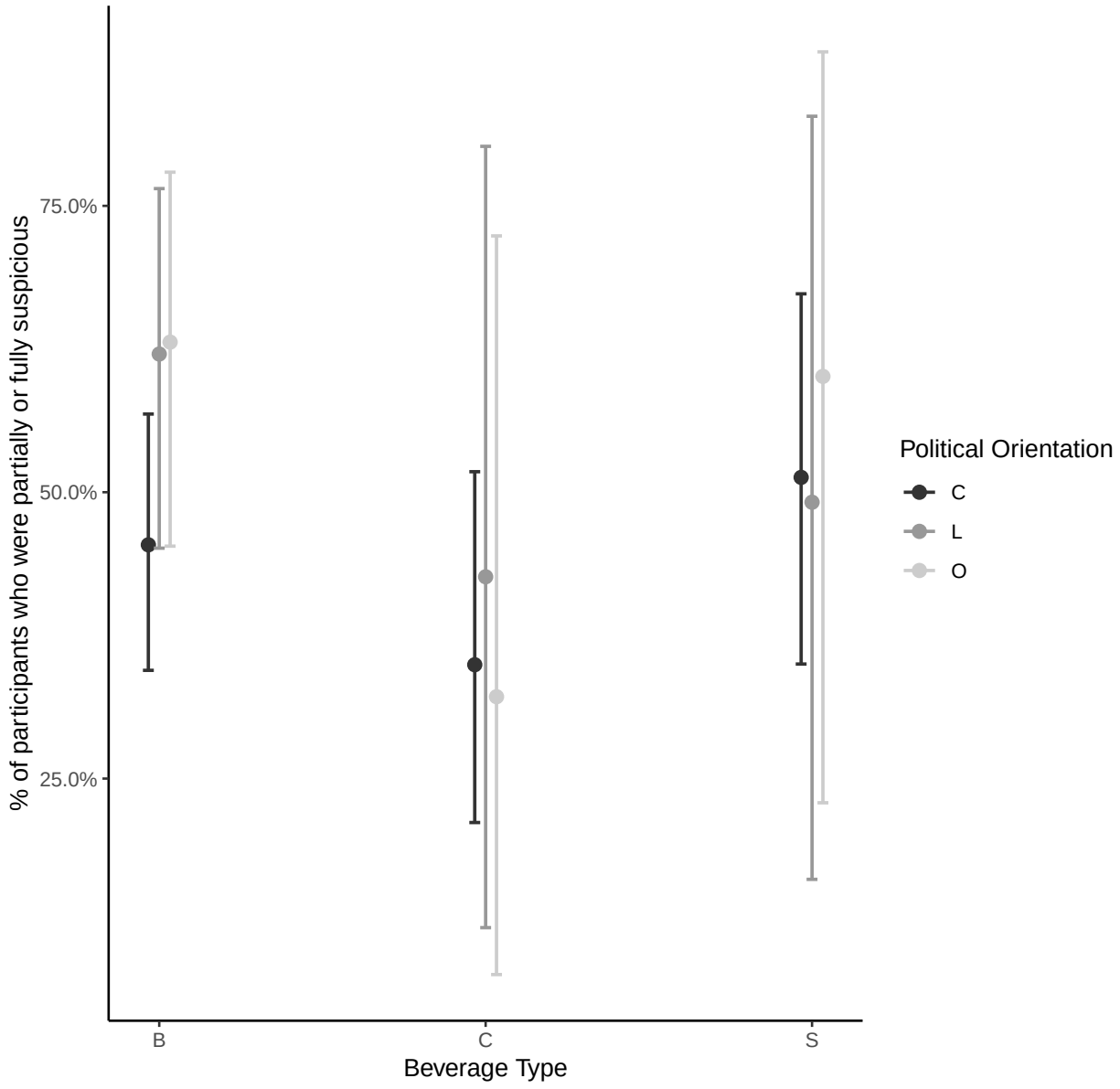

Figure A1. Predicted probabilities of being partially or fully suspicious by beverage type and political orientation obtained in a generalized linear mixed-effect logistic regression. For beverage type, B = bitter, C = control, S = sweet. For political orientation, C = conservative, L = liberal, O = other. Error bars represent the 95% confidence interval.

One-sided Tests: Small Telescopes Within Studies

For each study, we computed one-sided tests to determine whether the effect for each of the two contrasts of interest (bitter versus control and bitter versus sweet)

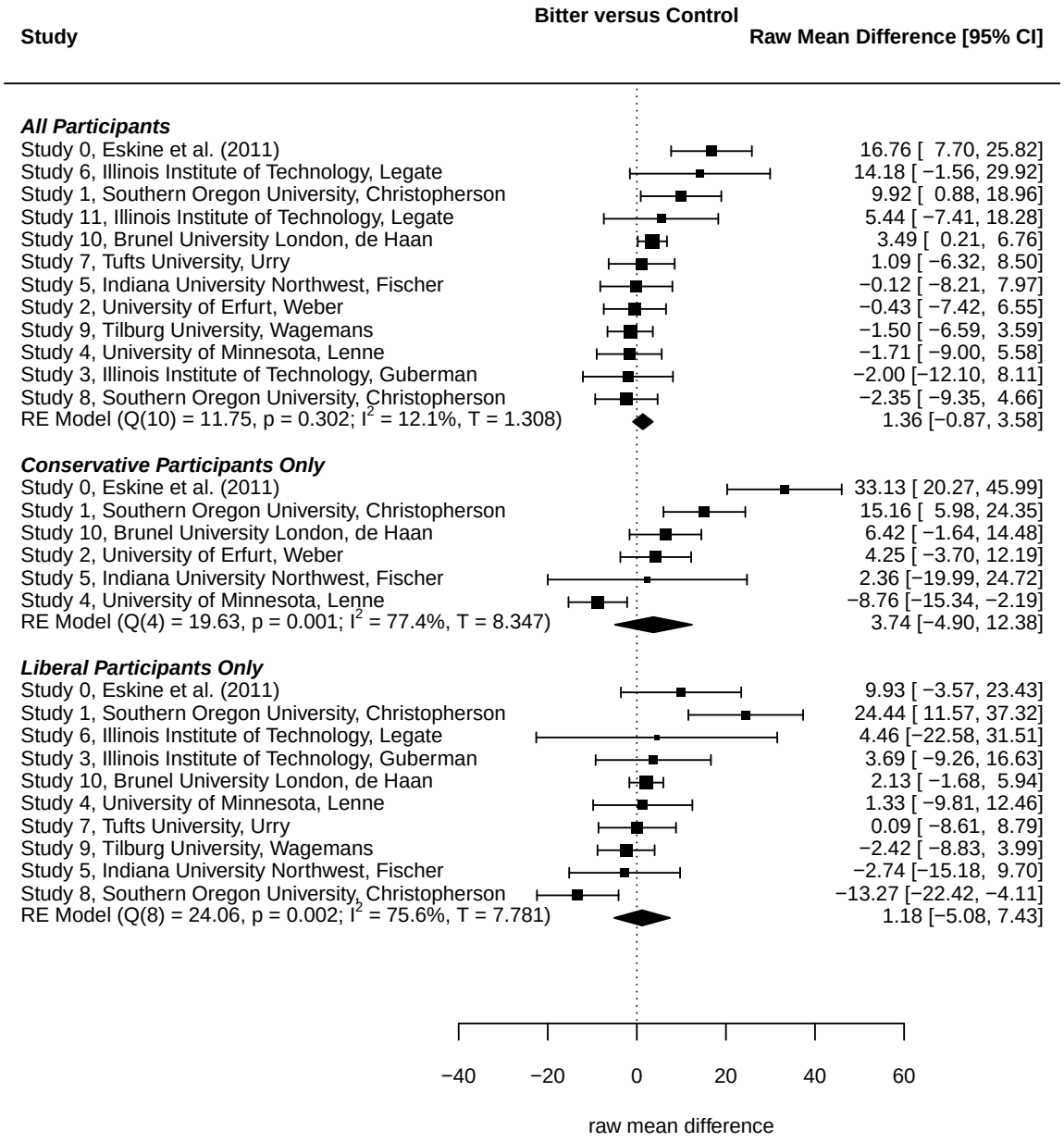

Figure A2. Effect sizes (raw mean difference) for the original Eskine et al. (2011) study (Bitter versus Control) and multiple replications across all participants, and for conservative and liberal participants subgroups. Within each subgroup, we present studies in descending order by point estimate. Error bars represent 95% confidence intervals. RE Model refers to the overall estimate across replication studies based on a random effects model; it excludes the original effect size.

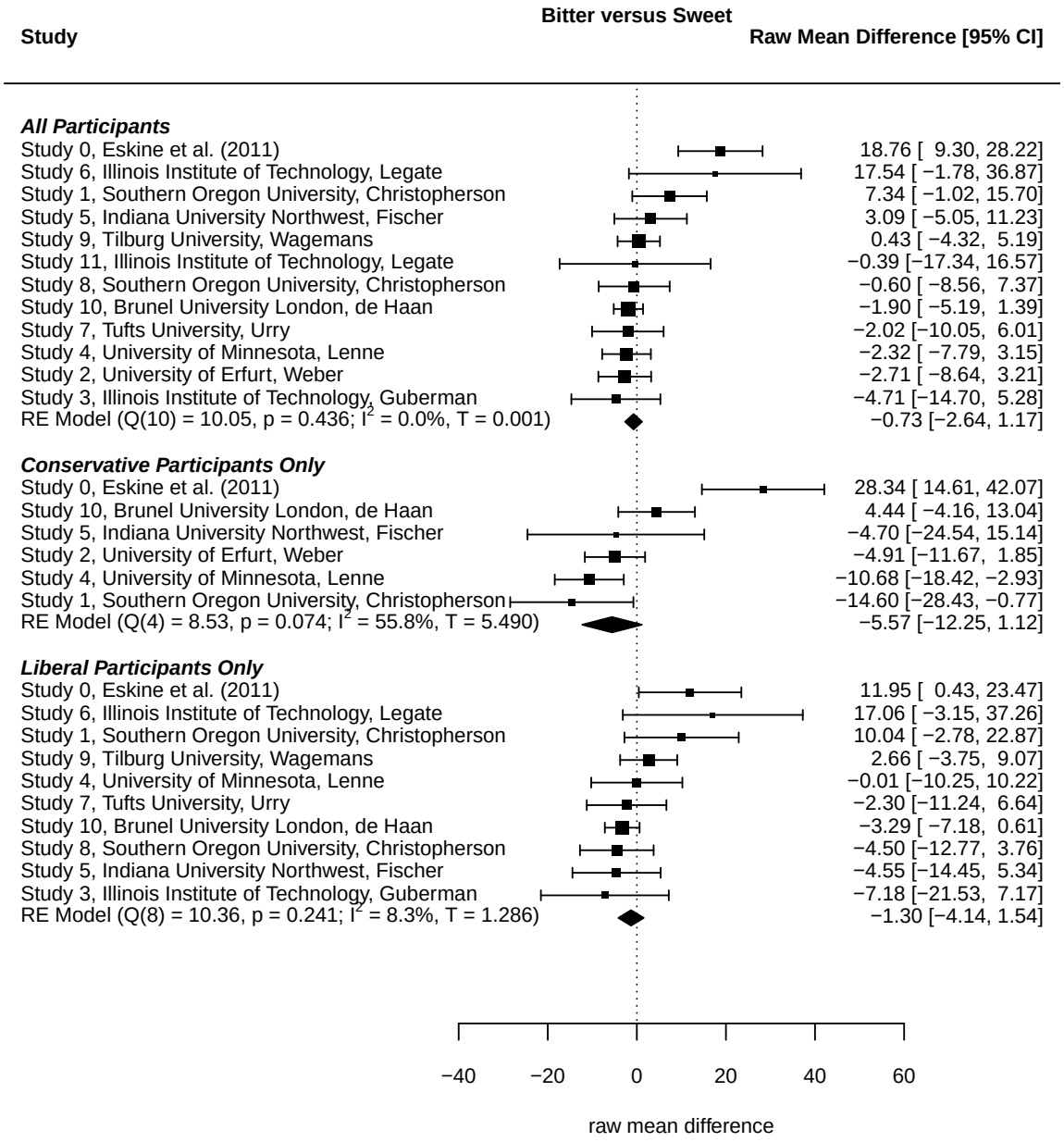

Figure A3. Effect sizes (raw mean difference) for the original Eskine et al. (2011) study (Bitter versus Sweet) and multiple replications across all participants, and for conservative and liberal participants subgroups. Within each subgroup, we present studies in descending order by point estimate. Error bars represent 95% confidence intervals. RE Model refers to the overall estimate across replication studies based on a random effects model; it excludes the original effect size.

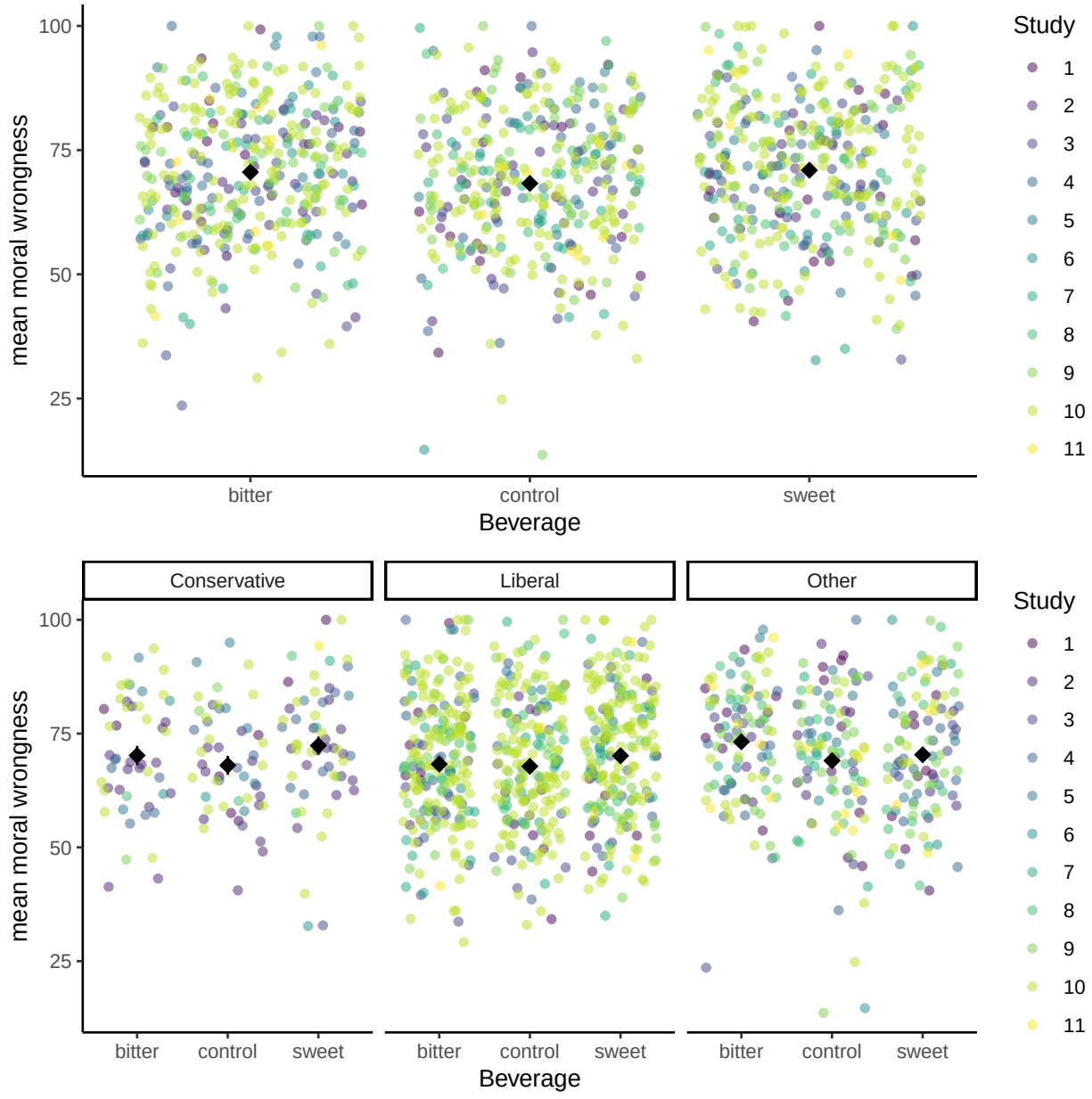

Figure A4. This figure shows mean moral wrongness (y axis) for each participant as a function of beverage condition (x axis) across (top panel) and within political orientations (bottom panel). The black vertical lines represent +/- 1 SE around the predicted estimate (black diamond) after accounting for all modeled sources of variation in Model 1. The color of each data point represents study.

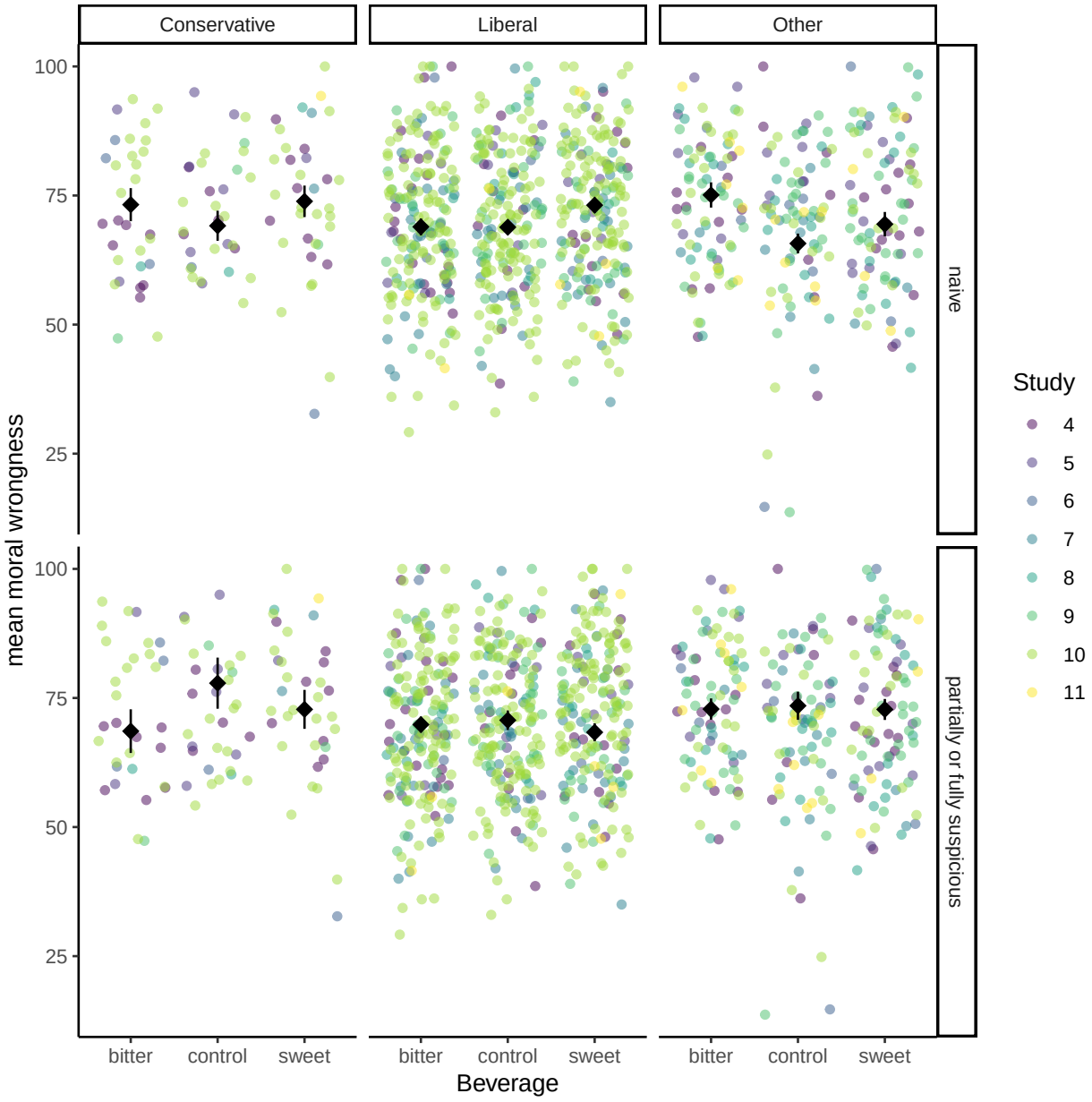

Figure A5. This figure shows mean moral wrongness (y axis) for each participant as a function of beverage condition (x axis), political orientation, and level of knowledge of the hypthesis. The black vertical lines represent +/- 1 SE around the predicted mean value (black diamond) after accounting for all modeled sources of variation in Model 2. The color of each data point represents study.

was significantly smaller than the effect size that Eskine et al. (2011) had 33% power to detect (d33%). Table A5 presents these results, which we summarize below.[11]

Table A5

Results of one-sided tests of the moral wrongness composite in each replication study: Is the observed effect size smaller than $d_{33\%}$?

study	Bitter versus Control	Bitter versus Sweet
All participants		
1	$t(30.39) = 0.51$, $p = 0.693$	$t(33.97) = -0.01$, $p = 0.495$
2	$t(41.44) = -1.87$, $p = 0.034$	$t(39.33) = -2.7$, $p = 0.005$
3	$t(35.13) = -2.02$, $p = 0.026$	$t(34.18) = -2.59$, $p = 0.007$
4	$t(38.95) = -2.44$, $p = 0.01$	$t(72.6) = -3.22$, $p = < .001$
5	$t(29.9) = -1.64$, $p = 0.056$	$t(29.78) = -0.9$, $p = 0.188$
6	$t(12.72) = 0.64$, $p = 0.734$	$t(6.98) = 0.79$, $p = 0.773$
7	$t(47.09) = -1.62$, $p = 0.056$	$t(49.58) = -2.47$, $p = 0.009$
8	$t(41.95) = -2.43$, $p = 0.01$	$t(39.66) = -2.01$, $p = 0.026$
9	$t(103.44) = -3.3$, $p = < .001$	$t(102.77) = -2.63$, $p = 0.005$
10	$t(286.81) = -2.41$, $p = 0.008$	$t(294.31) = -5.86$, $p = < .001$
11	$t(12.23) = -0.28$, $p = 0.393$	$t(15.79) = -1.21$, $p = 0.123$
Conservatives		
1	$t(3.8) = 1.96$, $p = 0.937$	$t(1.14) = -3.01$, $p = 0.089$
2	$t(26.69) = -1.48$, $p = 0.076$	$t(26.42) = -4.1$, $p = < .001$
4	$t(11.01) = -4.35$, $p = < .001$	$t(13.72) = -4.66$, $p = < .001$
5	$t(3.17) = -1.1$, $p = 0.175$	$t(2.01) = -1.61$, $p = 0.124$

[11] Across all participants, Cohen's d33% = 0.53 for the bitter versus control contrast, and Cohen's d33% = 0.55 for the bitter versus sweet contrast. For conservative participants, Cohen's d33% = 0.94 for the bitter versus control contrast, and Cohen's d33% = 0.94 for the bitter versus sweet contrast. For liberal participants, Cohen's d33% = 0.80 for the bitter versus control contrast, and Cohen's d33% = 0.80 for the bitter versus sweet contrast.

10	$t(30.67) = -1.14, p = 0.132$	$t(34.76) = -1.84, p = 0.037$
Liberals		
1	$t(9.54) = 2.34, p = 0.979$	$t(13.74) = -0.06, p = 0.477$
3	$t(13.2) = -1.13, p = 0.14$	$t(11.52) = -2.63, p = 0.011$
4	$t(17.66) = -1.85, p = 0.04$	$t(23.03) = -2.23, p = 0.018$
5	$t(12.81) = -1.97, p = 0.035$	$t(8.42) = -2.39, p = 0.021$
6	$t(3.96) = -0.65, p = 0.275$	$t(2.04) = 0.68, p = 0.717$
7	$t(37.13) = -2.5, p = 0.009$	$t(39.89) = -3.11, p = 0.002$
8	$t(12.62) = -4.62, p = < .001$	$t(10.69) = -2.7, p = 0.011$
9	$t(58.9) = -3.85, p = < .001$	$t(57) = -2.25, p = 0.014$
10	$t(214.2) = -4.79, p = < .001$	$t(217.75) = -7.57, p = < .001$

All Participants.

Bitter versus Control.

For the bitter versus control contrast, there were 11 studies to consider. Of those, 6 suggested that the effect in the replication was significantly smaller than d33% with alpha = .05. Of these, 1 survived Bonferroni correction with alpha = .05 / 11 = .005.

Bitter versus Sweet.

For the bitter versus sweet contrast among all participants, there were 11 studies to consider. Of those, 7 suggested that the effect in the replication was significantly smaller than d33% with alpha = .05. Of these, 2 survived Bonferroni correction with alpha = .05 / 11 = .005.

Conservative Participants.

Bitter versus Control.

For the bitter versus control contrast among conservative participants, there were 5 studies to consider. Of those, 1 suggested that the effect in the replication was significantly smaller than d33% with alpha = .05. Of these, 1 survived Bonferroni correction with alpha = .05 / 5 = .010.

Bitter versus Sweet.

For the bitter versus sweet contrast among conservative participants, there were 5 studies to consider. Of those, 3 suggested that the effect in the replication was significantly smaller than d33% with alpha = .05. Of these, 2 survived Bonferroni correction with alpha = .05 / 5 = .010.

Liberal Participants.

Bitter versus Control.

For the bitter versus control contrast among liberal participants, there were 9 studies to consider. Of those, 6 suggested that the effect in the replication was significantly smaller than d33% with alpha = .05. Of these, 3 survived Bonferroni correction with alpha = .05 / 9 = .006.

Bitter versus Sweet.

For the bitter versus sweet contrast among liberal participants, there were 9 studies to consider. Of those, 7 suggested that the effect in the replication was significantly smaller than d33% with alpha = .05. Of these, 2 survived Bonferroni correction with alpha = .05 / 9 = .006.

One-sided Tests: Equivalence Within Studies

We next determined whether the observed effect in each study was equivalent to 0 +/- d33%. Table A6 presents these results; we summarize them below.

Table A6

Results of one-sided tests of the moral wrongness composite in each replication study: Is the observed effect size equivalent to 0 +/- $d_{33\%}$?

study	Bitter versus Control	Bitter versus Sweet
All participants		
1	$t(30.39) = 0.51$, $p = 0.693$	$t(33.97) = -0.01$, $p = 0.495$
2	$t(41.44) = 1.63$, $p = 0.055$	$t(39.33) = 0.9$, $p = 0.186$
3	$t(35.13) = 1.24$, $p = 0.111$	$t(34.18) = 0.75$, $p = 0.23$
4	$t(38.95) = 1.52$, $p = 0.068$	$t(72.6) = 1.55$, $p = 0.063$
5	$t(29.9) = 1.58$, $p = 0.062$	$t(29.78) = -0.9$, $p = 0.188$
6	$t(12.72) = 0.64$, $p = 0.734$	$t(6.98) = 0.79$, $p = 0.773$
7	$t(47.09) = -1.62$, $p = 0.056$	$t(49.58) = 1.48$, $p = 0.072$
8	$t(41.95) = 1.11$, $p = 0.136$	$t(39.66) = 1.71$, $p = 0.047$
9	$t(103.44) = 2.15$, $p = 0.017$	$t(102.77) = -2.63$, $p = 0.005$
10	$t(286.81) = -2.41$, $p = 0.008$	$t(294.31) = 3.59$, $p = < .001$
11	$t(12.23) = -0.28$, $p = 0.393$	$t(15.79) = 1.12$, $p = 0.14$
Conservatives		
1	$t(3.8) = 1.96$, $p = 0.937$	$t(1.14) = -1.13$, $p = 0.779$
2	$t(26.69) = -1.48$, $p = 0.076$	$t(26.42) = 1.25$, $p = 0.11$
4	$t(11.01) = -0.87$, $p = 0.8$	$t(13.72) = -0.75$, $p = 0.766$
5	$t(3.17) = -1.1$, $p = 0.175$	$t(2.01) = 0.68$, $p = 0.282$
10	$t(30.67) = -1.14$, $p = 0.132$	$t(34.76) = -1.84$, $p = 0.037$
Liberals		
1	$t(9.54) = 2.34$, $p = 0.979$	$t(13.74) = -0.06$, $p = 0.477$
3	$t(13.2) = -1.13$, $p = 0.14$	$t(11.52) = 0.67$, $p = 0.258$
4	$t(17.66) = -1.85$, $p = 0.04$	$t(23.03) = 2.22$, $p = 0.018$
5	$t(12.81) = 1.11$, $p = 0.144$	$t(8.42) = 0.59$, $p = 0.286$
6	$t(3.96) = -0.65$, $p = 0.275$	$t(2.04) = 0.68$, $p = 0.717$

7	$t(37.13) = -2.5, p = 0.009$	$t(39.89) = 2.1, p = 0.021$
8	$t(12.62) = -1.06, p = 0.846$	$t(10.69) = 0.57, p = 0.291$
9	$t(58.9) = 2.37, p = 0.011$	$t(57) = -2.25, p = 0.014$
10	$t(214.2) = -4.79, p = < .001$	$t(217.75) = 4.26, p = < .001$

All Participants.

Bitter versus Control.

For the bitter versus control contrast among all participants, there were 11 studies to consider. Of those, 2 suggested that the effect in the replication was equivalent to 0 +/- d33% with alpha = .05. Of these, 0 survived Bonferroni correction with alpha = .05 / 11 = .005.

Bitter versus Sweet.

For the bitter versus sweet contrast among all participants, there were 11 studies to consider. Of those, 3 suggested that the effect in the replication was equivalent to 0 +/- d33% with alpha = .05. Of these, 1 survived Bonferroni correction with alpha = .05 / 11 = .005.

Conservative Participants.

Bitter versus Control.

For the bitter versus control contrast among conservative participants, there were 5 studies to consider. Of those, 0 suggested that the effect in the replication was equivalent to 0 +/- d33% with alpha = .05. Of these, 0 survived Bonferroni correction with alpha = .05 / 5 = .010.

Bitter versus Sweet.

For the bitter versus sweet contrast among conservative participants, there were

5 studies to consider. Of those, 1 suggested that the effect in the replication was equivalent to 0 +/- d33% with alpha = .05. Of these, 0 survived Bonferroni correction with alpha = .05 / 5 = .010.

Liberal Participants.

Bitter versus Control.

For the bitter versus control contrast among liberal participants, there were 9 studies to consider. Of those, 4 suggested that the effect in the replication was equivalent to 0 +/- d33% with alpha = .05. Of these, 1 survived Bonferroni correction with alpha = .05 / 9 = .006.

Bitter versus Sweet.

For the bitter versus sweet contrast among liberal participants, there were 9 studies to consider. Of those, 4 suggested that the effect in the replication was equivalent to 0 +/- d33% with alpha = .05. Of these, 1 survived Bonferroni correction with alpha = .05 / 9 = .006.

LMER Results for Participants Rating All Six Vignettes

Table A7 shows results for linear mixed-effects regressions examining the independent and interactive effects of beverage condition and political orientation on ratings of moral wrongness only among participants who rated all six vignettes. Study is modeled as a random effect in all.

Estimates in these models are all rather similar in magnitude and direction to the models that include participants who rated three or more vignettes; none of them reveal hypothesized elevations of moral wrongness for the bitter group relative to both the control and sweet groups across everyone or as a function of political orientation.

Table A7

Linear mixed-effects models examining moral wrongness including only participants who rated all six vignettes.

	(1)	(2)	(3)
(Intercept)	69.858*** (0.902)	70.758*** (1.030)	70.759*** (2.677)
bitter - control (BvC)	3.664** (1.416)	5.795** (2.137)	5.795** (2.137)
bitter - sweet (BvS)	−2.157 (1.383)	−2.520 (2.250)	−2.520 (2.250)
conservative - liberal (CvL)	3.186* (1.288)	1.461 (1.853)	1.461 (1.853)
conservative - other (CvO)	−2.691 (1.413)	0.782 (2.228)	0.782 (2.228)
Level of Knowledge (0 = naive)		1.140 (1.345)	1.140 (1.345)
BvC × CvL	3.227 (3.396)	7.112 (4.940)	7.112 (4.940)
BvC × CvO	1.783 (3.347)	6.776 (5.191)	6.776 (5.191)
BvS × CvL	1.295 (3.832)	−3.864 (5.832)	−3.865 (5.833)
BvS × CvO	−7.695* (3.763)	−9.265 (6.243)	−9.264 (6.244)
BvC × Knowledge		−8.611* (3.936)	−8.611* (3.936)
BvS × Knowledge		3.272 (3.573)	3.272 (3.573)
CvL × Knowledge		5.211 (3.022)	5.211 (3.022)
CvO × Knowledge		−3.848 (3.409)	−3.848 (3.409)
BvC × CvL × Knowledge		−11.302 (8.943)	−11.302 (8.944)
BvC × CvO × Knowledge		−10.073 (8.244)	−10.074 (8.244)
BvS × CvL × Knowledge		0.311 (10.065)	0.311 (10.065)
BvS × CvO × Knowledge		8.444 (9.317)	8.444 (9.317)
Observations	1,047	847	5,082
Log Likelihood	−4,223.109	−3,387.079	−23,556.750
Akaike Inf. Crit.	8,468.218	6,814.157	47,157.500
Bayesian Inf. Crit.	8,522.709	6,908.991	47,301.230

Note: *p<0.05; **p<0.01; ***p<0.001

LMER Results by Vignette

Table A8 shows results for linear mixed effects regressions examining the independent and interactive effects of beverage condition and political orientation on ratings of moral wrongness for each vignette separately. Study is modeled as a random effect in all. None of these models reveal hypothesized elevations of moral wrongness for the bitter group relative to both the control and sweet groups across everyone or as a function of political orientation.

Table A9 shows results for two linear mixed effects regressions examining the independent and interactive effects of beverage condition and political orientation on ratings of moral wrongness for the average of the two vignettes featuring purity violations by characters Bob, who has a sexual relationship with his second cousin, and Frank, who cooks and eats his dead dog. The second analysis echoes the first but includes level of knowledge of the hypothesis (0 = naive, 1 = partially or fully suspicious) as an additional fixed effect. Study is modeled as a random effect in both. Neither model reveals hypothesized elevations of moral wrongness for the bitter group relative to both the control and sweet groups across everyone or as a function of political orientation.

Bayes Factors Within Studies

Table A10 shows results for the Bayes Factors tests across all participants, and for conservative and liberal subgroups for each study. Below we summarize the degree to which we observed nonanecdotal evidence for and against the replication hypothesis using a reporting strategy similar to Wagenmakers et al. (2016).

Table A8

Linear mixed-effects models examining moral wrongness for each of six individual vignettes.

	Bob	Frank	George	Arnold	Robert	Tim
	(1)	(2)	(3)	(4)	(5)	(6)
(Intercept)	58.764*** (3.265)	71.755*** (1.882)	68.277*** (2.046)	78.877*** (0.986)	70.541*** (1.819)	70.855*** (1.731)
bitter - control (BvC)	1.603 (3.089)	4.516 (2.827)	3.660 (2.440)	1.063 (1.970)	4.663 (2.415)	4.604 (2.386)
bitter - sweet (BvS)	−4.279 (3.040)	−5.662* (2.761)	−2.739 (2.384)	1.770 (1.920)	−0.882 (2.364)	−0.434 (2.328)
conservative - liberal (CvL)	2.115 (2.852)	−2.204 (2.563)	2.084 (2.229)	−1.832 (1.738)	7.151** (2.195)	8.181*** (2.165)
conservative - other (CvO)	−5.078 (3.139)	−3.281 (2.833)	0.096 (2.473)	1.285 (1.947)	−2.409 (2.438)	−1.713 (2.401)
BvC × CvL	−5.634 (7.403)	8.303 (6.760)	2.538 (5.828)	0.775 (4.704)	8.196 (5.775)	3.369 (5.699)
BvC × CvO	−0.144 (7.333)	−6.112 (6.648)	2.525 (5.730)	2.914 (4.614)	5.452 (5.688)	5.378 (5.600)
BvS × CvL	−2.760 (8.391)	7.834 (7.661)	0.524 (6.617)	0.542 (5.349)	−9.088 (6.560)	2.501 (6.471)
BvS × CvO	1.105 (8.277)	−14.201 (7.549)	−2.624 (6.547)	−4.242 (5.251)	−8.078 (6.472)	−11.283 (6.373)
Observations	1,089	1,117	1,125	1,130	1,132	1,129
Log Likelihood	−5,263.426	−5,302.797	−5,181.533	−4,961.170	−5,208.594	−5,176.015
Akaike Inf. Crit.	10,548.850	10,627.590	10,385.070	9,944.339	10,439.190	10,374.030
Bayesian Inf. Crit.	10,603.780	10,682.800	10,440.350	9,999.669	10,494.540	10,429.350

Note: *p<0.05; **p<0.01; ***p<0.001

Table A9

Linear mixed-effects models examining moral wrongness for the average of the two purity violation vignettes.

	(1)	(2)
(Intercept)	65.176*** (2.062)	65.314*** (2.559)
bitter - control (BvC)	3.255 (2.307)	1.565 (3.478)
bitter - sweet (BvS)	−5.103* (2.265)	−6.091 (3.677)
conservative - liberal (CvL)	0.017 (2.129)	−3.873 (3.041)
conservative - other (CvO)	−4.838* (2.330)	−0.112 (3.652)
Level of Knowledge (0 = naive)		3.417 (2.192)
BvC × CvL	1.591 (5.527)	4.691 (8.045)
BvC × CvO	−3.106 (5.465)	−7.438 (8.461)
BvS × CvL	3.525 (6.252)	1.301 (9.486)
BvS × CvO	−7.090 (6.159)	−5.577 (10.139)
BvC × Knowledge		−6.254 (6.278)
BvS × Knowledge		7.023 (5.819)
CvL × Knowledge		11.195* (4.901)
CvO × Knowledge		−6.983 (5.531)
BvC × CvL × Knowledge		−23.594 (14.325)
BvC × CvO × Knowledge		12.481 (13.429)
BvS × CvL × Knowledge		−8.960 (16.225)
BvS × CvO × Knowledge		9.846 (15.197)
Observations	1,074	873
Log Likelihood	−4,868.599	−3,925.406
Akaike Inf. Crit.	9,759.197	7,890.812
Bayesian Inf. Crit.	9,813.968	7,986.250

Note: *p<0.05; **p<0.01; ***p<0.001

Table A10

Results of Bayes Factors tests of the moral wrongness composite in each replication study.

	Bitter versus Control				Bitter versus Sweet			
study	JZSBayes	RepBayes	EffectSizeBayes	MetaBayes	JZSBayes	RepBayes	EffectSizeBayes	MetaBayes
All participants								
1	1.63	4.69	2.91	59.38	0.82	0.98	1.48	43.49
2	0.22	0.04	0.34	0.74	0.32	0.01	0.04	0.39
3	0.25	0.04	0.27	0.63	0.35	0.01	0.05	0.47
4	0.21	0.02	0.17	0.34	0.24	0.00	0.03	0.20
5	0.24	0.06	0.44	0.99	0.31	0.11	0.62	5.84
6	1.03	3.31	2.92	37.82	1.08	3.24	2.63	108.74
7	0.21	0.06	0.50	1.08	0.23	0.01	0.07	0.49
8	0.27	0.02	0.16	0.38	0.22	0.02	0.14	0.92
9	0.18	0.01	0.10	0.16	0.15	0.01	0.09	0.48
10	0.76	0.27	0.54	5.27	0.17	0.00	0.01	0.04
11	0.43	0.65	1.95	8.52	0.32	0.08	0.50	3.95
Conservatives								
1	2.83	14.08	1.81	959.11	1.18	0.03	0.10	1.46
2	0.42	0.02	0.09	4.43	0.59	0.00	0.02	0.13
4	3.07	0.00	0.00	0.23	3.78	0.00	0.00	0.15

5	0.43	0.03	0.24	5.22	0.54	0.06	0.41	2.41
10	0.70	0.05	0.13	11.30	0.37	0.04	0.22	2.21
Liberals								
1	11.72	18.70	0.72	9.42	0.81	2.11	2.56	3.24
3	0.37	0.66	2.39	0.51	0.47	0.13	0.52	0.21
4	0.27	0.38	2.04	0.29	0.26	0.16	1.10	0.31
5	0.37	0.34	1.43	0.24	0.48	0.15	0.60	0.26
6	0.49	0.82	2.08	0.58	0.99	2.43	1.76	3.16
7	0.23	0.27	1.78	0.20	0.25	0.09	0.69	0.16
8	5.05	0.38	0.11	0.07	0.51	0.12	0.47	0.20
9	0.25	0.17	1.16	0.09	0.26	0.25	1.50	0.53
10	0.19	0.27	2.13	0.25	0.39	0.07	0.36	0.03

Note. JZSBayes = Jeffreys-Zellner-Siow (JZS) Bayes Factor; RepBayes = Replication test Bayes Factor; EffectSizeBayes = equality-of-effect-size Bayes Factor; MetaBayes = fixed-effect meta-analysis Bayes Factor.

All Participants. There were 11 studies for the contrasts in this section.

Bitter versus Control.

For the JZS prior, 0 Bayes factors provide nonanecdotal support for the replication hypothesis (i.e., $BF_{10} > 3$) and 7 provide nonanecdotal support for the null hypothesis (i.e., $BF_{10} < 1/3$).

For the replication prior, 2 Bayes factors provide nonanecdotal support for the replication hypothesis (i.e., $BF_{10} > 3$) and 8 provide nonanecdotal support for the null hypothesis (i.e., $BF_{10} < 1/3$).

For the equality-of-effect-size approach, 0 Bayes factors provide nonanecdotal support for the null hypothesis, signaling replication of original results (i.e., $BF_{01} > 3$) and 4 provide nonanecdotal support for the alternative hypothesis, signaling nonreplication of original results (i.e., $BF_{01} < 1/3$).

For the fixed-effect meta-analysis approach, 4 Bayes factors provide nonanecdotal support for the replication hypothesis (i.e., $BF_{10} > 3$) and 1 provides nonanecdotal support for the null hypothesis (i.e., $BF_{10} < 1/3$).

Bitter versus Sweet.

For the JZS prior, 0 Bayes factors provide nonanecdotal support for the replication hypothesis and 8 provide nonanecdotal support for the null hypothesis.

For the replication prior, 1 Bayes factor provides nonanecdotal support for the replication hypothesis and 9 provide nonanecdotal support for the null hypothesis.

For the equality-of-effect-size approach, 0 Bayes factors provide nonanecdotal support for the null hypothesis, signaling replication of original results and 7 provide nonanecdotal support for the alternative hypothesis, signaling nonreplication of original results.

For the fixed-effect meta-analysis approach, 4 Bayes factors provide nonanecdotal support for the replication hypothesis and 2 provide nonanecdotal support for the null hypothesis.

Conservative Participants. Turning to the conservative subgroup, there were 5 studies for contrasts in this section.

Bitter versus Control.

For the JZS prior, 1 Bayes factor provides nonanecdotal support for the replication hypothesis and 0 provide nonanecdotal support for the null hypothesis.

For the replication prior, 1 Bayes factor provides nonanecdotal support for the replication hypothesis and 4 provide nonanecdotal support for the null hypothesis.

For the equality-of-effect-size approach, 0 Bayes factors provide nonanecdotal support for the null hypothesis, signaling replication of original results and 4 provide nonanecdotal support for the alternative hypothesis, signaling nonreplication of original results.

For the fixed-effect meta-analysis approach, 4 Bayes factors provide nonanecdotal support for the replication hypothesis and 1 provides nonanecdotal support for the null hypothesis.

Bitter versus Sweet.

For the JZS prior, 1 Bayes factor provides nonanecdotal support for the replication hypothesis and 0 provide nonanecdotal support for the null hypothesis.

For the replication prior, 0 Bayes factors provide nonanecdotal support for the replication hypothesis and 5 provide nonanecdotal support for the null hypothesis.

For the equality-of-effect-size approach, 0 Bayes factors provide nonanecdotal support for the null hypothesis, signaling replication of original results and 4 provide

nonanecdotal support for the alternative hypothesis, signaling nonreplication of original results.

For the fixed-effect meta-analysis approach, 0 Bayes factors provide nonanecdotal support for the replication hypothesis and 2 provide nonanecdotal support for the null hypothesis.

Liberal Participants. Finally, turning to the liberal subgroup, there were 9 studies for the contrasts in this section.

Bitter versus Control.

For the JZS prior, 2 Bayes factors provide nonanecdotal support for the replication hypothesis and 4 provide nonanecdotal support for the null hypothesis.

For the replication prior, 1 Bayes factor provides nonanecdotal support for the replication hypothesis and 3 provide nonanecdotal support for the null hypothesis.

For the equality-of-effect-size approach, 0 Bayes factors provide nonanecdotal support for the null hypothesis, signaling replication of original results and 1 provides nonanecdotal support for the alternative hypothesis.

For the fixed-effect meta-analysis approach, 1 Bayes factor provides nonanecdotal support for the replication hypothesis and 6 provide nonanecdotal support for the null hypothesis.

Bitter versus Sweet.

For the JZS prior, 0 Bayes factors provide nonanecdotal support for the replication hypothesis and 3 provide nonanecdotal support for the null hypothesis.

For the replication prior, 0 Bayes factors provide nonanecdotal support for the replication hypothesis and 7 provide nonanecdotal support for the null hypothesis.

For the equality-of-effect-size approach, 0 Bayes factors provide nonanecdotal support for the null hypothesis, signaling replication of original results and 0 provide nonanecdotal support for the alternative hypothesis, signaling nonreplication of original results.

For the fixed-effect meta-analysis approach, 2 Bayes factors provide nonanecdotal support for the replication hypothesis and 6 provide nonanecdotal support for the null hypothesis.

Session Information

Following is the output of R's sessionInfo() command, which reveals the information necessary to ensure analytic reproducibility of our work.

R version 3.5.0 (2018-04-23) Platform: x86_64-w64-mingw32/x64 (64-bit) Running under: Windows 10 x64 (build 17134)

Matrix products: default

locale: [1] LC_COLLATE=English_United States.1252 [2] LC_CTYPE=English_United States.1252

[3] LC_MONETARY=English_United States.1252 [4] LC_NUMERIC=C

[5] LC_TIME=English_United States.1252

attached base packages: [1] stats graphics grDevices utils datasets methods base

other attached packages: [1] SDMTools_1.1-221.1 sjlabelled_1.0.17 sjPlot_2.6.3

[4] lmerTest_3.1-0 compute.es_0.2-4 emmeans_1.3.4

[7] gridExtra_2.3 sjstats_0.17.4 stargazer_5.2.2

[10] viridis_0.5.1 viridisLite_0.3.0 ggplot2_3.1.1

[13] polspline_1.1.14 R2WinBUGS_2.1-21 boot_1.3-20

[16] BayesFactor_0.9.12-4.2 dplyr_0.8.1 tidyr_0.8.3

[19] MBESS_4.5.1 pwr_1.2-2 BSDA_1.2.0

[22] lattice_0.20-35 MCMCpack_1.4-4 MASS_7.3-49

[25] coda_0.19-2 TOSTER_0.3.4 psych_1.8.12

[28] metafor_2.1-0 effsize_0.7.4 lme4_1.1-21

[31] Matrix_1.2-14 pacman_0.5.1 papaja_0.1.0.9842

loaded via a namespace (and not attached): [1] nlme_3.1-137 mcmc_0.9-6 pbkrtest_0.4-7

[4] insight_0.3.0 numDeriv_2016.8-1 tools_3.5.0

[7] TMB_1.7.15 backports_1.1.4 R6_2.4.0

[10] DBI_1.0.0 lazyeval_0.2.2 colorspace_1.4-1

[13] nnet_7.3-12 withr_2.1.2 tidyselect_0.2.5

[16] mnormt_1.5-5 compiler_3.5.0 performance_0.1.0 [19] quantreg_5.38 SparseM_1.77 labeling_0.3

[22] effects_4.1-0 bookdown_0.10 bayestestR_0.1.0

[25] scales_1.0.0 mvtnorm_1.0-10 pbapply_1.4-0

[28] stringr_1.4.0 digest_0.6.19 foreign_0.8-70

[31] minqa_1.2.4 R.utils_2.8.0 rmarkdown_1.12

[34] pkgconfig_2.0.2 htmltools_0.3.6 rlang_0.3.4

[37] generics_0.0.2 gtools_3.8.1 R.oo_1.22.0

[40] magrittr_1.5 Rcpp_1.0.1 munsell_0.5.0

[43] R.methodsS3_1.7.1 stringi_1.4.3 yaml_2.2.0

[46] carData_3.0-2 plyr_1.8.4 grid_3.5.0

[49] parallel_3.5.0 sjmisc_2.7.9 forcats_0.4.0

[52] crayon_1.3.4 ggeffects_0.10.0 haven_2.1.0

[55] splines_3.5.0 hms_0.4.2 knitr_1.22

[58] pillar_1.4.1 estimability_1.3 reshape2_1.4.3

[61] codetools_0.2-15 glue_1.3.1 evaluate_0.13

[64] mitools_2.4 modelr_0.1.4 nloptr_1.2.1

[67] MatrixModels_0.4-1 gtable_0.3.0 purrr_0.3.2

[70] assertthat_0.2.1 xfun_0.7 xtable_1.8-4

[73] broom_0.5.2 survey_3.36 e1071_1.7-1

[76] survival_2.41-3 class_7.3-14 tibble_2.1.1

[79] glmmTMB_0.2.3

Appendix References

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., 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.

Hawkins, R. X., Smith, E. N., Au, C., Arias, J. M., Catapano, R., Hermann, E., ... Reynolds, J. (2018). Improving the replicability of psychological science through pedagogy. *Advances in Methods and Practices in Psychological Science*, *1*(1), 7–18.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R., ... others. (2016). Registered replication report: Strack, martin, & stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928.